Machine Learning: A Probabilistic Perspective 习题解答(1)

李方圻

2017年10月22日

目录

1 导言

1.1 文档构成

本来这一章节应该作为正文的第一部分进行第一章的习题解答，但是介于MLAPP一书第一章并无习题，故将此处篇幅作为习题解答的导言。

本文档提供了Machine Learning: A Probabilistic Perspective一书第一章到第十四章的大部分习题的详细解答。题目本身一般没有重述，比较有理论意义的题目之核心内容表明在了小标题中，也方便在目录中查找。后半本书的习题解答预计会在习题解答（2）文档中给出。

MLAPP一书的习题一般分为两种：理论习题和实践习题，本文档给出了绝大多数理论习题的解答，除了部分过于简单的和两个尚未得解的。实践习题基于一个MATLAB的工具包。实践习题的解答目前尚未给出。

1.2 关于Machine Learning: A Probabilistic Perspective

即便是专业的机器学习课程也很难界定“机器学习”到底是什么。

一方面，诸多人工智能领域的新晋进展震惊着社会，而高校中相关课程的选修人数也大幅增长；一方面，仍有诸多的理论工作者对于和学习相关的技术抱有怀疑态度，这其中尤其以深度学习所受到的评论最为两极分化。

机器学习在最近呈现出的非凡成果常常使人忘记这是一门已经经历了漫长发展的学科，它的启蒙至少可以追溯到上个世纪四十年代关于“电子脑”的研究。但即便如此，也没有人能界定机器学习是一门完全成熟的理论，一个例子就是即便在研究者社区，也有很多人将机器学习冠以“玄学”的称呼。我个人认为被称作“玄学”是许多理论分支在未发展成熟阶段所共同经受过的经历。

机器学习作为一个学科的理论体系还处在一个逐渐构成的过程中，而其中相对最为成功的就是基于概率论的理论体系构成，就如同MLAPP一书封底上普林斯顿大学的David Blei评价的一样“In Machine Learning, the language of probability and statistics reveals important connections between seemingly disparate algorithms and strategies. Thus, its readers will become articulate in a holistic view of the state-of-art and poised to build the next generation of machine learning algorithms.”

MLAPP的核心思想很简单：机器学习等同于贝叶斯统计学（Bayesian Statistics），而这种统计方法能够广泛地联系起诸多看似独立的算法。不过贝叶斯统计本身的历史要远远长于机器学习本身（可以上溯到拉普拉斯的年代），而在这一核心思想上，MLAPP也并非首创，Pattern Recognition and Machine Learning就是一个更鲜明的例子。两本书都可以算作机器学习理论方面比较经典的教材。

总体上而言，MLAPP以牺牲了部分推导的完整性来降低了整本书的难度，但也涵盖了更广泛的模型，更适合有所理论基础的学习者入门。其和概率论紧密联系的经典模型相关的章节（譬如2，3,4,5,7,8,11,12章）精彩程度和PRML不相上下，但是也因论述的顺序不同而值得PRML的阅读者再次细读。但是也有部分章节的写作或者习题部分比较敷衍，使得整个成书质量显得不够稳定，不过这也是因为有些章节的篇幅实在不足以涵盖一整个比较成熟的模型理论（譬如10章，和PRML的第8章相比高下立判）。所以本文档暂时并未给出第6,10,14章的习题解答。

1.3 文档性质

本文档的写作动机是作者（我）在选修机器学习课程时需要阅读MLAPP一教材，但是搜索习题解答无果。虽有几个Github上的项目似乎已经开始着手进行解答的写作，但是无奈其进展实在过于缓慢，而且我更想关注这本教材的理论部分而不是实践代码。

是故写作本文档以做参考，本文档的写作完成是在正式学期开始前的两周间，因为时间比较仓促，难免有所纰漏，所以一方面建议读者采取批判地眼光阅读，不要我写什么就信什么；一方面希望读者提出修改意见，除了错误解答的修正意见以外，擅长使用MATLAB、擅长Latex排版或愿意参与文档改进的读者，欢迎主动与我联系。

2017年10月22日

慕尼黑

2 概率

2.1

分别记两个孩子为A和B。

$Event_1 = A$为男孩，B为女孩；$Event_2 = B$为男孩，A为女孩；$Event_3 = A$为男孩，B为男孩。

$$P(Event_1) = P(Event_2) = P(Event_3) = \frac{1}{4}$$

$$P(onegirl|oneboy) = \frac{P(Event_1) + P(Event_2)}{P(Event_1) + P(Event_2) + P(Event_3)} = \frac{2}{3}$$

在b题中，不妨设看到的是孩子A，则：

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

2.2

归一化后易知谬误。

2.3 随机变量和的方差

根据定义，从形式上得到：

$$var(X+Y) = \mathbb{E}((X+Y)^2) - \mathbb{E}^2(X+Y) = \mathbb{E}(X^2) - \mathbb{E}^2(X) + \mathbb{E}(Y^2) - \mathbb{Y}^2(X) + 2\mathbb{E}(XY) - 2\mathbb{E}^2(XY)$$

$$= var(X) + var(Y) + 2cov(X, Y)$$

2.4

直接使用贝叶斯公式：

$$P(ill|positive) = \frac{P(ill)P(positive|ill)}{P(ill)P(positive|ill) + P(health)P(positive|health)}$$

$$= 0.0098$$

2.5 三门问题

经典的三门问题，答案为b，即换一个选项，分子和分母中的信息量差异导致了最后的失衡，直接使用贝叶斯公式：

$$P(prize_1|choose_1, open_3) = \frac{P(choose_1)P(prize_1)P(choose_3|prize_1, choose_1)}{P(choose_1)P(open_3|choose_1)}$$

$$= \frac{P(prize_1)P(choose_3|prize_1, choose_1)}{P(open_3|choose_1)}$$

$$= \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1} = \frac{1}{3}$$

最后一步在分母上使用了贝叶斯全概率公式，对隐含变量（奖品的真实所在地）进行了求和。

2.6 条件独立性

a选ii，b选择i，ii，iii。直接使用贝叶斯公式即可。

2.7

问题在于当给定的条件变量变多时，两两独立性不能保证此时的独立性。

2.8 联合分布可分解时的条件独立性

证明方向1：记$g(x,z) = p(x|z)$，$h(y,z) = p(y|z)$。
证明方向2：记$p(x|z) = \sum_y g(x,z)h(y,z)$，$p(y|z) = \sum_x g(x,z)h(y,z)$。

这道题目给出了一个重要的结论：条件独立的贝叶斯网络结构可以直接转化为马尔科夫场结构而不需在图的层面进行改变。

2.9 条件独立性

从马尔科夫场的视角，两个判断均为正确。

2.10 逆伽马分布

直接使用结论：

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

即得结果。

2.11

不另做证明。

2.12 互信息公式推导

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

$$= H(X) - H(X|Y)$$

在贝叶斯公式中使用$p(y)$来除可以得到另一半结论，也可以使用$H(X,Y) = H(X) + H(X|Y)$来进行对称的变换。

2.13 协变正态随机变量的互信息

利用$I(X_1;X_2) = H(X_1) + H(X_2) - H(X_1,X_2)$，以及高斯分布的边缘概率：

$$p(X_1) = p(X_2) = N(0,\sigma^2)$$

得到：

$$I(X_1;X_2) = -\frac{1}{2}\log_2(1-\rho^2)$$

多元正态分布的微分熵求法可见《信息论基础》8.4节。

2.14

利用2.12的结论，结合互信息的物理意义可理解。

2.15 S-投影

见《概率图模型——原理与技术》8.5.2节

2.16 贝塔分布形式

Beta分布的mode通过求导数直接可得，其均值和方差的计算通过利用Gamma函数的性质：

$$\frac{\Gamma(a+1)}{\Gamma(a)} = a$$

易得。

2.17

基础的概率问题，以m记最右边的点，则：

$$p(m > x) = (1-x)^2$$

$$\mathbb{E}(m) = \int x \cdot p(m = x)dx = \int p(m > x)dx = \int_0^1 (1-x)^2 dx = \frac{1}{3}$$

3 离散数据的生成模型

3.1 伯努利分布的最大似然估计

最大似然概率的表达式为：

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

其对数形式为：

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

求导数并置零：

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

得到所要求的原书3.22式：

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

3.2 贝塔-伯努利模型的边缘似然

Beta-Bernoulli/Binomial模型对于这样一个问题场景建模：一个随机发生器从一个有限的离散状态空间取值。本题面向状态空间大小为2的情况并求归一化因子$\frac{1}{p(D)}$。

似然概率为：

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

先验概率为：

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

后验概率为：

$$p(\theta|D) \propto p(D|\theta) \cdot p(\theta|a,b) = \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

$$= Beta(\theta|N_1 + a, N_0 + b)$$

预测分布：

$$p(x_{new} = 1|D) = \int p(x_{new} = 1|\theta) \cdot p(\theta|D) d\theta$$

$$= \int \theta p(\theta|D)d\theta = \mathbb{E}(\theta) = \frac{N_1 + a}{N_1 + a + N_0 + b}$$

来求$p(D)$，其中$D = 1, 0, 0, 1, 1$：

$$p(D) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)...p(X_N|x_{N-1}, X_{N-2}, ...X_1)$$

$$= \frac{a}{a+b}\frac{b}{a+b+1}\frac{b+2}{a+b+2}\frac{a+1}{a+b+3}\frac{a+2}{a+b+4}$$

记$\alpha = a + b, \alpha_1 = a, \alpha_0 = b$，得到式3.83。
再利用$[(\alpha_1)..(\alpha_1 + N_1 - 1)] = \frac{(\alpha_1+N_1-1)!}{(\alpha_1-1)!} = \frac{\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_1)}$
可得式3.80。

3.3 贝塔-伯努利模型的后验预测

Straightforward algebra.

3.4

首先由：

$$p(\theta|X < 3) \propto p(\theta)p(X < 3|\theta)$$

对于$p(X < 3|\theta)$，分两种情况$(X = 1, X = 2)$计算即可。

3.5

由：

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

以及：

$$p(\theta) = p(\phi)|\frac{d\phi}{d\theta}| \propto \frac{1}{\theta(1-\theta)} \propto Beta(\theta|0, 0)$$

3.6 泊松分布的最大似然估计

似然概率为：

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

对其对数求导置零：

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

得到：

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

即速率的最大似然估计是数据集的均值。

3.7 泊松分布的贝叶斯分析

因为：

$$p(\lambda|D)\propto p(\lambda)p(D|\lambda)\propto\exp(-\lambda(N+b))\cdot\lambda^{\sum_{n=1}^{N}x_n+a-1}=Ga(a+\sum x,N+b)$$

可以认为，这个先验分布意味着引入b个先验观测量，其均值为$\frac{a}{b}$。

3.8 均匀分布的最大似然估计

显然地，如果$a<max(x_n)$，则似然概率为0，故首先要求$a\geq max(x_n)$，此时似然概率为：

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

这是和a单调负相关的函数，所以必有最大似然估计为$a=max(x_n)$（就像在下一题题干中给出的一样），对于新x_{n+1}的估计在$[-a,a]$间平均分布。

从模型的意义上而言，平均分布的最大似然估计通过一个max算子来确定性地计算，鲁棒性相当糟糕。其根本原因是$p(x|a)$的值在a变化时有可能突变。

3.9 均匀分布的贝叶斯分析

均匀分布的共轭先验分布定义为Pareto分布：

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

记$m=max(x_n)$，联合分布为：

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

观测集置信为：

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

记$\mu = max(m, b)$，后验分布为：

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

3.10

Straigutforward calculation.

3.11 指数分布的贝叶斯分析

最大似然估计：

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

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

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

从形式上容易验证：指数分布本身不是指数分布的共轭先验分布。指数分布的共轭先验分布是Gamma分布：

$$p(\theta|D) \propto p(\theta)p(D|\theta) = Gam(\theta|a, b)p(D|\theta) = Gam(\theta|N+a, b+\sum x_n)$$

可以认为Gamma分布的先验意义是：引入$a-1$个先验观测量，其和为b。

选择形式错误的先验分布是c问题中专家的问题所在。

3.12

Straightforward calculation.

3.13

Unsolved

3.14

Straightforward calculation.

3.15

Straightforward algebra.

3.16 贝特分布参数设定

对于Beta分布的双参数α_1和α_2，通过期望联系起来：

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

其次进行积分：

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

通过数值方法取α_1使得$u(\alpha_1) \to 0.95$即可。

3.17 均匀先验分布下贝塔-二项模型的边缘似然

本题要求Beta-Binomial模型中的边缘分布：

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

题干给出条件：

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

$$p(\theta) = Beta(1, 1)$$

所以：

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

$$= \binom{N}{N_1}B(N_1+1, N-N_1+1) = \frac{N!}{N_1!(N-N_1)!}\frac{N_1!(N-N_1)!}{(N+1)!}$$

$$= \frac{1}{N+1}$$

3.18

Straightforward calculation.

3.19 朴素贝叶斯分类器中的特征关联强度

本题考虑朴素贝叶斯分类器（NBC）中关于分类不敏感词汇（例如NLP中的停用词）的相关事宜。

以x_{iw}为示性变量表示单词w是否出现在文档i中，以θ_{cw}表示在文档类型c中出现单词w的概率。此时对于一个给定的类型，一篇文档的似然概率为：

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

其对数形式为：

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

可以记成：

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

其中：

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

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

对于b中的二类分类问题：

$$p(c_1|\mathbf{x}_i)=\frac{p(c_1)p(\mathbf{x}_i|c_1)}{p(\mathbf{x}_i)}$$

故：

$$\log\frac{p(c_1|\mathbf{x}_i)}{p(c_2|\mathbf{x}_i)}=\log\frac{p(c_1)}{p(c_2)}+\log p(x_i|c_1)-\log p(x_i|c_2)$$

代入均匀分布先验假设：

$$\log\frac{p(c_1|\mathbf{x}_i)}{p(c_2|\mathbf{x}_i)}=\phi(\mathbf{x}_i)^T(\beta_{c_1}-\beta_{c_2})$$

在问题b中，考察停用词，它们的缺失应当不影响后验概率之比$\frac{p(c_1|\mathbf{x})}{p(c_2|\mathbf{x})}$。因为NBC假设各个单词独立，所以我们只需要考虑对应某个单词w_j导致的$\log\frac{p(c_1\mathbf{x}_i)}{p(c_2|\mathbf{x}_i)}$中的分量，即：

$$x_{ij}(\beta_{c_1,j}-\beta_{c_2,j})$$

显然，只要$\beta_{c_1,j} = \beta_{c_2,j}$，则这个单词的存在与否不影响分类，这一条件蕴含：

$$\theta_{c_1,w_j} = \theta_{c_2,w_j}$$

考察在赋予先验分布$p(\theta_{cw} = Beta(1,1))$时的后验分布，记c类文档总数为$N_c$：

$$p(\theta_{cw}|\mathbf{X}) = p(\theta_{cw}) \prod_{i=1}^{N_c} p(\mathbf{i}|\theta_{cw})$$

$$= \prod_{i=1}^{N_c} \theta_{cw}^{x_{iw}} (1-\theta^{cw})^{1-x_{iw}} = \theta_{cw}^{\sum x_{iw}} (1-\theta_{cw})^{N_c - \sum x_{iw}}$$

即服从Beta$(\sum x_{iw} + 1, N_c - \sum x_{iw} + 1)$，其均值为：

$$mean(\theta_{cw}) = \frac{\sum x_{iw} + 1}{N_c + 2}$$

考虑一个单词w_c出现在所有的文档中，并且c_1和c_2类各有N_1与N_2篇文档。此时：

$$mean(\theta_{c_1,w_c}) = 1 - \frac{1}{2 + N_1}$$

$$mean(\theta_{c_2,w_c}) = 1 - \frac{1}{2 + N_2}$$

可见如果$N_1 \neq N_2$，则使用平均值估计的话，单词w_c仍旧会对分类产生影响，但这影响本身会随着样本数量的增加（即N的增加）而降低。

3.20

NBC的实质是将作为特征的结点之间的所有连接在PGM的层面切断，而一个全连接的图包含$O(D^2)$数量的边，以及$O(2^D)$级别的参数。相对应的分析可见概率图模型部分。

3.21 朴素贝叶斯分类器中的互信息

互信息式用来度量一个特征X和分类Y的关联性：

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

在特征为二进制表示时，第一项求和仅包含两种情况，代入$\pi_c = p(y = c), \theta_{jc} = p(x_j = 1|y = c), \theta_j = p(x_j = 1)$：

$$I_j = \sum_c p(x_j = 1, c) \log \frac{p(x_j = 1, c)}{p(x_j = 1)p(c)} + \sum_c p(x_j = 0, c) \log \frac{p(x_j = 0, c)}{p(x_j = 0)p(c)}$$

$$= \sum_c \pi_c \theta_{jc} \log \frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc})\pi_c \log \frac{1 - \theta_{jc}}{1 - \theta_j}$$

得到式3.76。

3.22

Straightforward calculation.

4 高斯模型

4.1

$$cov(X,Y) = \int\int (X - \mathbb{E}(X))(Y - \mathbb{E}(Y))p(X,Y)dXdY$$

$$= \int_{-1}^{1} X(X^2 - \frac{1}{\sqrt{3}})dX = 0$$

最后一步是因为积分的是一个奇函数。

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

4.2

求随机变量Y的分布密度函数：

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

利用X关于0的对称性，得到Y也服从$(0,1)$的正态分布。
利用题干信息：

$$cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X) - \mathbb{E}(Y) = \mathbb{E}_W(\mathbb{E}(XY|W))$$

$$= 0.5 \cdot (\mathbb{E}(X^2) + \mathbb{E}(-X^2)) = 0$$

4.3

不失一般性，假设：

$$\mathbb{E}(X) = \mathbb{E}(Y) = 0$$

$$\mathbb{E}(X^2) = \mathbb{E}(Y^2) = 1$$

即：

$$\int X^2 p(X)dX = \int Y^2 p(Y)dY = 1$$

则：

$$\int\int X^2Y^2 p(X,Y)dXdY = \int X^2(\int Y^2 p(Y)p(X|Y)dY)dX$$

又：

$$p(X) = \int p(X,Y)dY = \int p(Y)p(X|Y)dY \geq \int Y^2 p(Y)p(X|Y)dY$$

所以得证：

$$\mathbb{E}(X^2Y^2) \leq \int X^2 p(X)dX = 1$$

4.4

当$Y = aX + b$时：

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

$$var(Y) = a^2 var(X)$$

所以：

$$cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = a\mathbb{E}(X^2) + b\mathbb{E}(X) - a\mathbb{E}^2(X) - b\mathbb{E}(X)$$

$$= a \cdot var(X)$$

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

所以：

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

4.5

SVD对角化后对于每个维度独立地积分（此时转化为一维正态分布的归一化）证明即可。

4.6

Straightforward algebra.

4.7

利用4.69的结论，Straightforward algebra.

4.8

Practical...

4.9 已知方差时的两个传感器度数

分别以$Y^{(1)}$和$Y^{(2)}$记两个观测集，大小分别为N_1, N_2首先求似然概率：

$$p(Y^{(1)}, Y^{(2)}|\mu) = \prod_{n_1=1}^{N_1} p(Y_{n_1}^{(1)}|\mu) \prod_{n_2=1}^{N_2} p(Y_{n_2}^{(2)}|\mu)$$

$$\propto \exp\left\{A \cdot \mu^2 + B \cdot \mu\right\}$$

其中：

$$A = -\frac{N_1}{2v_1} - \frac{N_2}{2v_2}$$

$$B = \frac{1}{v_1} \sum_{n_1=1}^{N_1} Y_{n_1}^{(1)} + \frac{1}{v_2} \sum_{n_2=1}^{N_2} Y_{n_2}^{(2)}$$

通过求导数置零可得最大似然估计：

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

该模型的共轭先验分布应该具有正比于$\exp\{A \cdot \mu^2 + B \cdot \mu\}$的形式，故取正态分布为先验分布：

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

后验分布的形式为：

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

最大后验估计为：

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

容易发现，当A和B的绝对值随着观测次数的增加而上升时：

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

后验分布也是正态分布，通过变换幂函数中μ各次项的系数可得：

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

$$mean_{MAP} = \sigma_{MAP}^2 (B + b)$$

4.10 正态分布的边缘和条件分布

见PRML第二章。

4.11 正态-逆Wishart分布

本题考虑多维正态分布（MVN）的贝叶斯估计。

MVN的似然估计为：

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

4.195的处理为：

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

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

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

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

第二个等式利用了平均值条件消去了一项。

MVN的参数(μ,Σ)的共轭先验分布为Normal-inverse-Wishart（NIW）分布，定义为：

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

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

此时已经很容易给出后验分布的形式：

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

其中：

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

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

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

通过对齐$|\Sigma|$，$\mu^T\Sigma^{-1}\mu$和μ^T的次数可得。

再利用$A^T\Sigma^{-1}A = Tr\left\{A^T\Sigma^{-1}A\right\} = Tr\left\{\Sigma^{-1}AA^T\right\}$对齐幂函数中的常数项可得：

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

所以：

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

利用平均值的定义可以得到4.214的结果，因为：

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

所以MVN的后验分布为$NIW(\mathbf{m}_{\mathbf{X}}, k_{\mathbf{X}}, v_{\mathbf{X}}, \mathbf{S}_{\mathbf{X}})$

4.12

Straightforward calculation.

本题的关键点在于模型判别条件4.273。它指出：更复杂的模型（参数更多）虽然能够带来更大的似然性，但是贝叶斯方法会对于其过多的参数本身进行惩罚，惩罚的度量和似然性的关系式由4.273给出，因此更深刻的讨论放在5.3.2.4节中进行。

4.13 高斯后验分布的置信区间

根据题意，对于一个一维的正态分布，假设其均值服从先验分布：

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

而实际观测变量服从分布：

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

观测n个变量，要求μ的后验分布概率密度函数在一个长度为1的区间的概率和为0.95。

首先计算μ的后验分布：

$$p(\mu|D) \propto p(\mu)p(D|\mu) = N(\mu|\mu_0, \sigma_0^2) \prod_{i=1}^{n} N(x_n|\mu, \sigma^2)$$

$$\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}$$

$$= \exp\left\{(-\frac{1}{2\sigma_0^2} - \frac{n}{2\sigma^2})\mu^2 + ...\right\}$$

所以后验方差为：

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

由于正态分布0.95的质量集中在-1.96σ到1.96σ之间，我们得到：

$$n \geq 611$$

4.14 一维高斯分布的最大后验估计

考察一维正态分布后验估计的一些直觉性质，首先假设分布的方差σ^2已知，均值μ服从确定均值m、方差s^2的正态分布，首先类似上一题给出后验分布的形式：

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

类似上一题，我们解出后验分布的形式依旧为正态分布，其参数通过考察幂函数内μ的各次系数得到，其中二次项系数为：

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

一次项系数为：

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

所以此时的方差和均指分别为：

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

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

最大似然估计为：

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

可以发现当N增大时，μ_{post}趋向于μ_{ML}。

如果只考虑s^2的变化，当其增大时，后验估计趋向于最大似然估计；当其减小时，后验估计趋向于先验均值。先验方差的物理意义是我们对于这个先验假设的信任程度，直觉上，先验方差越大代表对于先验假设越不信任，所以结果趋向于似然估计，否则趋向于先验估计。

4.15

利用关系：

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

剩余部分为Straightforward algebra.

4.16 高斯分布的似然比值

考虑一个二类分类器，两个类的生成概率均为正态分布$p(x|y = C_i) = N(x|\mu_i, \Sigma_i)$，贝叶斯公式直接给出：

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

上式右侧的第二项是先验概率比的对数，先考虑似然概率比的对数。

当两类协方差矩阵任意时：

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

没有可化约的项，不过值得注意的是此时似然估计给出的决策边界是一个D维空间中的二次曲面。

当两个协方差矩阵都等于Σ时：

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

此时决策边界变为一个平面。

对于协方差矩阵为同一个对角阵的情况，解析形式和上式相仿，不过一些矩阵乘法可以转换成向量内积。

对于协方差矩阵为同一个单位矩阵的倍数时，上式中的矩阵乘法可以继续转化为数乘。

4.17

Practise by youself.

4.18

Straightforward calculation.

4.19

直接计算（参数条件是共有的，所以略去）：

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

假设先验分布相同，上式变为：

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

$$= \frac{1}{k^{\frac{D}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0) + \frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)\right\} + 1}$$

$$= \frac{1}{k^{\frac{D}{2}} \exp\left\{-\frac{1}{2}(1-\frac{1}{k})\mathbf{x}^T\Sigma_0^{-1}\mathbf{x} + \mathbf{x}^T\mathbf{u} + c\right\} + 1}$$

这里利用了：

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

这里的决策边界依然为一个二次曲面，当$k = 1$时退化为平面，当k增大时，决策面形成一个包络μ_0的曲面，当k趋向于无穷大时，决策面退化为$y = 0$类型的等概率面，此时意味着所有以外的区域均从属于一个方差无穷大的正态分布。

4.20

我们直接利用这个结论“最大似然估计能够过拟合数据，过拟合的程度一般而言正相关于模型复杂度。”来进行定性分析。

GaussI假设协方差矩阵正比于单位矩阵；

GaussX对于协方差矩阵不做假设；

LinLog等价于假设多个协方差矩阵相等；

QuadLog对于协方差矩阵不做假设。

四种方案同时假设给定类型时的生成概率服从正态分布，显然地，从复杂程度来看：

QuadLog =GaussX > LinLog > GaussI

所以最大似然估计的准确率按照上述顺序排列。

对于e问，一般而言不一定成立，因为乘积更大并不蕴含和也更大。

4.21

Straightforward albegra.

4.22

Straightforward calculation.

4.23

Practice by yourself.

5 贝叶斯统计

5.1 共轭先验分布的混合仍是共轭先验分布

式5.69和式5.70的推导只需要从贝叶斯公式出发，形式地进行：

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

其中：

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

5.2 概率分类的最佳阈值

此时的后验损失期望为：

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

$$= L(\hat{y}, 1) + p_0(L(\hat{y}, 0) - L(\hat{y}, 1))$$

解得选取两种结论导致同等损失期望时：

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

当$p_0 \geq \hat{p_0}$时，应该估计为$\hat{y} = 0$。

5.3 分类中的拒绝选项

后验损失期望为：

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

我们记此时后验概率最大的类别为$\hat{c}$，即：

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

则此时的决策行为只有两种：一是$a = \hat{c}$，二是$a = reject$。
选择$a = \hat{c}$时，损失期望：

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

选择reject行为时，损失期望：

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

选择$a = \hat{c}$而并非reject的条件是：

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

也即：

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

5.4

Straightforward calculation.

5.5

根据：

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

可得：

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

置其为零，并利用$\int_0^Q f(D)fD + \int_Q^{+\infty} f(D)dD = 1$得到平衡条件：

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

5.6

Practise by yourself.

5.7 贝叶斯模型平均提高预测准确率

将5.127左右两式分别形式展开并交换求和次序，易得：

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

而：

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

所以左式减去右式所得差为：

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

所以左式恒小于等于右式。

5.8 二维离散分布的最大似然估计和模型选择

首先易得联合分布$p(x, y|\theta_1, \theta_2)$：

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

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

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

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

归纳为：

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

最大似然估计很容易给出：

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

其中联合分布如上所示，可得：

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

当$\mathbf{X}, \mathbf{Y}$独立给出时，两个参数可以独立地估计。
如果进一步将联合分布化为形式：

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

则：

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

再利用归一化条件可进行最大似然估计。

只进行拟合度比较时，参数更多的模型（4参数模型）明显占优，因为其可以过拟合。但进行贝叶斯模型估计时，还需要利用BIC定量计算该过拟合的似然度增加和自由度惩罚的大小。

The rest is straightforward algebra.

5.9 L1损失函数最优化

此时的后验损失期望（此处我们不引起异议地略去条件中的D）：

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

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

求导：

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

解得：

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

5.10

给定条件：

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

此时5.115式的临界条件为：

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

再利用：

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

可得概率阈值为$\frac{c}{1+c}$。

6 Frequentist统计

本节介绍了一种在贝叶斯统计以外的统计思路，和贝叶斯流派的机器学习思想有所出入，习题较基本，可从任何一本概率与统计教材中寻找参考，故不做详细解答。

7 线性回归

7.1

一开始训练集较少时，训练出来的模型是过拟合于当前数据集的，所以正确率能达到很高。在训练集增大时，模型不得不去学习适应更一般场合的参数，所以侧面地降低了过拟合效应，从而导致准确率降低。

正如7.5.4节所指出的，增大训练集合是除了增加正则项以外的一种重要的扼制过拟合的方法。

7.2

Straightforward calculation.

7.3

我们将$\mathbf{x}$改成$(\mathbf{x}^T, 1)^T$使得没有必要额外设置w_0，此时NLL的形式为：

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

所以：

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

故：

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

7.4 线性回归中噪声方差的最大似然估计

首先给出似然函数：

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

$$= \prod_{n=1}^{N} N(y_n|\mathbf{w}^T\mathbf{x}_n, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2\right\}$$

对于σ^2：

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

取得：

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

7.5

计算NLL：

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

分别对两个参数求导：

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

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

对于$\mathbf{X}$和$\mathbf{y}$分别进行中心化处理：

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

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

此时中心化的数据集均值都为0，所以此时的线性回归模型中没有w_0这一项，同时可得：

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

7.6

直接利用习题7.5的结论，Straightforward algebra.

7.7 在线线性回归的充分统计量

a和b题如提示所示。

c将提示中的x替代为y同理可证。

d题中我们证明：

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

将两侧的C_{xy}展开并代入$\bar{x}^{(n+1)} = \bar{x}^{(n)} + \frac{1}{n+1}(x_{n+1} - \bar{x}^n)$易证。

e和f题：Practice by yourself.

7.8 已知噪声方差时的一维贝叶斯线性回归

a题：Practice by yourself.

b选择先验分布：

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

将其化约为：

$$p(\mathbf{w}) = N(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0) \propto$$
$$\exp\left\{-\frac{1}{2}\mathbf{V}_{0,11}^{-1}(w_0 - w_{00})^2 - \frac{1}{2}\mathbf{V}_{0,22}^{-1}(w_1 - w_{01})^2 - \mathbf{V}_{0,12}^{-1}(w_0 - w_{00})(w_1 - w_{01})\right\}$$

为了使得其形式符合给出的变分先验分布形式，取：

$$w_{01} = 0$$

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

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

$$w_{00} = arbitrary$$

注意到虽然给出了precision matrix的形式，但是这是一个奇异（半正定）矩阵，所以这不是一个良好定义的正态分布。

在c中我们考虑参数的后验分布：

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

考察上式右侧的幂函数中w_1的二次项和一次项系数，分别为：

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

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

可得参数w_1的后验方差和后验均值为：

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

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

可以看出样本的积累对于后验方差的缩减效果。

7.9 线性回归的生成模型

为了方便起见，我们仅考虑已经中心化的数据集（暂不改变符号），此时：

$$w_0 = 0$$

$$\mu_x = \mu_y = 0$$

根据协方差的定义：

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

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

利用4.3.1的结论：

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

其中：

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

一般而言，生成模型和判别模型的区别在于前者的参数数量更多、更复杂，但是有独立生成原始数据的能力。

7.10 g-先验时的贝叶斯线性回归

首先回顾一下线性回归中Ridge回归模型的推理，似然函数为：

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

先验分布取正态-逆伽马分布：

$$p(\mathbf{w}, \sigma^2) = NIG(\mathbf{w}, \sigma^2|\mathbf{w}_0, \mathbf{V}_0, a_0, b_0) = N(\mathbf{w}|\mathbf{w}_0, \sigma^2\mathbf{V}_0)IG(\sigma^2|a_0, b_0)$$

$$= \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\sigma^2\mathbf{V}_0|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T(\sigma^2\mathbf{V}_0)^{-1}(\mathbf{w} - \mathbf{w}_0)\right\} \cdot$$

$$\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}\exp\left\{-\frac{b_0}{\sigma^2}\right\}$$

$$= \frac{b_0^{a_0}}{(2\pi)^{\frac{D}{2}}|\mathbf{V}_0|^{\frac{1}{2}}\Gamma(a_0)}(\sigma^2)^{-(a_0+\frac{D}{2}+1)} \cdot \exp\left\{-\frac{(\mathbf{w} - \mathbf{w}_0)^T\mathbf{V}_0^{-1}(\mathbf{w} - \mathbf{w}_0) + 2b_0}{2\sigma^2}\right\}$$

后验分布形式：

$$p(\mathbf{w},\sigma^2|D) \propto p(\mathbf{w},\sigma^2)p(D|\mathbf{w},\sigma^2)$$

$$\propto \frac{b_0^{a_0}}{(2\pi)^{\frac{D}{2}}|\mathbf{V}_0|^{\frac{1}{2}}\Gamma(a_0)}(\sigma^2)^{-(a_0+\frac{D}{2}+1)}\cdot\exp\left\{-\frac{(\mathbf{w}-\mathbf{w}_0)^T\mathbf{V}_0^{-1}(\mathbf{w}-\mathbf{w}_0)+2b_0}{2\sigma^2}\right\}\cdot$$

$$(\sigma^2)^{-\frac{N}{2}}\cdot\exp\left\{-\frac{\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2}{2\sigma^2}\right\}$$

比较σ^2的系数得到：

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

比较$\mathbf{w}^T\mathbf{w}$的系数得到：

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

比较$\mathbf{w}$的系数得到：

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

所以：

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

最后对齐幂函数里的常数项（和$\mathbf{w}$无关）得到：

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

以上得到了7.70到7.73式，最后可得7.69的重要结论：

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

本题中的所有结论只需按题干代入参数的先验值即可得到。

8 逻辑回归

8.1

Practice by yourself.

8.2

Practice by yourself.

8.3 逻辑回归损失函数的梯度与海森矩阵

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

$$g(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} NLL(\mathbf{w}) = \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]$$

$$= \sum_{n=1}^{N} y_i \frac{1}{\sigma}\sigma(1-\sigma) - \mathbf{x}_i + (1-y_i)\frac{-1}{1-\sigma}\sigma(1-\sigma) - \mathbf{x}_i$$

$$= \sum_{n=1}^{N} (\sigma(\mathbf{w}^T\mathbf{x}_i) - y_i)\mathbf{x}_i$$

对于一个任意的（形状正确的）非零向量$\mathbf{u}$：

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

因为$\mathbf{S}$为正定矩阵，所以对于任何形状正确的非零$\mathbf{v}$：

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

因为$\mathbf{X}$满秩，所以$\mathbf{X}\mathbf{u}$非零，所以：

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

所以$\mathbf{X}^T\mathbf{S}\mathbf{X}$为正定矩阵。

8.4 多元逻辑回归损失函数的梯度与海森矩阵

通过每次仅考虑一个独立的分量，我们可以排除张量积带来的形式复杂性，考虑对个特定的类型对应的向量$\mathbf{w}^*$：

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

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

将对于所有类别的解合并到一个矩阵并对n求和给出8.38。

求解海森矩阵的形式时，我们考虑依次对于$\mathbf{w}_1$和$\mathbf{w}_2$两个分量求梯度：

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

当$\mathbf{w}_1$与$\mathbf{w}_2$实际上是同一个类型时：

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

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

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

当$\mathbf{w}_1$与$\mathbf{w}_2$不表示同一个类型时：

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

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

得到8.44。

$\sum_c y_{nc} = 1$的条件在8.34到8.35的过程中使用。

8.5

一个值得注意的核心思想是：加入正则项一方面等价于进行一次最大后验估计，即加入一个先验分布；一方面等价于引入一个拉格朗日乘子，进行一个额外的约束。本题中引入的先验分布为一个协方差为单位矩阵倍数的正态分布，等价的约束为$w_{cj}=0$。

在取到最优解时式8.47中的梯度为0，此时我们近似认为$\hat{\mu}_{cj} = y_{cj}$，则$g(\mathbf{W})=0$，额外的正则项条件即$\lambda\sum_{c=1}^{C}\mathbf{w}_c=0$，它等价于一组由$D$个线性方程组表达的$D$个约束，约束形式为：对于$j=1...D$，$\sum_{c=1}^{C}\hat{w}_{cj}=0$。

8.6 l2正则项的基本性质

$J(\mathbf{w})$的第一项之海森矩阵为正定矩阵（8.7），第二项的海森矩阵仍为正定矩阵（$\lambda>0$），所以这个函数的海森矩阵正定，故其有全局最优解。

考察后验分布的形式：

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

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

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

所以：

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

全剧最优解中零的数量和λ的取值有关，λ值负相关于**w**的先验不确定性，其不确定性越小，则**w**越趋向于零向量，所以最后取得的最终解中零也会越多。

如果$\lambda=0$，即先验不确定性无穷大，后验估计变为最大似然估计。因为此时对于**w**没有限制，所以有可能有**w**的分量趋近于无穷。

当λ增大时，意味着先验的不确定性减少，所以整个模型的过拟合性质被减弱，一般而言这会导致训练集上的准确率下降。

同时，过拟合性质的减弱一般而言能带来在测试集上的准确率上升，不过这并不一定发生。

8.7

Practice by yourself.

9 泛型线性模型与幂分布族

9.1 一元高斯分布的共轭先验分布属于幂分布族

一维正态分布的形式为：

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

整理为：

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

记$\theta=(-\frac{\lambda}{2},\lambda\mu)^T$，$A(\theta)=\frac{\lambda\mu^2}{2}+\frac{\ln(2\pi)}{2}-\frac{\ln\lambda}{2}$，$\phi(x)=(x^2,x)^T$考虑对于数据集$D$的似然函数：

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

根据先验分布的意义，我们设置一个观测场景来定义一个先验分布，依据幂分布族的形式，我们只需预设充分统计量，下面设先验统计了M次，其中平方项的均值和一次项的均值分别为v_2和v_1，则先验分布的形式为：

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

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

注意到这个先验分布的参数数量为三，下面证明它等价于$p(\mu,\lambda)=N(\mu|\gamma,\frac{1}{\lambda(2\alpha-1)})Ga(\lambda|\alpha,\beta)$，将这一分布展开到幂函数内并略去和$\mu,\lambda$无关的变量：

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

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

分别对齐$\lambda\mu^2,\lambda\mu,\lambda,\ln\lambda$这几项的系数，得到下列等式：

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

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

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

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

第一个与第四个等价，联立解得：

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

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

$$\gamma=v_2$$

可见进行参数变换后，这两个分布等价。

9.2 多元高斯分布属于幂分布族

MVN的充分统计量为$s_{ij}=x_ix_j$和$s_i=x_i$，其中$1\leq i,j\leq D$，相应的系数可从4.87中直接得到。

10 有向图模型

这章写得太垃圾了，不想做习题解答。

11 混合模型与期望最大算法

11.1 学生分布作为混合高斯分布

Student-t分布的一维形式为：

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

考察11.61的右侧：

$$\int N(x|\mu,\frac{\sigma^2}{z})Ga(z|\frac{v}{2},\frac{v}{2})dz$$

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

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

此时积分内的函数是伽马分布$Ga(z|\frac{v+1}{2},\frac{(x-\mu)^2}{2\sigma^2}+\frac{v}{2})$中和$z$相关的项，所以积分的结果是这一分布的归一化系数的倒数。

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

代入可证得两侧相等。

11.2 混合高斯分布的EM算法

混合高斯模型的最大似然估计为最优化如下函数：

$$Q(\theta,\theta^{old})=\mathbb{E}p(z|D,\theta^{old})[\sum_{n=1}^{N}\log(\mathbf{x}_n,\mathbf{z}_n|\theta)]$$

$$=\sum_{n=1}^{N}\mathbb{E}[\log\prod_{k=1}^{K}(\pi_k p(\mathbf{x}_n|z_k,\theta))^{z_{nk}}]$$

$$=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\log(\pi_k p(\mathbf{x}_n|z_k,\theta))$$

其中：

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

对于激发概率$p(\mathbf{x}|z,\theta)$为高斯分布的情况，首先考虑$Q(\theta,\theta^{old})$中与μ_k有关的项：

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

对其求导数并置零，得到平衡条件为：

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

解得11.31：

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

考虑$Q(\theta,\theta^{old})$中和Σ_k有关的项：

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

使用和4.1.3.1相同的手段：

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

它的平衡条件为：

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

得到11.32：

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

11.28由$Q(\theta,\theta^{old})$中的相关项加上附加的约束项$\lambda(1-\sum_k \pi_k)$后求导置零得到。

11.3 混合伯努利分布的EM算法

混合伯努利分布的最大似然估计中，考虑（其中$D=2$是可选元素的数量）：

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

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

引入拉格朗日乘子来约束$\sum_j \mu_{kj} = 1$，得到此时的导数为零条件：

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

对于所有的j求和：

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

得到：

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

代入得到11.116。

引入先验分布：

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

此时导数为零条件变化为：

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

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

而：

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

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

代入得到11.117。

11.4 混合学生分布的EM算法

学生分布模型中，完整数据项的似然对数为：

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

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

和式中关于v的项之和为：

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

在全数据集上的似然为：

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

求导数并置零得到平衡条件，也即v的M-step最大似然估计，通过数值求解：

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

对于μ和Σ而言：

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

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

所以对于μ和Σ的最大似然估计和在多元正正态分布中类似推导即可。

11.5 混合高斯分布的梯度下降算法

根据题干：

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

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

直接对μ_k求偏导数：

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

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

对π_k求导数：

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

我们使用拉格朗日乘子法而不是题干中的softmax归一化方法，得到平衡条件：

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

对k求和并归一化，得到：

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

对于Σ_k而言：

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

其中：

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

$$\left\{\nabla_{\Sigma_k}(-\frac{1}{2}(\mathbf{x}-\mu_k)\Sigma_k^{-1}(\mathbf{x}-\mu_k)) - \Sigma_k^{-1}\nabla_{\Sigma_k}|\Sigma_k|\right\}$$
$$= N(\mathbf{x}|\mu_k, \Sigma_k)\nabla(\log N(\mathbf{x}|\mu_k, \Sigma_k))$$

之后利用正态对数似然对协方差的求解公式即可，此时解出：

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

11.6 混合异方差高斯分布的EM算法

注意到J和K的选取是相互独立的，直接利用贝叶斯公式（略去条件中的参数θ）：

$$p(J_n = j, K_n = k|x_n) = \frac{p(J_n = j, K_n = k, x_n)}{p(x_n)} = \frac{p(J_n = j)p(K_n = k)p(x_n|J_n = j, K_n = k)}{\sum_{J_n, K_n} p(J_n, K_n, x_n)}$$

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

也可以当做底层的发生分布为ml个独立的分量，则上式的意义更加明显。

下面求出辅助函数$Q(\theta^{new}, \theta^{old})$的形式：

$$Q(\theta^{new}, \theta^{old}) = \mathbb{E}_{\theta^{old}} \sum_{n=1}^{N} \log p(x_n, J_n, K_n|\theta^{new})$$

$$= \sum_{n=1}^{N} \mathbb{E}[\log(\prod_{j=1}^{m}\prod_{k=1}^{l} p(x_n, J_n, K_n|\theta^{new})^{\mathbb{I}(J_n=j,K_n=k)})]$$

$$= \sum_{n=1}^{N}\sum_{j=1}^{m}\sum_{k=1}^{l} \mathbb{E}(\mathbb{I}(J_n = j, K_n = k))(\log p_j + \log q_k + \log N(x_n|\mu_j, \sigma_k^2))$$

$$= \sum_{n,j,k} r_{njk} \log p_j + \sum_{n,j,k} r_{njk} \log q_k + \sum_{njk} r_{njk} \log N(x_n|\mu_j, \sigma_k^2)$$

我们理论上要优化四种参数p, q, μ, σ^2，观察在Q中的形式可发现p和q可以独立地优化而不考虑和其他变量的协变关系，现考虑在σ^2固定的情况下优化μ：

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

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

解得：

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

11.7

Practise by yourself.

11.8 混合高斯分布的矩

混合正态分布的期望：

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

$$= \sum_k \pi_k \mu_k$$

利用$cov(\mathbf{x}) = \mathbb{E}(\mathbf{x}\mathbf{x}^T) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^T$，我们求：

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

其中：

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

$$= \Sigma_k + \mu_k \mu_k^T$$

所以：

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

11.9

Practise by yourself.

11.10 K-means的损失函数

对于k进行求和时的每一项，对内侧和外侧求和记号依次应用11.134：

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

$$= n_k^2 s^2 + n_k(n_k s^2) = 2n_k s_k$$

而11.131中右式对应于k的求和结果：

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

故11.131成立。

11.11 混合高斯分布的全显似然属于幂分布族

首先，以热洞形式编码隐含元，即$z_c = \mathbb{I}(x由c对应的分布产生)$。此时（为简洁起见略去条件中的$\theta$）：

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

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

写出联合分布的对数：

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

$$= \sum_{c=1}^{C} z_c(\log\pi_c - \frac{1}{2}\log 2\pi\sigma_c^2 - \frac{1}{2\sigma_c^2}(x-\mu_c)^2)$$

上式从形式上是一些内积的代数和，所以服从幂分布族的形式，故充分统计量为$\mathbf{z}$，$\mathbf{z}x$和$\mathbf{z}x^2$的线性组合。

11.12

利用11.4.5节得出的学生分布中关于μ的全数据集上似然函数：

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

对其求偏导数并置零：

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

解得：

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

11.13

对于每个分别的j，式5.90各自不同（这一步骤等价于E-step）：

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

这一积分消去了隐元θ_j，故此时的边缘似然对数为：

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

下面分别最优化（等价于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满足：

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

11.14

Unsolved.

11.15 截断高斯分布的后验均值与方差

我们记$A = \frac{c_i-\mu_i}{\sigma}$，对于均值有：

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

而：

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

最后一步代入了11.141和11.139，向上代入得到：

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

下面求平方的期望：

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

为求$\mathbb{E}[\epsilon_i^2|\epsilon_i \geq A]$，我们延续题干中提示的思路：

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

得到：

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

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

再代入第一问的结论得到：

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

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

12 隐含元线性模型

12.1 FA模型的M-step

此处完整重复FA（Fator-Analysis）的EM求解过程，首先我们有基本的（将$\mathbf{X}$中心化消去变量μ）：

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

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

以及：

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

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

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

以$\mathbf{z}_n$为$\mathbf{x}_n$对应的隐含变量，则全数据集$\{\mathbf{x}, \mathbf{z}\}$的似然对数为：

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

其中$\log p(\mathbf{z})$是先验项，参数为0和I，所以可以直接略去，故：

$$Q(\theta, \theta^{old}) = \mathbb{E}_{\theta^{old}}[\sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{z}_n, \theta)]$$

$$= \mathbb{E}[\sum_{n=1}^{N} c - \frac{1}{2}\log|\Psi| - \frac{1}{2}(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)^T\Psi^{-1}(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)]$$

$$= C - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)^T\Psi^{-1}(\mathbf{x}_n - \mathbf{W}\mathbf{z}_n)]$$

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

$$= C - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N}\mathbf{x}_n\Psi^{-1}\mathbf{x}_n - \frac{1}{2}\sum_{n=1}^{N}Tr\left\{\mathbf{W}^T\Psi^{-1}\mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T]\right\} + \sum_{n=1}^{N}\mathbf{x}_n^T\Psi^{-1}\mathbf{W}\mathbb{E}[\mathbf{z}_n]$$

根据$p(\mathbf{z}|\mathbf{x}, \theta^{old}) = N(\mathbf{z}|\mathbf{m}, \Sigma)$，有：

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

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

以下在求期望时略去条件中的$\mathbf{x}$和旧参数θ^{old}。

首先对$\mathbf{W}$优化：

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

将其置为零得到：

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

对Ψ^{-1}优化：

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

代入上述$\mathbf{W}$的表达式得到：

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

维持Ψ为对角矩阵的假设，取：

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

本题的解答来自论文“The EM Algorithm for Mixtures of Factor Analyzers, Zoubin Gharamani, Geoffrey E.Hinton, 1996”，文中也提供了混合FA模型的EM算法。

12.2 FA模型的最大后验估计

假设先验分布$p(\mathbf{W})$和$p(\Psi)$，和上一题的算法相比，只需要在M-step中：

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

获得解析即可。

12.3

Need pictures for illustration here!

12.4 第二主成分的导出

对于：

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

考察对$\mathbf{z}_2$一个分量的导数：

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

利用假设$\mathbf{v}_2^T\mathbf{v}_2 = 1$和$\mathbf{v}_2^T\mathbf{v}_1 = 0$，得到：

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

利用$\mathbf{C}$是对称矩阵的性质，以及$\mathbf{v}_1$和$\mathbf{v}_2$上的约束，我们首先对$\mathbf{C}$进行奇异值分解：

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

其中：

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

依次为$\mathbf{C}$的特征值从大到小排列（均为非负）。

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

依次为相对应的模为1的特征向量，它们两两正交$\mathbf{u}_i^T\mathbf{u}_j = \mathbb{I}(i = h)$，其中$\mathbf{u}_1 = \mathbf{v}_1$。

我们想要在$\mathbf{v}_2^T\mathbf{v}_2 = 1$以及$\mathbf{v}_2^T\mathbf{v}_1 = 0$的条件下最小化：

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

注意到$\mathbf{O}\mathbf{v}_2$的意义是$\mathbf{v}_2$进行一次正交变换，所以其模仍然为1，而$(\mathbf{O}\mathbf{v}_2)^T\Lambda(\mathbf{O}\mathbf{v}_2)$度量的是该向量的各个分量的平方与$\Lambda$中特征值相乘以后的和，所以此时的最优解是将所有的模集中到对应最大特征值的分量上，这就意味着：

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

所以：

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

通过矩阵微分引入拉格朗日乘子后的函数一样可得。

12.5 PCA的残差

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

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

代入$\mathbf{v}_i^T\mathbf{v}_j = \mathbb{I}(i=j)$，$z_{nj} = \mathbf{x}_n^T\mathbf{v}_j$，得到a的结论：

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

再代入$\mathbf{v}_j^T\mathbf{C}\mathbf{v}_j = \lambda_j$并对$n$求和可直接得到b的结论。

将$K = d$代入b的结论，得到：

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

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

代入一般的情况：

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

12.6 Fisher判别法

Straightforward algebra.

Fisher判别算法在处理多类分类问题时比较复杂，需要利用到一些额外的假设和近似。（need reference）

12.7

本题使用的思路就是本解答在12.4给出的思路，故此处不加赘述。

12.8

Practice by yourself.

12.9

wtf$\mathbf{x}_v$?
wtf$\mathbf{x}_h$?

12.10

因为：

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

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

利用第4章的结论：

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

12.2.4节中最大似然估计的推导方法见“Probabilistic Principal Component Analysis,Michael E.Tipping, Christopher M.Bishop,1999”。

代入最大似然估计，此时协方差矩阵（$D * D$）的逆可以如下计算：

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

其中只需要求一个$L * L$矩阵的逆。

12.11

Practice by youself.

13 稀疏线性模型

13.1 RSS的偏导数

定义：

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

直接地：

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

$$= -\sum_{n=1}^{N}2(x_{nj}y_n - x_{nj}\sum_{i=1}^{D}w_i x_{ni}) = -\sum_{n=1}^{N}2(x_{nj}y_n - x_{nj}\sum_{i\neq j}^{D}w_i x_{ni} - x_{nj}^2 w_j)$$

其中w_j的系数为：

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

其他无关项合并为：

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

最终：

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

13.2 线性回归的经验贝叶斯方法的M-step

此处给出自动关联判别（Automatic Relevance Determination，ARD）的EM推导，在一个线性回归场景中：

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

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

$$A = diag(\alpha)$$

在E-step要求出隐含元$\mathbf{w}$的期望，运用线性高斯关系得到：

$$p(\mathbf{w}|\mathbf{y},\alpha,\beta) = N(\mu,\Sigma)$$

$$\Sigma^{-1} = \mathbf{A} + \beta \mathbf{X}^T \mathbf{X}$$

$$\mu = \Sigma(\beta \mathbf{X}^T \mathbf{y})$$

即：

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

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

现求辅助函数：

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

$$= \mathbb{E}[\log p(\mathbf{y}|\mathbf{w}, \beta) + \log p(\mathbf{w}|)]$$

$$= \frac{1}{2}\mathbb{E}[N \log \beta - \beta(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \sum_j \log \alpha_j - \mathbf{w}^T \mathbf{A}^{-1} \mathbf{w}]$$

E-step取期望时，需要$\mathbb{E}[\mathbf{w}]$和$\mathbb{E}[\mathbf{w}\mathbf{w}^T]$，之前已完成计算。
引入α分量和β的先验分布：

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

故后验形式的辅助函数：

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

在M-step中，首先对α_i优化：

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

置零得到：

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

对β优化：

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

得到：

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

将期望展开可得到13.168式。

13.3

Unsolved.

本题同样考虑ARD的参数选取，但是使用EB而不是EM方案，首先在似然概率中通过积分消去隐变量$\mathbf{w}$：

$$p(\mathbf{y}|\alpha,\beta) = \int p(\mathbf{y}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

$$= \int N(\mathbf{y}|\mathbf{X}\mathbf{w},\beta\mathbf{I})N(\mathbf{w}|0,\mathbf{A}^{-1})d\mathbf{w}$$

$$= N(\mathbf{y}|0,\beta\mathbf{I}+\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T)$$

记此时的协方差矩阵为$\mathbf{C}$。

13.4

Straightforward algebra.

13.5 将Elastic net算法统一到lasso中

直接将13.196两侧展开，左侧：

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

$$= \mathbf{y}^T\mathbf{y} - c^2\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + c^2\lambda_2\mathbf{w}^T\mathbf{w} + \lambda_1|\mathbf{w}|_1$$

右侧有：

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

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

$$= \mathbf{y}^T\mathbf{y} + c^2\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + c^2\lambda_2\mathbf{w}^T\mathbf{w} + c\lambda_1|\mathbf{w}|_1$$

故13.196和13.195恒等。

这显示Elastic Net正则化方法，即选取l_1和l_0的线性组合为正则项的效果等效于一个lasso正则项。

13.6 稀疏如何导致线性回归中的参数收缩

回顾最大似然估计的过程，对于一般的最小二乘法：

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

引入条件$\mathbf{X}^T\mathbf{X} = I$：

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

求导：

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

得到：

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

Ridge回归中：

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

求导得到：

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

即：

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

Lasso回归利用次微分方法的求解在13.3.2节很详细描述，其最终估计为13.63：

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

考察图13.24，易发现黑色实线为OLS，灰色线为Ridge，虚线为lasso，并且$\lambda_1 = \lambda_2 = 1$。从该图中可以看出，Ridge造成最大似然估计向水平线的收缩，而lasso继而造成在相关性小于一个阈值时收缩到零。

13.7 Spike and Slab模型中的伯努利先验

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

积分消去中间变量π_d：

$$p(\gamma_d|\alpha_1,\alpha_2)=\frac{1}{B(\alpha_1,\alpha_2)}\int p(\gamma_d|\pi_d)p(\pi_d|\alpha_1,\alpha_2)d\pi_d$$

$$=\frac{1}{B(\alpha_1,\alpha_2)}\int \pi_d^{\gamma_d}(1-\pi_d)^{(1-\gamma_d)}\pi_d^{\alpha_1-1}(1-\pi_d)^{\alpha_2-1}d\pi_d$$

$$=\frac{1}{B(\alpha_1,\alpha_2)}\int \pi_d^{\alpha_1+\gamma_d-1}(1-\pi_d)^{\alpha_2+1-\gamma_d-1}d\pi_d$$

$$=\frac{B(\alpha_1+\gamma_d,\alpha_2+1-\gamma_d)}{B(\alpha_1,\alpha_2)}=\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{\Gamma(\alpha_1+\gamma_d)\Gamma(\alpha_2+1-\gamma_d)}{\Gamma(\alpha_1+\alpha_2+1)}$$

所以（N_1为γ中1的个数）：

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

$$=\frac{(\alpha_1+1)^{N_1}(\alpha_2+1)^{N-N_1}}{(\alpha_1+\alpha_2+1)^N}$$

而：

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

13.8

GSM和拉普拉斯分布的联系是一个重要的思想：

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

推荐的证明方法是对两边同时进行拉普拉斯变换/矩母函数变换。
所要求的：

$$\mathbb{E}[\frac{1}{\tau_j^2}|w_j]=\int\frac{1}{\tau_j^2}p(\tau_j^2|w_j)d\tau_j^2=\int\frac{1}{\tau_j^2}\frac{p(w_j|\tau_j^2)p(\tau_j^2)}{p(w_j)}d\tau_j^2$$

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

根据题干提示13.200，上式化为：

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

因为：

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

得到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。

13.9 先验分布服从拉普拉斯分布的Probit回归的EM算法

（第一个完整解出这道题的哥们发了篇Trans）

直接的Probit回归过程没有隐变量参与其中，引入拉普拉斯分布为线性因子$\mathbf{w}$的先验分布得到其lasso版本，因为拉普拉斯分布可表示为高斯分布的连续性混合，所以引入与$\mathbf{w}$维数相同的隐变量τ^2。Probit回归的PGM为：

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

所有变量的联合分布为：

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

为简单起见，我们将γ设置为常数。根据13.86：

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

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

所以：

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

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

在辅助函数$Q(\theta^{new},\theta^{old})$中，我们对$\theta^{old}$求期望，在本场景中参数为$\mathbf{w}$，隐变量为$\tau^2$，所以：

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

为构造辅助函数，取出$\log p(\tau^2, \mathbf{w}, \mathbf{y})$中所有和$\mathbf{w}$相关的项：

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

可见在E-step需要求的期望仍旧只有一项：

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

求法和13.4.4.3节相同，因为Probit和线性回归在这部分为止的PGM是相同的。

M-step的最优化和引入正态先验分布的Probit回归过程相同，这里不加复述。

13.10

Follow the hints and straightforward algebra.

13.11 投影梯度下降法

本题所要求的投影梯度下降法遵从这样的设计原理：

一般而言，可以通过在$\mathbf{w}$上进行梯度下降来进行优化。但是当模型在$\mathbf{w}$上有一些约束条件，而梯度下降可能打破这些约束时，就必须将每一次迭代的增量投影到一个使得新$\mathbf{w}$仍服从约束的空间中。

现在我们要求：

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

我们在线性回归的语境下讨论这个优化，即：

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

因为$\lambda||\mathbf{w}||_1$中有绝对值运算，故需要使用非平凡的优化方案，题干给出的方式是表达为：

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

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

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

即$\mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}$。此时有：

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

则原始的问题转化为：

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

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

改写成：

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

则上问题可重新表示为：

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

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

其中：

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

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

求梯度：

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

在最一般的梯度下降场景中，设置学习速率α，一般的梯度更新公式为：

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

在投影梯度下降法中，取$\mathbf{g}^k$：

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

并在迭代步骤中：

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

在原始论文中使用了更精细的梯度下降法来调整学习速率，可参考原文“Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems, Mario A.T.Figueiredo”。

13.12 分段损失函数的次微分

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

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

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

13.13

本题的讨论内容从理论体系上比较偏离机器学习概率论的主题，此处不做详细解答，推荐参考“Rigorous Affine Lower Bound Functions for Multivariate Polynomials and Their Use in Global Optimisation”。

14 核方法

这一页空着有点尴尬，我是不是该写点什么。
算了不写了，后半本再见。