Machine Learning: A Probabilistic Perspective Solution Manual Version 1.2

Fangqi Li, SJTU

Contents

1 Introduction

1.1 A Brief Introduction

Here we should have demonstrated the solution to problems in Chapter One in *Machine Learning, A Probabilistic Perspective*(MLAPP). Since the number of problem in Chapter is zero, we save this section as an introduction to this document, i.e.a solution manual.

This document provides detailed solution to almost all problems of textbook MLAPP from Chapter One to Chapter Fourteen(Chinese version) / Twenty-one(English version). We generally save the restatement of problems for readers themselves.

There are two class for problems in MLAPP: theortical inference and pratical projects. We provide solution to most inference problems apart from those which are nothing but straightforward algebra(and few which we failed to solve). Practical problems, which base on a Matlab toolbox, are beyond the scope of this document.

1.2 On *Machine Learning: A Probabilistic Perspective*

Booming studies and literatures have made the boundary of ”machine learning” vague.

On one hand, the rapid development of AI technology has kept the society shocked, which also results in sharply increase in number of students who would try to take related courses in colleges. On the other hand, some scholars are still uncertain in learning-related theories, especially deep learning.

The extraordinary achievements of machine learning in recent days often make one forget that this discipline has undergone a long evolution and whose establishment dates back at least to the studies of ”electronic brains” in the 1940s. Be that as it may, machine learning has not been defined as a ”closed” theory. Even in the some community of researchers, machine learning are crowned metaphysics or alchemistry. Personally, I believe tha being called metaphysics is a common experience shared by many branches of theory which are undergoing the most rapid development.

To be a completed theory, machine learning is still looking for a way to conclude itself in a closed system. The most successful attempt so far has been the one based on probability. As commented by David Blei from Princton on the back of MLAPP: ”In Machine Learning, the language of probability and statistics reveals important connections between seemingly disparate algorithms and strategies. Thus, its readers will become articulate in a holistic view of the state-of-art and poised to build the next generation of machine learning algorithms.”

The crucial idea in MLAPP is: machine learning is tantamount to Bayesian Statistics, which draws connections between numerous ”indepedent” algorithms. But the history of Bayesian statistics(which can be traced back to days of Laplace) outlengths the one of machinea learning a lot. MLAPP is not noval in holding such an idea. C.M.Bishop’s *Pattern Recognition and Machine Learning* is another typical example. Both of them are considered as classical textbooks in elementary machine learning.

In general, MLAPP reduces the difficulty of the entire book at the expense of partially deduced completeness(for the first seven chapters). It covers a wider range of models and is suitable for those with background in mathemathcal tools. The chapters that concerning classical probabilistic models (e.g, chapter 2, 3, 4, 5, 7, 8, 11, 12) is comparable to PRML. But due to the reordering and more details, they worth a read for one who have finished reading PRML.

1.3 Constitutions of this Document

The motivation for writing this document is that I need to read textbook MLAPP after selecting machine learning course, but I failed to find any free compiled solution manuals. Although several Github projects have started working on it, the velocity has been too slow. Also I want to focus more on the theoretical part of the text rather than the implementation code.

Hence I began working on this document. It is completed(first version, Chapter One to Chapter Fourteen) within the first two weeks before the official semester. Bacase of the hurry process, it is suggested that readers should read from a critical perspective and not hesitate to believe in everything I have written down. In the end, I hope that readers can provide comments and revise opinions. Apart from correcting the wrong answers, those who good at using MATLAB, Latex typesetting or those who are willing to participate in the improvement of the document are always welcome to contact me.

22/10/2017
Fangqi Li
Munich, Germany
solour_lfq@sjtu.edu.cn
ge72bug@tum.de

1.4 Updating log

22/10/2017(First Chinese compilation)
02/03/2018(English compilation)
06/03/2018(Begin Revising)
24/03/2018(First Revision)

2 Probability

2.1 Probability are sensitive to the form of the question that was used to generate the answer

Denote two children by A and B.

We use the following denotations:

$$E_1 : \text{A is a boy, B is a girl}$$

$$E_2 : \text{B is a boy, A is a girl}$$

$$E_3 : \text{A is a boy, B is a boy}$$

In **question a**:

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{4}$$

$$P(\text{one girl}|\text{one boy}) = \frac{P(E_1) + P(E_2)}{P(E_1) + P(E_2) + P(E_3)} = \frac{2}{3}$$

For **question b**,w.l.o.g, assume child A is observed:

$$P(\text{B is a girl}|\text{A is a boy}) = \frac{1}{2}$$

2.2 Legal reasoning

Let E_1 denote the event "the defendant commited the crime" and E_2 denotes "the defendant has special blood type" respectively. Thus:

$$\begin{aligned} p(E_1|E_2) =& \frac{p(E_1, E_2)}{p(E_2)} = \frac{p(E_2|E_1)p(E_1)}{p(E_2)} \\ =& \frac{1 \cdot \frac{1}{800000}}{\frac{1}{8000}} = \frac{1}{100} \end{aligned}$$

2.3 Variance of a sum

By straightforward calculation:

$$\begin{aligned} \text{var}[X+Y] =& \mathbb{E}[(X+Y)^2] - \mathbb{E}^2[X+Y] \\ =& \mathbb{E}[X^2] - \mathbb{E}^2[X] + \mathbb{E}[Y^2] - \mathbb{E}^2[Y] + 2\mathbb{E}[XY] - 2\mathbb{E}^2[XY] \\ =& \text{var}[X] + \text{var}[Y] + 2\text{cov}[X, Y] \end{aligned}$$

2.4 Bayes rule for medical diagnosis

We use E_i, E_h and E_p denotes whether one is ill or health, and one has been detected as positive. Applying Bayes's rules:

$$\begin{aligned} P(E_i|E_p) =& \frac{P(E_i)P(E_p|E_i)}{P(E_i)P(E_p|E_i) + P(E_h)P(E_p|E_h)} \\ =& 0.0098 \end{aligned}$$

2.5 The Monty Hall problem(The dilemma of three doors)

The answer is **b**. We use $E_{a,i}$ denotes the event that something happens to the ith box, a can be p(prize is in ith box), c(the gamer pick ith box), o(the host opens ith box). We assumes the participant choose the first box and the host reveals the third one. Applying Bayes's rules:

$$\begin{aligned} P(E_{p,1}|E_{c,1}, E_{o,3}) =& \frac{P(E_{c,1})P(E_{p,1})P(E_{c,3}|E_{p,1}, E_{c,1})}{P(E_{c,1})P(E_{o,3}|E_{c,1})} \\ =& \frac{P(E_{p,1})P(E_{c,3}|E_{p,1}, E_{c,1})}{P(E_{o,3}|E_{c,1})} \\ =& \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1} = \frac{1}{3} \end{aligned}$$

In the last step we summarize over the potential location of the prize. We conclude that it is always better to switch to another box after the host revealing one.

2.6 Conditional Independence

In **question a**, we have:

$$p(H|e_1, e_2) = \frac{p(H)p(e_1, e_2|H)}{p(e_1, e_2)}$$

Thus the answer is (ii).

For **question b**, we have further decomposition:

$$p(H|e_1, e_2) = \frac{p(H)p(e_1|H)p(e_2|H)}{p(e_1, e_2)}$$

So both (i) and (ii) are sufficient obviously. Since:

$$\begin{aligned} p(e_1, e_2) =& \sum_H p(e_1, e_2, H) \\ =& \sum_H p(H)p(e_1|H)p(e_2|H) \end{aligned}$$

(iii) is sufficint as well since we can calculate $p(e_1, e_2)$ independently.

2.7 Pairwise independence does not imply mutual independence

Let's assmue three boolean variables x_1, x_2, x_3. x_1 and x_2 have values of 0 or 1 with equal possibility independently:

$$p(x_1, x_2) = p(x_1)p(x_2)$$

$$p(x_1 = 0) = p(x_2 = 0) = \frac{1}{2}$$

And $x_3 = XOR(x_1, x_2)$. Now we have:

$$\begin{aligned} p(x_3 = 1) =& \sum_{x_1, x_2} p(x_1, x_2)p(x_3 = 1|x_1, x_2) \\ =& \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 1 \\ =& \frac{1}{2} \end{aligned}$$

Also:

$$\begin{aligned} p(x_3 = 1, x_1 = 1) =& \sum_{x_2} p(x_2)p(x_3 = 1, x_1 = 1|x_2) \\ =& \sum_{x_2} p(x_2)p(x_1 = 1)p(x_3 = 1|x_1 = 1, x_2) \\ =& \frac{1}{4} = p(x_3 = 1) \cdot p(x_1 = 1) \end{aligned}$$

Thus x_3 is pairwise independent w.r.t x_1 and x_2. But given x_1 and x_2, x_3 is uniquely determined and mutual independence failes.

2.8 Conditional independence iff joint factorizes

We prove 2.129 is equal to 2.130.

Firstly, by denoting:

$$g(x,z) = p(x|z)$$

$$h(y,z) = p(y|z)$$

We have the first half of proof.

Secondly we have:

$$\begin{aligned} p(x|z) =& \sum_y p(x,y|z) \\ =& \sum_y g(x,z)h(y,z) \\ =& g(x,z)\sum_y h(y,z) \\ p(y|z) =& h(y,z)\sum_x g(x,z) \end{aligned}$$

And:

$$\begin{aligned} 1 =& \sum_{x,y} p(x,y|z) \\ =& (\sum_x g(x,z))(\sum_y h(y,z)) \end{aligned}$$

Thus:

$$\begin{aligned} p(x|z)p(y|z) =& g(x,z)h(y,z)(\sum_x g(x,z))(\sum_y h(y,z)) \\ =& g(x,z)h(y,z) \\ =& p(x,y|z) \end{aligned}$$

2.9 Conditional independence

From a graphic view, both arguments are correct. But from a general view, both of them do not have a decomposition form, thus false.

2.10 Deriving the inverse gamma density

According to:

$$p(y) = p(x)|\frac{dx}{dy}|$$

We easily have:

$$\begin{aligned}IG(y) =& Ga(x) \cdot y^{-2} \\ =& \frac{b^a}{\Gamma(a)}(\frac{1}{y})^{(a-1)+2}e^{-\frac{b}{y}} \\ =& \frac{b^a}{\Gamma(a)}(y)^{-(a+1)}e^{-\frac{b}{y}}\end{aligned}$$

2.11 Normalization constant for a 1D Gaussian

This proof should be found around any textbook about multivariable calculus.Omitted here.

2.12 Expressing mutual information in terms of entropies

$$\begin{aligned}I(X;Y) =& \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ =& \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ =& \sum_{x,y} p(x,y) \log p(x|y) - \sum_{x}(\sum_{y} p(x,y)) \log p(x) \\ =& -H(X|Y) + H(X)\end{aligned}$$

Inversing X and Y yields to another formula.

2.13 Mutual information for correlated normals

We have:

$$\begin{aligned}
I(X_1;X_2) =& H(X_1) - H(X_1|X_2) \\
=& H(X_1) + H(X_2) - H(X_1,X_2) \\
=& \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\log(2\pi)^2\sigma^4(1-\rho^2) \\
=& -\frac{1}{2}\log(1-\rho^2)
\end{aligned}$$

(refer to **Elements of Information Theory**,Example 8.5.1)

We also give the derivation of Gaussian's entropy here:

$$\begin{aligned}
h =& -\int p(\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{x} \\
=& -\int p(\mathbf{x})\left\{-\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{x}-\mu)^{\mathbb{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right\}\mathrm{d}\mathbf{x} \\
=& \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\int p(\mathbf{x})(\mathbf{x}-\mu)^{\mathbb{T}}\Sigma^{-1}(\mathbf{x}-\mu)\mathrm{d}\mathbf{x} \\
=& \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\mathbb{E}_p[\sum_{i,j}(x_i-\mu_i)\Sigma_{ij}^{-1}(x_j-\mu_j)] \\
=& \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\sum_{i,j}\mathbb{E}_p[(x_i-\mu_i)(x_j-\mu_j)]\Sigma_{ij}^{-1} \\
=& \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\sum_{i,j}\Sigma_{ji}\Sigma_{ij}^{-1} \\
=& \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\mathrm{tr}\left\{\Sigma\Sigma^{-1}\right\} \\
=& \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma| + \frac{n}{2} \\
=& \frac{1}{2}\ln(2\pi e)^n|\Sigma|
\end{aligned}$$

The trick here is to use the defintion of covariance and the trace mark.

2.14 A measure of correlation

In question a:

$$\begin{aligned} r =& 1 - \frac{H(Y|X)}{H(X)} = \frac{H(X) - H(Y|X)}{H(X)} \\ =& \frac{H(Y) - H(Y|X)}{H(X)} \\ =& \frac{I(X;Y)}{H(X)} \end{aligned}$$

We have $0 \leq r \leq 1$ in question b for $I(X;Y) > 0$ and $H(X|Y) < H(X)$(properties of entropy).

$r = 0$ iff X and Y are independent.

$r = 1$ iff X is determined(not necassary equal) by Y.

2.15 MLE minimizes KL divergence to the empirical distribution

Expand the KL divergence:

$$\begin{aligned} \theta &= \arg\min_{\theta} \left\{ \mathbb{KL}(p_{\text{emp}}||q(\theta)) \right\} \\ &= \arg\min_{\theta} \left\{ \mathbb{E}_{p_{\text{emp}}}[\log \frac{p_{\text{emp}}}{q(\theta)}] \right\} \\ &= \arg\min_{\theta} \left\{ -H(p_{\text{emp}}) - \int p_{\text{emp}}(\mathbf{x}) \log q(\mathbf{x};\theta) \mathrm{d}\mathbf{x} \right\} \\ &\approx \arg\min_{\theta} \left\{ -H(p_{\text{emp}}) - \sum_{\mathbf{x} \in \text{dataset}} (\log q(\mathbf{x};\theta)) \right\} \\ &= \arg\max_{\theta} \left\{ \sum_{\mathbf{x} \in \text{dataset}} \log q(\mathbf{x};\theta) \right\} \end{aligned}$$

We end up in MLE. We use the weak form of the law of large numbers in the fourth step and drop the entropy of empirical distribution in the last step.

2.16 Mean, mode, variance for the beta distribution

Firstly, derive the mode for beta distribution by differentiating the pdf:

$$\frac{\mathrm{d}}{\mathrm{d}x} x^{a-1}(1-x)^{b-1} = [(1-x)(a-1) - (b-1)x] x^{a-2}(1-x)^{b-2}$$

Setting this to zero yields:

$$\text{mode} = \frac{a-1}{a+b-2}$$

Secondly, derive the moment in beta distribution:

$$\begin{aligned}\mathbb{E}[x^N] =& \frac{1}{B(a,b)}\int x^{a+N-1}(1-x)^{b-1}dx \\ =& \frac{B(a+N,b)}{B(a,b)} \\ =& \frac{\Gamma(a+N)\Gamma(b)}{\Gamma(a+N+b)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\end{aligned}$$

Setting $N = 1, 2$:

$$\mathbb{E}[x] = \frac{a}{a+b}$$

$$\mathbb{E}[x^2] = \frac{a(a+1)}{(a+b)(a+b+1)}$$

Where we have used the property of Gamma function. Straightforward algebra gives:

$$\text{mean} = \mathbb{E}[x] = \frac{a}{a+b}$$

$$\text{variance} = \mathbb{E}[x^2] - \mathbb{E}^2[x] = \frac{ab}{(a+b)^2(a+b+1)}$$

2.17 Expected value of the minimum

Let m denote the location of the left most point, we have:

$$\begin{aligned}p(m > x) =& p([X > x]\textbf{and}[Y > x]) \\ =& p(X > x)p(Y > x) \\ =& (1-x)^2\end{aligned}$$

Therefore:

$$\begin{aligned}\mathbb{E}[m] =& \int x \cdot p(m = x)dx \\ =& \int p(m > x)dx \\ =& \int_0^1 (1-x)^2 dx \\ =& \frac{1}{3}\end{aligned}$$

3 Generative models for discrete data

3.1 MLE for the Beroulli/binomial model

Likelihood:

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Log-Likelihood:

$$\ln p(D|\theta) = N_1 \ln \theta + N_0 \ln(1-\theta)$$

Setting the derivative to zero:

$$\frac{\partial}{\partial\theta} \ln p(D|\theta) = \frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0$$

This ends in 3.22:

$$\theta = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}$$

3.2 Marginal likelihood for the Beta-Bernoulli model

Likelihood:

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Prior distribution:

$$p(\theta|a,b) = \text{Beta}(\theta|a,b) = \theta^{a-1}(1-\theta)^{b-1}$$

Posterior distribution:

$$\begin{aligned} p(\theta|D) \propto & p(D|\theta) \cdot p(\theta|a,b) \\ = & \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \\ = & \text{Beta}(\theta|N_1+a, N_0+b) \end{aligned}$$

Predictive distribution:

$$\begin{aligned} p(x_{\text{new}} = 1|D) = & \int p(x_{\text{new}} = 1|\theta) \cdot p(\theta|D) \mathrm{d}\theta \\ = & \int \theta p(\theta|D) \mathrm{d}\theta \\ = & \mathbb{E}(\theta) = \frac{N_1 + a}{N_1 + a + N_0 + b} \end{aligned}$$

Calcualte $p(D)$ where $D = 1, 0, 0, 1, 1$:

$$\begin{aligned} p(D) =& p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)...p(X_N|x_{N-1}, X_{N-2}, ...X_1) \\ =& \frac{a}{a+b}\frac{b}{a+b+1}\frac{b+2}{a+b+2}\frac{a+1}{a+b+3}\frac{a+2}{a+b+4} \end{aligned}$$

Denote $\alpha = a + b, \alpha_1 = a, \alpha_0 = b$, we have 3.83. To derive 3.80, we make use of:

$$[(\alpha_1)..(\alpha_1 + N_1 - 1)] = \frac{(\alpha_1 + N_1 - 1)!}{(\alpha_1 - 1)!} = \frac{\Gamma(\alpha_1 + N_1)}{\Gamma(\alpha_1)}$$

3.3 Posterior predictive for Beta-Binomial model

Straightforward algebra:

$$\begin{aligned} \text{Bb}(\alpha_1', \alpha_0', 1) =& \frac{B(\alpha_1' + 1, \alpha_0')}{B(\alpha_1', \alpha_0')} \\ =& \frac{\Gamma(\alpha_0' + \alpha_1')}{\Gamma(\alpha_0' + \alpha_1' + 1)}\frac{\Gamma(\alpha_1' + 1)}{\Gamma(\alpha_1')} \\ =& \frac{\alpha_1'}{\alpha_1' + \alpha_0'} \end{aligned}$$

3.4 Beta updating from censored likelihood

The derivation is straightforward:

$$\begin{aligned} p(\theta, X < 3) =& p(\theta)p(X < 3|\theta) \\ =& p(\theta)(p(X = 1|\theta) + p(X = 2|\theta)) \\ =& \text{Beta}(\theta|1, 1)(\text{Bin}(1|5, \theta) + \text{Bin}(2|5, \theta)) \end{aligned}$$

3.5 Uninformative prior for log-odds ratio

Since:

$$\phi = \log\frac{\theta}{1-\theta}$$

By using change of variables formula:

$$p(\theta) = p(\phi)|\frac{d\phi}{d\theta}| \propto \frac{1}{\theta(1-\theta)}$$

Hence:

$$p(\theta) = \text{Beta}(\theta|0, 0)$$

3.6 MLE for the Poisson distribution

Likelihood (we use Poi in condition to represent the fact that we have knowledge about the form of the distribution):

$$p(D|\text{Poi},\lambda) = \prod_{n=1}^{N} \text{Poi}(x_n|\lambda) = \exp(-\lambda N) \cdot \lambda^{\sum_{n=1}^{N} x_n} \cdot \frac{1}{\prod_{n=1}^{N} x_n!}$$

Setting the derivative of Log-Likelihood to zero:

$$\frac{\partial}{\partial\lambda} p(D|\text{Poi},\lambda) = \exp(-\lambda N) \cdot \lambda^{\sum x - 1} \cdot \left\{ -N\lambda + \sum_{n=1}^{N} x_n \right\} = 0$$

Thus:

$$\lambda = \frac{\sum_{n=1}^{N} x_n}{N}$$

3.7 Bayesian analysis of the Poisson distribution

We have:

$$\begin{aligned} p(\lambda|D) \propto & p(\lambda)p(D|\lambda) \\ \propto & \exp(-\lambda(N+b)) \cdot \lambda^{\sum_{n=1}^{N} x_n + a - 1} \\ = & \text{Gamma}(a + \sum x, N + b) \end{aligned}$$

This prior distribution equals introduing b prior observations with mean $\frac{a}{b}$.

3.8 MLE for the uniform distrbution

The likelihood goes to zero if $a < \max(x_n)$, so we must have $\hat{a} \geq \max(x_n)$, likelihood lookes like:

$$p(D|a) = \prod_{n=1}^{N} \frac{1}{2a}$$

Which has a negative correlation with a, so:

$$\hat{a} = \max \{x_i\}_{i=1}^{n}$$

This model assign $p(x_{n+1}) = 0$ if $x_{n+1} > \max \{x_i\}_{i=1}^{n}$, which cause discontinuity in predictive distribution, and is not an adorable feature.

3.9 Bayesian analysis of the uniform distribution

The conjugate prior for uniform distribution if Pareto distribution:

$$p(\theta) = \text{Pareto}(\theta|K, b) = Kb^K\theta^{-(K+1)}[\theta \geq b]$$

Let $m = \max\{x_i\}_{i=1}^n$, the joint distribution is:

$$p(\theta, D) = p(\theta)p(D|\theta) = Kb^K\theta^{-(K+N+1)}[\theta \geq b][\theta \geq m]$$

The marginal likelihood/evidence is:

$$p(D) = \int p(D, \theta)\mathrm{d}\theta = \frac{Kb^K}{(N+K)\max(m, b)^{N+K}}$$

Let $\mu = \max\{m, b\}$, the posterior distribution is again the form of a Parato distribution:

$$p(\theta|D) = \frac{p(\theta, D)}{p(D)} = \frac{(N+K)\mu^{N+K}[\theta \geq \mu]}{\theta^{N+K+1}} = \text{Pareto}(\theta|N+K, \mu)$$

3.10 Taxicab problem

For **question a**, we have $D = \{100\}$, $m = \max\{D\} = 100$, $N = 1$. Using a prior $K = 0$, $b = 0$, we have the posterior:

$$\text{Pareto}(\theta|1, 100)$$

A for **question b**, with posterior mode given by ($k = 1$, thus mean and variance does not exist):

$$\text{mode} = 100$$

To calculate the median:

$$\int_m^{\text{median}} km^k x^{-(k+1)}\mathrm{d}x = \frac{1}{2}$$

Plug in figures and using the fact $\int x^{-2}\mathrm{d}x = -x^{-1} + C$, solve for

$$\text{median} = 200$$

In **question c**, we already had (from **exercise 3.9**) the predictive distribution, in this case $\alpha = (b = 0, K = 0)$, $\beta = (c = m, N + K = 1)$. Plug them into the form of evidence:

In case when $x > m$(the number of new taxi is larger than the one we saw):

$$p(x|D,\alpha) = \frac{m}{2x^2}$$

If $x < m$:

$$p(x|D,\alpha) = \frac{m}{2m^2} = \frac{1}{2m}$$

We plug in $m = 100$ into the equations above to solve for **question d**:

$$p(x = 50|m = 100) = \frac{1}{200}$$

$$p(x = 100|m = 100) = \frac{1}{200}$$

$$p(x = 150|m = 100) = \frac{1}{450}$$

(Do please reason on this result!)

We omit **question e** as an open quesion.

3.11 Bayesian analysis of the exponential distribution

Log-Likelihood for an exponential distribution is:

$$\ln p(D|\theta) = N \ln \theta - \theta \sum_{n=1}^{N} x_n$$

The derivative is:

$$\frac{\partial}{\partial \theta} \ln p(D|\theta) = \frac{N}{\theta} - \sum_{n=1}^{N} x_n$$

Thus in question a:

$$\theta_{ML} = \frac{N}{\sum_{n=1}^{N} x_n}$$

We skip other questions and state that the conjugate prior for exponential distribution is Gamma distribution:

$$\begin{aligned} p(\theta|D) \propto & p(\theta)p(D|\theta) \\ = & \text{Gamma}(\theta|a,b)p(D|\theta) \\ = & \text{Gamma}(\theta|N+a, b+\sum x_n) \end{aligned}$$

A Gamma prior introduces $a - 1$ prior observation with the sum b.

3.12 MAP estimation for the Bernoulli with non-conjugate priors

In **question a**, we have:

$$p(\theta = 0.5|D) \propto p(\theta = 0.5)p(D|\theta = 0.5) = \frac{1}{2}^{1+N_1+N_0}$$

$$p(\theta = 0.4|D) \propto p(\theta = 0.4)p(D|\theta = 0.4) = \frac{1}{2}^{1} \cdot \frac{2}{5}^{N_1} \cdot \frac{3}{5}^{N_0}$$

If the MAP estimation is $\theta = 0.5$, i.e:

$$\ln \frac{p(\theta = 0.5|D)}{p(\theta = 0.4|D)} = N_1 \ln \frac{5}{4} + N_0 \ln \frac{5}{6} > 0$$

Then this must be held:

$$\frac{N_1}{N_0} > 0.817$$

Else MAP estimation gives $\theta = 0.4$.

For **question b**, in case N is small, prior in **question a** is able to yield a fairly good estimation (the prior is not conjugate yet close to truth). But as N grows, it can only getting close to 0.4, while Beta-prior tends to yield the true parameter with less error.

3.13 Posterior predictive distribution for a batch of data with the dirichlet-multinomial model

Since we already have 3.51:

$$p(X = j|D, \alpha) = \frac{\alpha_j + N_j}{\alpha_0 + N}$$

We can easily derive:

$$\begin{aligned} p(\tilde{D}|D, \alpha) &= \prod_{x \in \tilde{D}} p(x|D, \alpha) \\ &= \prod_{j=1}^{C} (\frac{\alpha_j + N_j^{\text{old}}}{\alpha_0 + N^{\text{old}}})^{N_j^{\text{new}}} \end{aligned}$$

3.14 Posterior predictive for Dirichlet-multinomial

Solutions to this exercise can be obtained from conclusions drawn from **exercise 3.13**.

3.15 Setting the hyper-parameters I

We already have:

$$\begin{cases} \text{mean} = \dfrac{\alpha_1}{\alpha_1 + \alpha_2} \\ \text{var} = \dfrac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)} \end{cases}$$

Using notation $A = \alpha_1$ and $B = \alpha_1 + \alpha_2$, we have:

$$m = \frac{A}{B}$$

$$v = \frac{A(B - A)}{B^2(B + 1)}$$

Cancell A:

$$B = \frac{m(1 - m)}{v} - 1$$

$$A = mB$$

For given figures, we calculate:

$$B = 130$$

$$A = 91$$

Hence:

$$\alpha_1 = 91$$

$$\alpha_2 = 39$$

3.16 Setting the beta hyper-parameters II

For paremeters of a Beta distribution α_1 and α_2 are connected through:

$$\alpha_2 = \alpha_1(\frac{1}{m} - 1) = f(\alpha_1)$$

Calculate this intergral:

$$\int_l^u \frac{1}{B(\alpha_1, f(\alpha_1))} \theta^{\alpha_1}(1 - \theta)^{f(\alpha_1)} = u(\alpha_1)$$

Setting this intergral $u(\alpha_1) \rightarrow 0.95$ by altering α_1 through numerical method will do.

3.17 Marginal likelihood for beta-binomial under uniform prior

The marginal likelihood is given by:

$$p(N_1|N) = \int_0^1 p(N_1, \theta|N)\mathrm{d}\theta = \int_0^1 p(N_1|\theta, N)p(\theta)\mathrm{d}\theta$$

We already have:

$$p(N_1|\theta, N) = \mathrm{Bin}(N_1|\theta, N)$$

$$p(\theta) = \mathrm{Beta}(1, 1)$$

Thus:

$$\begin{aligned} p(N_1|N) = & \int_0^1 \binom{N}{N_1}\theta^{N_1}(1-\theta)^{N-N_1}\mathrm{d}\theta \\ = & \binom{N}{N_1}B(N_1+1, N-N_1+1) \\ = & \frac{N!}{N_1!(N-N_1)!}\frac{N_1!(N-N_1)!}{(N+1)!} \\ = & \frac{1}{N+1} \end{aligned}$$

Where B is the regulizer for a Beta distribution:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

3.18 Bayes factor for coin tossing*

Straightforward calculation.

3.19 Irrelevant features with naive Bayes

Log-Likelihood is given by:

$$\log p(\mathbf{x}_i|c, \theta) = \sum_{w=1}^{W} x_{iw} \log \frac{\theta_{cw}}{1-\theta_{cw}} + \sum_{w=1}^{W} \log(1-\theta_{cw})$$

In a succint way:

$$\log p(\mathbf{x}_i|c, \theta) = \phi(\mathbf{x}_i)^T \beta_c$$

Where:

$$\phi(\mathbf{x}_i) = (\mathbf{x}_i, 1)^T$$

$$\beta_c = (\log \frac{\theta_{c1}}{1-\theta_{c1}}, ... \sum_{w=1}^{W} \log(1-\theta_{cw}))^T$$

For **question a**:

$$\begin{aligned}
\log \frac{p(c=1|\mathbf{x}_i)}{p(c=2|\mathbf{x}_i)} =& \log \frac{p(c=1)p(\mathbf{x}_i|c=1)}{p(c=2)p(\mathbf{x}_i|c=2)} \\
=& \log \frac{p(\mathbf{x}_i|c=1)}{p(\mathbf{x}_i|c=1)} \\
=& \phi(\mathbf{x}_i)^T(\beta_1 - \beta_2)
\end{aligned}$$

For **question b**, in a binary context:

$$p(c=1|\mathbf{x}_i) = \frac{p(c=1)p(\mathbf{x}_i|c=1)}{p(\mathbf{x}_i)}$$

Thus:

$$\log \frac{p(c=1|\mathbf{x}_i)}{p(c=2|\mathbf{x}_i)} = \log \frac{p(c=1)}{p(c=2)} + \phi(\mathbf{x}_i)^T(\beta_1 - \beta_2)$$

A word w will not affect this posterior measure as long as:

$$x_{iw}(\beta_{1,w} - \beta_{2,w}) = 0$$

Hence:

$$\theta_{c=1,w} = \theta_{c=2,w}$$

So the chance that word w appears in both class of documents are equal.

In question c, we have:

$$\hat{\theta}_{1,w} = 1 - \frac{1}{2+N_1}$$

$$\hat{\theta}_{2,w} = 1 - \frac{1}{2+N_2}$$

They do not equal when $N_1 \neq N_2$ so the bias effect remains. However, this effect reduces when N grows large.

3.20 Class conditional densities for binary data

In **question a**, we have:

$$p(\mathbf{x}|y=c) = \prod_{i=1}^{D} p(x_i|y=c, x_1, ..., x_{i-1})$$

The number of parameter is:

$$C \cdot \sum_{i=1}^{D} 2^i = C \cdot (2^{D+1} - 2) = O(C \cdot 2^D)$$

For **question b** and **question c**, we generally think that naive models fit better when N is large, because delicate models have problems of overfitting.

In **question d**, **question e** and **question f**, it is assumed that looking up for a value according to a D-dimensional index cost O(D) time. It is easy to calculate the fitting complexity: $O(ND)$ for a naive model and $O(N \cdot 2^D)$ for a full model, and the applying complexity is $O(CD)$ and $O(C \cdot 2^D)$ respectively.

For **question f**:

$$p(y|\mathbf{x}_v) \propto p(\mathbf{x}_v|y) = \sum_{\mathbf{x}_h} p(\mathbf{x}_v, \mathbf{x}_h|y)$$

Thus the complexity is multiplied by an extra const $2^{|\mathbf{x}_h|}$.

3.21 Mutual information for naive Bayes classifiers with binary features

By definition:

$$I(X;Y) = \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}$$

For binary features, consider the value of x_j to be zero and one, given

$\pi_c = p(y = c), \theta_{jc} = p(x_j = 1|y = c), \theta_j = p(x_j = 1)$:

$$\begin{aligned} I_j &= \sum_c p(x_j = 1, c) \log \frac{p(x_j = 1, c)}{p(x_j = 1)p(c)} \\ &+ \sum_c p(x_j = 0, c) \log \frac{p(x_j = 0, c)}{p(x_j = 0)p(c)} \\ &= \sum_c \pi_c \theta_{jc} \log \frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc})\pi_c \log \frac{1 - \theta_{jc}}{1 - \theta_j} \end{aligned}$$

Which ends in 3.76.

3.22 Fitting a naive Bayesian spam filter by hand*

Straightforward calculation.

4 Gaussian models

4.1 Uncorrelated does not imply independent

We first calculate the covariance of X and Y:

$$\begin{aligned}\mathrm{cov}(X,Y) =& \int\int(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))p(X,Y)\mathrm{d}X\mathrm{d}Y\\ =& \int_{-1}^{1}X(X^2-\frac{1}{\sqrt{3}})\mathrm{d}X=0\end{aligned}$$

The intergral ends in zero since we are intergrating an odd function in range [-1,1], hence:

$$\rho(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}=0$$

4.2 Uncorrelated and Gaussian does not imply independent unless jointly Gaussian

The pdf for Y is:

$$p(Y=a)=0.5\cdot p(X=a)+0.5\cdot p(X=-a)=p(X=a)$$

The pdf of X is symetric with 0 as the core, so Y subject to a normal distribution $(0,1)$.

For question b, we have:

$$\begin{aligned}\mathrm{cov}(X,Y)=&\mathbb{E}(XY)-\mathbb{E}(X)-\mathbb{E}(Y)\\ =&\mathbb{E}_W(\mathbb{E}(XY|W))-0\\ =&0.5\cdot\mathbb{E}(X^2)+0.5\cdot\mathbb{E}(-X^2)=0\end{aligned}$$

4.3 Correlation coefficient is between -1 and 1

The statement:

$$-1\leq\rho(X,Y)\leq 1$$

Equals:

$$|\rho(X,Y)|\leq 1$$

Hence we are to prove:

$$|\mathrm{cov}(X,Y)|^2 \leq \mathrm{var}(X) \cdot \mathrm{var}(Y)$$

Which can be drawn straightforwardly from Cauchy–Schwarz inequality in R^2.

4.4 Correlation coefficient for linearly related variables is 1 or -1

When $Y = aX + b$:

$$\mathbb{E}(Y) = a\mathbb{E}(x) + b$$

$$\mathrm{var}(Y) = a^2\mathrm{var}(X)$$

Therefore:

$$\begin{aligned}\mathrm{cov}(X,Y) =& \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\\ =& a\mathbb{E}(X^2) + b\mathbb{E}(X) - a\mathbb{E}^2(X) - b\mathbb{E}(X)\\ =& a \cdot \mathrm{var}(X)\end{aligned}$$

We also have:

$$\mathrm{var}(X)\mathrm{var}(Y) = a^2 \cdot \mathrm{var}(X)$$

These two make:

$$\rho(X,Y) = \frac{a}{|a|}$$

4.5 Normalization constant for a multidimensional Gaussian

Applying SVD on the precision matrix $\Sigma^{-1} = Q^{\mathbb{T}}\Lambda Q$, using the fact that Q is orthonomal ($|Q| = 1$) and we can set $\mu = 0$ w.l.o.g. Hence the integral:

$$\int \exp\left\{-\frac{1}{2}(Q\mathbf{x})^{\mathbb{T}}\Lambda(Q\mathbf{x})\right\} \mathrm{d}Q\mathbf{x} = \int \exp\left\{-\frac{1}{2}\sum_{i=1}^{d}(Q\mathbf{x})_i^2\lambda_i\right\} \mathrm{d}Q\mathbf{x}$$

Can be factorized into independent components, with each component a zero-mean one-dimensional Gaussian. Thus the normalization constant is given by the product of all the independent normalization constants.

$$\prod_{i=1}^{d}\sqrt{\frac{2\pi}{\lambda_i}} = (2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}$$

4.6 Bivariate Gaussian

This can be solved through straightforward algebra.

4.7 Conditioning a bivariate Gaussian

Reasoning on Gaussian distribution mostly bases on a standard procedure named "completing the square". Which has been demonstrated throughly in PRML Chapter 2, and the solution to this exercise can be obtained by plugging figures into formula derived in those sections directly.

4.8 Whitening vs standardizing*

Practical by yourself.

4.9 Sensor fusion with known variances in 1d

Denate the two observed datasets by $Y^{(1)}$ and $Y^{(2)}$, with size N_1, N_2, the likelihood is:

$$\begin{aligned} p(Y^{(1)}, Y^{(2)}|\mu) = \prod_{n_1=1}^{N_1} p(Y_{n_1}^{(1)}|\mu) \prod_{n_2=1}^{N_2} p(Y_{n_2}^{(2)}|\mu) \\ \propto \exp\left\{A\cdot\mu^2 + B\cdot\mu\right\} \end{aligned}$$

Where we have used:

$$\begin{aligned} A =& -\frac{N_1}{2v_1} - \frac{N_2}{2v_2} \\ B =& \frac{1}{v_1}\sum_{n_1=1}^{N_1} Y_{n_1}^{(1)} + \frac{1}{v_2}\sum_{n_2=1}^{N_2} Y_{n_2}^{(2)} \end{aligned}$$

Differentiate the likelihood and set it to zero, we have:

$$\mu_{\text{ML}} = -\frac{B}{2A}$$

The conjugate prior of this model must have form proporitional to $\exp\{A \cdot \mu^2 + B \cdot \mu\}$, namely a normal distribution:

$$p(\mu|a,b) \propto \exp\left\{a \cdot \mu^2 + b \cdot \mu\right\}$$

The posterior distribution is:

$$p(\mu|Y) \propto \exp\left\{(A+a) \cdot \mu^2 + (B+b) \cdot \mu\right\}$$

Hence we have the MAP estimation:

$$\mu_{\text{MAP}} = -\frac{B+b}{2(A+a)}$$

It is noticable that the MAP converges to ML estimation when observation times grow:

$$\mu_{\text{MAP}} \rightarrow \mu_{\text{ML}}$$

The posterior distribution is another normal distribution, with:

$$\sigma^2_{\text{MAP}} = -\frac{1}{2(A+a)}$$

4.10 Derivation of information form formulae for marginalizing and conditioning

Please refer to PRML chapter 2.

4.11 Derivation of the NIW posterior

The likelihood for a MVN is given by:

$$p(\mathbf{X}|\mu,\Sigma) = (2\pi)^{-\frac{ND}{2}} |\Sigma|^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_i - \mu)^{\text{T}}\Sigma^{-1}(\mathbf{x}_i - \mu)\right\}$$

According to 4.195:

$$\sum_{n=1}^{N}(\mathbf{x}_i-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\mu)=\sum_{n=1}^{N}(\bar{\mathbf{x}}-\mu+(\mathbf{x}_i-\bar{\mathbf{x}}))^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu+(\mathbf{x}_i-\bar{\mathbf{x}}))$$

$$=N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\sum_{n=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})$$

$$=N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\mathrm{tr}\left\{\Sigma^{-1}\sum_{n=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}\right\}$$

$$=N(\bar{\mathbf{x}}-\mu)^{T}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\mathrm{tr}\left\{\Sigma^{-1}\mathbf{S}_{\bar{\mathbf{x}}}\right\}$$

The conjugate prior for MVN's parameters (μ,Σ) is Normal-inverse-Wishart(NIW) distribution, defined by:

$$\mathrm{NIW}(\mu,\Sigma|\mathbf{m}_0,k_0,v_0,\mathbf{S}_0)=\mathcal{N}(\mu|\mathbf{m}_0,\frac{1}{k_0}\Sigma)\cdot\mathrm{IW}(\Sigma|\mathbf{S}_0,v_0)$$

$$=\frac{1}{Z}|\Sigma|^{-\frac{v_0+D+2}{2}}\exp\left\{-\frac{k_0}{2}(\mu-\mathbf{m}_0)^{\mathrm{T}}\Sigma^{-1}(\mu-\mathbf{m}_0)-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\mathbf{S}_0\right\}\right\}$$

Hence the posterior:

$$p(\mu,\Sigma|\mathbf{X})\propto|\Sigma|^{-\frac{v_{\mathbf{X}}+D+2}{2}}\exp\left\{-\frac{k_{\mathbf{X}}}{2}(\mu-\mathbf{m}_{\mathbf{X}})^{\mathrm{T}}\Sigma^{-1}(\mu-\mathbf{m}_{\mathbf{X}})-\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}\mathbf{S}_{\mathbf{X}}\right\}\right\}$$

Where we have:

$$k_{\mathbf{X}}=k_0+N$$

$$v_{\mathbf{X}}=v_0+N$$

$$\mathbf{m}_{\mathbf{X}}=\frac{N\bar{\mathbf{x}}+k_0\mathbf{m}_0}{k_{\mathbf{X}}}$$

By comparing the exponential for $|\Sigma|$,$\mu^{\mathrm{T}}\Sigma^{-1}\mu$ and μ^{T}.

Making use of $A^{\mathrm{T}}\Sigma^{-1}A=\mathrm{tr}\left\{A^{\mathrm{T}}\Sigma^{-1}A\right\}=\mathrm{tr}\left\{\Sigma^{-1}AA^{\mathrm{T}}\right\}$ and comparing the constant term inside the exponential function:

$$N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}+\mathbf{S}_{\bar{\mathbf{X}}}+k_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}}+\mathbf{S}_0=k_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^{\mathrm{T}}+\mathbf{S}_{\mathbf{X}}$$

Hence

$$\mathbf{S}_{\mathbf{X}}=N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}+\mathbf{S}_{\bar{\mathbf{X}}}+k_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}}+\mathbf{S}_0-k_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^{\mathrm{T}}$$

Use the definition for mean we ends in 4.214 since:

$$\mathbf{S}=\sum_{n=1}^{N}\mathbf{x}_i\mathbf{x}_t^{\mathrm{T}}=\mathbf{S}_{\bar{\mathbf{X}}}+N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}$$

Hence the posterior distribution for MVN takes the form:NIW$(\mathbf{m}_{\mathbf{X}},k_{\mathbf{X}},v_{\mathbf{X}},\mathbf{S}_{\mathbf{X}})$

4.12 BIC for Gaussians

Straightforward calculation.

4.13 Gaussian posterior credible interval

Assume a prior distribution for an 1d normal distribution:

$$p(\mu) = N(\mu|\mu_0, \sigma_0^2 = 9)$$

And the observed variable subjects to:

$$p(x) = N(x|\mu, \sigma^2 = 4)$$

Having observed n variables, it is vital that the probability mass of μ's posterior distribution is no less than 0.95 in an interval no longer than 1.

Posterior for μ is:

$$\begin{aligned} p(\mu|D) \propto p(\mu)p(D|\mu) =& N(\mu|\mu_0, \sigma_0^2)\prod_{i=1}^{n} N(x_n|\mu, \sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}\prod_{i=1}^{n}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \\ &= \exp\left\{(-\frac{1}{2\sigma_0^2}-\frac{n}{2\sigma^2})\mu^2 + ...\right\} \end{aligned}$$

Hence the posterior variance is given by:

$$\sigma_{\text{post}}^2 = \frac{\sigma_0^2\sigma^2}{\sigma^2 + n\sigma_0^2}$$

Since 0.95 of probability mass for a normal distribution lies within -1.96σ and 1.96σ, we have:

$$n \geq 611$$

4.14 MAP estimation for 1d Gaussians

Assume the variance for this distribution σ^2 is known, the mean μ subject to a normal distribution with mean m and variance s^2, similiar to the question before, the posterior takes the form:

$$p(\mu|X) \propto p(\mu)p(X|\mu)$$

The posterior is another normal distribution, by comparing the coefficient for μ^2:

$$-\frac{1}{2s^2} - \frac{N}{2\sigma^2}$$

And that for μ:

$$\frac{m}{s^2} + \frac{\sum_{n=1}^{N} x_n}{\sigma^2}$$

We have the posterior mean and variance by the technique "completing the square":

$$\sigma^2_{\text{post}} = \frac{s^2\sigma^2}{\sigma^2 + Ns^2}$$

$$\mu_{\text{post}} = (\frac{m}{s^2} + \frac{\sum_{n=1}^{N} x_n}{\sigma^2}) \cdot \sigma^2_{\text{post}}$$

Already we knew the MLE is:

$$\mu_{\text{ML}} = \frac{\sum_{n=1}^{N} x_i}{N}$$

When N increases, μ_{post} converges to μ_{ML}.

Consider the variance s^2. When it increases, the MAP goes to MLE, when in decreases, ,the MAP goes to prior mean. Prior variance quantify our confidence in the prior guess. Intuitively, the larger the prior variance, the less we trust the prior mean.

4.15 Sequential(recursive) updating of covariance matrix

Making use of:

$$\mathbf{m}_{n+1} = \frac{n\mathbf{m}_n + \mathbf{x}_{n+1}}{n+1}$$

What left is straightforward algebra.

4.16 Likelihood ratio for Gaussians

Consider a classifier for two classes, the generative distribution for them are two normal distributions$p(x|y = C_i) = N(x|\mu_i, \Sigma_i)$, by Bayes formula:

$$\log \frac{p(y = 1|x)}{p(y = 0|x)} = \log \frac{p(x|y = 1)}{p(x|y = 0)} + \log \frac{p(y = 1)}{p(y = 0)}$$

The second term is the ratio of likelihood probability.

When we have arbitrary covariance matrix:

$$\frac{p(x|y=1)}{p(x|y=0)} = \sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}} \exp\left\{-\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma_0^{-1}(x-\mu_0)\right\}$$

This can not be reduced further. However, it is noticable that the decision boundary is a quardric curve in D-dimension space.

When both covariance matrixes are given by Σ:

$$\frac{p(x|y=1)}{p(x|y=0)} = \exp\left\{x^{\mathrm{T}}\Sigma^{-1}(\mu_1-\mu_0) - \frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}(\mu_1\mu_1^{\mathrm{T}} - \mu_0\mu_0^{\mathrm{T}})\right\}\right\}$$

The decision boundary becomes a plate.

If we assume the covariance matrix to be a diagnoal matrix, the closed form of answer have a similiar look, with some matrix multiplation changed into inner product or arthimatic multiplation.

4.17 LDA/QDA on height/weight data*

Practise by youself.

4.18 Naive Bayes with mixed features

We now have:

$$\begin{cases} p(y=1) =& 0.5 \\ p(y=2) =& 0.25 \\ p(y=3) =& 0.25 \end{cases}$$

For **question a**:

$$\begin{aligned} p(y=1|x_1=0,x_2=0) =& \frac{p(y=1)p(x_1=0|y=1)p(x_2=0|y=1)}{\sum_{i=1}^{3}p(y=i)p(x_1=0|y=i)p(x_2=0|y=i)} \\ =& \frac{0.5\cdot 0.5\cdot \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\right\}}{0.5\cdot 0.5\cdot \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\right\} + 0.25\cdot 0.5\cdot \frac{1}{\sqrt{2\pi}}\exp\left\{0\right\} + 0.25\cdot 0.5\cdot \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\right\}} \\ =& 0.43 \end{aligned}$$

Consequently:

$$p(y=2|x_1=0,x_2=0) = 0.35$$

$$p(y=3|x_1=0,x_2=0)=0.22$$

For **question b**, we know the component x_1 provides no information w.r.t y, hence $p(y|x_1)$ reduces to the prior of y, i.e. $(0.5, 0.25, 0.25)$.

For **question c**, the answer is tantamount to that to **quesition a**. Because in **question a**, we have cancelling the dependence on component x_1.

4.19 Decision boundary for LDA with semi tied covariances

Omitting the shared parameters ends in:

$$p(y=1|\mathbf{x})=\frac{p(y=1)p(\mathbf{x}|y=1)}{p(y=0)p(\mathbf{x}|y=0)+p(y=1)p(\mathbf{x}|y=1)}$$

Consider a uniform prior, this can be reduced to:

$$\begin{aligned}\frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)+p(\mathbf{x}|y=1)}\\ &=\frac{1}{k^{\frac{D}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_0)^{\mathrm{T}}\Sigma_0^{-1}(\mathbf{x}-\mu_0)+\frac{1}{2}(\mathbf{x}-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{x}-\mu_1)\right\}+1}\\ &=\frac{1}{k^{\frac{D}{2}}\exp\left\{-\frac{1}{2}(1-\frac{1}{k})\mathbf{x}^{\mathrm{T}}\Sigma_0^{-1}\mathbf{x}+\mathbf{x}^{\mathrm{T}}\mathbf{u}+c\right\}+1}\end{aligned}$$

Where we have used:

$$|\Sigma_1|=|k\Sigma_0|=k^D|\Sigma_0|$$

The decision boundary is still a quardric curve. It reduces to a plate when $k=1$. When k increases, the decision boundary becomes a curve that surrenders μ_0. When k goes to infinity, the decision boundary degenerates to a $y=0$ curve, which implies that every space out of it belongs to a normal distribution with infinite variance.

4.20 Logistic regression vs LDA/QDA

We give a qualitative answer according to the argument ”overfitting arises from MLE, and is in a positive correlation with the complexity of the model(namely the number of independent parameters in the model)”.

GaussI assumes a covariance matrix propoetional to identity matrix;

GaussX has not prior assumption on covariance matrix;

LinLog assumes that different classes have the same covariance matrix;

QuadLog has not prior assumption on covariance matrix;

From the perspective of complexity:

QuadLog =GaussX $>$ LinLog $>$ GaussI

The accuracy of MLE follows the same order.

The argument in e is not true in general, a larger product does not necessarily imply a larger sum.

4.21 Gaussian decision boundaries*

[Need illustration here]

4.22 QDA with 3 classes*

This follows a straightforward calculation.

4.23 Scalar QDA*

This follows a straightforward calculation.

5 Bayesian statistics

5.1 Proof that a mixture of conjugate priors is indeed conjugate

For 5.69 and 5.70, formly:

$$p(\theta|D) = \sum_k p(\theta, k|D) = \sum_k p(k|D)p(\theta|k, D)$$

Where:

$$p(k|D) = \frac{p(k, D)}{p(D)} = \frac{p(k)p(D|k)}{\sum_{k'} p(k')p(D|k')}$$

5.2 Optimal threshold on classification probability

The posterior loss expectation is given by:

$$\begin{aligned}\rho(\hat{y}|x) = &\sum_y L(\hat{y}, y)p(y|x) = p_0 L(\hat{y}, 0) + p_1 L(\hat{y}, 1)\\ =&L(\hat{y}, 1) + p_0(L(\hat{y}, 0) - L(\hat{y}, 1))\end{aligned}$$

When two classficied result yield to the same loss:

$$\hat{p_0} = \frac{\lambda_{01}}{\lambda_{01} + \lambda_{10}}$$

Hence when $p_0 \geq \hat{p_0}$, we estimete $\hat{y} = 0$。

5.3 Reject option in classifiers

The posterior loss expectation is given by:

$$\rho(a|x) = \sum_c L(a, c)p(c|x)$$

Denote the class with max posterior confidence by $\hat{c}$:

$$\hat{c} = \arg\max_c \{p(c|x)\}$$

Now we have two applicable actions: $a = \hat{c}$ or $a =$ reject.
When $a = \hat{c}$, the posterior loss expectation is:

$$\rho_{\hat{c}} = (1 - p(\hat{c}|x)) \cdot \lambda_s$$

When reject, the posterior loss expectation is:

$$\rho_{\text{reject}} = \lambda_r$$

Thus the condition that we choose $a = \hat{c}$ instead of reject is:

$$\rho_{\hat{c}} \geq \rho_{\text{reject}}$$

Or:

$$p(\hat{c}|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

5.4 More reject options*

Straightforward calculation.

5.5 Newsvendor problem

By:

$$\mathbb{E}(\pi|Q) = P\int_0^Q Df(D)\mathrm{d}D - CQ\int_0^Q f(D)\mathrm{d}D + (P-C)Q\int_Q^{+\infty} f(D)\mathrm{d}D$$

We have:

$$\frac{\partial}{\partial Q}\mathbb{E}(\pi|Q) = PQf(Q) - C\int_0^Q f(D)\mathrm{d}D - CQf(Q) + (P-C)\int_Q^{+\infty} f(D)\mathrm{d}D - (P-C)Qf(Q)$$

Set it to zero by making use of $\int_0^Q f(D)fD + \int_Q^{+\infty} f(D)\mathrm{d}D = 1$:

$$\int_0^{Q^*} = F(Q^*) = \frac{P-C}{P}$$

5.6 Bayes factors and ROC curves*

Practise by yourself.

5.7 Bayes model averaging helps predictive accuracy

Expand both side of 5.127 and exchange the integral sequence:

$$\mathbb{E}[L(\Delta, p^{\text{BMA}})] = H(p^{\text{BMA}})$$

We also have:

$$\mathbb{E}[L(\Delta, p^m)] = \mathbb{E}_{p^{\text{BMA}}}[-\log(p^m)]$$

Substract the right side from the left side ends in:

$$-\mathbb{KL}(p^{\text{BMA}}||p^m) \leq 0$$

Hence the left side is always smaller than the right side.

5.8 MLE and model selection for a 2d discrete distribution

The joint distribution $p(x, y|\theta_1, \theta_2)$ is given by:

$$\begin{aligned}
p(x = 0, y = 0) =& (1 - \theta_1)\theta_2 \\
p(x = 0, y = 1) =& (1 - \theta_1)(1 - \theta_2) \\
p(x = 1, y = 0) =& \theta_1(1 - \theta_2) \\
p(x = 1, y = 1) =& \theta_1\theta_2
\end{aligned}$$

Which can be concluded as:

$$p(x, y|\theta_1, \theta_2) = \theta_1^x(1 - \theta_1)^{(1-x)}\theta_2^{\mathbb{I}(x=y)}(1 - \theta_2)^{(1-\mathbb{I}(x=y))}$$

The MLE is:

$$\theta_{\text{ML}} = \arg\max_\theta(\sum_{n=1}^N \ln p(x_n, y_n|\theta))$$

Hence:

$$\theta_{\text{ML}} = \arg\max_\theta(N \ln(\frac{1 - \theta_1}{1 - \theta_2}) + N_x \ln(\frac{\theta_1}{1 - \theta_1}) + N_{\mathbb{I}(x=y)} \ln(\frac{\theta_2}{1 - \theta_2}))$$

Two parameters can be estimated independently given $\mathbf{X}$ and $\mathbf{Y}$.

We can further rewrite the joint distribution into:

$$p(x, y|\theta) = \theta_{x,y}$$

Then

$$\theta_{\text{ML}} = \arg\max_\theta(\sum_{x,y} N_{x,y} \ln \theta_{x,y})$$

MLE can de done by using regularization condition.

The rest is straightforward algebra.

5.9 Posterior median is optimal estimate under L1 loss

The posterior loss expectation is(where we have omitted D w.l.o.g):

$$\begin{aligned}\rho(a) =& \int |y-a|p(y)\mathrm{d}y\\ =& \int_{-\infty}^{a}(a-y)p(y)\mathrm{d}y + \int_{a}^{+\infty}(y-a)p(y)\mathrm{d}y\\ =& a\cdot\left\{\int_{-\infty}^{a}p(y)\mathrm{d}y - \int_{a}^{+\infty}p(y)\mathrm{d}y\right\} - \int_{-\infty}^{a}yp(y)\mathrm{d}y + \int_{a}^{+\infty}yp(y)\mathrm{d}y\end{aligned}$$

Differentiate and we have:

$$\frac{\partial}{\partial a}\rho(a) = \left\{\int_{-\infty}^{a}p(y)\mathrm{d}y - \int_{a}^{+\infty}p(y)\right\} + 2a\cdot p(a) - 2a\cdot p(a)$$

Set it to zero and:

$$\int_{-\infty}^{a}p(y)\mathrm{d}y = \int_{a}^{+\infty}p(y) = \frac{1}{2}$$

5.10 Decision rule for trading off FPs and FNs

Given:

$$L_{\mathrm{FN}} = cL_{\mathrm{FP}}$$

The critical condition for 5.115 is:

$$\frac{p(y=1|x)}{p(y=2|x)} = c$$

Using:

$$p(y=1|x) + p(y=0|x) = 1$$

We get the threshold $\frac{c}{1+c}$.

6 Frequentist statistics*

The philosophy behind this chapter is out of the scope of probabilistic ML, you should be able to find solutions to the four listed problems in any decent textbook on mathematics statistics.

GL.

7 Linear regression

7.1 Behavior of training set error with increasing sample size

When the training set is small at the beginning, the trained model is over-fitted to the current data set, so the correct rate can be relatively high. As the training set increases, the model has to learn to adapt to more general-purpose parameters, thus reducing the overfitting effect laterally, resulting in lower accuracy.

As pointed out in Section 7.5.4, increasing the training set is an important method of countering over-fitting besides adding regulizer.

7.2 Multi-output linear regression

Straightforward calculation.

7.3 Centering and ridge regression

By rewriting $\mathbf{x}$ into $(\mathbf{x}^{\mathrm{T}}, 1)^{\mathrm{T}}$ to eliminate w_0, then NLL is given by:

$$\mathrm{NLL}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

So:

$$\frac{\partial}{\partial\mathbf{w}}\mathrm{NLL}(\mathbf{w}) = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\lambda\mathbf{w}$$

Therefore:

$$\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

7.4 MLE for σ^2 for linear regression

Firstly, we give the likelihood:

$$\begin{aligned}
p(D|\mathbf{w},\sigma^2) =& p(\mathbf{y}|\mathbf{w},\sigma^2,\mathbf{X}) \\
=& \prod_{n=1}^{N} p(y_n|\mathbf{x}_n,\mathbf{w},\sigma^2) \\
=& \prod_{n=1}^{N} N(y_n|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n,\sigma^2) \\
=& \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2\right\}
\end{aligned}$$

As for σ^2:

$$\frac{\partial}{\partial\sigma^2}\log p(D|\mathbf{w},\sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2$$

We have:

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2$$

7.5 MLE for the offset term in linear regression

NLL:

$$\mathrm{NLL}(\mathbf{w},w_0) \propto \sum_{n=1}^{N}(y_n-w_0-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2$$

Differentiate with two parameters:

$$\frac{\partial}{\partial w_0}\mathrm{NLL}(\mathbf{w},w_0) \propto -Nw_0 + \sum_{n=1}^{N}(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)$$

$$w_{0,\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) = \bar{y}-\mathbf{w}^{\mathrm{T}}\bar{\mathbf{x}}$$

Centering within $\mathbf{X}$ and $\mathbf{y}$:

$$\mathbf{X}_c = \mathbf{X} - \hat{\mathbf{X}}$$

$$\mathbf{y}_c = \mathbf{y} - \hat{\mathbf{y}}$$

The centered datasets have zero-mean, thus regression model have w_0 as zero, by the same time:

$$\mathbf{w}_{\mathrm{ML}} = (\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c)^{-1}\mathbf{X}_c^{\mathrm{T}}\mathbf{y}_c$$

7.6 MLE for simple linear regression*

Using the conclusion from problem 7.5. What left is straightforward algebra.

7.7 Sufficient statistics for online linear regression

Problem a and **Problem b**can be solved according to hints.

For **Problem c**, substituting the x in hint by y yields to the conclusion.

In d we are to prove:

$$(n+1)C_{xy}^{(n+1)} = nC_{xy}^{(n)} + x_{n+1}y_{n+1} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}$$

Expand the C_{xy} in two sides and use $\bar{x}^{(n+1)} = \bar{x}^{(n)} + \frac{1}{n+1}(x_{n+1} - \bar{x}^{n})$.

Problem e and **Problem f**: practice by yourself.

7.8 Bayesian linear regression in 1d with known σ^2

Problem a: practice by yourself.

For **Problem b**, choose the prior distribution:

$$p(\mathbf{w}) \propto N(w_1|0,1) \propto \exp\left\{-\frac{1}{2}w_1^2\right\}$$

Reduce it into:

$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0) \propto$$

$$\exp\left\{-\frac{1}{2}\mathbf{V}_{0,11}^{-1}(w_0 - w_{00})^2 - \frac{1}{2}\mathbf{V}_{0,22}^{-1}(w_1 - w_{01})^2 - \mathbf{V}_{0,12}^{-1}(w_0 - w_{00})(w_1 - w_{01})\right\}$$

Formly, we take:

$$\begin{aligned} w_{01} =& 0 \\ \mathbf{V}_{0,22}^{-1} =& 1 \\ \mathbf{V}_{0,11}^{-1} =& 0 \\ \mathbf{V}_{0,12}^{-1} =& 0 \\ w_{00} =& \text{arbitrary} \end{aligned}$$

In problem c, we consider the posterior distribution for parameters:

$$p(\mathbf{w}|D,\sigma^2) = N(\mathbf{w}|\mathbf{m}_0,\mathbf{V}_0)\prod_{n=1}^{N} N(y_n|w_0 + w_1 x_n, \sigma^2)$$

The coefficients for w_1^2 and w_1 in the exponential are:

$$-\frac{1}{2} - \frac{1}{2\sigma^2}\sum_{n=1}^{N} x_n^2$$

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n(w_0 - y)$$

Hence the posterior mean and variance are given by:

$$\sigma_{\text{post}}^2 = \frac{\sigma^2}{\sigma^2 + \sum_{n=1}^{N} x_n^2}$$

$$\mathbb{E}[w_1|D,\sigma^2] = \sigma_{\text{post}}^2(-\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n(w_0 - y))$$

It can be noticed that accumulation of samples reduces the posterior variance.

7.9 Generative model for linear regression

For sake of convinence, we consider a centered dataset(without changing symbols):

$$\begin{aligned} w_0 =& 0 \\ \mu_x =& 0 \\ \mu_y =& 0 \end{aligned}$$

By covariance's definition:

$$\Sigma_{XX} = X^{\mathrm{T}}X$$

$$\Sigma_{YX} = Y^{\mathrm{T}}X$$

Using the conclusion from section 4.3.1:

$$p(Y|X=x) = N(Y|\mu_{Y|X}, \Sigma_{Y|X})$$

Where:

$$\mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X) = Y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X = \mathbf{w}^{\mathrm{T}}X$$

7.10 Bayesian linear regression using the g-prior

Recall ridge regression model, where we have likelihood:

$$p(D|\mathbf{w},\sigma^2) = \prod_{n=1}^{N}\mathcal{N}(y_n|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n, \sigma^2)$$

The prior distribution is Gaussian-Inverse Gamma distribution:

$$\begin{aligned}
p(\mathbf{w},\sigma^2) =& \mathrm{NIG}(\mathbf{w},\sigma^2|\mathbf{w}_0,\mathbf{V}_0,a_0,b_0) \\
=& \mathcal{N}(\mathbf{w}|\mathbf{w}_0,\sigma^2\mathbf{V}_0)\mathrm{IG}(\sigma^2|a_0,b_0) \\
=& \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\sigma^2\mathbf{V}_0|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\mathrm{T}}(\sigma^2\mathbf{V}_0)^{-1}(\mathbf{w}-\mathbf{w}_0)\right\}\cdot \\
& \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}\exp\left\{-\frac{b_0}{\sigma^2}\right\} \\
=& \frac{b_0^{a_0}}{(2\pi)^{\frac{D}{2}}|\mathbf{V}_0|^{\frac{1}{2}}\Gamma(a_0)}(\sigma^2)^{-(a_0+\frac{D}{2}+1)}\cdot\exp\left\{-\frac{(\mathbf{w}-\mathbf{w}_0)^{\mathrm{T}}\mathbf{V}_0^{-1}(\mathbf{w}-\mathbf{w}_0)+2b_0}{2\sigma^2}\right\}
\end{aligned}$$

The posterior distribution takes the form:

$$\begin{aligned}
p(\mathbf{w},\sigma^2|D) \propto& p(\mathbf{w},\sigma^2)p(D|\mathbf{w},\sigma^2) \\
\propto& \frac{b_0^{a_0}}{(2\pi)^{\frac{D}{2}}|\mathbf{V}_0|^{\frac{1}{2}}\Gamma(a_0)}(\sigma^2)^{-(a_0+\frac{D}{2}+1)}\cdot \\
& \exp\left\{-\frac{(\mathbf{w}-\mathbf{w}_0)^{\mathrm{T}}\mathbf{V}_0^{-1}(\mathbf{w}-\mathbf{w}_0)+2b_0}{2\sigma^2}\right\}\cdot \\
& (\sigma^2)^{-\frac{N}{2}}\cdot\exp\left\{-\frac{\sum_{n=1}^{N}(y_n-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2}{2\sigma^2}\right\}
\end{aligned}$$

Comparing the coefficient of σ^2:

$$a_N = a_0 + \frac{N}{2}$$

Comparing the coefficient of $\mathbf{w}^{\mathrm{T}}\mathbf{w}$:

$$\mathbf{V}_N^{-1} = \mathbf{V}_0^{-1} + \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} = \mathbf{V}_0^{-1} + \mathbf{X}^{\mathrm{T}}\mathbf{X}$$

Comparing the coefficient of $\mathbf{w}$:

$$\mathbf{V}_N^{-1}\mathbf{w}_N = \mathbf{V}_0^{-1}\mathbf{w}_0 + \sum_{n=1}^{N} y_n \mathbf{x}_n$$

Thus:

$$\mathbf{w}_N = \mathbf{V}_N(\mathbf{V}_0^{-1}\mathbf{w}_0 + \mathbf{X}^{\mathrm{T}}\mathbf{y})$$

Finally, comparing the constant term inside the exponential:

$$b_N = b_0 + \frac{1}{2}(\mathbf{w}_0^{\mathrm{T}}\mathbf{V}_0^{-1}\mathbf{w}_0 + \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{w}_N^{\mathrm{T}}\mathbf{V}_N^{-1}\mathbf{w}_N)$$

We have obtained 7.70 to 7.73, which can be concluded into 7.69:

$$p(\mathbf{w}, \sigma^2 | D) = \mathrm{NIG}(\mathbf{w}, \sigma^2 | \mathbf{w}_N, \mathbf{V}_N, a_N, b_N)$$

8 Logistic regression

8.1 Spam classification using logistic regression*

Practice by yourself.

8.2 Spam classification using naive Bayes*

Practice by yourself.

8.3 Gradient and Hessian of log-likelihood for logistic regression

$$\frac{\partial}{\partial a}\sigma(a) = \frac{\exp(-a)}{(1+\exp(-a))^2} = \frac{1}{1+e^{-a}}\frac{e^{-a}}{1+e^{-a}} = \sigma(a)(1-\sigma(a))$$

$$\begin{aligned}
g(\mathbf{w}) =& \frac{\partial}{\partial \mathbf{w}} NLL(\mathbf{w}) \\
=& \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)] \\
=& \sum_{n=1}^{N} y_i \frac{1}{\sigma}\sigma(1-\sigma) - \mathbf{x}_i + (1-y_i)\frac{-1}{1-\sigma}\sigma(1-\sigma) - \mathbf{x}_i \\
=& \sum_{n=1}^{N} (\sigma(\mathbf{w}^T\mathbf{x}_i) - y_i)\mathbf{x}_i
\end{aligned}$$

For an arbitrary non-zero vector$\mathbf{u}$(with proper shape):

$$\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{S}\mathbf{X}\mathbf{u} = (\mathbf{X}\mathbf{u})^{\mathrm{T}}\mathbf{S}(\mathbf{X}\mathbf{u})$$

Since $\mathbf{S}$ is positive definite, for arbitrary non-zero $\mathbf{v}$:

$$\mathbf{v}^{\mathrm{T}}\mathbf{S}\mathbf{v} > 0$$

Assume $\mathbf{X}$ is a full-rank matrix, $\mathbf{X}\mathbf{u}$ is not zero, thus:

$$(\mathbf{X}\mathbf{u})^{\mathrm{T}}\mathbf{S}(\mathbf{X}\mathbf{u}) = \mathbf{u}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{S}\mathbf{X})\mathbf{u} > 0$$

So $\mathbf{X}^{\mathrm{T}}\mathbf{S}\mathbf{X}$ is positive definite.

8.4 Gradient and Hessian of log-likelihood for multinomial logistic regression

By considering one independent component each time, the complexity in form caused by tensor product is reduced. For a specific $\mathbf{w}^*$:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}^*}\mathrm{NLL}(\mathbf{W}) &= -\sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}^*}[y_{n*}\mathbf{w}^{*\mathrm{T}}\mathbf{x}_n - \log(\sum_{c=1}^{C}\exp(\mathbf{w}_c^{\mathrm{T}}\mathbf{x}_n))]\\ &= \sum_{n=1}^{N} -y_{n*}\mathbf{x}_n + \frac{\exp(\mathbf{w}^{*\mathrm{T}}\mathbf{x}_n)}{\sum_{c=1}^{C}\exp(\mathbf{w}_c^{T}\mathbf{x}_n)}\mathbf{x}_n\\ &= \sum_{n=1}^{N}(\mu_{n*} - y_{n*})\mathbf{x}_n\end{aligned}$$

Combine the independent solutions for all classes into one matrix yield 8.38.

On soloving for Hessian matrix, consider to take gradient w.r.t $\mathbf{w}_1$ and $\mathbf{w}_2$:

$$\mathbf{H}_{1,2} = \nabla_{\mathbf{w}_2}\nabla_{\mathbf{w}_1}NLL(\mathbf{W}) = \frac{\partial}{\partial \mathbf{w}_2}\sum_{n=1}^{N}(\mu_{n1} - y_{n1})\mathbf{x}_n$$

When $\mathbf{w}_1$ and $\mathbf{w}_2$ are the same vector:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}_1}\sum_{n=1}^{N}(\mu_{n1} - y_{n1})\mathbf{x}_n^{\mathrm{T}} &= \sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}_1}\mu_{n1}\mathbf{x}_n^{\mathrm{T}}\\ &= \sum_{n=1}^{N}\frac{\exp(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_n)(\sum\exp)\mathbf{x}_n - \exp(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_n)^2\mathbf{x}_n}{(\sum\exp)^2}\mathbf{x}_n^{\mathrm{T}}\\ &= \sum_{n=1}^{N}\mu_{n1}(1 - \mu_{n1})\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\end{aligned}$$

When $\mathbf{w}_1$ and $\mathbf{w}_2$ are different:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}_2}\sum_{n=1}^{N}\mu_{n1}\mathbf{x}_n^{\mathrm{T}} &= \sum_{n=1}^{N}\frac{-\exp(\mathbf{w}_2^{\mathrm{T}}\mathbf{x}_n)\exp(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}_n)\mathbf{x}_n}{(\sum\exp)^2}\mathbf{x}_n^{\mathrm{T}}\\ &= \sum_{n=1}^{N} -\mu_{n1}\mu_{n2}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\end{aligned}$$

Ends in 8.44。

The condition $\sum_c y_{nc} = 1$ is used from 8.34 to 8.35.

8.5 Symmetric version of l2 regularized multinomial logistic regression

Adding a regularizer equals doing a posterior estimationg, which equals introducing a languarge multipler for a new constraint. In this problem a Gaussian prior distribution with a homogeneous diagonal matrix is introduced, this leads to the constraint $w_{cj} = 0$.

At optima, the gradient in 8.47 goes to zero. Assume that $\hat{\mu}_{cj} = y_{cj}$, then $g(\mathbf{W}) = 0$. The extra regularization is $\lambda \sum_{c=1}^{C} \mathbf{w}_c = 0$, which equals D independent linear constraints, with form of: for $j = 1...D$, $\sum_{c=1}^{C} \hat{w}_{cj} = 0$.

8.6 Elementary properties of l2 regularized logistic regression

The first term of $J(\mathbf{w})$'s Hessian is positive definite(8.7), the second term's Hessian is positive definite as well($\lambda > 0$). Therefore this function has a positive definite Hessian, it has a global optimum.

The form of posterior distribution takes:

$$\begin{aligned}
p(\mathbf{w}|D) \propto & p(D|\mathbf{w})p(\mathbf{w}) \\
p(\mathbf{w}) =& N(\mathbf{w}|\mathbf{0}, \sigma^{-2}\mathbf{I}) \\
\mathrm{NLL}(\mathbf{w}) =& -\log p(\mathbf{w}|D) \\
=& -\log p(D|\mathbf{w}) + \frac{1}{2\sigma^2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + c
\end{aligned}$$

Therefore:

$$\lambda = \frac{1}{2\sigma^2}$$

The number of zero in global optimun is related to the value of λ, which is in a negative correlationship with the prior uncertainty of $\mathbf{w}$. The less the uncertainty is, the more that $\mathbf{w}$ converges to zero, which ends in more zeros in answer.

If $\lambda = 0$, which implies prior uncertainty goes to infnity. Then posterior estimation converges to MLE. As long as there is no constraint on $\mathbf{w}$, it is possible that some component of $\mathbf{w}$ goes to infinity.

When λ increase, the prior uncertainty reduces, hence the over-fitting effect reduces. Generally this implide a decrease on training-set accuracy.

At the same time, this also increases the accuracy of model on test-set, but it does not always happen.

8.7 Regularizing separate terms in 2d logistic regression*

Practice by yourself.

9 Generalized linear models and the exponential family

9.1 Conjugate prior for univariate Gaussian in exponential family form

The 1d Gaussian distribution is:

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Rewrite it into:

$$p(x|\mu,\sigma^2)=\exp\left\{-\frac{1}{2\sigma^2}x^2+\frac{1}{\sigma^2}x-\left\{\frac{\mu^2}{2\sigma^2}+\frac{\ln(2\pi\sigma^2)}{2}\right\}\right\}$$

Denote $\theta=(-\frac{\lambda}{2},\lambda\mu)^{\mathrm{T}}$, $A(\theta)=\frac{\lambda\mu^2}{2}+\frac{\ln(2\pi)}{2}-\frac{\ln\lambda}{2}$, $\phi(x)=(x^2,x)^{\mathrm{T}}$. Consider the likelihood with dataset$D$:

$$\log p(D|\theta)=\exp\left\{\theta^{\mathrm{T}}(\sum_{n=1}^{N}\phi(x_n))-N\cdot A(\theta)\right\}$$

According to the meaning of prior distribution, we set a observation background in order to define a prior distribution. The sufficient statistics is the only thing matters by the form of exponential family. Assume that we have M prior observations. The mean of them and their square are v_1 and v_2 respectively, then the prior distribution takes the form:

$$\begin{aligned}p(\theta|M,v_1,v_2)=&\exp\left\{\theta_1\cdot Mv_1+\theta_2\cdot Mv_2-M\cdot A(\theta)\right\}\\=&\exp\left\{-\frac{\lambda}{2}Mv_1+\lambda\mu Mv_2-\frac{M}{2}\lambda\mu^2-\frac{M}{2}\ln 2\pi+\frac{M}{2}\ln\lambda\right\}\end{aligned}$$

It has three independent parameters. We are to prove that is equals $p(\mu,\lambda)=\mathcal{N}(\mu|\gamma,\frac{1}{\lambda(2\alpha-1)})\mathrm{Gamma}(\lambda|\alpha,\beta)$. Expand it into exponential form and ignore the terms independent with μ,λ:

$$\begin{aligned}p(\mu,\lambda)=&\exp\left\{(\alpha-1)\ln\lambda-\beta\lambda-\frac{\lambda(2\alpha-1)}{2}\mu^2-\frac{\lambda(2\alpha-1)}{2}\gamma^2\right\}\\&\cdot\exp\left\{\lambda(2\alpha-1)\mu\gamma+\frac{1}{2}\ln\lambda\right\}\end{aligned}$$

Compare the coefficients for $\lambda\mu^2, \lambda\mu, \lambda, \ln\lambda$, we obtain:

$$\begin{aligned}
-\frac{(2\alpha-1)}{2} =& -\frac{M}{2} \\
\gamma(2\alpha-1) =& M v_2 \\
\frac{(2\alpha-1)}{2}\gamma^2 - \beta =& -\frac{1}{2} M v_1 \\
(\alpha-1) + \frac{1}{2} =& \frac{M}{2}
\end{aligned}$$

Combining them ends in:

$$\begin{aligned}
\alpha =& \frac{M+1}{2} \\
\beta =& \frac{M}{2}(v_2^2 + v_1) \\
\gamma =& v_2
\end{aligned}$$

Thus two distributions are equal with naive change of variables' names.

9.2 The MVN is in the exponential family

Here you can find a comprehensive solution:
https://stats.stackexchange.com/questions/231714/sufficient-statistic-for-multivari

10 Directed graphical models(Bayes nets)

...

11 Mixture models and the EM algorithm

11.1 Student T as infinite mixture of Gaussian

The 1d Student-t distribution takes the form:

$$\text{St}(x|\mu,\sigma^2,v) = \frac{\Gamma(\frac{v}{2}+\frac{1}{2})}{\Gamma(\frac{v}{2})}(\frac{1}{\pi v\sigma^2})^{\frac{1}{2}}(1+\frac{(x-\mu)^2}{v\sigma^2})^{-\frac{v+1}{2}}$$

Consider the left side of 11.61:

$$\begin{aligned}
&\int N(x|\mu,\frac{\sigma^2}{z})\text{Gamma}(z|\frac{v}{2},\frac{v}{2})\text{d}z \\
=&\int \frac{\sqrt{z}}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{z}{2\sigma^2}(x-\mu)^2\right\}\frac{(\frac{v}{2})^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}z^{\frac{v}{2}-1}\exp\left\{-\frac{v}{2}z\right\}\text{d}z \\
=&\frac{1}{\sqrt{2\pi\sigma^2}}\frac{(\frac{v}{2})^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\int z^{\frac{v-1}{2}}\exp\left\{-(\frac{v}{2}+\frac{(x-\mu)^2}{2\sigma^2})z\right\}\text{d}z
\end{aligned}$$

The integrated function is the terms related to z in Gamma distribution $\text{Gamma}(z|\frac{v+1}{2},\frac{(x-\mu)^2}{2\sigma^2}+\frac{v}{2})$, which gives to the normalized term's inverse.

$$\int z^{\frac{v-1}{2}}\exp\left\{-(\frac{v}{2}+\frac{(x-\mu)^2}{\sigma^2})z\right\}\text{d}z = \Gamma(\frac{v+1}{2})(\frac{(x-\mu)^2}{2\sigma^2}+\frac{v}{2})^{-\frac{v+1}{2}}$$

Plug in can derive 11.61.

11.2 EM for mixture of Gaussians

We are to optimize:

$$\begin{aligned}
Q(\theta,\theta^{old}) =&\mathbb{E}_{p(z|D,\theta^{\text{old}})}[\sum_{n=1}^{N}\log(\mathbf{x}_n,\mathbf{z}_n|\theta)] \\
=&\sum_{n=1}^{N}\mathbb{E}[\log\prod_{k=1}^{K}(\pi_k p(\mathbf{x}_n|z_k,\theta))^{z_{nk}}] \\
=&\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\log(\pi_k p(\mathbf{x}_n|z_k,\theta))
\end{aligned}$$

Where:

$$r_{nk} = p(z_{nk}=1|\mathbf{x}_n,\theta^{\text{old}})$$

When the emission distribution $p(\mathbf{x}|z,\theta)$ is Gaussian, consider the terms involveμ_k in $Q(\theta,\theta^{\text{old}})$ first:

$$\sum_{n=1}^{N} r_{nk} \log p(\mathbf{x}_n|z_k,\theta) = \sum_{n=1}^{N} r_{nk}(-\frac{1}{2})(\mathbf{x}_n-\mu_k)^{\text{T}}\Sigma^{-1}(\mathbf{x}_n-\mu_k)+C$$

Setting the derivative to zero results in:

$$\sum_{n=1}^{N} r_{nk}(\mu_k-\mathbf{x}_n)=0$$

And we obtain 11.31:

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

For terms involve Σ_k in $Q(\theta,\theta^{\text{old}})$:

$$\sum_{n=1}^{N} r_{nk} \log p(\mathbf{x}_n|z_k,\theta) = \sum_{n=1}^{N} r_{nk}(-\frac{1}{2})(\log|\Sigma_k|+(\mathbf{x}_n-\mu_k)^{\text{T}}\Sigma^{-1}(\mathbf{x}_n-\mu_k))+C$$

Using the same way as in 4.1.3.1:

$$L(\Sigma^{-1}=\Lambda) = (\sum_{n=1}^{N} r_{nk})\log|\Lambda| - \text{tr}\left\{(\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^{\text{T}})\Lambda\right\}$$

The balance condition is:

$$(\sum_{n=1}^{N} r_{nk})\Lambda^{-\text{T}} = \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^{\text{T}}$$

Obtain 11.32:

$$\Sigma_k = \frac{\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^{\text{T}}}{\sum_{n=1}^{N} r_{nk}}$$

11.3 EM for mixtures of Bernoullis

During the MLE for mixtures of Bernoullis, consider($D=2$ marks the number of potential elements):

$$\begin{aligned}\frac{\partial}{\partial\mu_{kj}}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\log p(\mathbf{x}_n|\theta,k) &= \sum_{n=1}^{N} r_{nk}\frac{\partial}{\partial\mu_{kj}}(\sum_{i}^{D} x_{ni}\log\mu_{ki}) \\ &= \sum_{n=1}^{N} r_{nk}x_{nj}\frac{1}{\mu_{kj}}\end{aligned}$$

Introduce a multipler to constrain $\sum_j \mu_{kj} = 1$, then condition for the derivative to be zero is:

$$\mu_{kj} = \frac{\sum_{n=1}^{N} r_{nk} x_{nj}}{\lambda}$$

Summer over all j:

$$1 = \sum_{j=1}^{D} \mu_{kj} = \frac{1}{\lambda} \sum_{j=1}^{D} \sum_{n=1}^{N} r_{nk} x_{nj} = \frac{1}{\lambda} \sum_{n=1}^{N} r_{nk} \sum_{j=1}^{D} x_{nj} = \frac{\sum_{n=1}^{N} r_{nk}}{\lambda}$$

Results in:

$$\lambda = \sum_{n=1}^{N} r_{nk}$$

Hence 11.116。

Introduce a prior:

$$p(\mu_{k0}) \propto \mu_{k0}^{\alpha-1} \mu_{k1}^{\beta-1}$$

The zero-derivative condition becomes:

$$\mu_{k0} = \frac{\sum_{n=1}^{N} r_{nk} x_{n0} + \alpha - 1}{\lambda}$$

$$\mu_{k1} = \frac{\sum_{n=1}^{N} r_{nk} x_{n1} + \beta - 1}{\lambda}$$

And:

$$1 = \mu_{k0} + \mu_{k1} = \frac{1}{\lambda}(\sum_{n=1}^{N} r_{nk}(x_{n0} + x_{n1}) + \alpha + \beta - 2)$$

$$\lambda = \sum_{n=1}^{N} r_{nk} + \alpha + \beta - 2$$

Hence 11.117。

11.4 EM for mixture of Student distributions

The log-likelihood for complete data set is:

$$\begin{aligned} l_c(\mathbf{x}, z) =& \log(\mathcal{N}(\mathbf{x}|\mu, \frac{\Sigma}{z})\mathrm{Gamma}(z|\frac{\lambda}{2}, \frac{\lambda}{2})) \\ =& -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| + \frac{D}{2}\log(z) \\ & -\frac{z}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu) \\ & +\frac{v}{2}\log(\frac{v}{2}) - \log(\Gamma(\frac{v}{2})) + (\frac{v}{2}-1)\log(z) - \frac{v}{2}z \end{aligned}$$

Sum the terms involving v:

$$l_v(\mathbf{x}, z) = \frac{v}{2}\log(\frac{v}{2}) - \log(\Gamma(\frac{v}{2})) + \frac{v}{2}(\log(z) - z)$$

The likelihood w.r.t v on complete data set is:

$$L_v = \frac{vN}{2}\log(\frac{v}{2}) - N\log(\Gamma(\frac{v}{2})) + \frac{v}{2}\sum_{n=1}^{N}(\log(z_n) - z_n)$$

Setting derivative to zero gives:

$$\frac{\nabla\Gamma(\frac{v}{2})}{\Gamma(\frac{v}{2})} - 1 - \log(\frac{v}{2}) = \frac{\sum_{n=1}^{N}(\log(z_n) - z_n)}{N}$$

For μ and Σ:

$$l_{\mu,\Sigma}(\mathbf{x}, z) = -\frac{1}{2}\log|\Sigma| - \frac{z}{2}(\mathbf{x} - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \mu)$$

$$L_{\mu,\Sigma} = \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{n=1}^{N} z_n(\mathbf{x}_n - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n - \mu)$$

Hence equals the MLE used for MVN.

11.5 Gradient descent for fitting GMM

From the given information:

$$p(\mathbf{x}|\theta) = \sum_k \pi_k N(\mathbf{x}|\mu_k, \Sigma_k)$$

$$l(\theta) = \sum_{n=1}^{N}\log p(\mathbf{x}_n|\theta)$$

Deriavte w.r.t μ_k:

$$\begin{aligned}\frac{\partial}{\partial\mu_k}l(\theta) =& \sum_{n=1}^{N}\frac{\pi_k N(\mathbf{x}_n|\mu_k, \Sigma_k)\nabla_{\mu_k}\left\{-\frac{1}{2}(\mathbf{x}_n - \mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}_n - \mu_k)\right\}}{\sum_{k'=1}^{K}\pi_{k'}N(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})} \\ =& \sum_{n=1}^{N} r_{nk}\Sigma_k^{-1}(\mathbf{x}_n - \mu_k)\end{aligned}$$

w.r.t π_k:

$$\frac{\partial}{\partial\pi_k}l(\theta) = \sum_{n=1}^{N}\frac{N(\mathbf{x}_n|\mu_k, \Sigma^k)}{\sum_{k'=1}^{K}\pi_{k'}N(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})} = \frac{1}{\pi_k}\sum_{n=1}^{N} r_{nk}$$

Using Languarge multipler ends in:

$$\pi_k = \frac{\sum_{n=1}^{N} r_{nk}}{\lambda}$$

Sum over k and normalize:

$$\pi_k = \frac{\sum_{n=1}^{N} r_{nk}}{N}$$

For Σ_k:

$$\frac{\partial}{\partial \Sigma_k} l(\theta) = \sum_{n=1}^{N} \frac{\pi_k \nabla_{\Sigma_k} \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\mu_{k'}, \Sigma_{k'})}$$

Where:

$$\begin{aligned}\nabla_{\Sigma_k} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) =& \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{x}-\mu_k)\right\} \nabla_{\Sigma_k} \\ &\cdot \left\{\nabla_{\Sigma_k}(-\frac{1}{2}(\mathbf{x}-\mu_k)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{x}-\mu_k)) - \Sigma_k^{-1} \nabla_{\Sigma_k} |\Sigma_k|\right\} \\ =& \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \nabla(\log \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k))\end{aligned}$$

Thus we have:

$$\Sigma_k = \frac{\sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^{\mathrm{T}}}{\sum_{n=1}^{N} r_{nk}}$$

11.6 EM for a finite scale mixture of Gaussians

J and K are independent, using Bayes' rules(we have omitted θ in condition w.l.o.g):

$$\begin{aligned}p(J_n = j, K_n = k|x_n) =& \frac{p(J_n = j, K_n = k, x_n)}{p(x_n)} \\ =& \frac{p(J_n = j)p(K_n = k)p(x_n|J_n = j, K_n = k)}{\sum_{J_n, K_n} p(J_n, K_n, x_n)} \\ =& \frac{p_j q_k N(x_n|\mu_j, \sigma_k^2)}{\sum_{J_n=1}^{m} \sum_{K_n=1}^{l} p_{J_n} q_{K_n} \mathcal{N}(x_n|\mu_{J_n}, \sigma_{K_n}^2)}\end{aligned}$$

Derive the form of auxiliary fucntion $Q(\theta^{\text{new}}, \theta^{\text{old}})$:

$$\begin{aligned}
Q(\theta^{\text{new}}, \theta^{\text{old}}) =& \mathbb{E}_{\theta^{\text{old}}} \sum_{n=1}^{N} \log p(x_n, J_n, K_n|\theta^{\text{new}}) \\
=& \sum_{n=1}^{N} \mathbb{E}[\log(\prod_{j=1}^{m}\prod_{k=1}^{l} p(x_n, J_n, K_n|\theta^{\text{new}})^{\mathbb{I}(J_n=j,K_n=k)})] \\
=& \sum_{n=1}^{N}\sum_{j=1}^{m}\sum_{k=1}^{l} \mathbb{E}(\mathbb{I}(J_n = j, K_n = k))(\log p_j + \log q_k + \log \mathcal{N}(x_n|\mu_j, \sigma_k^2)) \\
=& \sum_{n,j,k} r_{njk} \log p_j + \sum_{n,j,k} r_{njk} \log q_k + \sum_{njk} r_{njk} \log \mathcal{N}(x_n|\mu_j, \sigma_k^2)
\end{aligned}$$

We are to optimize parameters p, q, μ, σ^2. It is noticealbe that p and q can be optimized independently. Now fix σ^2 and optimize μ:

$$\begin{aligned}
\frac{\partial}{\partial \mu_j} \sum_{n,j',k} r_{nj'k} \mathcal{N}(x_n|\mu_j, \sigma_k^2) =& \sum_{n,k} r_{njk} \nabla_{\mu_k} \mathcal{N}(x_n|\mu_j, \sigma_k^2) \\
=& \sum_{n,k} r_{njk} \mathcal{N}(x_n|\mu_j, \sigma_k^2) \frac{x_n - \mu_j}{\sigma_k^2}
\end{aligned}$$

And we ends in:

$$\mu_j = \frac{\sum_{n,k} r_{njk} \mathcal{N}(x_n|\mu_j, \sigma_k^2) \frac{x_n}{\sigma_k^2}}{\sum_{n,k} r_{njk} \mathcal{N}(x_n|\mu_j, \sigma_k^2) \frac{1}{\sigma_k^2}}$$

11.7 Manual calculation of the M step for a GMM*

Practise by yourself.

11.8 Moments of a mixture of Gaussians

For the expectation of mixture distribution:

$$\begin{aligned}
\mathbb{E}(\mathbf{x}) =& \int \mathbf{x} \sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \mathrm{d}\mathbf{x} \\
=& \sum_k \pi_k \left\{ \int \mathbf{x} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \mathrm{d}\mathbf{x} \right\} \\
=& \sum_k \pi_k \mu_k
\end{aligned}$$

Using $\text{cov}(\mathbf{x}) = \mathbb{E}(\mathbf{x}\mathbf{x}^{\text{T}}) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^{\text{T}}$, we have:

$$\begin{aligned}\mathbb{E}(\mathbf{x}\mathbf{x}^{\text{T}}) =& \int \mathbf{x}\mathbf{x}^{\text{T}} \sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)\text{d}\mathbf{x} \\ =& \sum_k \pi_k \int \mathbf{x}\mathbf{x}^{\text{T}} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)\text{d}\mathbf{x}\end{aligned}$$

Where:

$$\begin{aligned}\int \mathbf{x}\mathbf{x}^{\mathcal{T}} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)\text{d}\mathbf{x} =& \mathbb{E}_{\mathcal{N}(\mu_k,\Sigma_k)}(\mathbf{x}\mathbf{x}^{\text{T}}) \\ =& \text{cov}_{\mathcal{N}(\mu_k,\Sigma_k)}(\mathbf{x}) + \mathbb{E}_{\mathcal{N}(\mu_k,\Sigma_k)}(\mathbf{x})\mathbb{E}_{\mathcal{N}(\mu_k,\Sigma_k)}(\mathbf{x})^{\mathcal{T}} \\ =& \Sigma_k + \mu_k\mu_k^{\mathcal{T}}\end{aligned}$$

Therefore:

$$\text{cov}(\mathbf{x}) = \sum_k \pi_k(\Sigma_k + \mu_k\mu_k^{\text{T}}) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^{\text{T}}$$

11.9 K-means clustering by hand*

Practise by yourself.

11.10 Deriving the K-means cost function

For every term sum over k, apply 11.134 onto the inner and outer sum process:

$$\begin{aligned}\sum_{i:z_i=k}\sum_{i':z_{i'}=k}(x_i - x_{i'})^2 =& \sum_{i:z_i=k} n_k s^2 + n_k(\bar{x}_k - x_i)^2 \\ =& n_k^2 s^2 + n_k(n_k s^2) \\ =& 2n_k s_k\end{aligned}$$

The right side of 11.131 equals to sum over k:

$$n_k \sum_{i:z_i=k}(x_i - \bar{x}_k)^2 = n_k(n_k s^2 + n(\hat{x}_n - \hat{x}_n))$$

Thus 11.131.

11.11 Visible mixtures of Gaussians are in exponential family

Encode latent variable with hot-pot code, $z_c = \mathbb{I}(x$ is generated from the c distribution), then(omit θ in condition w.l.o.g):

$$\begin{aligned}p(\mathbf{z}) =& \prod_{c=1}^{C} \pi_c^{z_c} \\ p(x|\mathbf{z}) =& \prod_{c=1}^{C} (\frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left\{ -\frac{1}{2\sigma_c^2}(x-\mu_c)^2 \right\})^{z_c}\end{aligned}$$

The log for joint distribution is:

$$\begin{aligned}\log p(x, \mathbf{z}) =& \log \prod_{c=1}^{C} (\frac{\pi_c}{\sqrt{2\pi\sigma_c^2}} \exp\left\{ -\frac{1}{2\sigma_c^2}(x-\mu_c)^2 \right\})^{z_c} \\ =& \sum_{c=1}^{C} z_c(\log \pi_c - \frac{1}{2}\log 2\pi\sigma_c^2 - \frac{1}{2\sigma_c^2}(x-\mu_c)^2)\end{aligned}$$

Which is a sum of some inner products, hence an exponential family.The sufficient statics are linear combinations of $\mathbf{z}$, $\mathbf{z}x$ and $\mathbf{z}x^2$.

11.12 EM for robust linear regression with a Student t likelihood

Using the complete data likelihood w.r.t μ derived in 11.4.5:

$$L_N(\mu) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} z_n(y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2$$

Set the deriavte to zero:

$$\mathbf{w}^{\mathrm{T}} \sum_{n=1}^{N} z_n \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} = \sum_{n=1}^{N} z_n y_n \mathbf{x}_n^{\mathrm{T}}$$

This means:

$$\mathbf{w}^{\mathrm{T}} = (\sum_{n=1}^{N} z_n y_n \mathbf{x}_n^{\mathrm{T}})(\sum_{n=1}^{N} z_n \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}})^{-1}$$

11.13 EM for EB estimation of Gaussian shrinkage model

For every j, 5.90 takes different forms(this equals E-step):

$$p(\bar{x}_i|\mu, t^2, \sigma^2) = \mathcal{N}(\bar{x}_j|\mu, t^2 + \sigma_j^2)$$

Integrate out θ_j, the marginal likelihood is given by:

$$\log \prod_{j=1}^{D} N(\bar{x}_j|\mu, t^2 + \sigma_j^2) = (-\frac{1}{2}) \sum_{j=1}^{D} \log 2\pi(t^2 + \sigma_j^2) + \frac{1}{t^2 + \sigma_j^2}(\bar{x}_j - \mu)^2$$

Then we optimize respectively(this equals M-step):

$$\mu = \frac{\sum_{j=1}^{D} \frac{\bar{x}_j}{t^2+\sigma_j^2}}{\sum_{j=1}^{D} \frac{1}{t^2+\sigma_j^2}}$$

t^2 satisfies:

$$\sum_{j=1}^{D} \frac{(t^2 + \sigma^2) - (\bar{x}_j - \mu)^2}{(t^2 + \sigma_j^2)^2}$$

11.14 EM for censored linear regression*

Unsolved.

11.15 Posterior mean and variance of a truncated Gaussian

We denote $A = \frac{c_i - \mu_i}{\sigma}$, for mean:

$$\mathbb{E}[z_i|z_i \geq c_i] = \mu_i + \sigma\mathbb{E}[\epsilon_i|\epsilon_i \geq A]$$

And we have:

$$\mathbb{E}[\epsilon_i|\epsilon_i = \frac{1}{p(\epsilon_i \geq A)} \int_A^{+\infty} \epsilon_i \mathcal{N}(\epsilon_i|0, 1)\mathrm{d}x = \frac{\phi(A)}{1 - \Phi(A)} = H(A)$$

In the last step we use 11.141 and 11.139, plug it up:

$$\mathbb{E}[z_i|z_i \geq c_i] = \mu_i + \sigma H(A)$$

Now to calculate the expectation for square term:

$$\mathbb{E}[z_i^2|z_i \geq c_i] = \mu_i^2 + 2\mu_i\sigma\mathbb{E}[\epsilon_i|\epsilon_i \geq A] + \sigma^2\mathbb{E}[\epsilon_i^2|\epsilon_i \geq A]$$

To address $\mathbb{E}[\epsilon_i^2|\epsilon_i \geq A]$, expand the hint from question:

$$\frac{\mathrm{d}}{\mathrm{d}w}(wN(w|0,1)) = \mathcal{N}(w|0,1) - w^2\mathcal{N}(w|0,1)$$

We have:

$$\int_b^c w^2\mathcal{N}(w|0,1)\mathrm{d}w = \Phi(c) - \Phi(b) - c \cdot \mathcal{N}(c|0,1) + b \cdot \mathcal{N}(b|0,1)$$

$$\mathbb{E}[\epsilon_i^2|\epsilon_i \geq A] = \frac{1}{p(\epsilon_i \geq A)} \int_A^{+\infty} w^2\mathcal{N}(w|0,1)\mathrm{d}w = \frac{1 - \Phi(A) + A\phi(A)}{1 - \Phi(A)}$$

Plug it into the conclusion drawn from question a:

$$\begin{aligned}\mathbb{E}[z_i^2|z_i \geq c_i] &= \mu_i^2 + 2\mu_i\sigma H(A) + \sigma^2\frac{1 - \Phi(A) + A\phi(A)}{1 - \Phi(A)} \\ &= \mu_i^2 + \sigma^2 + H(A)(\sigma c_i + \sigma\mu_i)\end{aligned}$$

12 Latent linear models

12.1 M-step for FA

Review the EM for FA(Fator-Analysis) first. Basically, we have(centralize $\mathbf{X}$ to cancel μ w.l.o.g):

$$\begin{aligned} p(\mathbf{z}) =&\mathcal{N}(\mathbf{z}|0, I) \\ p(\mathbf{x}|\mathbf{z}) =&\mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z}, \Psi) \end{aligned}$$

And:

$$\begin{aligned} p(\mathbf{z}|\mathbf{x}) =&\mathcal{N}(\mathbf{z}|\mathbf{m}, \Sigma) \\ \Sigma =&(I + \mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W})^{-1} \\ \mathbf{m} =&\Sigma\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n \end{aligned}$$

Denote $\mathbf{x}_n$'s latent variable as $\mathbf{z}_n$. The log-likelihood for complete data set$\{\mathbf{x}, \mathbf{z}\}$ is:

$$\log \prod_{n=1}^{N} p(\mathbf{x}_n, \mathbf{z}_n) = \sum_{n=1}^{N} \log p(\mathbf{z}_n) + \log p(\mathbf{x}_n|\mathbf{z}_n)$$

With prior $\log p(\mathbf{z})$ that can be omitted with parameter 0 and $\mathbf{I}$, hence:

$$\begin{aligned} Q(\theta, \theta^{\text{old}}) =&\mathbb{E}_{\theta^{\text{old}}}[\sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{z}_n, \theta)] \\ =&\mathbb{E}[\sum_{n=1}^{N} c - \frac{1}{2}\log|\Psi| - \frac{1}{2}(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)^{\mathrm{T}}\Psi^{-1}(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)] \\ =&C - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)^{\mathrm{T}}\Psi^{-1}(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)] \\ =&C - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W}\mathbf{z}_n] + \sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\Psi^{-1}\mathbf{W}\mathbb{E}[\mathbf{z}_n] \\ =&C - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\Psi^{-1}\mathbf{x}_n - \frac{1}{2}\sum_{n=1}^{N}\operatorname{tr}\left\{\mathbf{W}^{\mathrm{T}}\Psi^{-1}\mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right\} + \sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\Psi^{-1}\mathbf{W}\mathbb{E}[\mathbf{z}_n] \end{aligned}$$

As long as $p(\mathbf{z}|\mathbf{x},\theta^{\text{old}}) = \mathcal{N}(\mathbf{z}|\mathbf{m},\Sigma)$, we have:

$$\begin{aligned}\mathbb{E}[\mathbf{z}_n|\mathbf{x}_n] =&\Sigma\mathbf{W}^{\text{T}}\Psi^{-1}\mathbf{x}\\ \mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T|\mathbf{x}_n] =&\text{cov}(\mathbf{z}_n|\mathbf{x}_n)+\mathbb{E}[\mathbf{z}_n|\mathbf{x}_n]\mathbb{E}[\mathbf{z}_n|\mathbf{x}_n]^{\text{T}}\\ =&\Sigma+(\Sigma\mathbf{W}^{\text{T}}\Psi^{-1}\mathbf{x})(\Sigma\mathbf{W}^{\text{T}}\Psi^{-1}\mathbf{x})^{\text{T}}\end{aligned}$$

From now on, the $\mathbf{x}$ and θ^{old} are omitted from conditions when calculating expectation.

Optimize w.r.t $\mathbf{W}$:

$$\frac{\partial}{\partial\mathbf{W}}Q = \sum_{n=1}^{N}\Psi^{-1}\mathbf{x}_n\mathbb{E}[\mathbf{z}_n]^{\text{T}} - \sum_{n=1}^{N}\Psi^{-1}\mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\text{T}}]$$

Set it to zero:

$$\mathbf{W} = (\sum_{n=1}^{N}\mathbf{x}_n\mathbb{E}[\mathbf{z}_n]^{\text{T}})(\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\text{T}}])^{-1}$$

Optimize w.r.t Ψ^{-1}:

$$\frac{\partial}{\partial\Psi^{-1}}Q = \frac{N}{2}\Psi - \frac{1}{2}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\text{T}} - \frac{1}{2}\sum_{n=1}^{N}\mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\text{T}}]\mathbf{W}^T + \sum_{n=1}^{N}\mathbf{W}\mathbb{E}[\mathbf{z}_n]\mathbf{x}_n$$

Plug in the expression of $\mathbf{W}$:

$$\Psi = \frac{1}{N}(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\text{T}} - \mathbf{W}\mathbb{E}[\mathbf{z}_n]\mathbf{x}_n^{\text{T}})$$

Assume Ψ to be a diagnal matrix:

$$\Psi = \frac{1}{N}\text{diag}(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\text{T}} - \mathbf{W}\mathbb{E}[\mathbf{z}_n]\mathbf{x}_n^{\text{T}})$$

This solution comes from "The EM Algorithm for Mixtures of Factor Analyzers, Zoubin Gharamani, Geoffrey E.Hinton, 1996", where the EM for mixtures of FA is given as well.

12.2 MAP estimation for the FA model

Assume prior $p(\mathbf{W})$ and $p(\Psi)$. Compare with the question before, the M-step needs to be moderated:

$$\frac{\partial}{\partial\mathbf{W}}(Q+\log p(\mathbf{W})) = 0$$

$$\frac{\partial}{\partial\Psi}(Q+\log p(\Psi)) = 0$$

12.3 Heuristic for assessing applicability of PCA*

Need pictures for illustration here!

12.4 Deriving the second principal component

For:

$$J(\mathbf{v}_2, \mathbf{z}_2) = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - z_{n1}\mathbf{v}_1 - z_{n2}\mathbf{v}_2)^{\mathrm{T}}(\mathbf{x}_n - z_{n1}\mathbf{v}_1 - z_{n2}\mathbf{v}_2)$$

Consider the derivative w.r.t one component of $\mathbf{z}_2$:

$$\frac{\partial}{\partial z_{m2}}J = \frac{1}{N}(2z_{m2}\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2 - 2\mathbf{v}_2^{\mathrm{T}}(\mathbf{x}_m - z_{m1}\mathbf{v}_1)) = 0$$

Using $\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2 = 1$ and $\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 = 0$ yields to:

$$z_{m2} = \mathbf{v}_2^{\mathrm{T}}\mathbf{x}_m$$

Since $\mathbf{C}$ is symmitric, use the constrain on $\mathbf{v}_1$ and $\mathbf{v}_2$. We apply SVD onto $\mathbf{C}$ first:

$$\mathbf{C} = \mathbf{O}^{\mathrm{T}}\Lambda\mathbf{O}$$

Where:

$$\Lambda = \mathrm{diag}\left\{\lambda_1, \lambda_2, ...\right\}$$

Are $\mathbf{C}$'s eigenvalues from the largest to the smallest.

$$\mathbf{O}^{\mathrm{T}} = \{\mathbf{u}_1, \mathbf{u}_2, ...\}$$

Are eigenvectors, that are vertical to each other $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \mathbb{I}(i = h)$. With$\mathbf{u}_1 = \mathbf{v}_1$.

Under constrains $\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2 = 1$ and $\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 = 0$, we are to minimize:

$$(\mathbf{O}\mathbf{v}_2)^{\mathrm{T}}\Lambda(\mathbf{O}\mathbf{v}_2)$$

Notice $\mathbf{O}\mathbf{v}_2$ means a transform on $\mathbf{v}_2$, with its length unchanged. And $(\mathbf{O}\mathbf{v}_2)^{\mathrm{T}}\Lambda(\mathbf{O}\mathbf{v}_2)$ measures the sum of the vector's components' square timed by Λ's eigenvalues. Hence the optimum is reached with all length converges to the component associated to the largest eigenvalue, which means:

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_2 = \mathbb{I}(i = 2)$$

Therefore:

$$\mathbf{v}_2 = \mathbf{u}_2$$

12.5 Deriving the residual error for PCA

$$\begin{aligned}||\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}\mathbf{v}_j||^2 =& (\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}\mathbf{v}_j)^{\mathrm{T}}(\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}\mathbf{v}_j) \\ =& \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n + \sum_{j=1}^{N} z_{nj}^2 - 2\mathbf{x}_n^T \sum_{j=1}^{N} z_{nj}\mathbf{v}_j\end{aligned}$$

Use $\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = \mathbb{I}(i=j)$, $z_{nj} = \mathbf{x}_n^{\mathrm{T}}\mathbf{v}_j$. We ends in the conclusion of a.

$$||\mathbf{x}_n - \sum_{j=1}^{K} z_{nj}\mathbf{v}_j||^2 = \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - 2\sum_{j=1}^{K} \mathbf{v}_j^{\mathrm{T}}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\mathbf{v}_j$$

Plug in $\mathbf{v}_j^{\mathrm{T}}\mathbf{C}\mathbf{v}_j = \lambda_j$ and sum over n can draw the conclusion in b.

Plug $K = d$ into the conclusion in b, we have:

$$J_{K=d} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{d}\lambda_j = 0$$

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - \sum_{j=1}^{d}\lambda_j = 0$$

In general cases:

$$J_K = \sum_{j=1}^{d}\lambda_j - \sum_{j=1}^{K}\lambda_j = \sum_{j=d+1}^{K}\lambda_j$$

12.6 Derivation of Fisher's linear discriminant*

Straightforward algebra.

（need reference）

12.7 PCA via successive deflation*

This problem involves the same technique used in solving 12.4, hence omitted.

12.8 Latent semantic indexing*

Practice by yourself.

12.9 Imputation in a FA model*

wtf$\mathbf{x}_v$?

wtf$\mathbf{x}_h$?

12.10 Efficiently evaluating the PPCA density

With:

$$\begin{aligned} p(\mathbf{z}) =& \mathcal{N}(\mathbf{z}|0,\mathbf{I}) \\ p(\mathbf{x}|\mathbf{z}) =& \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z},\sigma^2\mathbf{I}) \end{aligned}$$

Use the conclusion from chapter 4.

$$\mathcal{N}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0,\sigma^2\mathbf{I}+\mathbf{W}\mathbf{W}^{\mathrm{T}})$$

Deriavtion for MLE in 12.2.4 can be found in "Probabilistic Principal Component Analysis,Michael E.Tipping, Christopher M.Bishop,1999".

Plug in the MLE, thence the covariance matrix$(D * D)$'s inverse can be computed:

$$(\sigma^2\mathbf{I}+\mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1} = \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{W}(\frac{1}{\sigma^{-2}}\mathbf{W}^{\mathrm{T}}\mathbf{W}+\sigma^{-2}\mathbf{I})^{-1}\mathbf{W}^{\mathrm{T}}\sigma^{-2}$$

Which involves only inversing a $L * L$ matrix.

12.11 PPCA vs FA*

Practice by youself.

13 Sparse linear models

13.1 Partial derivative of the RSS

Define:

$$\text{RSS}(\mathbf{w}) = \sum_{n=1}^{N}(y_n - \mathbf{w}^{\text{T}}\mathbf{x}_n)^2$$

Straightforwardly:

$$\begin{aligned}
\frac{\partial}{\partial w_j}\text{RSS}(\mathbf{w}) =& \sum_{n=1}^{N} 2(y_n - \mathbf{w}^{\text{T}}\mathbf{x}_n)(-x_{nj}) \\
=& -\sum_{n=1}^{N} 2(x_{nj}y_n - x_{nj}\sum_{i=1}^{D} w_i x_{ni}) \\
=& -\sum_{n=1}^{N} 2(x_{nj}y_n - x_{nj}\sum_{i \neq j}^{D} w_i x_{ni} - x_{nj}^2 w_j)
\end{aligned}$$

With w_j's coefficient:

$$a_j = 2\sum_{n=1}^{N} x_{nj}^2$$

Other irrelevent terms can be absorbed into:

$$c_j = 2\sum_{n=1}^{N} x_{nj}(y_n - \mathbf{w}_{-j}^T \mathbf{x}_{n,-j})$$

In the end:

$$w_j = \frac{c_j}{a_j}$$

13.2 Derivation of M-step for EB for linear regression

We give the EM for Automatic Relevance Determination(ARD). For linear regression scene:

$$\begin{aligned}
p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) =& \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1}) \\
p(\mathbf{w}) =& \mathcal{N}(\mathbf{w}|0, \mathbf{A}^{-1}) \\
A =& \text{diag}\left\{\alpha\right\}
\end{aligned}$$

In E-step, we are to estimate expectation of $\mathbf{w}$. Using linear Gaussian relationship:

$$\begin{aligned} p(\mathbf{w}|\mathbf{y},\alpha,\beta) =& \mathcal{N}(\mu,\Sigma) \\ \Sigma^{-1} =& \mathbf{A}+\beta\mathbf{X}^{\mathrm{T}}\mathbf{X} \\ \mu =& \Sigma(\beta\mathbf{X}^{\mathrm{T}}\mathbf{y}) \end{aligned}$$

Then:

$$\begin{aligned} \mathbb{E}_{\alpha,\beta}[\mathbf{w}] =& \mu \\ \mathbb{E}_{\alpha,\beta}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] =& \Sigma+\mu\mu^{\mathrm{T}} \end{aligned}$$

For auxiliay function:

$$\begin{aligned} Q(\alpha,\beta,\alpha^{\mathrm{old}},\beta^{\mathrm{old}}) =& \mathbb{E}_{\alpha^{\mathrm{old}},\beta^{\mathrm{old}}}[\log p(\mathbf{y},\mathbf{w}|\alpha,\beta)] \\ =& \mathbb{E}[\log p(\mathbf{y}|\mathbf{w},\beta)+\log p(\mathbf{w}|)] \\ =& \frac{1}{2}\mathbb{E}[N\log\beta-\beta(\mathbf{y}-\mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\mathbf{w})+\sum_j\log\alpha_j-\mathbf{w}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{w}] \end{aligned}$$

In E-step, we need $\mathbb{E}[\mathbf{w}]$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]$, which have been computed: Introduce a prior for component in α and β:

$$p(\alpha,\beta)=\prod_j \mathrm{Gamma}(\alpha_j|a+1,b)\cdot\mathrm{Gamma}(\beta|c+1,d)$$

Hence the posterior auxiliary function is:

$$Q'=Q+\log p(\alpha,\beta)=Q+\sum_j(a\log\alpha_j-b\alpha_j)+(c\log\beta-d\beta)$$

In M-step, optimize w.r.t α_i:

$$\frac{\partial}{\partial\alpha_i}Q'=\frac{1}{2\alpha_i}-\frac{\mathbb{E}[w_i^2]}{2}+\frac{a}{\alpha_i}-b$$

Set it to zero:

$$\alpha_i=\frac{1+2a}{\mathbb{E}[w_i^2]-b}$$

Optimize w.r.t β:

$$\frac{\partial}{\partial\beta}Q'=\frac{N}{2\beta}-\mathbb{E}[||\mathbf{y}-\mathbf{X}\mathbf{w}||^2]+\frac{c}{\beta}-d$$

End in:

$$\beta=\frac{N+2c}{\mathbb{E}[||\mathbf{y}-\mathbf{X}\mathbf{w}||^2]+2d}$$

Expand the expectation ends in 13.168.

13.3 Derivation of fixed point updates for EB for linear regression*

Unsolved.

13.4 Marginal likelihood for linear regression*

Straightforward algebra.

13.5 Reducing elastic net to lasso

Expand both sides of 13.196, the right side:

$$\begin{aligned}J_1(c\mathbf{w}) =&(\mathbf{y}-c\mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y}-c\mathbf{X}\mathbf{w})+c^2\lambda_2\mathbf{w}^{\mathrm{T}}\mathbf{w}+\lambda_1|\mathbf{w}|_1\\ =&\mathbf{y}^{\mathrm{T}}\mathbf{y}-c^2\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}+c^2\lambda_2\mathbf{w}^{\mathrm{T}}\mathbf{w}+\lambda_1|\mathbf{w}|_1\end{aligned}$$

The left side:

$$\begin{aligned}J_2(\mathbf{w}) =&\begin{pmatrix}\mathbf{y}-c\mathbf{X}\mathbf{w}\\ -c\sqrt{\lambda_2}\mathbf{w}\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\mathbf{y}-c\mathbf{X}\mathbf{w}\\ -c\sqrt{\lambda_2}\mathbf{w}\end{pmatrix}+c\lambda_1|\mathbf{w}|_1\\ =&(\mathbf{y}-c\mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y}-c\mathbf{X}\mathbf{w})+c^2\lambda_2\mathbf{w}^{\mathrm{T}}\mathbf{w}+c\lambda_1|\mathbf{w}|_1\\ =&\mathbf{y}^{\mathrm{T}}\mathbf{y}+c^2\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}+c^2\lambda_2\mathbf{w}^{\mathrm{T}}\mathbf{w}+c\lambda_1|\mathbf{w}|_1\end{aligned}$$

Hence 13.196 and 13.195 are equal.

This shows elastic net regularization, which pick a regularing term as a linear combination of l_1 andl_0 equals a lasso one.

13.6 Shrinkage in linear regression

For ordinary least square:

$$\mathrm{RSS}(\mathbf{w})=(\mathbf{y}-\mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\mathbf{w})$$

Using $\mathbf{X}^{\mathrm{T}}\mathbf{X}=I$:

$$\mathrm{RSS}(\mathbf{w})=c+\mathbf{w}^{\mathrm{T}}\mathbf{w}-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}$$

Take the derivative:

$$\frac{\partial}{\partial w_k}\mathrm{RSS}(\mathbf{w})=2w_k-2\sum_{n=1}^{N}y_nx_{nk}$$

We have:

$$\hat{w}_k^{\text{OLS}} = \sum_{n=1}^{N} y_n x_{nk}$$

In ridge regression:

$$\text{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^{\text{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^{\text{T}}\mathbf{w}$$

Take the derivative:

$$(2 + 2\lambda)w_k = 2\sum_{n=1}^{N} y_n x_{nk}$$

Thus

$$\hat{w}_k^{\text{ridge}} = \frac{\sum_{n=1}^{N} y_n x_{nk}}{1 + \lambda}$$

Solution for lasso regression using subderivative is exploited in 13.3.2, which concludes in 13.63:

$$\hat{w}_k^{\text{lasso}} = \text{sign}(\hat{w}_k^{\text{OLS}})(|\hat{w}_k^{\text{OLS}}| - \frac{\lambda}{2})_+$$

Observe picture 13.24, it is easy to address the black line as OLS, gray one Ridge and dotted one lasso. And $\lambda_1 = \lambda_2 = 1$. It is noticeable that ridge cause a shrinkage to horizontal axis while lasso cause a sharp shrinkage to zero under certain threshold.

13.7 Prior for the Bernoulli rate parameter in the spike and slab model

$$p(\gamma|\alpha_1, \alpha_2) = \prod_{d=1}^{D} p(\gamma_d|\alpha_1, \alpha_2)$$

Integrate out π_d:

$$\begin{aligned}
p(\gamma_d|\alpha_1, \alpha_2) =& \frac{1}{B(\alpha_1, \alpha_2)} \int p(\gamma_d|\pi_d) p(\pi_d|\alpha_1, \alpha_2) \mathrm{d}\pi_d \\
=& \frac{1}{B(\alpha_1, \alpha_2)} \int \pi_d^{\gamma_d}(1 - \pi_d)^{(1-\gamma_d)} \pi_d^{\alpha_1 - 1}(1 - \pi_d)^{\alpha_2 - 1} \mathrm{d}\pi_d \\
=& \frac{1}{B(\alpha_1, \alpha_2)} \int \pi_d^{\alpha_1 + \gamma_d - 1}(1 - \pi_d)^{\alpha_2 + 1 - \gamma_d - 1} \mathrm{d}\pi_d \\
=& \frac{B(\alpha_1 + \gamma_d, \alpha_2 + 1 - \gamma_d)}{B(\alpha_1, \alpha_2)} = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + \gamma_d)\Gamma(\alpha_2 + 1 - \gamma_d)}{\Gamma(\alpha_1 + \alpha_2 + 1)}
\end{aligned}$$

Therefore(N_1 marks the number of 1 in γ):

$$\begin{aligned}p(\gamma|\alpha_1,\alpha_2) =&\frac{\Gamma(\alpha_1+\alpha_2)^N}{\Gamma(\alpha_1)^N\Gamma(\alpha_2)^N}\frac{\Gamma(\alpha_1+1)^{N_1}\Gamma(\alpha_2+1)^{N-N_1}}{\Gamma(\alpha_1+\alpha_2+1)^N}\\ =&\frac{(\alpha_1+1)^{N_1}(\alpha_2+1)^{N-N_1}}{(\alpha_1+\alpha_2+1)^N}\end{aligned}$$

And:

$$\log p(\gamma|\alpha_1,\alpha_2) = N\log\frac{\alpha_2+1}{\alpha_1+\alpha_2+1} + N_1\log\frac{\alpha_1+1}{\alpha_2+1}$$

13.8 Deriving E step for GSM prior

$$\text{Laplace}(w_j|0,\frac{1}{\gamma}) = \int N(w_j|0,\tau_j^2)Ga(\tau_j^2|1,\frac{\gamma^2}{2})\text{d}\tau_j^2$$

Take Laplace transform/generating transform to both sides:

To calculate:

$$\begin{aligned}\mathbb{E}[\frac{1}{\tau_j^2}|w_j] =& \int\frac{1}{\tau_j^2}p(\tau_j^2|w_j)\text{d}\tau_j^2 = \int\frac{1}{\tau_j^2}\frac{p(w_j|\tau_j^2)p(\tau_j^2)}{p(w_j)}\text{d}\tau_j^2\\ =&\frac{1}{p(w_j)}\int\frac{1}{\tau_j^2}N(w_j|0,\tau_j^2)p(\tau_j^2)\text{d}\tau_j^2\end{aligned}$$

According to 13.200, it reduces to:

$$\frac{1}{p(w_j)}\frac{-1}{|w_j|}\frac{\text{d}}{\text{d}w_j}\int N(w_j|0,\tau_j^2)p(\tau_j^2)\text{d}\tau_j^2$$

Because:

$$\frac{\text{d}}{\text{d}w}\log p(w) = \frac{1}{p(w)}\frac{\text{d}}{\text{d}w}p(w)$$

This gives 13.197:

$$\frac{1}{p(w_j)}\frac{-1}{|w_j|}\frac{\text{d}}{\text{d}w_j}p(w_j) = \frac{1}{|w_j|}\frac{\text{d}}{\text{d}w_j} - \log p(w_j)$$

! 此题存疑，Hint 1和Hint 2中可能均有印刷错误。

13.9 EM for sparse probit regression with Laplace prior

Straightforward Probit regression involves no latent variable. Introducing Laplace prior for linear factor **w** results in its lasso version. Since

Laplace distribution is a continuous mixture of Gaussian, a latent variable τ^2 with the same dimension as $\mathbf{w}$ is introduced. The PGM for Probit regression looks like:

$$\gamma \to \tau^2 \to \mathbf{w} \to \mathbf{y} \leftarrow \mathbf{X}$$

The joint distribution is:

$$p(\gamma, \tau^2, \mathbf{w}, \mathbf{y}|\mathbf{X}) = p(\gamma) \prod_{d=1}^{D} p(\tau_d^2|\gamma) \prod_{d=1}^{D} p(w_d|\tau_d^2) \prod_{n=1}^{N} \Phi(\mathbf{w}^T \mathbf{x}_n)^{y_n} (1-\Phi(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n))^{1-y_n}$$

For concise, we set γ as constant, according to 13.86:

$$\begin{aligned} p(\tau^2|\gamma) =& \mathrm{Gamma}(\tau_d^2|1, \frac{\gamma^2}{2}) \\ p(w_d|\tau_d^2) =& \mathcal{N}(w_d|0, \tau_d^2) \end{aligned}$$

Hence:

$$\begin{aligned} p(\tau^2, \mathbf{w}, \mathbf{y}|\mathbf{X}, \gamma) \propto \exp\left\{ -\frac{1}{2} \sum_{d=1}^{D} (\gamma^2 \tau_d^2 + \frac{w_d^2}{\tau_d^2}) \right\} \cdot \prod_{d=1}^{D} \frac{1}{\tau_d} \\ \cdot \prod_{n=1}^{N} \Phi(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n)^{y_n} (1 - \Phi(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n))^{1-y_n} \end{aligned}$$

In $Q(\theta^{\text{new}}, \theta^{\text{old}})$, we take expectation of θ^{old}. We have assumed $\mathbf{w}$ as parameter and τ^2 as latent variable, thus:

$$Q(\mathbf{w}, \mathbf{w}^{\text{old}}) = \mathbb{E}_{\mathbf{w}^{\text{old}}}[\log p(\mathbf{y}, \tau^2|\mathbf{w})]$$

Now extract terms involve $\mathbf{w}$ from $\log p(\tau^2, \mathbf{w}, \mathbf{y})$:

$$\log p(\mathbf{y}, \tau^2|\mathbf{w}) = c - \frac{1}{2} \sum_{d=1}^{D} \frac{w_d^2}{\tau_d^2} + \sum_{n=1}^{N} y_n \log \Phi(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n) + (1-y_n)(1-\Phi(\mathbf{w}^{\mathrm{T}} \mathbf{x}_n))$$

Thus we only need to calculate one expectation in E-step:

$$\mathbb{E}[\frac{1}{\tau_d^2}|\mathbf{w}^{\text{old}}]$$

Which can be done as in 13.4.4.3, because Probit and linear regression share the same PGM up to this stage.

The M-step is the same as Gaussian-prior Probit regression hence omitted.

13.10 GSM representation of group lasso*

Follow the hints and straightforward algebra.

13.11 Projected gradient descent for l1 regularized least squares

Generally, we take gradient on $\mathbf{w}$ and optimize. When there are constrains on $\mathbf{w}$ that could be broken by gradient descent, the increment has to be moderated to fit in the constrains.

To calculate:

$$\min_{\mathbf{w}} \{\text{NLL}(\mathbf{w}) + \lambda||\mathbf{w}||_1\}$$

Consider under a linear regression context:

$$\text{NLL}(\mathbf{w}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2$$

For $\lambda||\mathbf{w}||_1$ can not be differentiate, we need a non-trivial solution, it is suggest:

$$\mathbf{w} = \mathbf{u} - \mathbf{v}$$

$$u_i = (x_i)_+ = \max\{0, x_i\}$$

$$v_i = (-x_i)_+ = \max\{0, -x_i\}$$

With $\mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}$, then:

$$||\mathbf{w}||_1 = \mathbf{1}_n^{\text{T}}\mathbf{u} + \mathbf{1}_n^{\text{T}}\mathbf{v}$$

The original problem is changed to:

$$\min_{\mathbf{w}} \left\{ \frac{1}{2}||\mathbf{y} - \mathbf{X}(\mathbf{u} - \mathbf{v})||_2^2 + \lambda\mathbf{1}_n^{\text{T}}\mathbf{u} + \lambda\mathbf{1}_n^{\text{T}}\mathbf{v} \right\}$$

$$s.t. \mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}$$

Denote:

$$\mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$$

Rewrite the original target:

$$\min_{\mathbf{z}} \left\{ f(\mathbf{z}) = \mathbf{c}^{\text{T}}\mathbf{z} + \frac{1}{2}\mathbf{z}^{\text{T}}\mathbf{A}\mathbf{z} \right\}$$

$$s.t.\mathbf{z} \geq \mathbf{0}$$

Where:

$$\mathbf{c} = \begin{pmatrix} \lambda \mathbf{1}_n - \mathbf{y}\mathbf{X} \\ \lambda \mathbf{1}_n + \mathbf{y}\mathbf{X} \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & -\mathbf{X}^{\mathrm{T}}\mathbf{X} \\ -\mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{X} \end{pmatrix}$$

The gradient is given by:

$$\nabla f(\mathbf{z}) = \mathbf{c} + \mathbf{A}\mathbf{z}$$

For ordinary gradient descent:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha \nabla f(\mathbf{z}^k)$$

For projected case, take $\mathbf{g}^k$:

$$\mathbf{g}_i^k = \min \left\{ \mathbf{z}_i^k, \alpha \nabla f(\mathbf{z}^k)_i \right\}$$

During iteration:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \mathbf{g}^k$$

The original paper suggest more delicate method to moderate the learning rate, refer to ”Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems, Mario A.T.Figueiredo”.

13.12 Subderivative of the hinge loss function

$$\begin{aligned} &\text{if}(\theta < 1)\partial f(\theta) = \{-1\} \\ &\text{if}(\theta = 1)\partial f(\theta) = [-1, 0] \\ &\text{if}(\theta > 1)\partial f(\theta) = \{0\} \end{aligned}$$

13.13 Lower bounds to convex functions

Refer to ”Rigorous Affine Lower Bound Functions for Multivariate Polynomials and Their Use in Global Optimisation”.

14 Kernels

****

15 Gaussian processes

15.1 Reproducing property

We denote $\kappa(\mathbf{x}_1, \mathbf{x})$ by $f(\mathbf{x})$ and $\kappa(\mathbf{x}_2, \mathbf{x})$ by $g(\mathbf{x})$. From definition:

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} f_i \phi(\mathbf{x})$$

$$\kappa(\mathbf{x}_1, \mathbf{x}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}_1) \phi_i(\mathbf{x})$$

Since $\mathbf{x}$ can be chosen arbitrarily, we have the properties hold(the one for g is obtained similarly):

$$f_i = \lambda_i \phi_i(\mathbf{x}_1)$$

$$g_i = \lambda_i \phi_i(\mathbf{x}_2)$$

Therefore:

$$\begin{aligned} < \kappa(\mathbf{x}_1, .), \kappa(\mathbf{x}_2, .) >= & < f, g > \\ = & \sum_{i=1}^{\infty} \frac{f_i g_i}{\lambda_i} \\ = & \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}_1) \phi_i(\mathbf{x}_2) \\ = & \kappa(\mathbf{x}_1, \mathbf{x}_2) \end{aligned}$$

16 Adaptive basis function models*

16.1 Nonlinear regression for inverse dynamics

Practise by yourself.

17 Markov and hidden Markov models

17.1 Derivation of Q function for HMM

Firstly, we estimate the distribution of $\mathbf{z}_{1:T}$ w.r.t θ^{old}, for auxiliay function, we are to calculate the log-likelihood w.r.t θ and $\mathbf{z}_{1:T}$.

$$\begin{aligned}
Q(\theta, \theta^{\text{old}}) =& \mathbb{E}_{p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}, \theta^{\text{old}})}[\log p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}|\theta)] \\
=& \mathbb{E}_p[\log \left\{ \prod_{i=1}^{N} \left\{ p(z_{i,1}|\pi) \prod_{t=2}^{T_i} p(z_{i,t}|z_{i,t-1}, \mathbf{A}) \prod_{t=1}^{T_i} p(x_{i,t}|z_{i,t}, \mathbf{B}) \right\} \right\}] \\
=& \mathbb{E}_p[\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}[z_{i,1} = k] \log \pi_k + \sum_{i=1}^{N} \sum_{t=2}^{T_i} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbb{I}[z_{i,t} = k, z_{i,t-1} = j] \log \mathbf{A}(j,k) \\
&+ \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{k=1}^{K} \mathbb{I}[z_{i,t} = k] \log p(x_{i,t}|z_{i,t} = k, \mathbf{B})]
\end{aligned}$$

Further we have 17.98, 17.99, 17.100, using the definition of expectation yields to 17.97.

17.2 Two filter approach to smoothing in HMMs

For $r_t(i) = p(z_t = i|x_{t+1:T})$, we have:

$$\begin{aligned}
p(z_t = i|x_{t+1:T}) =& \sum_j p(z_t = i, z_{t+1} = j|x_{t+1:T}) \\
=& \sum_j p(z_{t+1} = j|x_{t+1:T}) p(z_t = i|z_{t+1} = j, x_{t+1:T}) \\
=& \sum_j p(z_{t+1} = j|x_{t+1:T}) p(z_t = i|z_{t+1} = j) \\
=& \sum_j p(z_{t+1} = j|x_{t+1:T}) \Psi^{-}(j, i)
\end{aligned}$$

Where Ψ^{-} denotes the transform matrix in an inverse sense, we further have:

$$\begin{aligned}
p(z_{t+1} = j|x_{t+1:T}) =& p(z_{t+1} = j|x_{t+1}, x_{t+2:T}) \\
\propto& p(z_{t+1} = j, x_{t+1}, x_{t+2:T}) \\
=& p(x_{t+2:T}) p(z_{t+1} = j|x_{t+2:T}) p(x_{t+1}|z_{t+1} = j, x_{t+2:T}) \\
\propto& r_{t+1}(j) \phi_{t+1}(j)
\end{aligned}$$

Therefore we can calculate $r_t(i)$ recursively:

$$r_t(i) \propto \sum_j r_{t+1}(j)\phi_{t+1}(j)\Psi^-(j,i)$$

And initial element $p(z_T)$ is given by $\prod_T(i)$.

To rewrite $\gamma_t(i)$ in terms of new factors:

$$\begin{aligned}
\gamma_t(i) \propto & p(z_t = i|x_{1:T}) \\
\propto & p(z_t = i, x_{1:T}) \\
= & p(z_t = i)p(x_{1:T}|z_t = i) \\
= & p(z_t = i)p(x_{1:t}|z_t = i)p(x_{t+1:T}|z_t = i, x_{1:t}) \\
= & p(z_t = i)p(x_{1:t}|z_t = i)p(x_{t+1:T}|z_t = i) \\
= & \frac{1}{p(z_t = i)}p(x_{1:t}, z_t = i)p(x_{t+1:T}, z_t = i) \\
\propto & \frac{1}{p(z_t = i)}p(z_t = i|x_{1:t})p(z_t = i|x_{t+1:T}) \\
= & \frac{\alpha_t(i) \cdot r_t(i)}{\prod_t(i)}
\end{aligned}$$

17.3 EM for HMMs with mixture of Gaussian observations

Using mixture of Gaussians as the emission distribution does not the evaluation of γ and ϵ, hence the E-step does not change compared to the one in exercise 17.1.

As long as $\mathbf{A}$ and $\mathbf{B}$ are estimated independently, we are now focus on estimating $\mathbf{B} = (\pi, \mu, \Sigma)$ during M-step, the involved target function is:

$$\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T_i} \gamma_{i,t}(k) \log p(x_{i,t}|\mathbf{B})$$

Since the parameters are independent w.r.t k, we delve into a case where k is given. We also denote the iteration through $i = 1$ to N and $t = 1$ to T_i by $n = 1$ to $T = \sum_{i=1}^{N} T_i$, now the log-likelihood takes the form:

$$\sum_{n=1}^{T} \gamma_n(k) \log p(x_n|\pi_k, \mu_k, \Sigma_k)$$

It can be seen as a weighted form of log-likelihood for a mixture of Gaussian, assume the mixture contains C(it should be C_k, but this notation causes no contradiction as long as we take k for granted) Gaussians. We are to apply another EM procedure during the M-step for this HMM. Denote the latent variable corresponding to x_n by $h_{n,k}$. Estimate the distribution of $p(h_{n,k}|z_n, \pi_k, \mu_k, \Sigma_k)$ is tantamount to the E-step used in handling traditional mixture of Gaussians. Denote the expectation of $h_{n,k}$'s components by $\gamma'_{c,n}(k)$.

Now applying the M-step of mixture of Gaussians, recall that auxiliay takes the form:

$$\sum_{n=1}^{T} \gamma_n(k) \sum_{c=1}^{C} \gamma'_{c,n}(k) \left\{ \log \pi_{k,c} + \log \mathcal{N}(x_n|\mu_{k,c}, \Sigma_{k,c}) \right\}$$

Hence this HMM reweighted a traditional mixture of Gaussians, with the weight changed from $\gamma'_{c,n}(k)$ into $\gamma_n(k) \cdot \gamma'_{c,n}(k)$. The rest estimation is trivially the application of M-step in mixture of Gaussians using new weights.

17.4 EM for HMMs with tied mixtures

Recall the conclusion from exercise 17.3, the last M-step inside M-step takes the form:

$$\sum_{k=1}^{K} \sum_{n=1}^{T} \sum_{c=1}^{C} \gamma_{c,n}(k) \left\{ \log \pi_{k,c} + \log \mathcal{N}(x_n|\mu_c, \Sigma_c) \right\}$$

Where we accordingly update the meaning of γ, and we also remove k from the footnotes of μ and Σ given the conditions in this exercise.

It is easy to notice that this target function again takes the form of M-step target for a traditional mixture of Gaussians. Taking independent k and update π_k gives the learning process of K mixing weights. Sum out k and C independent Gaussian parameters can be updated.

18 State space models

18.1 Derivation of EM for LG-SSM

We directly work on the auxiliary function:

$$\begin{aligned}
Q(\theta, \theta^{\text{old}}) =& \mathbb{E}_{p(\mathbf{Z}|\mathbf{Y},\theta^{\text{old}})}[\log \prod_{n=1}^{N} p(z_{n,1:T_n}, y_{n,1:T_n}|\theta)] \\
=& \mathbb{E}[\sum_{n=1}^{N} \log p(z_{n,1}) \prod_{i=2}^{T_n} p(z_{n,i}|z_{n,i-1}) \prod_{i=1}^{T_n} p(y_{n,i}|z_{n,i})] \\
=& \mathbb{E}[\sum_{n=1}^{N} \log \mathcal{N}(z_{n,1}|\mu_0, \Sigma_0) + \sum_{i=2}^{T_n} N(z_{n,i}|A_i z_{n,i-1} + B_i u_i, Q_i) \\
&+ \sum_{i=1}^{T_n} \mathcal{N}(y_{n,i}|C_i z_{n,i} + D_i u_i, R_i)] \\
=& \mathbb{E}[N \log \frac{1}{|\Sigma_0|^{\frac{1}{2}}} + \left\{ -\frac{1}{2} \sum_{n=1}^{N} (z_{n,1} - \mu_0)^{\mathrm{T}} \Sigma_0^{-1} (z_{n,1} - \mu_0) \right\} \\
&+ \sum_{i=2}^{T} N_i \log \frac{1}{|Q_i|^{\frac{1}{2}}} \\
&+ \left\{ -\frac{1}{2} \sum_{n=1}^{N_i} (z_{n,i} - A_i z_{n,i-1} - B_i u_i)^{\mathrm{T}} Q_i^{-1} (z_{n,i} - A_i z_{n,i-1} - B_i u_i) \right\}] \\
&+ \sum_{i=2}^{T} N_i \log \frac{1}{|R_i|^{\frac{1}{2}}} \\
&+ \left\{ -\frac{1}{2} \sum_{n=1}^{N_i} (y_{n,i} - C_i z_{n,i} - D_i u_i)^{\mathrm{T}} R_i^{-1} (y_{n,i} - C_i z_{n,i} - D_i u_i) \right\}]
\end{aligned}$$

When exchanging the order of sum over data, we have $T = \max_n \{T_n\}$ and N_i denotes the number of data set with size no more than i.

To estimate μ_0, take the related terms:

$$\mathbb{E}[-\frac{1}{2} \sum_{n=1}^{N} (z_{n,1} - \mu_0) \Sigma_0^{-1} (z_{n,1} - \mu_0)]$$

Take derivative w.r.t μ_0:

$$\mathbb{E}[\sum_{n=1}^{N} -\frac{1}{2} \mu_0^{\mathrm{T}} \Sigma_0^{-1} \mu_0 + z_{n,1} \Sigma_0^{-1} \mu_0]$$

Setting it to zero yields:

$$\mu_0 = \frac{1}{N}\mathbb{E}[z_{n,1}]$$

It is obvious that such estimation is similar to that for MVN with x_n replaced by $\mathbb{E}[z_{n,1}]$. This similarity works for other parameters as well. For example, estimate Σ_0 is tantamount to estimate the covariance of MVN with data terms replaced.

Such analysis works for Q_i and R_i as well. To estimate coefficient matrix, we consider A_i firstly. The related term is:

$$\mathbb{E}[\sum_{n=1}^{N_i}\left\{z_{n,i}^{\mathrm{T}}A_i^{\mathrm{T}}Q_i^{-1}A_i z_{n,i} - 2z_{n,i-1}^{\mathrm{T}}A_i^{\mathrm{T}}Q_i^{-1}(z_{n,i} - B_i u_i)\right\}]$$

Setting derivative to zero yields a solution similar to that for μ_0, the same analysis can be applied for B_i, C_i, D_i as well.

18.2 Seasonal LG-SSM model in standard form

From Fig.18.6(a), we have:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & \mathbf{0}_{S-1}^{\mathrm{T}} \\ 0 & 1 & 0 & \mathbf{0}_{S-1}^{\mathrm{T}} \\ 0 & 0 & 1 & \mathbf{0}_{S-1}^{\mathrm{T}} \\ \mathbf{0}_{S-1} & \mathbf{0}_{S-1} & \mathbf{I} & \mathbf{0}_{S-1} \end{pmatrix}$$

$$\mathbf{Q} = \begin{pmatrix} Q_a & \mathbf{0}_{S+1}^{\mathrm{T}} & & \\ 0 & Q_b & \mathbf{0}_{S}^{\mathrm{T}} & \\ 0 & 0 & Q & \mathbf{0}_{S-1}^{\mathrm{T}} \\ \mathbf{0}_{(S-1)*(S+2)} & & & \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & \mathbf{0}_{S-1}^{\mathrm{T}} \end{pmatrix}$$

Where we use $\mathbf{0}_n$ to denote a colomn vector of 0 with length n, and $\mathbf{0}_{m*n}$ to denote a $m*n$ matrix of 0.

19 Undirected graphical models(Markov random fields)

19.1 Derivation of the log partition function

According to the definition:

$$Z(\theta) = \sum_{\mathbf{y}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c)$$

It is straightforward to give:

$$\begin{aligned}
\frac{\partial \log Z(\theta)}{\partial \theta_{c'}} =& \frac{\partial}{\partial \theta_{c'}} \log \sum_{\mathbf{y}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \frac{\partial}{\partial \theta_{c'}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \prod_{c \in C, c \neq c'} \psi_c(\mathbf{y}_c|\theta_c) \frac{\partial}{\partial \theta_{c'}} \psi_{c'}(\mathbf{y}_{c'}|\theta_{c'}) \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \prod_{c \in C, c \neq c'} \psi_c(\mathbf{y}_c|\theta_c) \frac{\partial}{\partial \theta_{c'}} \exp\left\{\theta_{c'}^{\mathrm{T}} \phi_{c'}(\mathbf{y}_{c'})\right\} \\
=& \frac{1}{Z(\theta)} \sum_{\mathbf{y}} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta_c) \phi_{c'}(\mathbf{y}_{c'}) \\
=& \sum_{\mathbf{y}} \phi_{c'}(\mathbf{y}_{c'}) \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(\mathbf{y}_c|\theta) \\
=& \sum_{\mathbf{y}} \phi_{c'}(\mathbf{y}_{c'}) p(\mathbf{y}|\theta) \\
=& \mathbb{E}[\phi_{c'}(\mathbf{y}_{c'})|\theta]
\end{aligned}$$

19.2 CI properties of Gaussian graphical models

Problem a:

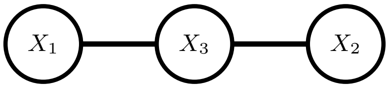
We have:

$$\Sigma = \begin{pmatrix} 0.75 & 0.5 & 0.25 \\ 0.5 & 1.0 & 0.5 \\ 0.25 & 0.5 & 0.75 \end{pmatrix}$$

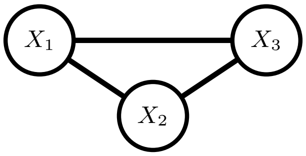
And:

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

Thus we have independency: $X_1 \perp X_2 | X_3$. This introduces a MRF like:



Problem b: The inverse of Σ contains no zero element, hence no conditional independency. Therefore there have to be edges between any two vertexes.



This model also cancels the marginal independency $X_1 \perp X_3$. But it is possible to model this set of properties by Bayesian network with two directed edges $X_1 \rightarrow X_2$ and $X_3 \rightarrow X_2$.

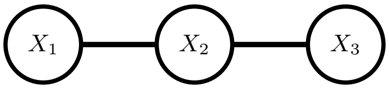
Problem c: Consider the terms inside the exponential:

$$-\frac{1}{2}\left\{x_1^2 + (x_2 - x_1)^2 + (x_3 - x_2^2)\right\}$$

It is easy to see the precision matrix and covariance matrix take:

$$\Lambda = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

Problem d: The only independency is $X_1 \perp X_3 | X_2$:



19.3 Independencies in Gaussian graphical models

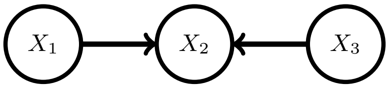
Problem a and b:

This PGM implies $X_1 \perp X_3 | X_2$, hence we are looking for a precision matrix with $\Lambda_{1,3} = 0$, thus C and D meet the condition. On the other hand, $(A^{-1})_{1,3} = (B^{-1})_{1,3} = 0$. So A and B are candidates for covariance matrix.

Problem c and d:

This PGM tells that $X_1 \perp X_3$. Hence C and D can be covariance matrix, A and B can be precision matrix.

The only possible PGM is:



Problem e:

The answer can be derived from the conclusion of marginal Gaussian directly, A is true while B not.

19.4 Cost of training MRFs and CRFs

The answer are generally:

$$O(r(Nc+1))$$

and

$$O(r(Nc+N))$$

19.5 Full conditional in an Ising model

Straightforwardly(we have omitted θ from condition w.l.o.g):

$$\begin{aligned} p(x_k = 1|\mathbf{x}_{-k}) =& \frac{p(x_k = 1, \mathbf{x}_{-k})}{p(\mathbf{x}_{-k})} \\ =& \frac{p(x_k = 1, \mathbf{x}_{-k})}{p(x_k = 0, \mathbf{x}_{-k}) + p(x_k = 1, \mathbf{x}_{-k})} \\ =& \frac{1}{1 + \frac{p(x_k = 0, \mathbf{x}_{-k})}{p(x_k = 1, \mathbf{x}_{-k})}} \\ =& \frac{1}{1 + \frac{\exp(h_k \cdot 0) \prod_{<k,i>} \exp(J_{k,i} \cdot 0)}{\exp(h_k \cdot 1) \prod_{<k,i>} \exp(J_{k,i} \cdot x_i)}} \\ =& \sigma(h_k + \sum_{i=1, i \neq k}^{n} J_{k,i} x_i) \end{aligned}$$

When using denotation $x = \{0, 1\}$, the full conditional becomes:

$$p(x_k = 1|\mathbf{x}_{-k}) \sigma(2 \cdot (h_k + \sum_{i=1, i \neq k}^{n} J_{k,i} x_i))$$

20 Exact inference for graphical models

20.1 Variable elimination*

Where tf is the figure?!

20.2 Gaussian times Gaussian is Gaussian

We have:

$$\begin{aligned}
&\mathcal{N}(x|\mu_1, \lambda_1^{-1}) \times \mathcal{N}(x|\mu_2, \lambda_2^{-1}) \\
=&\frac{\sqrt{\lambda_1\lambda_2}}{2\pi} \exp\left\{-\frac{\lambda_1}{2}(x-\mu_1)^2 - \frac{\lambda_2}{2}(x-\mu_2)^2\right\} \\
=&\frac{\sqrt{\lambda_1\lambda_2}}{2\pi} \exp\left\{-\frac{\lambda_1+\lambda_2}{2}x^2 + (\lambda_1\mu_1+\lambda_2\mu_2)x - \frac{\lambda_1\mu_1^2+\lambda_2\mu_2^2}{2}\right\}
\end{aligned}$$

By completing the square:

$$\begin{aligned}
&\exp\left\{-\frac{\lambda_1+\lambda_2}{2}x^2 + (\lambda_1\mu_1+\lambda_2\mu_2)x - \frac{\lambda_1\mu_1^2+\lambda_2\mu_2^2}{2}\right\} \\
=&c \cdot \exp -\frac{\lambda}{2}(x-\mu)^2
\end{aligned}$$

Where:

$$\lambda = \lambda_1 + \lambda_2$$

$$\mu = \lambda^{-1}(\lambda_1\mu_1 + \lambda_2\mu_2)$$

The constant factor c can be obtained by computing the constant terms inside the exponential.

20.3 Message passing on a tree

Problem a:

It is easy to see after variable elimination:

$$p(X_2 = 50) = \sum_{G_1}\sum_{G_2} p(G_1)p(G_2|G_1)p(X_2 = 50|G_2)$$

$$p(G_1 = 1, X_2 = 50) = p(G_1)\sum_{G_2} p(G_2|G_1 = 1)p(X_2 = 50|G_2)$$

Thus:

$$p(G_1 = 1|X_2 = 50) = \frac{0.45 + 0.05 \cdot \exp(-5)}{0.5 + 0.5 \cdot \exp(-5)} \approx 0.9$$

Problem b(here X denotes X_2 or X_3):

$$\begin{aligned}
&p(G_1 = 1|X_2 = 50, X_3 = 50)\\
=&\frac{p(G_1 = 1, X_2 = 50, X_3 = 50)}{p(X_2 = 50, X_3 = 50)}\\
=&\frac{p(G_1 = 1)p(X_2|G_1 = 1)p(X_3|G_1 = 1)}{p(G_1 = 0)p(X_2|G_1 = 0)p(X_3|G_1 = 0) + p(G_1 = 1)p(X_2|G_1 = 1)p(X_3|G_1 = 1)}\\
=&\frac{p(X = 50|G_1 = 1)^2}{p(X = 50|G_1 = 0)^2 + p(X = 50|G_1 = 1)^2}\\
\approx&\frac{0.9^2}{0.1^2 + 0.9^2} \approx 0.99
\end{aligned}$$

Extra evidence makes the belief in $G_1 = 1$ firmer.

Problem c:

The answer to problem c is symmetric to that to problem b, $p(G_1 = 0|X_2 = 60, X_3 = 60) \approx 0.99$.

Problem d:

Using the same pattern of analysis from Problem b, we have:

$$p(G_1 = 1|X_2 = 50, X_3 = 60)$$

$$= \frac{p(X = 50|G_1 = 1)p(X = 60|G_1 = 1)}{p(X = 50|G_1 = 0)p(X = 60|G_1 = 0) + p(X = 50|G_1 = 1)p(X = 60|G_1 = 1)}$$

Notice we have:

$$p(X = 50|G_1 = 1) = p(X = 60|G_1 = 0)$$

$$p(X = 50|G_1 = 0) = p(X = 60|G_1 = 1)$$

Hence:

$$P(G_1 = 1|X_2 = 50, X_3 = 60) = 0.5$$

In this case, X_2 and X_3 have equal strength as evidence and their effects achieve a balance so they provide not enough information to distort the prior knowledge.

20.4 Inference in 2D lattice MRFs

Please refer to PGM:principals and techniques 11.4.1.

21 Variational inference

21.1 Laplace approximation to $p(\mu, \log\sigma|D)$ for a univariate Gaussian

Laplace approximation equals representing $f(\mu, l) = \log p(\mu, l = \log\sigma|D)$ with second-order Taylor expansion. We have:

$$\begin{aligned}
\log p(\mu, l|D) =& \log p(\mu, l, D) - \log p(D) \\
=& \log p(\mu, l) + \log p(D|\mu, l) + c \\
=& \log p(D|\mu, l) + c \\
=& \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mu)^2\right\} + c \\
=& - N \log\sigma + \sum_{n=1}^{N} -\frac{1}{2\sigma^2}(y_n - \mu)^2 + c \\
=& - N \cdot l + \frac{1}{2}\frac{1}{\exp\{2 \cdot l\}} \sum_{n=1}^{N}(y_n - \mu)^2 + c
\end{aligned}$$

Thus we derive:

$$\begin{aligned}
\frac{\partial \log p(\mu, l|D)}{\partial\mu} =& \frac{1}{2}\frac{1}{\exp\{2 \cdot l\}} \sum_{n=1}^{N} 2 \cdot (y_n - \mu) \\
=& \frac{N}{\sigma^2} \cdot (\bar{y} - \mu) \\
\frac{\partial \log p(\mu, l|D)}{\partial l} =& - N + \frac{1}{2}\sum_{n=1}^{N}(y_n - \mu)^2 \cdot (-2) \cdot \frac{1}{\exp\{2 \cdot l\}} \\
=& - N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \mu)^2 \\
\frac{\partial^2 \log p(\mu, l|D)}{\partial\mu^2} =& - \frac{N}{\sigma^2} \\
\frac{\partial^2 \log p(\mu, l|D)}{\partial l^2} =& - \frac{2}{\sigma^2}\sum_{n=1}^{N}(y_n - \mu)^2 \\
\frac{\partial^2 \log p(\mu, l|D)}{\partial\mu\partial l} =& N \cdot (\bar{y} - \mu) \cdot (-2) \cdot \frac{1}{\sigma^2}
\end{aligned}$$

For approximation, $p(\mu, l) \approx N(\mu, \Sigma)$ with:

$$\Sigma = \begin{pmatrix} \frac{\partial^2 \log p(\mu,l|D)}{\partial \mu^2} & \frac{\partial^2 \log p(\mu,l|D)}{\partial l^2} \\ \frac{\partial^2 \log p(\mu,l|D)}{\partial l^2} & \frac{\partial^2 \log p(\mu,l|D)}{\partial \mu \partial l} \end{pmatrix}^{-1}$$

$$\mu = \Sigma \begin{pmatrix} \frac{\partial \log p(\mu,l|D)}{\partial \mu} \\ \frac{\partial \log p(\mu,l|D)}{\partial l} \end{pmatrix}$$

21.2 Laplace approximation to normal-gamma

This is the same with exercise 21.1 when the prior is uniformative. We formally substitute:

$$\begin{aligned} \sum_{n=1}^{N}(y_n - \mu)^2 =& \sum_{n=1}^{N}((y_n - \bar{y}) - (\mu - \bar{y}))^2 \\ =& \sum_{n=1}^{N}(y_n - \bar{y})^2 + \sum_{n=1}^{N}(\mu - \bar{y})^2 + 2(\mu - \bar{y}) \cdot \sum_{n=1}^{N}(y_n - \bar{y}) \\ =& Ns^2 + N(\mu - \bar{y})^2 \end{aligned}$$

Where $s^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2$

Conclusions in all problems a, b and c are included in the previous solution.

21.3 Variational lower bound for VB for univariate Gaussian

What left in section 21.5.1.6 is the derivation for 21.86 to 21.91. We omit the derivation for entropy for Gaussian and moments, which can be found in any information theory textbook. Now we derive the $\mathbb{E}[\ln x | x \sim \text{Gamma}(a, b)]$, which can therefore yields to the entropy for a Gamma distribution.

We know that Gamma distribution is an exponential family distribution:

$$\begin{aligned} \text{Gamma}(x|a, b) =& \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-b \cdot x\} \\ \propto& \exp\{-b \cdot x + (a-1)\ln x\} \\ =& \exp\left\{\phi(x)^{\mathrm{T}}\theta\right\} \end{aligned}$$

The sufficient statistics is $\phi(x) = (x, \ln x)^T$ and natural parameter is given by $\theta = (-b, a-1)^T$. Thus Gamma distribution can be seen as the maximum entropy distribution under constraints on x and $\ln x$.

The culumant function is given by:

$$\begin{aligned}
A(\theta) &= \log Z(\theta) \\
&= \log \frac{\Gamma(a)}{b^a} \\
&= \log \Gamma(a) - a \log b
\end{aligned}$$

The expectation of sufficient statistics is given by the derivative of cumulant function, therefore:

$$\mathbb{E}[\ln x] = \frac{\partial A}{\partial(a-1)} = \frac{\Gamma'(a)}{\Gamma(a)} - \log b$$

According to defintion $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$:

$$\mathbb{E}[\ln x] = \psi(a) - \log b$$

The rest derivations are completed or trivial.

21.4 Variational lower bound for VB for GMMs

The lower bound is given by:

$$\begin{aligned}
\mathbb{E}_q[\log \frac{p(\theta, D)}{q(\theta)}] =& \mathbb{E}_q[\log p(\theta, D)] - \mathbb{E}_q[q(\theta)] \\
=& \mathbb{E}_q[\log p(D|\theta)] + \mathbb{E}_q[\log p(\theta)] + \mathbb{E}_q[\log q(\theta)] \\
=& \mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda, \pi)] + \mathbb{E}[\log p(\mathbf{z}, \mu, \Lambda, \pi)] \\
& - \mathbb{E}[\log q(\mathbf{z}, \mu, \Lambda, \pi)] \\
=& \mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda, \pi)] + \mathbb{E}[\log p(\mathbf{z}|\pi)] + \mathbb{E}[\log p(\pi)] + \mathbb{E}[\log p(\mu, \Lambda)] \\
& + \mathbb{E}[\log q(\mathbf{z})] + \mathbb{E}[\log q(\pi)] + \mathbb{E}[\log q(\mu, \Lambda)]
\end{aligned}$$

We are now showing 21.209 to 21.215.

For 21.209:

$$\begin{aligned}
\mathbb{E}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda)] =& \mathbb{E}_{q(\mathbf{z})q(\mu,\Lambda)}[\log p(\mathbf{x}|\mathbf{z}, \mu, \Lambda)] \\
=& \sum_n \sum_k \mathbb{E}_{q(\mathbf{z})q(\mu,\Lambda)}[-\frac{D}{2}\log 2\pi + \frac{1}{2}\log|\Lambda_k| - \frac{1}{2}(x_n - \mu_k)^T \Lambda_k (x_n - \mu_k)]
\end{aligned}$$

Using 21.132 and converting summing by average $\bar{x}_k$ yields to solution.

For 21.210:

$$\begin{aligned}\mathbb{E}[\log p(\mathbf{z}|\pi)] =&\mathbb{E}_{q(\mathbf{z})q(\pi)}[\log p(\mathbf{z}|\pi)]\\ =&\mathbb{E}_{q(\mathbf{z})q(\pi)}[\log \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}]\\ =&\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{q(\mathbf{z})q(\pi)}[z_{nk}\log \pi_k]\\ =&\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{q(\mathbf{z})}[z_{nk}]\mathbb{E}_{q(\pi)}[\log \pi_k]\\ =&\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\log \bar{\pi}_k\end{aligned}$$

For 21.211:

$$\begin{aligned}\mathbb{E}[\log p(\pi)] =&\mathbb{E}_{q(\pi)}[\log p(\pi)]\\ =&\mathbb{E}_{q(\pi)}[\log(C\cdot\prod_{k=1}^{K}\pi_k^{\alpha_0-1})]\\ =&\ln C+(\alpha_0-1)\sum_{k=1}^{K}\log \bar{\pi}_k\end{aligned}$$

For 21.212:

$$\begin{aligned}\mathbb{E}[\log p(\mu,\Lambda)] =&\mathbb{E}_{q(\mu,\Lambda)}[\log p(\mu,\Lambda)]\\ =&\mathbb{E}_{q(\mu,\Lambda)}[\log \prod_{k=1}^{K}\mathrm{Wi}(\Lambda_k|L_0,v_0)\cdot\mathcal{N}(\mu_k|m_0,(\beta_0\Lambda_k)^{-1}]\\ =&\sum_{k=1}^{K}\mathbb{E}_{q(\mu,\Lambda)}[\log C+\frac{1}{2}(v_0-D-1)\log|\Lambda_k|-\frac{1}{2}tr\left\{\Lambda_k L_0^{-1}\right\}\\ &-\frac{D}{2}\log 2\pi-\frac{1}{2}\log|\beta_0\Lambda_k|-\frac{1}{2}(\mu_k-m_0)^{\mathrm{T}}(\beta_0\Lambda_k)(\mu_k-m_0)]\end{aligned}$$

Using 21.131 to expand the expected value of the quadratic form and using the fact that the mean of a Wi distribution is $v_k L_k$ and we are done.

For 21.213:

$$\begin{aligned}\mathbb{E}[\log q(\mathbf{z})] =& \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})]\\ =& \mathbb{E}_{q(\mathbf{z})}[\sum_i \sum_k z_{ik} \log r_{ik}]\\ =& \sum_i \sum_k \mathbb{E}_{q(\mathbf{z})}[z_{ik}] \log r_{ik}\\ =& \sum_i \sum_k r_{ik} \log r_{ik}\end{aligned}$$

For 21.214:

$$\begin{aligned}\mathbb{E}[\log q(\pi)] =& \mathbb{E}_{q(\pi)}[\log q(\pi)]\\ =& \mathbb{E}_{q(\pi)}[\log C + \sum_{k=1}^{K}(\alpha_k - 1)\log \pi_k]\\ =& \log C + \sum_k (\alpha_k - 1)\log \bar{\pi}_k\end{aligned}$$

For 21.215:

$$\begin{aligned}\mathbb{E}[\log q(\mu, \Lambda)] =& \mathbb{E}_{q(\mu,\Lambda)}[\log q(\mu, \Lambda)]\\ =& \sum_k \mathbb{E}_{q(\mu,\Lambda)}[\log q(\Lambda_k) - \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\beta_k \Lambda_k|\\ &- \frac{1}{2}(\mu_k - m_k)^{\mathrm{T}}(\beta_k \Lambda_k)(\mu_k - m_k)]\end{aligned}$$

Using 21.132 to expand the quadratic form to give $\mathbb{E}[(\mu_k - m_k)^{\mathrm{T}}(\beta_k \Lambda_k)(\mu_k - m_k)] = D$

21.5 Derivation of $\mathbb{E}[\log \pi_k]$

under a Dirichlet distribution Dirichlet distribution is an exponential family distribution, we have:

$$\phi(\pi) = (\log \pi_1, \log \pi_2, ... \log \pi_K)$$

$$\theta = \alpha$$

The cumulant function is:

$$A(\alpha) = \log B(\alpha) = \sum_{i=1}^{K} \log \Gamma(\alpha_i) - \log \Gamma(\sum_{i=1}^{K} \alpha_i)$$

And:

$$\mathbb{E}[\log \pi_k] = \frac{\partial A(\alpha)}{\partial \alpha_k} = \frac{\Gamma'(\alpha_k)}{\Gamma(\alpha_k)} - \frac{\Gamma'(\sum_{i=1}^K \alpha_k)}{\Gamma(\sum_{i=1}^K \alpha_k)} = \psi(\alpha_k) - \psi(\sum_{i=1}^K \alpha_i)$$

Take exponential on both sides:

$$\exp(\mathbb{E}[\log \pi_k]) = \exp(\psi(\alpha_k) - \psi(\sum_{i=1}^K \alpha_k)) = \frac{\exp(\alpha_k)}{\exp(\sum_{i=1}^K \alpha_i)}$$

21.6 Alternative derivation of the mean field updates for the Ising model

This is no different than applying the procedure in section 21.3.1 before derivating updates, hence omitted.

21.7 Forwards vs reverse KL divergence

We have:

$$\begin{aligned}
\mathbb{KL}(p(x,y)||q(x,y)) =& \mathbb{E}_{p(x,y)}[\log \frac{p(x,y)}{q(x,y)}] \\
=& \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \log q(x) - \sum_{x,y} p(x,y) \log q(y) \\
=& \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x} (\sum_{y} p(x,y)) \log q(x) - \sum y(\sum_{x} p(x,y)) \log q(q) \\
=& H(p(x,y)) - H(p(x)) - H(p(y)) + \mathbb{KL}(p(x)||q(x)) + \mathbb{KL}(p(y)||q(y)) \\
=& \text{constant} + KL(p(x)||q(x)) + \mathbb{KL}(p(y)||q(y))
\end{aligned}$$

Thus the optimal approximation is $q(x) = p(x)$ and $q(y) = p(y)$.

We skip the practical part.

21.8 Derivation of the structured mean field updates for FHMM

According to the conclusion from mean-field varitional methods, we have:

$$E(\mathbf{x}_m) = \mathbb{E}_{q/m}[E(\bar{p}(\mathbf{x}_m))]$$

Thus:

$$-\sum_{t=1}^{T}\sum_{k=1}^{K} x_{t,m,k}\tilde{\epsilon}_{t,m,k} = \frac{1}{2}\mathbb{E}[\sum_{t=1}^{\mathrm{T}}(\mathbf{y}_t - \sum_{l\neq m}^{M} W_l\mathbf{x}_{t,m})^T\Sigma^{-1}(\mathbf{y}_t - \sum_{l\neq m}^{M} W_l\mathbf{x}_{t,m})] + C$$

Comparing the coefficient of $x_{t,m,k}$ (i.e. setting $x_{t,m,k}$ to 1) ends in:

$$\tilde{\epsilon}_{t,m,k} = W_m^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}_t - \sum_{l\neq m} W_l\mathbb{E}[\mathbf{x}_{t,l}]) - \frac{1}{2}(W_m^{\mathrm{T}}\Sigma^{-1}W_m)_{k,k}$$

Write into matrix form yields to 21.62.

21.9 Variational EM for binary FA with sigmoid link

Refer to ”Probabilistic Visualisation of High-Dimensional Binary Data, Tipping, 1998”.

21.10 VB for binary FA with probit link

The major difference in using probit link is the uncontinuous likelihood caused by $p(y_i = 1|z_i) = \mathbb{I}(z_i > 0)$. In the context of hiding $\mathbf{X}$, we assume Gaussian prior on $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{Z}$. The approximation takes the form:

$$q(\mathbf{X}, \mathbf{Z}, \mathbf{W}) = \prod_{l=1}^{L} q(\mathbf{w}_l) \prod_{i=1}^{N} q(\mathbf{x}_i)q(z_i)$$

It is a mean-field approximation, hence in an algorithm similari to EM, we are to update the distribution of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ stepwise.

For variable $\mathbf{X}$, we have:

$$\begin{aligned}\log q(\mathbf{x}_i) =& \mathbb{E}_{q(\mathbf{z}_i)q(\mathbf{w})}[\log p(\mathbf{x}_i, \mathbf{w}, z_i, y_i)] \\ =& \mathbb{E}_{q(\mathbf{z}_i)q(\mathbf{w})}[\log p(\mathbf{x}_i) + \log p(\mathbf{w}) + \log p(z_i|\mathbf{w}_i, \mathbf{w}) + \log p(y_i|z_i)]\end{aligned}$$

Given the likelihood form, for i corresponding to $y_i = 1$, $q(z_i)$ have to be a truncated one, i.e. we only consider the expectations in the form $\mathbb{E}[z|z > \mu]$ and $\mathbb{E}[z^2|z > \mu]$.

$\log q(\mathbf{x}_i) = -\frac{1}{2}\mathbf{x}_i^{\mathrm{T}}\Lambda_1\mathbf{x}_i - \frac{1}{2}\mathbb{E}[z^2] - \frac{1}{2}\mathbf{x}_i^{\mathrm{T}}\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\mathbf{x}_i + \mathbb{E}[z]\mathbb{E}[\mathbf{w}]^{\mathrm{T}}\mathbf{x}_i$

Where Λ_1 is the covariance of $\mathbf{x}_i$'s prior distribution, $\mathbb{E}[\mathbf{w}\mathbf{w}^T]$ can be calculated given the Gaussian form of $q(\mathbf{w})$, and truncated expectations $\mathbb{E}[z]$

and $\mathbb{E}[z^2]$ can be obtained from solutions to exercise 11.15. It is obvious that $q(\mathbf{x}_i)$ is a Gaussian.

The update for $\mathbf{w}$ is similar to that for $\mathbf{x}_i$ as long as they play symmetric roles in likelihood. The only difference is we have to sum over i when updating $\mathbf{w}$.

At last we update z_i:

$$\log q(z_i) = \mathbb{E}_{q(\mathbf{x}_i)q(\mathbf{w})}[\log p(z_i|\mathbf{x}_i, \mathbf{w}) + \log p(y_i|z_i)]$$

Inside the expectation we have:

$$-\frac{1}{2}z_i^2 + \mathbb{E}[\mathbf{w}]^{\mathrm{T}}\mathbb{E}[\mathbf{x}]z_i + c$$

Therefore $q(z_i)$ again takes a Gaussian form.