Machine Learning: A Probabilistic Perspective Solution Manual Version 2.1

Fangqi Li,
Shanghai Jiao Tong University,
P. R. China.

Contents

1 Preface: The Second Edition

The tide of artificial intelligence (AI) has swept and reformed so many disciplines and pushed forward the borderline of the state-of-the-art. Such situation has resulted in a positive feedback that drives even more attention and effort into the study and research of AI, together with more unsettled regret and pities.

I have participated in this study, with equal passion that any student who has not formed an exclusive view of the world should have when engaging in a booming subject.

It has been three years and four months since I started the first edition of this solution manual. Although I have received no patron and have no intention of finding one, I received several grateful and advisory messages, from which I felt more pleased than having any of my technical papers published online. For the convenience of them, I would gladly edit this manuscript again, be the time goes back to 2017. In the second edition, I tried to be more concrete in deduction so readers can follow up easier. Some graphical or numerical examples were provided to increase the overall readability. At the beginning part of each chapter/after some exercises I left some remarks, which I thought that could be of help.

The purpose of this manuscript is, as its first edition, to complete the textbook *Machine Learning, A Probabilistic Perspective* as a closed collection of knowledge as far as I could, and to save those who are lost in the ocean of deduction and symbols in ML, whom any talent mind could have become for some times in his/her course with this textbook. I hope that this manuscript can help, be it ever so little, to any reader who purposedly or accidently finds it.

I personally take responsibility for any typo, or mistake in β-reductions.

Fangqi Li,
Shanghai Jiao Tong University,
Shanghai, P.R.China.
January the 4th, 2021.
Contact me by: solour_lfq@sjtu.edu.cn or 1524587011@qq.com.
My homepage is at: https://solour-lfq.github.io/.

第二版序

这篇文档的主体由英文书写，这一方面是因为英文是学术上最泛用的语言，有利于本文档的传播；另一方面是我认为有能力阅读 MLaPP 原教材的中国学生、汉语母语学生基本上也能畅通无阻地阅读本文档中的英文。

希望中国学者和其他可以以中文为语言写作的学者一同努力，提升中文期刊、会议的质量和中文区科研院所的硬实力、影响力，让越来越多的学者乐于用中文叙述自己的观点。

第二版修订了第一版的一些文本问题，补充了第一版在习题推理上比较缺少的理念连接，同时增加了一些示例以提升可读性。文档内的所有排版错误、推导错误由我一人负责。

李方圻

上海交通大学

2021 年 1 月 4 日

邮箱：solour_lfq@sjtu.edu.cn，1524587011@qq.com。

主页：https://solour-lfq.github.io/。

2 Preface: The First Edition

This document provides detailed solutions to almost all exercises in the textbook MLaPP from Chapter One to Chapter Twenty-one. A reader is assumed to find support from this document when he/she is teaching himself/herself an introductory or advanced course with MLaPP.

There are two class for problems in MLaPP: theoretical ones and pratical ones. We provide solution to most theoretical problems. Practical problems, which are based on a Matlab toolbox, are beyond the scope of this document.

I started reading MLaPP after selecting a machine learning course, but I failed to find any free compiled solution manuals. Although several publicly available projects have started working on it, the velocity has been too slow. 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

3 Updating log

22/10/2017 First Chinese compilation.

02/03/2018 English version.

06/01/2020 The second edition begins.

4 Probability

The probability theory for ML is usually a small subset of elementary probability. More involved topics in probability begining from Kolmogorov's theory to martigale and Markov process is usually beyond the scope of an ordinary statistical ML textbook. Readers are encouraged to refer to Shiryaev's *Probability, 3rd edition*, whose first chapter gives a comprehensive summary of elementary probability theory. For supplementary materials, readers can refer to Thomas's *Elements of Information Theory* for a solid introduction.

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

Denote two children by A and B. The space of all experiment results, Ω is composed of:

$$\omega_1 : A \text{ is a girl, }, B \text{ is a girl,}$$

$$\omega_2 : A \text{ is a boy, }, B \text{ is a boy,}$$

$$\omega_3 : A \text{ is a girl, }, B \text{ is a boy,}$$

$$\omega_4 : A \text{ is a boy, }, B \text{ is a girl.}$$

With uniform probability measure. Denote the σ-algebra on Ω as 2^{Ω}.

In question (a), with the knowledge *there is at least one boy*, Ω is modified into:

$$\Omega' = \{\omega_2, \omega_3, \omega_4\}.$$

The event that one child is a girl is $\{\omega_3, \omega_4\}$, whose probability is:

$$\frac{|\{\omega_3, \omega_4\}|}{|\Omega'|} = \frac{2}{3}.$$

In question (b), with the knowledge that A is a boy, the reduced experiment space is:

$$\Omega'' = \{\omega_2, \omega_4\}.$$

Then the probability that B is a girl is:

$$\frac{|\{\omega_4\}|}{|\Omega''|} = \frac{1}{2}.$$

The difference in the form of the question is reflected in that Ω is reduced to different forms and the desired events vary with our questions.

4.2 Legal reasoning

Given the assertation that the criminal has the special blood type, the space of all possibilities contains $800,000 \times \frac{1}{100} = 8,000$ samples. A sample $\omega_i \in \Omega$ denotes that the i-th person with this blood type committed the crime. Let the suspect be the j-th person with this special blood type, the event that he/she was the criminal is

$$\frac{1}{|\Omega|} = \frac{1}{8,000}.$$

For question (a): The probability that an innocent person has this blood type is almost 1%, whose opposite event is *the probability that an innocent person has another blood type*, which is 99%. This event is different from *the suspect is the crimimal*.

For question (b): Justice is not measured by probability. More evidence from forensics might increase this probability to unity or reduce it to zero.

4.3 Variance of a sum

Calculate this straightforwardly:

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

Using the definition of operators var, cov and the linearity of expectation should yield this result easily.

4.4 Bayes rule for medical diagnosis

Let ill and positive denote the event that you are infected and are tested positive for this disease respectively. Let health denote the opposite event

of ill. Apply Bayes's rules:

$$\begin{aligned}\Pr(\text{ill}|\text{positive}) =& \frac{\Pr(\text{ill}, \text{positive})}{\Pr(\text{positive})} \\ =& \frac{\Pr(\text{ill})\Pr(\text{positive}|\text{ill})}{\Pr(\text{ill})\Pr(\text{positive}|\text{ill})+\Pr(\text{health})\Pr(\text{positive}|\text{health})} \\ =& 0.0098\end{aligned}$$

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

The answer is (b). Use prize_i, choose_i, open_i to denote the event that the prize is in/the player chooses/the host opens the i-th box. Apply Bayes's rules:

$$\begin{aligned}\Pr(\text{prize}_1|\text{choose}_1, \text{open}_3) =& \frac{\Pr(\text{choose}_1)\Pr(\text{prize}_1)\Pr(\text{choose}_3|\text{prize}_1, \text{choose}_1)}{\Pr(\text{choose}_1)\Pr(\text{open}_3|\text{choose}_1)} \\ =& \frac{\Pr(\text{prize}_1)\Pr(\text{choose}_3|\text{prize}_1, \text{choose}_1)}{\Pr(\text{open}_3|\text{choose}_1)} \\ =& \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0+\frac{1}{3}\cdot 1} = \frac{1}{3}\end{aligned}$$

In the last step we summarize over the potential location of the prize. This is a classical example of anti-intuition results from probability.

4.6 Conditional Independence

In question (a), we have:

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

Thus the answer is (ii).

For question (b), we have the further decomposition:

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

So both (i) and (ii) and sufficient obviously. Moreover, we have:

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

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

4.7 Pairwise independence does not imply mutual independence

Consider three boolean variables ξ_1, ξ_2, ξ_3, ξ_1 and ξ_2 take values in 0 or 1 with equal possibility independently and $x_3 = \text{XOR}(x_1, x_2)$. It is easy to prove that x_3 is independent with x_1 or x_2, but given both x_1 and x_2, the value of x_3 is determined and thereby the mutual independence fails. For a detailed examination, denote the space of experiment outcomes by

$$\Omega = \{00, 01, 10, 11\},$$

with the first component denote the value of ξ_1, the second is for ξ_2. Then ξ_1 generates the σ-algebra:

$$\mathcal{F}_1 = \{\emptyset, \Omega, \{00, 01\}, \{10, 11\}\}.$$

ξ_2 generates:

$$\mathcal{F}_2 = \{\emptyset, \Omega, \{00, 10\}, \{01, 11\}\}.$$

ξ_3 generates:

$$\mathcal{F}_3 = \{\emptyset, \Omega, \{00, 11\}, \{01, 10\}\}.$$

One can easily check that each pair out of the triplet $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ is a pair of independent σ-algebra, hence meets the pairwise independence.

However, each pair out of the triplet $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ can span the entire 2^Ω. Then we would have counter examples, e.g., consider

$$A = \{00, 01\} \in \mathcal{F}_1,$$

$$B = \{11\} \in \sigma(\xi_2, \xi_3).$$

Then

$$\Pr(A \cap B) = 0 \neq \Pr(A) \cdot \Pr(B) = \frac{1}{8}.$$

Hence the mutual independence does not hold.

This example comes from cryptography. The only theoretical secure (defined by Shannon) encryption system is the one-pad cipher book. With the message denoted by m in binary code, the encryption is done by XOR m with a binary key k drawn from a cipher book. The result is the cipher text

c. From a statistical point of view, c is equally likely to be the ciphertext of any message, hence the adversary cannot break the security (since the cipher text can be equally understood as any message, so no specific message prevails). But the encryption is a deterministic process (by using a stricy one-pad cipher book, this cipher is resistent to chosen-plaintext attack as well), so the mutual independence fails.

4.8 Conditional independence iff joint factorizes

We prove that (2.129) is tantamount to (2.130). One direction is trivial by denoting:

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

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

Conversely, we have:

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

And vice versa,

$$p(y|z) = h(y,z)\sum_x g(x,z).$$

Moreover, for any z:

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

Thus:

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

4.9 Conditional independence

For question (a), the antecedent $(X \perp W|Z, Y)$ means that $\forall x \in \sigma(X)$, $\forall w \in \sigma(W)$ and $\forall v \in \sigma(Y, Z)$ we have

$$\Pr(x \cap w|v) = \Pr(x|v) \cdot \Pr(w|v).$$

The antecedent $(X \perp Y|Z)$ can be translated to that $\forall x \in \sigma(X)$, $\forall y \in \sigma(Y)$ and $\forall z \in \sigma(Z)$,

$$\Pr(x \cap y|z) = \Pr(x|z) \cdot \Pr(y|z).$$

What we desire to obtain is $\forall x \in \sigma(X)$, $\forall w \in \sigma(W)$ and $\forall z \in \sigma(Z)$,

$$\Pr(x \cap w|z) = \Pr(x|z) \cdot \Pr(w|z).$$

This is obviously correct by having v in the first equation taking values in $\sigma(Z)$ solely. Since $\sigma(Z) \subset \sigma(Y, Z)$.

For question (b), we have the premises: $\forall x \in \sigma(X)$, $\forall y \in \sigma(Y)$, $\forall z \in \sigma(Z)$ and $\forall w \in \sigma(W)$:

$$\Pr(x \cap y|z) = \Pr(x|z) \cdot \Pr(y|z).$$

$$\Pr(x \cap y|w) = \Pr(x|w) \cdot \Pr(y|w).$$

The desired result if $\forall v \in \sigma(Z, W)$,

$$\Pr(x \cap y|v) = \Pr(x|v) \cdot \Pr(y|v).$$

Let x and y be two disjoint events, z and w be another pair of disjoints, and none of $x \cap z, x \cap w, y \cap z, y \cap w$ is empty w.r.t. the underlying probability measure. Finally, let $v = w \cup z$. One can then check that the equation above does not hold for this setting, hence the deduction in (b) is false.

In fact, (b) is intuitively false. A straightforward example is a cryptography example: group signature with three participants. The group signature is a protocal for encryption/verfication that ensures a series of security requirements including:

- Any one participant soly cannot pass the verification.
- Two participant can pass the verification.

For example, let Alice and Bob each hold a half of the secret key denoted by Z and W respectively, Y denotes the ciphertext, and X denotes the plaintext. A good group encryption would meet both antecedents in (b) but fails the conclusion. An example of a naive group signature is to use a quadratic function as the secret key:

$$f(t) = x \cdot t^2 + y \cdot t + c.$$

And provide two different points $z = (t_1, f_1)$, $w = (t_2, f_2)$ from f to Alice and Bob. It is obvious that both antecedents in (b) are satisfied (by correctly translating the density) since the value of x yields no information about y. However, given both z and w then y is a determinstic function of x:

$$y = \frac{f_1 - f_2}{t_1 - t_2} - x \cdot (t_1 + t_2),$$

and the independence no longer holds. Note that a third participant is necessary to reveal the entire secret key (x, y, c). In practice, the calculation is usually done on an algebraic field/group, e.g. the elliptic curves, to deal with the problem with the density of x, y, c which is usually not uniform in the real number field.

4.10 Deriving the inverse gamma density

According to the change of variables formula:

$$p(y) = p(x)|\frac{\mathrm{d}x}{\mathrm{d}y}|.$$

We have:

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

The change of variables formula is a simplified version of Lebesgue-Rikdon Theorem, which formally address the transform between probability measures defined on the same space. In the simplified version, we take

the existence of the derivative $\frac{\mathrm{d}x}{\mathrm{d}y}$ for granted. In the general case, such differential is obtained by taking the limit of simple functions that meet the dominance condition. In fact, the Lebesgue Theorem was developed to properly define the *differential of one probability measure w.r.t. another probability measure*. The derived general differential is usually denoted by $\frac{\mathrm{d}\mu_1(x)}{\mathrm{d}\mu_2(x)}$ where μ_i are probability measures.

4.11 Normalization constant for a 1D Gaussian

We have:

$$\begin{aligned} C &= \int_0^{2\pi} \int_0^{\infty} r \cdot \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \mathrm{d}r\mathrm{d}\theta \\ &= 2\pi\sigma^2 \cdot \int_0^{\infty} \exp\{-u\}\,\mathrm{d}u \\ &= 2\pi\sigma^2. \end{aligned}$$

For multivariate Gaussian, the trick is to diagonalize the covariance matrix and integral each components independently.

4.12 Expressing mutual information in terms of entropies

We have:

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

Inversing X and Y yields another formula. One can proceed to show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

4.13 Mutual information for correlated normals

We have:

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

Here we incorporate a comprehensive deduction on (2.138) and (2.139), which shall not be taken for granted. The differential entropy for a 1D-Gaussian with density function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}$$

is

$$\begin{aligned} & -\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\right)\mathrm{d}x \\ &= \ln(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2}\mathbb{E}[x^2] \\ &= \frac{1}{2}\ln(2\pi e\sigma^2). \end{aligned}$$

For the multi-dimensional case, we begin by diagonalizing the covariance matrix/decoupling the components and integrating along each independent component. Under this new set of coordinates $v_1, \cdots, v_d$, the logarithm of the density can be decomposed into

$$C + \sum_{i=1}^{d} -\frac{v_i^2}{2\sigma_i^2},$$

where σ_i^2 is the i-th diagonal component in the transformed covariance matrix. The product of all diagonal components is exactly $\det\Sigma$, hence proving (2.138).

4.14 A measure of correlation

For question (a), we only have to borrow the conclusion from Exercise 2.12.:

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

For question (b), we have $0 \leq r \leq 1$ in question b for $I(X;Y) \geq 0$ and $H(X|Y) \geq 0$.

For question (c), $r = 0$ iff X and Y are independent so the distance between $p(x, y)$ and $p(x) \cdot p(y)$ is zero regarding KL-divergence.

For question (d), $r = 1$ iff X is determined by, but not necessarily equal to, Y.

4.15 MLE minimizes KL divergence to the empirical distribution

Expand the KL divergence:

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

We use the weak law of large numbers in the third step and drop the entropy of empirical distribution, which is independent of θ, in the last step. The other direction of optimization is $\arg\min_{\theta} \{\mathbb{KL}(q(\theta)||p_{\text{emp}})\}$. It contains an expectation term w.r.t. $q(\theta)$ and is harder to solve.

4.16 Mean, mode, variance for the beta distribution

Firstly, we 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 properties of the Gamma function. Finally:

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

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

4.17 Expected value of the minimum

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

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

Therefore:

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

For the second equation, note that:

$$p(m \geq t) = \int_t^{1-t} p(m = t')\mathrm{d}t',$$

therefore

$$\int_0^1 \int_t^{1-t} p(m = t')\mathrm{d}t'\mathrm{d}t = \int_0^1 \int_0^{t'} p(m = t')\mathrm{d}t\mathrm{d}t' = \int_0^1 t \cdot p(m = t')\mathrm{d}t'.$$

There are perhaps more intuitive solutions to this problem, for example, plotting the value of X, Y and $\min(X, Y)$ into one graph: The height of the

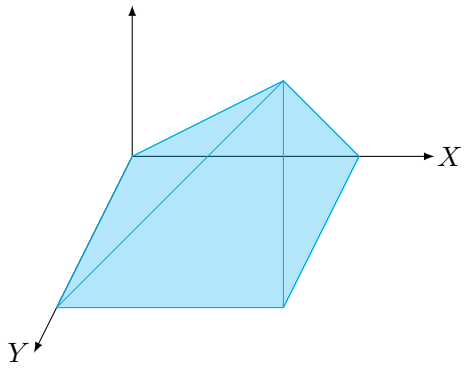


Figure. 1. Exercise 2.17, P1.

cyan pyramid at (x, y) marks the value of $\min(x, y)$, so the expectation of the statistics equals the average height, in this case also the volume of the pyramid, $\frac{1}{3}$. One can also graphically compute the average distance between X and Y from the following plot: Since the average distance between X and Y is $\frac{1}{3}$, so is that between $\min(X, Y)$ and $\max(X, Y)$. Moreover, we have $\mathbb{E}[\min(X, Y) + \max(X, Y)] = \mathbb{E}[X] + \mathbb{E}[Y] = 1$, hence $\min(X, Y) = \frac{1}{3}$.

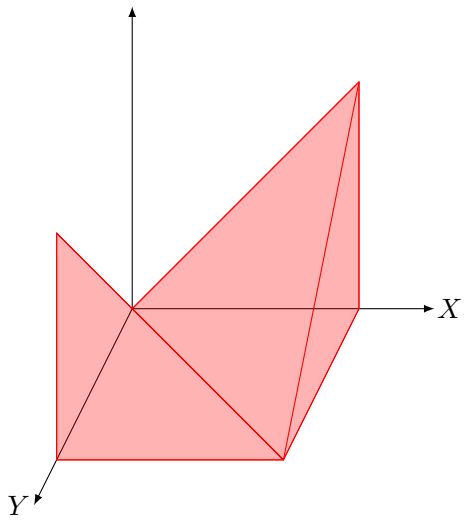


Figure. 2. Exercise 2.17, P2

However, the graphical method, although entertaining and inspiring, should not be considered as an reliable option in proving probability properties. The dependency can complicate the underlying topology (e.g., the $X - Y$ plance might have another geometry other than the Euclidean one, if X and Y are not independent), resulting in confusions and fallacies.

For example, if X is subject to a uniform distribution on $[0, 1]$ while Y is uniformly distributed in $[\max(0, X - 0.2), \min(1, X + 0.2)]$. Then the pyramid in Fig. 1 is left with the region along the diagonal line in X-Y plane. This generalization is insignificant since it only changes the region where the integral shall be done.

Consider the case where X and Y are independent random variables subject to truncated Gaussian centered at $\frac{1}{2}$ on $[0, 1]$. The plot for visualizing $\min(X, Y)$ is the same as Fig. 1. However, in order to compute the expectation of $\min(X, Y)$, one cannot simply calculcate the volume of the pyramid since the geometry of the X-Y plane has changed.

5 Generative models for discrete data

The Bayesian paradigm basically follows the following steps:

- Writing down the likelihood as a function of the parameters to be learned from the formulation of the problem.
- Choosing a corresponding prior distribution from the likelihood function such that the increment of data can be turned into easier operators.
- Writing down the posterior distribution as a function of the hyperparameters of the prior distribution and the observed data.
- If the task is to learn the (distribution of the) parameters: maximizing the posterior density at the observed data w.r.t. the hyperparameters.
- If the task is to predict: integrate out the posterior distribution conditioned on the observed data.

The Bayesian statiscs beginning from this chapter provides an interesting and intuitive perspective into understanding and predicting the world. I learned Bayesian statistics from Bishop's *Pattern Recognition and Machine Learning*.

In conducting Bayesian analysis to examples provided in the exercises, we compute an extra term, the *evidence*. The evidence is the probability that a dataset being generated from a set of hyperparameters, and is usually implicitly absorbed into the nomarlization term of the posterior distribution. The evidence is of practical value in empirical Bayesian, where we manage to select the optimal hyperparameters. In models such as variational inference, the evidence plays an role with even more weight.

5.1 MLE for the Beroulli/binomial model

We begin with (3.11), which is the likelihood function of a collection of the outcomes in a coin-toss experiment $\mathcal{D}$ w.r.t. the parameter θ, the probability of heads:

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

where N_0 and N_1 are the number of tails/heads respectively.

To decompose the differential into term-independent forms, taking logarithm:

$$\ln p(\mathcal{D}|\theta) = N_1 \ln \theta + N_0 \ln(1-\theta).$$

Setting its derivative to zero:

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

yields (3.22):

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

where N is the size of $\mathcal{D}$.

Of course one need not turn to the logarithmic field. Differentiating $p(\mathcal{D}|\theta)$ w.r.t. θ directly gives the same result. But taking logarithm almost always simplifies the form and the deduction procedure.

5.2 Marginal likelihood for the Beta-Bernoulli model

This exercise continues the discussion of the toy coin-toss experiment, so we borrow all symbols from the exercise above. The likelihood takes the form:

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

The prior distribution of θ takes the form:

$$p(\theta|a,b) = \text{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} = C_1(a,b) \cdot \theta^{a-1}(1-\theta)^{b-1},$$

where we adopt $C_1(a,b)$ in the hope of eliminating the ambiguity of using $\propto$, which, although simplifies the symbolization, results in countless errors.

The posterior distribution takes the form:

$$\begin{aligned} p(\theta|\mathcal{D},a,b) =& \frac{p(\theta|a,b) \cdot p(\mathcal{D}|\theta,a,b)}{p(\mathcal{D}|a,b)} \\ =& \frac{p(\theta|a,b) \cdot p(\mathcal{D}|\theta)}{p(\mathcal{D}|a,b)} \\ =& \frac{C_1(a,b)}{p(\mathcal{D}|a,b)} \cdot \theta^{N_1+a-1} \cdot (1-\theta)^{N_0+b-1}. \end{aligned}$$

The first step is the straightforward Bayesian rule, the second is the Markov property. In the last step we adopt the equations before. Since $p(\theta|\mathcal{D}, a, b)$ should be normalized w.r.t. θ, it has to be a Beta distribution with hyperparameters $N_1 + a, N_0 + b$. We can now derive the *evidence* of $\mathcal{D}$ w.r.t. a and b explicitly. The normalization of $p(\theta|\mathcal{D}, a, b)$ indicates that:

$$\frac{C_1(a,b)}{p(\mathcal{D}|a,b)} = C_1(N_1 + a, N_0 + b),$$

so:

$$p(\mathcal{D}|a,b) = \frac{C_1(a,b)}{C_1(N_1 + a, N_0 + b)},$$

where $C_1(\cdot,\cdot)$ is the normalization factor for the Beta distribution. This is enough for deriving (3.80) by recalling the normalization of Beta distribution. The value of $p(\mathcal{D}|a,b)$ can help us select proper hyperparametes.

As for prediction:

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

The first step is Bayesian rule, the second is Markov property. The rest is straightforward algebra.

Concretely, we calcualte $p(\mathcal{D})$ where $\mathcal{D} = \{1, 0, 0, 1, 1\}$:

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

Rename the variables $\alpha = a + b, \alpha_1 = a, \alpha_0 = b$, we have (3.83). To derive (3.80), we make use of:

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

5.3 Posterior predictive for Beta-Binomial model

Straightforward algebra (recall (2.61)):

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

The hint provided in the textbook is incorrect by mistaking

$$\Gamma(a) = (a-1) \cdot \Gamma(a-1)$$

for

$$\Gamma(a) = a \cdot \Gamma(a-1).$$

5.4 Beta updating from censored likelihood

The derivation is straightforward:

$$\begin{aligned}p(\theta, X<3) =& p(\theta) \cdot p(X<3|\theta) \\ =& p(\theta) \cdot \left(\sum_{i=0}^{2} p(X=i|\theta) \right) \\ =& \mathrm{Beta}(\theta|1,1) \cdot \left(\sum_{i=0}^{2} \mathrm{Bin}(i|5,\theta) \right),\end{aligned}$$

with

$$\mathrm{Bin}(m|n,\theta) = \binom{n}{m} \cdot \theta^m \cdot (1-\theta)^{n-m}$$

is the probability that m heads appear in n times of experimens with the probability of head θ. The posterior distribution over θ in this case becomes much more involved.

5.5 Uninformative prior for log-odds ratio

Since:

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

By using change of variables formula:

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

hence

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

That is to say, we can generate samples subject to a Beta distribution by transforming samples drawn from a uniform distribution. This trick is of significant practical value. For a direction sampling from Beta distribution requires inversing the cumulative probability function of it, which involves too much computation.

5.6 MLE for the Poisson distribution

The Poisson distribution plays a central role in stochastic process, e.g., the queueing theory. If data are assumed to be generated from a similar process then the Bayesian analysis of the Poisson distribution derived in this exercise and the next can be applied directly. The likelihood of data for a Poisson distribution is (assuming i.i.d.):

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

Setting the derivative of the likelihood w.r.t. λ to zero:

$$\frac{\partial}{\partial\lambda} p(\mathcal{D}|\lambda) = \frac{\exp(-\lambda N) \cdot \lambda^{(\sum_{n=1}^{N} x_n)-1}}{\prod_{n=1}^{N} x_n!} \left\{ -N\lambda + \sum_{n=1}^{N} x_n \right\}.$$

Thus:

$$\lambda_{\mathrm{MLE}} = \frac{\sum_{n=1}^{N} x_n}{N}.$$

The formulation could be made easier by taking logarithm (since the Poisson distribution can be considered an element of the exponential family as well):

$$\log p(\mathcal{D}|\lambda) = -\lambda \cdot N + \left(\sum_{n=1}^{N} x_n \right) \cdot \log \lambda,$$

where we have omitted the term independent of λ.

5.7 Bayesian analysis of the Poisson distribution

The conjugate prior for the Poisson distribution is the Gamma distribution:

$$\mathrm{Ga}(\lambda|a,b) = \frac{b^a}{\Gamma(a)} \cdot \lambda^{a-1} \cdot \exp(-\lambda \cdot b).$$

The posterior for a Bayesian Poisson model reads:

$$\begin{aligned} p(\lambda|\mathcal{D},a,b) =& \frac{p(\mathcal{D}|\lambda) \cdot p(\lambda|a,b)}{p(\mathcal{D}|a,b)} \\ =& \frac{b^a}{\prod_{n=1}^N x_n! \cdot p(\mathcal{D}|a,b) \cdot \Gamma(a)} \cdot \lambda^{a+\sum_{n=1}^N x_n - 1} \cdot \exp\left(-\lambda \cdot (N+b)\right) \\ =& \mathrm{Ga}(a + \sum_{n=1}^N x_n, N+b), \end{aligned}$$

in which the last step follows the normalization condition. We now have the evidence:

$$p(\mathcal{D}|a,b) = \frac{b^a \cdot \Gamma(a + \sum_{n=1}^N x_n)}{\prod_{n=1}^N x_n! \cdot (N+b)^{a+\sum_{n=1}^N x_n} \cdot \Gamma(a)}.$$

Finally, be a and b approximate zero, the posterior mean approaches $\frac{\sum_{n=1}^N x_n}{N}$, the same as the MLE, i.e., $a = 0, b = 0$ is a non-informative prior. One should note that this property does not hold for all Bayesian analysis, setting all hyperparameters to zero does not necessarily gracefully degenerate the posterior mean to the MLE. Since the names and definitions of those symbols might differ.

5.8 MLE for the uniform distribution

The Bayesian analysis for the uniform distribution seems to be of less significance since uniform distribution appears to appear less frequently than other continuous distributions. But the exercises remain good introductory examples.

The likelihood for the uniform distribution is a truncated function, whose domain is $[-a, a]$, so we must have $a \geq \max_i \{|x_i| \in \mathcal{D}\}$. Then the likelihood lookes like:

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

or generally:

$$p(\mathcal{D}|a) = \mathbb{I}[a \geq \max_i \{|x_i| \in \mathcal{D}\}] \cdot (2a)^{-n}.$$

For question (a), in order to maximize this value with $a \geq \max_n \{|x_n| \in \mathcal{D}\}$, the outcome is:

$$a_{\text{MLE}} = \max_i \{|x_i| \in \mathcal{D}\}.$$

For question (b), if $|x_{n+1}| > \max_{i=1}^n \{|x_i|\}$ then $p(x_{n+1})$ is zero. Otherwise the probability is $\frac{1}{2 \cdot a_{\text{MLE}}}$.

For question (c), we believe that MLE for the uniform distribution is *not fluent enough* since when x_{n+1} passes $\pm \max_{i=1}^n \{|x_i|\}$, the predicted probability drops as a step function, which is undesired for a continuous distribution.

5.9 Bayesian analysis of the uniform distribution

The conjugate prior for uniform distribution is the Pareto distribution, whose density function is defined by:

$$p(\theta|K, b) = \text{Pa}(\theta|K, b) = K \cdot b^K \cdot \theta^{-(K+1)} \cdot \mathbb{I}[\theta \geq b].$$

Let $m = \max \{|x_i|\}_{i=1}^n$, the joint distribution of θ and $\mathcal{D}$ is:

$$\begin{aligned} p(\theta, \mathcal{D}|K, b) =& p(\theta|K, b) \cdot p(\mathcal{D}|\theta) \\ =& K \cdot b^K \cdot \theta^{-(K+1)} \cdot \mathbb{I}[\theta \geq b] \cdot \mathbb{I}[\theta \geq m] \cdot (\theta)^{-n} \\ =& K \cdot b^K \cdot \theta^{-(K+n+1)} \cdot \mathbb{I}[\theta \geq \max(b, m)]. \end{aligned}$$

Now $p(\theta, \mathcal{D}|K, b) = p(\mathcal{D}|K, b) \cdot p(\theta|\mathcal{D}, K, b)$, hence the posterior distribution depends on θ through:

$$\theta^{-(K+n+1)} \cdot \mathbb{I}[\theta \geq \max(b, m)].$$

So the posterior distribution is another Pareto distribution with hyperparameters $K + n, \max(b, \max \{|x_i|\}_{i=1}^n)$.

The evidence is computed from the Bayesian rule:

$$\begin{aligned} p(\mathcal{D}|K, b) &= \int_0^\infty p(\mathcal{D}, \theta|K, b) \mathrm{d}\theta \\ &= \int_{\max(b,m)}^\infty \frac{K \cdot b^K}{\theta^{K+n+1}} \mathrm{d}\theta. \end{aligned}$$

The rest is trivial calculus.

5.10 Taxicab problem

Some similar entertaining problems are *guessing the number of piano tuners from the average time for a tuner to arrive in one guest's house*, etc.

For question (a), we begin with hyperparameters $K = 0$, $b = 0$, which is improper since the Pareto distribution cannot normalize. With $\mathcal{D} = \{100\}$, we have the posterior distribution another Pareto distribution with $K = 1$ and $b = 100$, i.e.,

$$p(\theta|\mathcal{D}) = \frac{100}{\theta^2} \cdot \mathbb{I}[\theta \geq 100].$$

For question (b), we firstly derive the distribution of the taxi index:

$$\begin{aligned} p(x|\mathcal{D}, K, b) &= \int_0^\infty p(x, \theta) \mathrm{d}\theta \\ &= \int_0^\infty p(x|\theta) \cdot p(\theta|\mathcal{D}, K, b) \mathrm{d}\theta \\ &= \int_{100}^\infty \mathbb{I}[x \leq \theta] \cdot \frac{1}{\theta} \cdot \frac{100}{\theta^2} \mathrm{d}\theta \\ &= \int_{\max(x,100)}^\infty \frac{100}{\theta^3} \mathrm{d}\theta \\ &= 50 \cdot \max(x, 100)^{-2}, \end{aligned}$$

whose plots looks very much similar to that of electrical potential along an axis that penetrates the center of a conductor sphere with radius 100, through declines exponentially faster.

The posterior mode of x is any number in $[0, 100]$.

The posterior mean of x is:

$$\mathbb{E}(x) = \sum_{x=0}^{100} \frac{x}{200} + \sum_{x=100}^{\infty} \frac{50}{x},$$

whose second term diverges, so the posterior mean does not exist.

The posterior median is 99.5, since:

$$\sum_{x=0}^{99} \frac{1}{200} < 0.5 < \sum_{x=0}^{100} \frac{1}{200}.$$

Question (c) is identical to (b), as we have adopted a Bayesian treatment for (b).

For question (d), we have:

$$p(x = 100|\mathcal{D}, K, b) = \frac{1}{200},$$
$$p(x = 50|\mathcal{D}, K, b) = \frac{1}{200},$$
$$p(x = 150|\mathcal{D}, K, b) = \frac{1}{450}.$$

For question (e), we might adopt better K and b with expert knowledge and collect more samples.

5.11 Bayesian analysis of the exponential distribution

The exponential distribution is also crucial for the queueing theory. The log-likelihood for an exponential distribution with density:

$$p(x|\theta) = \theta \cdot \exp(-\theta \cdot x)$$

is:

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

whose derivative is:

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

Thus for question (a), we have:

$$\theta_{\text{MLE}} = \frac{N}{\sum_{n=1}^{N} x_n}.$$

For question (b), $\theta_{\text{MLE}} = 5$.

For question (c), we begin with an exponential prior distribution:

$$p(\theta|\lambda) = \lambda \cdot \exp(-\lambda \cdot \theta),$$

whose expectation is:

$$\int_0^{\infty} \lambda \cdot \theta \cdot \exp(-\lambda \cdot \theta) \mathrm{d}\theta.$$

Integration by parts (or resort to the nomarlization term of the Gamma distribution) yields:

$$\mathbb{E}(\theta) = \frac{1}{\lambda}.$$

So $\hat{\lambda} = 3$.

For question (d), the posterior distribution is:

$$\begin{aligned} p(\theta|\mathcal{D}, \lambda) &= \frac{p(\mathcal{D}|\theta) \cdot p(\theta|\lambda)}{p(\mathcal{D}|\lambda)} \\ &= \frac{1}{p(\mathcal{D}|\lambda)} \cdot \theta^N \cdot \exp(-\theta \cdot \sum_{n=1}^{N} x_n) \cdot \lambda \cdot \exp(-\lambda \cdot \theta) \\ &= \frac{\theta^N \cdot \lambda}{p(\mathcal{D}|\lambda)} \cdot \exp\left(-\theta \cdot \left(\lambda + \sum_{n=1}^{N} x_n\right)\right). \end{aligned}$$

Hence the posterior is a Gamma distribution with hyperparameters:

$$a = N + 1,$$
$$b = \lambda + \sum_{n=1}^{N} x_n.$$

The evidence is given by: $\frac{\lambda \cdot \Gamma(a)}{b^a}$, a function of λ and $\mathcal{D}$. Hence the exponential distribution is not the conjugate distribution of itself, answering question (e).

For question (f), the posterior mean is the mean of the Gamma distribution:

$$\frac{a}{b} = \frac{N + 1}{\lambda + \sum_{n=1}^{N} x_n}.$$

Compared with the MLE, the posterior mean has additional terms for both the numerator and the denominator as basic knowledge when N is relatively small. The influence of using this prior is tantamount to introducing a prior sample with value λ.

5.12 MAP estimation for the Bernoulli with non-conjugate priors

For question (a), we adopt the different prior:

$$p(\theta) = \begin{cases} 0.5, \text{ if } \theta = 0.5, \\ 0.5, \text{ if } \theta = 0.4, \end{cases}$$

The posterior distribution now reads:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \cdot p(\theta)}{p(\mathcal{D})},$$

whose support is $\{0.4, 0.5\}$, so the MAP is:

$$\max_{\theta \in \{0.4, 0.5\}} \left\{ \theta^{N_1} \cdot (1-\theta)^{N_0} \right\}.$$

For question (b), it is intuitive that the non-conjugate has better performance when N is small. But the conjugate Bayesian method prevails with N grows. For a solid verification, consider the case where N is large, the probability that $\frac{N_1}{N}$ deviates ϵ from 0.41 can be bounded by the Chernoff bounding. Let $\{\xi_n\}_{n=1}^N$ be a collection of i.i.d. Bernoulli random variables with distribution:

$$\xi_n = \begin{cases} 1, & \text{with probability } 0.41, \\ 0, & \text{with probability } 0.59, \end{cases}$$

Denote $X = \sum_{n=1}^N \xi_n$ as the random variable marks their summation. Then:

$$\begin{aligned} \Pr(X \geq N \cdot (0.41 + \epsilon)) &= \Pr(\mathrm{e}^{\lambda \cdot X} \geq \mathrm{e}^{N\lambda(0.41+\epsilon)}) \\ &\leq \frac{\mathbb{E}[\mathrm{e}^{\lambda \cdot X}]}{\mathrm{e}^{N\lambda(0.41+\epsilon)}} \\ &= \frac{\left(\mathbb{E}[\mathrm{e}^{\lambda \cdot \xi_0}]\right)^N}{\mathrm{e}^{N\lambda(0.41+\epsilon)}} \\ &= \frac{\left(\mathbb{E}[\mathrm{e}^{\lambda \cdot \xi_0}]\right)^N}{\mathrm{e}^{N\lambda(0.41+\epsilon)}} \\ &= \frac{\left(0.41\mathrm{e}^{\lambda} + 0.59\right)^N}{\mathrm{e}^{N\lambda(0.41+\epsilon)}}. \end{aligned}$$

Where λ can be an arbitrary positive number. The proabability that X exceeds $N \cdot (0.41 + \epsilon)$, denoted by P_1 is bounded by the lower bound of the last line in the deduction above. With $N = 1000$, $\epsilon = 0.05$, the probability is numerically bounded by 0.00602. The other side of error $\Pr(X \leq N \cdot (0.41 - \epsilon))$, whose probability is P_2, can be derived in a similar way. The probability that the Bayesian way is dominated by the non-uniform prior is no higher than $P_1 + P_2$. Taking $\epsilon \to 0.1$ and $N \to \infty$, this bound remains negligable.

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

The likelihood for Dirichlet-multinomial model is:

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{N_k^{\text{old}}},$$

following the symbols defined in the textbook. The conjugate prior is the Dirichlet distribution:

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \cdot \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where θ is a K-dimension simplex. The (3.37) in the textbook mistake θ for $\mathbf{x}$.

The posterior distribution is another Dirichlet distribution with update:

$$\alpha_k + N_k^{\text{old}} \leftarrow \alpha_k.$$

To predict a new batch of data $\tilde{\mathcal{D}}$, we begin with one sample $x \in \tilde{\mathcal{D}}$:

$$\begin{aligned} p(x = k|\mathcal{D}, \alpha) &= \int_\theta p(x = k|\theta) \cdot p(\theta|\mathcal{D}, \alpha) \mathrm{d}\theta \\ &= \mathbb{E}_{\text{Dir}}[\theta_k], \end{aligned}$$

where the expectation is computed w.r.t. the posterior Dirichlet distribution, hence is:

$$\frac{\alpha_k + N_k^{\text{old}}}{\sum_{t=1}^{K} \alpha_t + N_t^{\text{old}}}.$$

Finally,

$$\begin{aligned} p(\tilde{\mathcal{D}}|\mathcal{D}, \alpha) &= \prod_{x \in \tilde{\mathcal{D}}} p(x|\mathcal{D}, \alpha) \\ &= \prod_{k=1}^{K} \left(\frac{\alpha_k + N_k^{\text{old}}}{\sum_{t=1}^{K} \alpha_t + N_t^{\text{old}}} \right)^{N_k^{\text{new}}}. \end{aligned}$$

5.14 Posterior predictive for Dirichlet-multinomial

For question (a). In this concrete case we have $K = 27$, $N = 2,000$, any component of α be 10, and $N_{\text{e}}^{\text{old}} = 260$. To derive $p(x_{2001} = \text{e}|\mathcal{D})$, we

resort to the deduction in Exercise 3.13:

$$p(x_{2001} = \mathrm{e}|\mathcal{D}) = \frac{10 + 260}{10 \times 27 + 2000} \approx 0.1189.$$

For question (b), the independence between characters is still ignored, bringing no significant change to the computation:

$$p(x_{2001} = \mathrm{p}, x_{2002} = \mathrm{a}|\mathcal{D}) = \frac{10 + 87}{2270} \cdot \frac{10 + 100}{2270} \approx 0.0021.$$

5.15 Setting the hyper-parameters I

Solve for:

$$\frac{\alpha_1}{\alpha_1 + \alpha_2} = m,$$
$$\frac{\alpha_1 \cdot \alpha_2}{(\alpha_1 + \alpha_2)^2 \cdot (\alpha_1 + \alpha_2 + 1)} = v.$$

We have:

$$\alpha_2 = \frac{m \cdot (1 - m)^2}{v} + m - 1,$$
$$\alpha_1 = \alpha_2 \cdot \frac{m}{1 - m}.$$

5.16 Setting the beta hyper-parameters II

```
import math
m=0.15
l=0.05
u=0.3
MC=1000
delta=(u-l)/MC
def pm(a2):
    a1=a2*m/(1-m)
    pivot=l
    mass=0
    B=math.gamma(a1)*math.gamma(a2)/math.gamma(a1+a2)
    for i in range(MC):
        pivot=pivot+delta
        mass=mass+pivot**(a1-1)*(1-pivot)**(a2-1)
    mass=mass*delta/B
    return mass
```

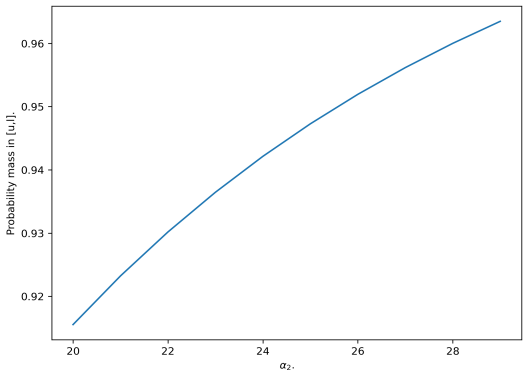


Figure. 3. Exercise 3.16.

The result of which is better demonstrated through the graph: So the optimal choice is $\alpha = 26$. This is tantamount to adopt 32 extra samples.

5.17 Marginal likelihood for beta-binomial under uniform prior

The marginal likelihood is given by:

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

Plug in:

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

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

Thus:

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

Where B is the regulizer for a Beta distribution:

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

The physics behind this setting is that if no prior information is introduced then all $N+1$ possibilities are equal likely to appear.

5.18 Bayes factor for coin tossing

The Bayes factor for hypothesis test is defined by:

$$\mathrm{BF}_{1,0} = \frac{p(\mathrm{data}|H_1)}{p(\mathrm{data}|H_0)},$$

where H_0 is the null hypothesis.

We have:

$$p(\mathrm{data}|H_0) = \mathrm{Bin}(9|0.5,10) = \binom{10}{9}\cdot 0.5^{10} \approx 0.00977.$$

And

$$p(\mathrm{data}|H_1) = \frac{1}{10+1} \approx 0.09091,$$

according to Exercise 3.17. The Bayes factor is approximately 9.3.

When $N = 100$ and $N_1 = 90$, the Bayes factor is:

$$\frac{\frac{1}{100+1}}{\binom{100}{90}\cdot 0.5^{100}} > \frac{2^{100}}{101^{11}} \approx 113622530.$$

When $\frac{N}{N_1}$ remains a constant deviated from 0.5, the larger N is, the more likely that the coin is biased. This is an intuitive conclusion from the law of large numbers.

5.19 Irrelevant features with naive Bayes

The log-likelihood is defined by:

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

In a succint way:

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

where:

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

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

For question (a):

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

For question (b), with:

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

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

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

Hence if:

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

then it cannot affect the classification decision. That is to say, w appear in class 1 and 2 with the same frequency.

For 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 are different when $N_1 \neq N_2$ so the bias effect remains. However, this bias reduces when N grows large.

For question (d), using information theory would be a solid option.

5.20 Class conditional densities for binary data

For 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 in this case is:

$$C \cdot \sum_{i=0}^{D-1} 2^i = C \cdot (2^{D+1} - 2) = \mathcal{O}(C \cdot 2^D).$$

For question (b) and (c), the overfitting is generally assmued to decline when N grows. The dependence within the variables generally hinder generalization when N is small, but it could correctly capture the dependence, be it exist.

For question (d), fitting each parameter for the naive Bayes model requires averaging all components in the samples belong to one class, which is of complexity order $\mathcal{O}(N)$. The total complexity is thus of order $\mathcal{O}(C \cdot D \cdot N)$. For the full Bayes classifier, the complexity is of order $\mathcal{O}(C \cdot 2^D \cdot N)$.

For question (e), the cost of inference in naive Bayes model is $\mathcal{O}(C \cdot D)$ for each sample. While that for full Bayes model is $\mathcal{O}(C \cdot 2^D)$.

For question (f), we have:

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

where we assumed a uniform prior on all classes since it is not the bottleneck of complexity. The complexity for the naive model is then:

$$\mathcal{O}(C \cdot v \cdot 2^h),$$

while that for the full model is:

$$\mathcal{O}(C \cdot 2^v) \cdot 2^h).$$

5.21 Mutual information for naive Bayes classifiers with binary features

By definition:

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

For binary features, the value of x_j is either zero or one. Given $\pi_c = p(y = c), \theta_{jc} = p(x_j = 1|y = c), \theta_j = p(x_j = 1)$, we have the mutual information between x_j and Y be:

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

which ends in (3.76).

5.22 Fitting a naive Bayesian spam filter by hand

θ_{spam} is $\frac{3}{3+4}$.
$\theta_{\text{secret|spam}}$ is $\frac{2}{3}$.
$\theta_{\text{secret|non-spam}}$ is $\frac{1}{4}$.
$\theta_{\text{sports|non-spam}}$ is $\frac{1}{2}$.
$\theta_{\text{dollar|spam}}$ is $\frac{1}{3}$.

6 Gaussian models

Though being a family of models for continuous variables, traditional Gaussian models do not cast a more important role than models covered in the chapter before. Since in most scenarios, the assumption that the data are subject to a normal distribution is unrealistic.

However, Gaussian models are crucial for latent space analysis. Even for complex objects such as image or video, the assumption that they can be encoded into features under a Gaussian distribution turns out to be reliable. Moreover, Gaussian models pave the way to general variational inference and Gaussian process, an important kernel machine.

For the reason above one cannot overestimate the significance of Gaussian models, be their history so long. Most of the mathematical difficulties with Gaussian model have been covered in sections (marked with stars) in the textbook.

6.1 Uncorrelated does not imply independent

The mean for Y is:

$$\int_{-1}^{1} X^2 \mathrm{d}X = \mathbb{E}[X^2].$$

Calculate the covariance of X and Y:

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

whose value is zero since we are intergrating an odd function in range $[-1, 1]$, hence:

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

Independence is a much stronger condition than uncorrelation. The former exerts constrains on the σ-algebra that random variables generates while the later only regulates the value of the expectation of a new random

variable. Decomposition $p(X,Y) = p(X) \cdot p(Y)$ is sufficient for reducing the covariance to zero, but not necessary.

6.2 Uncorrelated and Gaussian does not imply independent unless jointly Gaussian

For question (a). The p.d.f. for Y is:

$$p(Y \in [a, a+\mathrm{d}a]) = 0.5 \cdot p(X \in [a, a+\mathrm{d}a]) + 0.5 \cdot p(X \in [-a-\mathrm{d}a], -a) = p(X \in [a, a+\mathrm{d}a]),$$

since X is symmetric. So Y subject to a normal distribution $(0, 1)$.

For question (b), we have:

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

So they are uncorrelated.

To disprove dependence (in case of confusion), let:

$$a = \Phi^{-1}(\frac{1}{4}),$$

where Φ is the c.d.f of X, i.e.:

$$\int_{-\infty}^{a} \mathcal{N}(x|0,1)\mathrm{d}x = \frac{1}{4}.$$

Let $R_1 = (-\infty, a]$, $R_2 = (a, 0]$. The space of experiment results for $X \times Y$ is $\mathcal{R}^2$. Let $R_1 \times R_2$ be a Borel set in $\mathcal{R}^2$. Be X and Y independent, its probability measure should be $\frac{1}{16}$. However, when $X \in R_1$, it is impossible for Y to take a value from R_2. Hence the independency fails.

The rule of iterated expectation is but the Bayes rule:

$$\begin{aligned}\mathbb{E}[XY] &= \int_X \int_Y XY \cdot p(X,Y)\mathrm{d}X\mathrm{d}Y \\ &= \int_X \int_Y XY \cdot \left(\int_W p(X,Y,W)\mathrm{d}W \right) \mathrm{d}X\mathrm{d}Y \\ &= \int_W \left(\int_X \int_Y XY \cdot p(X,Y|W) \right) p(W)\mathrm{d}W.\end{aligned}$$

6.3 Correlation coefficient is between -1 and 1

With out loss of generality, assume $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. The statement:

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

is equal to

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

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. Let

$$g(t) = t^2 \cdot \mathrm{var}(X) + 2t \cdot \mathrm{var}(X, Y) + \mathrm{var}Y = \mathbb{E}[(tX + Y)^2] \geq 0.$$

Taking g's discriminator finishes the proof.

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

When $Y = aX + b$:

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

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

Therefore:

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

Meanwhile:

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

These two are sufficient to derive:

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

6.5 Normalization constant for a multidimensional Gaussian

Assume $\mu = \mathbf{0}$ w.l.o.g. If the covariance matrix already takes a diagonal form:

$$\Sigma = \begin{pmatrix} \lambda_1^{-1} & \cdots \\ \cdots & \cdots \\ \cdots & \lambda_d^{-1} \end{pmatrix},$$

then:

$$\begin{aligned} \int \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\Sigma\mathbf{x}\right)\mathrm{d}\mathbf{x} &= \int \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{d}\frac{x_i^2}{\lambda_i}\right)\right)\mathrm{d}\mathbf{x} \\ &= \prod_{i=1}^{d}\int \exp\left(-\frac{x_i^2}{2\cdot\lambda_i}\right)\mathrm{d}x_i \\ &= (2\pi\lambda_i)^{-\frac{d}{2}}. \end{aligned}$$

Plugging in $|\Sigma| = \prod_{i=1}^{d}\lambda_i^{-1}$ yields the desired normalization constant. In the second equation, using the distribution law (though somewhat intimidating).

For the general case, we begin by diagonalizing Σ into:

$$\Sigma = U^{\mathrm{T}}\Lambda U,$$

where Λ is a diagonal matrix with components $\lambda_1^{-1}\cdots\lambda_d^{-1}$ and U is a orthogonal matrix. The integral now becomes:

$$\int \exp\left(-\frac{1}{2}\left(U\mathbf{x}\right)^{\mathrm{T}}\Lambda\left(U\mathbf{x}\right)\right)\mathrm{d}\mathbf{x}.$$

Since $|U| = 1$ uniformly, we can directly rewrite the integral into:

$$\int \exp\left(-\frac{1}{2}\mathbf{u}^{\mathrm{T}}\Lambda\mathbf{u}\right)\mathrm{d}\mathbf{u}.$$

The rest is repeating the diagonal case.

6.6 Bivariate Gaussian

We have:

$$p(x_1, x_2) = \frac{1}{2\pi\sqrt{|\Sigma|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma(\mathbf{x}-\mu)\right)$$
$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right).$$

6.7 Conditioning a bivariate Gaussian

For question (a), we begin with the form from Exercise 4.6.:

$$p(x_1, x_2) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}.$$

By the Bayes rule:

$$p(x_2|x_1) = \frac{p(x_1, x_2)}{p(x_1)}.$$

So we only need to compute:

$$p(x_1) = \int p(x_1, x_2)\mathrm{d}x_2.$$

This can be done by *completing the square* inside $p(x_1, x_2)$ w.r.t. x_2:

$$2\pi\sigma_1\sigma_2\sqrt{1-\rho^2} \cdot p(x_1, x_2) = \exp\left(-\frac{1}{2(1-\rho^2)}\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right)$$
$$\cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{\rho^2(x_1-\mu_1)^2}{\sigma_1^2}\right)\right)$$
$$\cdot \exp\left(\frac{1}{2(1-\rho^2)}\frac{\rho^2(x_1-\mu_1)^2}{\sigma_1^2}\right)$$
$$= \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)$$
$$\cdot \exp\left(-\frac{1}{2\sigma_2^2(1-\rho^2)}\left((x_2-\mu_2) - \frac{\sigma_2\rho(x_1-\mu_1)}{\sigma_1}\right)^2\right).$$

Now we are ready to perform the integrating:

$$\int p(x_1, x_2)\mathrm{d}x_2 = \frac{\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \int \exp\left(-\frac{1}{2\sigma^2}(x_2-\mu)^2\right)\mathrm{d}x_2$$
$$= \frac{\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \sqrt{2\pi\sigma_2^2(1-\rho^2)},$$

where

$$\sigma^2 = \sigma_2^2(1-\rho^2),$$
$$\mu = \mu_2 + \frac{\sigma_2 \rho (x_1 - \mu_1)}{\sigma_1}.$$

Finally,

$$p(x_2|x_1) = \frac{\exp\left(\frac{(x_1-\mu_1)^2}{2\sigma^1}\right) \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right)}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}}.$$

For question (b), we can further simplify the numerator of $p(x_2|x_1)$ into:

$$p(x_2|x_1) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left(\rho(x_1-\mu_1) - (x_2-\mu_2)\right)^2\right)}{\sqrt{2\pi(1-\rho^2)}}.$$

6.8 Whitening vs standardizing

Standardizing is a gold standard step for complex data, e.g., the batch normalization layer. Whitening, although fancy, is harder to carry out since it involves solving an eigen problem.

6.9 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:

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

where we have dropped terms independent from μ and 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 w.r.t. μ and set it to zero, we have:

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

The conjugate prior of this model must have a form proportional to $\exp\{A \cdot \mu^2 + B \cdot \mu\}$, so it is 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:

$$\mu_{\text{MAP}} \to \mu_{\text{ML}}.$$

The posterior distribution is another normal distribution, with:

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

For non-informative prior, we have $a = b = 0$ so $p(\mu|a,b)$ is uniform in the domain, then the MAP estimation is the same as MLE.

6.10 Derivation of information form formulae for marginalizing and conditioning

Plugging (4.93) and (4.95) into proven lines of (4.69) yields the information form formulae.

6.11 Derivation of the NIW posterior

We begin with the likelihood for a MVN:

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

By (4.195), which can be proven by:

$$\begin{aligned}\sum_{n=1}^{N}(\mathbf{x}_i-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\mu) =& \sum_{n=1}^{N}(\bar{\mathbf{x}}-\mu+(\mathbf{x}_i-\bar{\mathbf{x}}))^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu+(\mathbf{x}_i-\bar{\mathbf{x}}))\\ =&N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\sum_{n=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})\\ =&N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\mathrm{tr}\left\{\Sigma^{-1}\sum_{n=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}}\right\}\\ =&N(\bar{\mathbf{x}}-\mu)^{\mathrm{T}}\Sigma^{-1}(\bar{\mathbf{x}}-\mu)+\mathrm{tr}\left\{\Sigma^{-1}\mathbf{S}_{\bar{\mathbf{x}}}\right\},\end{aligned}$$

where we have used the fact that $\mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Z})=\mathrm{tr}(\mathbf{Z}\mathbf{Y}^{\mathrm{T}})$, with $\mathbf{Y}$ the shifted design matrix and $\mathbf{Z}=\Sigma^{-1}\mathbf{Y}$.

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

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

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

Hence the posterior reads (where we have omitted the condition on hyper-parameters):

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

where $v_{\mathbf{X}}$, $k_{\mathbf{X}}$, $\mathbf{m}_{\mathbf{X}}$ and $\mathbf{S}_{\mathbf{X}}$ are variables whose values are to be decided. Only terms that dependent on μ and Σ can explicitly enter the terms on the r.h.s.

Firstly, by comparing the exponential for $|\Sigma|$, we have:

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

Secondly, compare the coefficient for the term $\mu^{\mathrm{T}}\Sigma^{-1}\mu$ inside the exponential and we have:

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

Thirdly, check the coefficient for μ^{T} so we have:

$$N\Sigma^{-1}\bar{\mathbf{x}}+k_0\Sigma^{-1}\mathbf{m}_0=k_{\mathbf{X}}\Sigma^{-1}\mathbf{m}_{\mathbf{X}},$$

therefore:

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

Finally, recall that for an arbitrary column vector A:

$$A^{\mathrm{T}}\Sigma^{-1}A = \mathrm{tr}(A^{\mathrm{T}}\Sigma^{-1}A) = \mathrm{tr}(\Sigma^{-1}AA^{\mathrm{T}}).$$

The terms that solely dependent on Σ^{-1} should equal to each other, so:

$$\mathrm{tr}(\Sigma^{-1}(k_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + \mathbf{S}_0)) + \mathrm{tr}(\Sigma^{-1}(N\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}\mathbf{S}_{\bar{\mathbf{x}}})) = \mathrm{tr}(\Sigma^{-1}(k_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^{\mathrm{T}} + \mathbf{S}_{\mathbf{X}})).$$

Having arrived in:

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

we obtain:

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

Recall the definition for mean we ends in (4.214) since:

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

This finishes proving that the posterior distribution for MVN takes the form: $\mathrm{NIW}(\mathbf{m}_{\mathbf{X}}, k_{\mathbf{X}}, v_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}})$.

6.12 BIC for Gaussians

For question (a), recall that the maximum likelihood estimation for a MVN model is:

$$\mu_{\mathrm{MLE}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n,$$

$$\Sigma_{\mathrm{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{MLE}})(\mathbf{x}_n - \mu_{\mathrm{MLE}})^{\mathrm{T}}.$$

So the likelihood reads:

$$\begin{aligned}
p(\mathcal{D}|\mu_{\mathrm{MLE}}, \Sigma_{\mathrm{MLE}}) &= \prod_{n=1}^{N} p(\mathbf{x}_n|\mu_{\mathrm{MLE}}, \Sigma_{\mathrm{MLE}}) \\
&= \prod_{n=1}^{N}(2\pi)^{-\frac{D}{2}} \cdot |\Sigma_{\mathrm{MLE}}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}_n - \mu_{\mathrm{MLE}})^{\mathrm{T}}\Sigma_{\mathrm{MLE}}^{-1}(\mathbf{x}_n - \mu_{\mathrm{MLE}})\right) \\
&= (2\pi)^{-\frac{ND}{2}} \cdot |\Sigma_{\mathrm{MLE}}|^{-\frac{N}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{MLE}})^{\mathrm{T}}\Sigma_{\mathrm{MLE}}^{-1}(\mathbf{x}_n - \mu_{\mathrm{MLE}})\right).
\end{aligned}$$

Denote:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{x}_1 - \mu_{\text{MLE}} & \cdots & \mathbf{x}_N - \mu_{\text{MLE}} \end{pmatrix},$$

then $\Sigma_{\text{MLE}} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^{\text{T}}$, while the term in the exponential of the likelihood is:

$$-\frac{1}{2} \cdot \text{tr}(\mathbf{Y}^{\text{T}}\Sigma_{\text{MLE}}^{-1}\mathbf{Y}) = -\frac{1}{2} \cdot \text{tr}(\Sigma_{\text{MLE}}^{-1}\mathbf{Y}\mathbf{Y}^{\text{T}}) = -\frac{ND}{2}.$$

Thus the BIC is:

$$-\frac{ND}{2} \cdot \log(2\pi\text{e}) - \frac{N}{2} \cdot \log|\Sigma_{\text{MLE}}| - \frac{D + \frac{D(D+1)}{2}}{2} \cdot \log N.$$

For question (b), the fitting of a diagonal MVN model is tantamount to fitting D independent 1d Gaussian models simultaneously, thus the d-th diagnoal component of $\Sigma_{\text{MLE}}^{\text{d}}$ is:

$$\frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_{n,d}^2,$$

where we have assumed $\mathbf{x} = \mathbf{0}$ w.l.o.g. Thus the term inside the exponential of the likelihood remains $-\frac{ND}{2}$. So the BIC in this case is:

$$-\frac{ND}{2} \cdot \log(2\pi\text{e}) - \frac{N}{2} \cdot \log|\Sigma_{\text{MLE}}^{\text{d}}| - D \cdot \log N.$$

We observe that if all D components are mutually independent, i.e., Σ_{MLE} is diagonal then the BIC for diagonal MVN model is strictly larger than that for general MVN, hence the diagonal version is always preferred. In cases there exists dependence among components, the BIC for general MVN is still not necessarily larger than that of disgonal MVN. This is a reflection of the trade-off between complexity and generalization.

The Bayesian information criterion is an approximation of a model's evidence, $p(\mathcal{D})$. Let us start from:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta) \cdot p(\theta)\text{d}\theta,$$

where θ is the collection of all parameters within the current model. The trick here is to expand $\log p(\mathbf{x}|\theta)$ as a function of θ and taking approximation to the second order at θ_{MAP} so the first order gradient vanishes:

$$\log p(\mathbf{x}|\theta) \approx \log p(\mathbf{x}|\theta_0) - \frac{1}{2}(\theta - \theta_0)^{\text{T}}\mathbf{H}(\theta - \theta_0),$$

where $\mathbf{H}$ is the Hessian matrix at $\log p(\mathbf{x}|\theta_0)$. Thus we have:

$$p(\mathbf{x}|\theta) \approx p(\mathbf{x}|\theta_0) \cdot \exp\left(-\frac{1}{2}(\theta-\theta_0)^{\mathrm{T}}\mathbf{H}(\theta-\theta_0)\right).$$

We are now ready to perform the integral, with:

$$p(\mathcal{D}|\theta) = p(\mathbf{x}|\theta_0)^N \cdot \exp\left(-\frac{N}{2}(\theta-\theta_0)^{\mathrm{T}}\mathbf{H}(\theta-\theta_0)\right),$$

conducting the integral on the neighbour of $\theta_0 = \theta_{\mathrm{MAP}}$:

$$\begin{aligned}\int p(\mathcal{D}|\theta)\cdot p(\theta)\mathrm{d}\theta &\approx p(\mathcal{D}|\theta_{\mathrm{MAP}})\cdot p(\theta_{\mathrm{MAP}})\cdot \int \exp\left(-\frac{N}{2}(\theta-\theta_{\mathrm{MAP}})^{\mathrm{T}}\mathbf{H}(\theta-\theta_{\mathrm{MAP}})\right)\mathrm{d}\theta \\ &= p(\mathcal{D}|\theta_{\mathrm{MAP}})\cdot p(\theta_{\mathrm{MAP}})\cdot (2\pi)^{\frac{d}{2}}|N^{-1}\mathbf{H}^{-1}|^{\frac{1}{2}} \\ &= p(\mathcal{D}|\theta_{\mathrm{MAP}})\cdot p(\theta_{\mathrm{MAP}})\cdot (2\pi)^{\frac{d}{2}}\cdot N^{-\frac{d}{2}}\cdot|\mathbf{H}^{-1}|^{\frac{1}{2}},\end{aligned}$$

where d is the number of components in θ. Taking the logarithm of both side of the evidence yields the BIC. One can see how many compromises and assumptions have been applied in deriving an analytic form of the evidence, which is arguably the most complex variable for Bayesian analysis.

See also *PRML*, Section 4.4.

6.13 Gaussian posterior credible interval

Assume the prior distribution for an 1d normal distribution:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2 = 9).$$

And the likelihood is:

$$p(x) = \mathcal{N}(x|\mu, \sigma^2 = 4).$$

Having observed n variables, we want that the probability measure of μ's posterior distribution is no less than 0.95 within an interval no longer than 1. The posterior for μ is:

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

where we have dropped the terms inrevelent with μ. The posterior variance of μ is determined by the coefficient of μ^2 in the exponential of the posterior distribution:

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

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

$$n \geq 611.$$

6.14 MAP estimation for 1d Gaussians

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

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

So the posterior is another normal distribution, by comparing the coefficient for μ^2 in the exponential:

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

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 completing the square:

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

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

This finishes question (a).

For question (b), we already knew that the MLE is:

$$\mu_{\text{MLE}} = \frac{\sum_{n=1}^N x_n}{N}.$$

As N increases, we have:

$$\begin{aligned}\lim_{N\to\infty} \mu_{\text{post}} &= \lim_{N\to\infty} \frac{\frac{\sigma^2}{s^2} \cdot m + \sum_{n=1}^N x_n}{\frac{\sigma^2}{s^2} + N} \\ &= \frac{\sum_{n=1}^N x_n}{N}.\end{aligned}$$

For question (c), when $s^2 \to \infty$, μ_{post} also converges to μ_{MLE} since $\frac{\sigma^2}{s^2} \to 0$.

For question (d), when $s^2 \to 0$, then $\frac{\sigma^2}{s^2} \to \infty$ and μ_{post} converges to m.

Both (c) and (d) are very intuitive. $s^2 \to \infty$ means a non-informative prior has been introduced so MAP is the same as MLE. $s^2 \to 0$ means that the knowledge that μ is close to m is very strong so that finite observations cannot modify this belief.

6.15 Sequential(recursive) updating of covariance matrix

For question (a), note that:

$$n\mathbf{C}_{n+1}-(n-1)\mathbf{C}_n = \sum_{i=1}^{n+1}(\mathbf{x}_i-\mathbf{m}_{n+1})(\mathbf{x}_i-\mathbf{m}_{n+1})^{\mathrm{T}} - \sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m}_n)(\mathbf{x}_i-\mathbf{m}_n)^{\mathrm{T}}.$$

Making use of:

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

we have:

$$\begin{aligned} n\mathbf{C}_{n+1} - (n-1)\mathbf{C}_n &= \mathbf{x}_{n+1}\mathbf{x}_{n+1}^{\mathrm{T}} - (n+1)\mathbf{m}_{n+1}\mathbf{m}_{n+1}^{\mathrm{T}} + n\mathbf{m}_n\mathbf{m}_n^{\mathrm{T}} \\ &= \frac{n}{n+1}(\mathbf{x}_{n+1} - \mathbf{m}_n)(\mathbf{x}_{n+1} - \mathbf{m}_n)^{\mathrm{T}}. \end{aligned}$$

For question (b), the complexity is $\mathcal{O}(d^2)$.

For question (c), plugging (4.281) directly into (4.279) yields (4.280).

For question (d), the complexity remains $\mathcal{O}(d^2)$.

6.16 Likelihood ratio for Gaussians

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

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

The first term on r.h.s. is the ratio of likelihood probability.

When we have arbitrary covariance matrices:

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

As Σ_0, Σ_1 are arbitrary matrices, this formulation cannot be reduced further:

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

Note that the decision boundary ($\log \frac{p(x|y=1)}{p(x|y=0)} = 0$) is a quardratic surface in D-dimension space.

When both covariance matrixes are given by Σ:

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

so:

$$\begin{aligned}\log \frac{p(x|y=1)}{p(x|y=0)} &= -\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0) \\ &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left[(x-\mu_1)(x-\mu_1)^{\mathrm{T}} - (x-\mu_2)(x-\mu_2)^{\mathrm{T}}\right]\right).\end{aligned}$$

When Σ is a diagonal matrix, we have:

$$\begin{aligned}\log \frac{p(x|y=1)}{p(x|y=0)} &= -\frac{1}{2}(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^{\mathrm{T}}\Sigma^{-1}(x-\mu_0) \\ &= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\left[(x-\mu_1)(x-\mu_1)^{\mathrm{T}} - (x-\mu_2)(x-\mu_2)^{\mathrm{T}}\right]\right) \\ &= \frac{1}{2}\mathrm{tr}\left(\Lambda^{-1}\Phi\right) \\ &= \frac{1}{2}\sum_{i=1}^{d}\lambda_i^{-1}\Phi_{i,i}.\end{aligned}$$

where:

$$\Phi = (x-\mu_1)(x-\mu_1)^{\mathrm{T}} - (x-\mu_2)(x-\mu_2)^{\mathrm{T}}.$$

Finally, if $\Sigma = \sigma^2 I$ then:

$$\log \frac{p(x|y=1)}{p(x|y=0)} = \frac{1}{2\sigma^2}\mathrm{tr}(\Phi).$$

Note that for the last three cases, the decision boundary is a linear plane in the space, since the quadratic term on x has been cancelled in Φ.

6.17 LDA/QDA on height/weight data

...

6.18 Naive Bayes with mixed features

For question (a):

$$p(y=1|x_1=0,x_2=0) = \frac{p(y=1)\cdot p(x_1|y=1)\cdot p(x_2=0|y=1)}{p(x_1=0,x_2=0)} = \frac{0.5\cdot 0.5\cdot \frac{\exp(-0.5)}{\sqrt{2\pi}}}{p(x_1=0,x_2=0)}.$$

Similarly,

$$p(y=2|x_1=0,x_2=0) = \frac{0.25*0.5*\frac{1}{\sqrt{2\pi}}}{p(x_1=0,x_2=0)},$$

$$p(y=3|x_1=0,x_2=0) = \frac{0.25*0.5*\frac{\exp(-0.5)}{\sqrt{2\pi}}}{p(x_1=0,x_2=0)}.$$

A normalization yields the final result by eliminating $p(x_1=0,x_2=0)$.

For question (b), we have:

$$p(y=1|x_1=0) = 0.5,$$

$$p(y=2|x_1=0) = 0.25,$$

$$p(y=3|x_1=0) = 0.25,$$

since x_1 yields no more information for the classification label.

For question (c), we have:

$$p(y=1|x_2=0) \propto 0.5*\frac{\exp(-0.5)}{\sqrt{2\pi}},$$

$$p(y=2|x_2=0) \propto 0.25*\frac{1}{\sqrt{2\pi}},$$

$$p(y=3|x_2=0) \propto 0.25*\frac{\exp(-0.5)}{\sqrt{2\pi}}.$$

One can observe that unlike $p(y|x_1=0)$, $p(y|x_1=0)$ is different from the prior on labels.

6.19 Decision boundary for LDA with semi tied covariances

We begin from the Bayes rule:

$$p(y=1|\mathbf{x},\theta) \propto p(y=1|\theta)\cdot p(\mathbf{x}|y=1,\theta),$$

where we have omitted the terms independent of y. With a uniform prior on two classes:

$$p(y=1|\mathbf{x},\theta) \propto \mathcal{N}(\mathbf{x}|\mu_1, k\Sigma_0),$$
$$p(y=0|\mathbf{x},\theta) \propto \mathcal{N}(\mathbf{x}|\mu_0, \Sigma_0).$$

The decision boundary, in which we are interested, is a curve depicted by $f(\mathbf{x}) = 0$, where:

$$f(x) = \log \frac{p(y=1|\mathbf{x},\theta)}{p(y=0|\mathbf{x},\theta)}.$$

Therefore the decision boundary is:

$$\operatorname{tr}\left(\Sigma_0^{-1}\left[\Phi(\mathbf{x})\right]\right) = d \cdot \ln k,$$

where:

$$\Phi(\mathbf{x}) = \frac{(\mathbf{x}-\mu_1)(\mathbf{x}-\mu_1)^{\mathrm{T}}}{k} - (\mathbf{x}-\mu_0)(\mathbf{x}-\mu_0)^{\mathrm{T}}.$$

The decision boundary is a quardratic curve unless $k = 1$, which is geometrically very intuitive.

Let us consider a 2d case, focusing on the transformed coordinates where Σ_0 is the identity matrix. With out loss of generality, let $\mu_0 = (0,0)$, $\mu_1 = (z,0)$, denote the distance between a point p in this place from μ_0 and μ_1 by $a(p)$ and $b(p)$. The decision boundary is exactly:

$$\left\{p : \frac{b(p)^2}{k} - a(p)^2 = -\frac{1}{2}\ln k\right\}.$$

Plugging in the Cartesian representation $p(x,y)$, we ends up with a conical curve. The linear transform of the space would not change its conical nature.

6.20 Logistic regression vs LDA/QDA

The underlying assumptions for all four classifiers are as follows:

- GaussI assumes a covariance matrix as identity matrix;
- GaussX has not prior assumption on the covariance matrix;
- LinLog assumes that different classes share the same covariance matrix;

- QuadLog has not prior assumption on covariance matrix, yet it assumes that all data from one class are subject to a normal distribution;

From the perspective of complexity we have the following order:

QuadLog=GaussX>LinLog>GaussI.

The MLE likelihood should follow the same order, this answers question (a)-(d).

For question (e), the argument is untrue in general. For example, model M predicts two samples belonging to the first class with probability vectors $(0.49, 0.51)$ and $(0.99, 0.1)$. While M' outputs $(0.51, 0.49)$ and $(0.51, 0.49)$. Now M' is correct on both samples so $R(M) > R(M')$, but:

$$\frac{\log(0.49) + \log(0.99)}{2} > \frac{\log(0.51) + \log(0.51)}{2},$$

so $L(M) > L(M')$, this is sufficient for disproving the argument.

6.21 Gaussian decision boundaries

We have:

$$p(x|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{(x - \mu_i^2)}{2\sigma_i^2}\right),$$

so:

$$p(x|\mu_1, \sigma_1^2) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right),$$

$$p(x|\mu_2, \sigma_2^2) = \frac{1}{10^3 \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(x-1)^2}{2 \times 10^6}\right).$$

The decision region satisfies:

$$\frac{p(x|\mu_1, \sigma_1^2)}{p(x|\mu_1, \sigma_1^2)} \geq 1,$$

this is tantamount to:

$$\frac{(x-1)^2}{10^6} - x^2 \geq -6 \cdot \ln 10.$$

Denote x_1 and x_2 as the zeros of this quardratic form, then

$$R_1 = [x_1, x_2].$$

When $\sigma^2 = \sigma^1$, R_1 is exactly $(\infty, \frac{1}{2})$.

One can solve for (a) and (b) by plugging in (4.289).

6.22 QDA with 3 classes

Solve (a) and (b) numerically:

```
import math
import numpy as np
mu1=np.array([0,0])
mu2=np.array([1,1])
mu3=np.array([1,-1])
s1=np.array([[0.7,0],[0,0.7]])
s2=np.array([[0.8,0.2],[0.2,0.8]])
s3=np.array([[0.8,0.2],[0.2,0.8]])
def p(x):
    f1=np.linalg.det(s1)**(-0.5)*math.exp(-0.5*(mu1-x)@np.
        linalg.inv(s1)@(mu1-x))
    f2=np.linalg.det(s2)**(-0.5)*math.exp(-0.5*(mu2-x)@np.
        linalg.inv(s2)@(mu2-x))
    f3=np.linalg.det(s3)**(-0.5)*math.exp(-0.5*(mu3-x)@np.
        linalg.inv(s3)@(mu3-x))
    return f1,f2,f3
x1=np.array([-0.5,0.5])
x2=np.array([0.5,0.5])
print(p(x1))
print(p(x2))
```

The outcome is:

```
(0.9995321962501862, 0.31309336105606445, 0.030361279346887253)
(0.9995321962501862, 1.005427487616282, 0.18990072283298057)
```

So the predicted labels are 1 and 2 respectively.

6.23 Scalar QDA

```
import math
hm=[67,79,71]
hf=[68,67,60]
def mu(h):
    return (h[0]+h[1]+h[2])/3
```

```
6  def sigma2(h):
7      m=mu(h)
8      return ((h[0]-m)**2+(h[1]-m)**2+(h[2]-m)**2)/3
9  # For question (a):
10 mu_m=mu(hm)
11 mu_f=mu(hf)
12 sigma2_m=sigma2(hm)
13 sigma2_f=sigma2(hf)
14 print(mu_m)
15 print(sigma2_m)
16 print(mu_f)
17 print(sigma2_f)
18 # \pi_{m}=\pi_{f}=0.5.
19 # For question (b):
20 temp_m=(2*math.pi*sigma2_m)**(-0.5)*math.exp(-(72-mu_m)**2/2/
       sigma2_m)
21 temp_f=(2*math.pi*sigma2_f)**(-0.5)*math.exp(-(72-mu_f)**2/2/
       sigma2_f)
22 print(temp_m/(temp_m+temp_f))
```

For question (c), using a naive Bayes is tantamount to adopting diagonal covariance matrices for both classes.

7 Bayesian statistics

7.1 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:

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

7.2 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 $\hat{y} = 0$。

7.3 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 $\hat{c}$:

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

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:

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

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

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

Or:

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

7.4 More reject options

Straightforward calculation.

7.5 Newsvendor problem

By:

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

We have:

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

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

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

7.6 Bayes factors and ROC curves

Practise by yourself.

7.7 Bayes model averaging helps predictive accuracy

Expand both side of 5.127 and exchange the integral sequence:

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

We also have:

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

Substract the right side from the left side ends in:

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

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

7.8 MLE and model selection for a 2d discrete distribution

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

$$p(x = 0, y = 0) = (1 - \theta_1)\theta_2$$

$$p(x = 0, y = 1) = (1 - \theta_1)(1 - \theta_2)$$

$$p(x = 1, y = 0) = \theta_1(1 - \theta_2)$$

$$p(x = 1, y = 1) = \theta_1\theta_2$$

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:

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

Hence:

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

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

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

MLE can de done by using regularization condition.

The rest is straightforward algebra.

7.9 Posterior median is optimal estimate under L1 loss

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

$$\rho(a) = \int |y-a|p(y)dy = \int_{-\infty}^{a}(a-y)p(y)dy + \int_{a}^{+\infty}(y-a)p(y)dy$$

$$= a\left\{\int_{-\infty}^{a}p(y)dy - \int_{a}^{+\infty}p(y)dy\right\} - \int_{-\infty}^{a}yp(y)dy + \int_{a}^{+\infty}yp(y)dy$$

Differentiate and we have:

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

Set it to zero and:

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

7.10 Decision rule for trading off FPs and FNs

Given:

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

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 $\frac{c}{1+c}$.

8 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 a decent textbook on mathematics statistics.

GL.

9 Linear regression

9.1 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.

9.2 Multi-output linear regression

Straightforward calculation.

9.3 Centering and ridge regression

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

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

So:

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

Therefore:

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

9.4 MlE for σ^2 for linear regression

Firstly, we give the likelihood:

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

As for σ^2:

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

We have:

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

9.5 MLE for the offset term in linear regression

NLL:

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

Differentiate with two parameters:

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

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

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:

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

9.6 MLE for simple linear regression

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

9.7 Sufficient statistics for online linear regression

a and b can be solved according to hints.

For 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 f: practice by yourself.

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

Problem a: practice by yourself.

For 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:

$$w_{01} = 0$$

$$\mathbf{V}_{0,22}^{-1} = 1$$

$$\mathbf{V}_{0,11}^{-1} = \mathbf{V}_{0,12}^{-1} = 0$$

$$w_{00} = arbitrary$$

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:

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

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

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

9.9 Generative model for linear regression

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

$$w_0 = 0$$

$$\mu_x = \mu_y = 0$$

By covariance's definition:

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

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

Using the conclusion from section 4.3.1:

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

Where:

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

9.10 Bayesian linear regression using the g-prior

Recall ridge regression model, where we have likelihood:

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

The prior distribution is Gaussian-Inverse Gamma distribution:

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

The posterior distribution takes the form:

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

Comparing the coefficient of σ^2:

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

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

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

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:

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

Finally, comparing the constant term inside the exponential:

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

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

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

10 Logistic regression

10.1 Spam classification using logistic regression

Practice by yourself.

10.2 Spam classification using naive Bayes

Practice by yourself.

10.3 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):

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

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

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

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

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

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

10.4 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 $\mathbf{w}^*$:

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

$$\sum_{n=1}^{N} -y_{n*}\mathbf{x}_n + \frac{\exp(\mathbf{w}^{*T}\mathbf{x}_n)}{\sum_{c=1}^{C} \exp(\mathbf{w}_c^T \mathbf{x}_n}) \mathbf{x}_n = \sum_{n=1}^{N} (\mu_{n*} - y_{n*}) \mathbf{x}_n$$

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:

$$\frac{\partial}{\partial \mathbf{w}_1} \sum_{n=1}^{N} (\mu_{n1} - y_{n1}) \mathbf{x}_n^T = \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}_1} \mu_{n1} \mathbf{x}_n^T$$

$$\sum_{n=1}^{N} \frac{\exp(\mathbf{w}_1^T \mathbf{x}_n)(\sum \exp)\mathbf{x}_n - \exp(\mathbf{w}_1^T \mathbf{x}_n)^2 \mathbf{x}_n}{(\sum \exp)^2} \mathbf{x}_n^T$$

$$= \sum_{n=1}^{N} \mu_{n1}(1 - \mu_{n1}) \mathbf{x}_n \mathbf{x}_n^T$$

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

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

$$= \sum_{n=1}^{N} -\mu_{n1} \mu_{n2} \mathbf{x}_n \mathbf{x}_n^T$$

Ends in 8.44。

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

10.5 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 $j = 1...D$, $\sum_{c=1}^{C} \hat{w}_{cj} = 0$.

10.6 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:

$$p(\mathbf{w}|D) \propto p(D|\mathbf{w})p(\mathbf{w})$$

$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{0}, \sigma^{-2}\mathbf{I})$$

$$NLL(\mathbf{w}) = -\log p(\mathbf{w}|D) = -\log p(D|\mathbf{w}) + \frac{1}{2\sigma^2}\mathbf{w}^T\mathbf{w} + c$$

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.

10.7 Regularizing separate terms in 2d logistic regression

Practice by yourself.

11 Generalized linear models and the exponential family

11.1 Conjugate prior for univariate Gaussian in exponential family form

The 1d Gaussian distribution is:

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

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

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

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 为 v_1 and v_2 respectively, then the prior distribution takes the form:

$$p(\theta|M,v_1,v_2)=\exp\left\{\theta_1\cdot Mv_1+\theta_2\cdot Mv_2-M\cdot A(\theta)\right\}$$

$$=\exp\left\{-\frac{\lambda}{2}Mv_1+\lambda\mu Mv_2-\frac{M}{2}\lambda\mu^2-\frac{M}{2}\ln 2\pi+\frac{M}{2}\ln\lambda\right\}$$

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

$$p(\mu,\lambda)=\exp\left\{(\alpha-1)\ln\lambda-\beta\lambda-\frac{\lambda(2\alpha-1)}{2}\mu^2-\frac{\lambda(2\alpha-1)}{2}\gamma^2\right\}$$

$$\cdot\exp\left\{\lambda(2\alpha-1)\mu\gamma+\frac{1}{2}\ln\lambda\right\}$$

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

$$-\frac{(2\alpha-1)}{2} = -\frac{M}{2}$$

$$\gamma(2\alpha-1) = Mv_2$$

$$\frac{(2\alpha-1)}{2}\gamma^2 - \beta = -\frac{1}{2}Mv_1$$

$$(\alpha-1) + \frac{1}{2} = \frac{M}{2}$$

Combining them ends in:

$$\alpha = \frac{M+1}{2}$$

$$\beta = \frac{M}{2}(v_2^2 + v_1)$$

$$\gamma = v_2$$

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

11.2 The MVN is in the exponential family

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

12 Directed graphical models(Bayes nets)

...

13 Mixture models and the EM algorithm

13.1 Student T as infinite mixture of Gaussian

The 1d Student-t distribution takes the form:

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

Consider the left side of 11.61:

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

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

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

Plug in can derive 11.61.

13.2 EM for mixture of Gaussians

We are to optimize:

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

Where:

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

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

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

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 $Q(\theta,\theta^{old})$:

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

Using the same way as in 4.1.3.1:

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

The balance condition is:

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

Obtain 11.32:

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

13.3 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。

13.4 EM for mixture of Student distributions

The log-likelihood for complete data set is:

$$l_c(\mathbf{x}, z) = \log(N(\mathbf{x}|\mu, \frac{\Sigma}{z}) Ga(z|\frac{\lambda}{2}, \frac{\lambda}{2}))$$

$$= -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| + \frac{D}{2} \log(z) - \frac{z}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) +$$

$$\frac{v}{2}\log(\frac{v}{2}) - \log(\Gamma(\frac{v}{2})) + (\frac{v}{2}-1)\log(z) - \frac{v}{2}z$$

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:

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

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

Hence equals the MLE used for MVN.

13.5 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:

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

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:

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

Where:

$$\nabla_{\Sigma_k} N(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\} \nabla_{\Sigma_k} \cdot$$

$$\left\{\nabla_{\Sigma_k}(-\frac{1}{2}(\mathbf{x}-\mu_k)\Sigma_k^{-1}(\mathbf{x}-\mu_k)) - \Sigma_k^{-1}\nabla_{\Sigma_k}|\Sigma_k|\right\}$$

$$= N(\mathbf{x}|\mu_k, \Sigma_k)\nabla(\log N(\mathbf{x}|\mu_k, \Sigma_k))$$

Thus we have:

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

13.6 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):

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

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

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

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:

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

And we ends in:

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

13.7 Manual calculation of the M step for a GMM

Practise by yourself.

13.8 Moments of a mixture of Gaussians

For the expectation of mixture distribution:

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

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

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

Where:

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

Therefore:

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

13.9 K-means clustering by hand

Practise by yourself.

13.10 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.

13.11 Visible mixtures of Gaussians are in exponential family

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

$$p(\mathbf{z}) = \prod_{c=1}^{C} \pi_c^{z_c}$$

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

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 $\mathbf{z}x^2$.

13.12 EM for robust linear regression with a Student t likelihood

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

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

Set the deriavte to zero:

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

This means:

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

13.13 EM for EB estimation of Gaussian shrinkage model

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

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

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:

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

13.14 EM for censored linear regression

Unsolved.

13.15 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:

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

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:

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

We have:

$$\int_b^c w^2 N(w|0,1)dw = \Phi(c) - \Phi(b) - cN(c|0,1) + bN(b|0,1)$$

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

Plug it into the conclusion drawn from question a:

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

$$= \mu_i^2 + \sigma^2 + H(A)(\sigma c_i + \sigma\mu_i)$$

14 Latent linear models

14.1 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):

$$p(\mathbf{z}) = N(\mathbf{z}|0, I)$$

$$p(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|\mathbf{W}\mathbf{z}, \Psi)$$

And:

$$p(\mathbf{z}|\mathbf{x}) = N(\mathbf{z}|\mathbf{m}, \Sigma)$$

$$\Sigma = (I + \mathbf{W}^T\Psi^{-1}\mathbf{W})^{-1}$$

$$\mathbf{m} = \Sigma\mathbf{W}^T\Psi^{-1}\mathbf{x}_n$$

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:

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

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

$$\mathbb{E}[\mathbf{z}_n|\mathbf{x}_n] = \Sigma\mathbf{W}^T\Psi^{-1}\mathbf{x}$$

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

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

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

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

Set it to zero:

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

Optimize w.r.t Ψ^{-1}:

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

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

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

Assume Ψ to be a diagnal matrix:

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

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.

14.2 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:

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

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

14.3 Heuristic for assessing applicability of PCA*

Need pictures for illustration here!

14.4 Deriving the second principal component

For:

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

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

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

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

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

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

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

Where:

$$\Lambda = diag\{\lambda_1, \lambda_2, ...\}$$

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

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

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

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

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

Notice $\mathbf{O}\mathbf{v}_2$ means a transform on $\mathbf{v}_2$, with its length unchanged. And $(\mathbf{O}\mathbf{v}_2)^T\Lambda(\mathbf{O}\mathbf{v}_2)$ 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:

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

Therefore:

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

14.5 Deriving the residual error for PCA

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

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

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

Plug in $\mathbf{v}_j^T\mathbf{C}\mathbf{v}_j = \lambda_j$ and sum over n can draw the conclusion in b.
Plug $K = d$ into the conclusion in b, we have:

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

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

In general cases:

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

14.6 Derivation of Fisher's linear discriminant

Straightforward algebra.
（need reference）

14.7 PCA via successive deflation

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

14.8 Latent semantic indexing

Practice by yourself.

14.9 Imputation in a FA model*

wtf$\mathbf{x}_v$?

wtf$\mathbf{x}_h$?

14.10 Efficiently evaluating the PPCA density

With:

$$p(\mathbf{z}) = N(\mathbf{z}|0, \mathbf{I})$$

$$p(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|\mathbf{W}\mathbf{z}, \sigma^2\mathbf{I})$$

Use the conclusion from chapter 4.

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

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:

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

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

14.11 PPCA vs FA

Practice by youself.

15 Sparse linear models

15.1 Partial derivative of the RSS

Define:

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

Straightforwardly:

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

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:

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

15.2 Derivation of M-step for EB for linear regression

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

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) = N(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1})$$

$$p(\mathbf{w}) = N(\mathbf{w}|0, \mathbf{A}^{-1})$$

$$A = diag(\alpha)$$

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

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

Then:

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

For auxiliay function:

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

In E-step, we need $\mathbb{E}[\mathbf{w}]$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^T]$, which have been computed:

Introduce a prior for component in α and β:

$$p(\alpha,\beta) = \prod_j Ga(\alpha_j|a+1,b)\cdot Ga(\beta|c+1,d)$$

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.

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

Unsolved.

15.4 Marginal likelihood for linear regression*

Straightforward algebra.

15.5 Reducing elastic net to lasso

Expand both sides of 13.196, the right side:

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

The left side:

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

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.

15.6 Shrinkage in linear regression

For ordinary least square:

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

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

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

Take the derivative:

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

We have:

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

In ridge regression:

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

Take the derivative:

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

Thus

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

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

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

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.

15.7 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:

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

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:

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

15.8 Deriving E step for GSM prior

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

Take Laplace transform/generating transform to both sides:
To calculate:

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

According to 13.200, it reduces to:

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

Because:

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

This gives 13.197:

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

! 此题存疑，Hint 1 和 Hint 2 中可能均有印刷错误。

15.9 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:

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

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

$$p(\tau^2|\gamma) = Ga(\tau_d^2|1, \frac{\gamma^2}{2})$$

$$p(w_d|\tau_d^2) = N(w_d|0, \tau_d^2)$$

Hence:

$$p(\tau^2, \mathbf{w}, \mathbf{y}|\mathbf{X}, \gamma) \propto \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}(\gamma^2\tau_d^2 + \frac{w_d^2}{\tau_d^2})\right\} \cdot \prod_{d=1}^{D}\frac{1}{\tau_d}$$

$$\cdot \prod_{n=1}^{N} \Phi(\mathbf{w}^T\mathbf{x}_n)^{y_n}(1 - \Phi(\mathbf{w}^T\mathbf{x}_n))^{1-y_n}$$

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

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

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

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

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

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

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.

15.10 GSM representation of group lasso*

Follow the hints and straightforward algebra.

15.11 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:

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

Consider under a linear regression context:

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

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:

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

The original problem is changed to:

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

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

Denote:

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

Rewrite the original target:

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

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

Where:

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

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

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".

15.12 Subderivative of the hinge loss function

$$if(\theta < 1)\partial f(\theta) = \{-1\}$$

$$if(\theta = 1)\partial f(\theta) = [-1, 0]$$

$$if(\theta > 1)\partial f(\theta) = \{0\}$$

15.13 Lower bounds to convex functions

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

16 Kernels

17 Gaussian processes

17.1 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:

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

18 Adaptive basis function models

18.1 Nonlinear regression for inverse dynamics

Practise by yourself.

19 Markov and hidden Markov models

19.1 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 $\mathbf{z}_{1:T}$.

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

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

19.2 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:

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

19.3 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:

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

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.

19.4 EM for HMMs with tied mixtures

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

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

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.

20 State space models

20.1 Derivation of EM for LG-SSM

We directly work on the auxiliary function:

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

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:

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

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:

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

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.

20.2 Seasonal LG-SSM model in standard form

From Fig.18.6(a), we have:

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

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

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

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.

21 Undirected graphical models(Markov random fields)

21.1 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:

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

21.2 CI properties of Gaussian graphical models

Problem a:

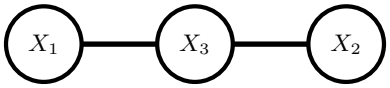
We have:

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

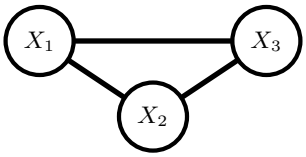
And:

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

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 $X_1 \to X_2$ and $X_3 \to X_2$.

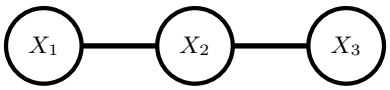
Problem c: Consider the terms inside the exponential:

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

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 $X_1 \perp X_3 | X_2$:



21.3 Independencies in Gaussian graphical models

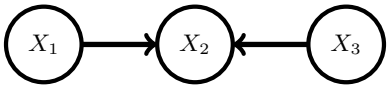
Problem a and b:

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.

21.4 Cost of training MRFs and CRFs

The answer are generally:

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

and

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

21.5 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:

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

22 Exact inference for graphical models

22.1 Variable elimination

Where tf is the figure?!

22.2 Gaussian times Gaussian is Gaussian

We have:

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

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.

22.3 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:

$$\begin{aligned}
&p(G_1 = 1|X_2 = 50, X_3 = 60) \\
=&\frac{p(X = 50|G_1 = 1)p(X = 60|G_1 = 1)}{p(X = 50|G_1 = 0)p(X = 60|G_1 = 0) + p(X = 50|G_1 = 1)p(X = 60|G_1 = 1)}
\end{aligned}$$

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.

22.4 Inference in 2D lattice MRFs

Please refer to PGM:principals and techniques 11.4.1.

23 Variational inference

23.1 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:

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

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

23.2 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.

23.3 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 $\mathbb{E}[\ln x|x \sim Ga(a,b)]$, which can therefore yields to the entropy for a Gamma distribution.

We know that Gamma distribution is an exponential family distribution:

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

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.

23.4 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:

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

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:

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

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

23.5 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:

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

23.6 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.

23.7 Forwards vs reverse KL divergence

We have:

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

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

We skip the practical part.

23.8 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:

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

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

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

Write into matrix form yields to 21.62.

23.9 Variational EM for binary FA with sigmoid link

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

23.10 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 $\mathbb{E}[z^2|z > \mu]$.

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

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:

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

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