Homework 1

PB20000296 郑滕飞

第一章：

1.

由于$det\ A = \sum_i A_{ij}\ C_{ij}$，其中$C_{ij}$代表代数余子式，因此类似分解可知$\left(\frac{\partial\ det\ A}{\partial A}\right)_{ij} = C_{ij}$，从而$\frac{\partial\ det\ A}{\partial A} = (A^*)^T$，其中$A^*$代表其伴随矩阵。由此，$\frac{\partial \ln det\ A}{\partial \mathrm{A}} = \frac{(A^*)^T}{det\ A} = (A^{-1})^T$。从而

$$\frac{\partial \ln det\ A}{\partial \mathrm{x}} = \sum_{\mathrm{ij}} \frac{\partial \ln det\ A}{\partial \mathrm{A_{ij}}} \frac{\partial \mathrm{A}_{ij}}{\partial \mathrm{x}} = \mathrm{tr}\left(\left(\frac{\partial \ln det\ A}{\partial \mathrm{A}}\right)^T \frac{\partial A}{\partial x}\right) = \mathrm{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)$$

2.

假设色泽、根蒂、敲声的取值各有 a，b，c 种可能，则总的取值有 abc 种可能。对于每种取值，可以判断其为好瓜或坏瓜，共有2^{abc}(包含好瓜集合为空集的情况)种情况。注意到，此时通配的情况已被展开成为单个式子，因此不可能存在重复。

对于抽象的 a，b，c，k，计算能在 k 个以下表达的情况会非常困难，因为需要考虑最简情况。此处按照书 P5，a=3，b=2，c=2 的情况进行估算。

此时，相当于有一个$3 \times 2 \times 2$的长方体，每次可以取全体(三个通配符)、一面(两个通配符)、一条边(一个通配符)，一个小正方体(无通配符)，问取 k 次时能取出多少种。

k 不超过 0 时，只能表示空集，1 种。

k 不超过 1 时，可以额外表示某个面、某条边、某个小正方体、整体，共 7+16+12+1+1=37 种。

k 超过 2 时，由于可能性过于复杂，组合已经变得无法计算(对小正方体的2^{12}种可能分布作程序验证的复杂程度也极高)，不过仍然可以证明出最大值的存在：

k 的最大值，可以证明为 6：3 个2×2的面中，无论每个面如何分布，一定可以通过两个析取的合取表示出来(小于等于两个时取小正方体，三个时取一条边与小正方体，四个时取整体)，而当需要取出的小正方体为(1,1,1) (1,2,2) (2,1,2) (2,2,1) (3,1,1) (3,2,2)时，必须取 6 次才能取出。因此，最多六个合取可以表示任何假设的分布情况。

很明显，当考虑通配符时，情况很快变得过于复杂以至无法计算。对一般的 a，b，c，哪怕是计算 k 的上限也是极为困难的， 不过容易发现，这个上限至少为$\frac{abc}{2}$，通过类似国际象棋棋盘黑白格的分布即可取到(事实上，改进分布会更加改进上限，基本可以确定上限可以表示为$abc - t_1 ab - t_2 b - t_3 ac$一类的形式)。

对于估算来说，抽象的 a，b，c 下，会发现通配符对上限的改进并不算明显，因此可以直接在通配符不存在的情况下作估算，此时至多 k 个可以表示的情况总数为$\sum_{i=0}^{k} C_{abc}^{i}$。

在通配符存在时，k 越小带来的新的可能性比起原来的可能性会越高，这也代表着这个总数需要存在某个随着 k 上升而减小的比例修正项。

3.

假设x为$m+n$维，x_1为m维，x_2为n维，由$N(\mu,\Sigma)$的定义积分可知x_1的分布为$N\left(\mu_m,\Sigma_m\right)$，

其中μ_m代表μ的前m个分量，Σ_m代表Σ左上角的$m\times m$子矩阵。

同理x_2的分布为$N\left(\mu_n,\Sigma_n\right)$，其中$\mu_n$代表μ的后$n$个分量，$\Sigma_n$代表Σ右下角的$n\times n$子矩阵。

由条件密度定义，$x_1|x_2$的分布为整体分布除以x_2的分布，即

$$\frac{\sqrt{(2\pi)^n\left|\Sigma_{\mathrm{n}}\right|}}{\sqrt{(2\pi)^{m+n}|\Sigma|}}\exp\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T+\frac{1}{2}(x_2-\mu_n)\Sigma_{\mathrm{n}}{}^{-1}(x_2-\mu_n)^T\right)$$

由于仍满足 e 指数上二次，且其为分布函数，这仍然是一个正态分布。

4.

范数满足$||\mathrm{x}||_P=\sqrt[p]{|x_1|^p+|x_2|^p+\cdots+|x_n|^p}$，由定义只需证明$0<t<1$时

$$||\mathrm{tx}+(1-\mathrm{t})\mathrm{y}||_{\mathrm{p}}\leq \mathrm{t}||\mathrm{x}||_{\mathrm{p}}+(1-t)||y||_p$$

由定义式与t的范围可知右边等于$||\mathrm{tx}||_p+||(1-t)\mathrm{y}||_p$，因此只需说明对任何向量 a，b，有$||a||_p+||b||_p\geq||a+b||_p$，而此即为闵可夫斯基不等式的向量形式。

以下设q满足$\frac{1}{p}+\frac{1}{q}=1$。

对这个不等式，先证明 Young 不等式：对非负的 a，b，有$ab\leq\frac{a^p}{p}+\frac{b^q}{q}$。

固定 a，将式子移到右侧后对 b 求导可知最小值在$b^q=a^p$时取到，从而得证。

接着证明赫尔德不等式：$||a||_p||b||_q\geq||ab||_1$

当 a，b 范数均不为 0 时，由于两边同除以$||a||_p||b||_q$不影响，可不妨设$||a||_p=||b||_q=1$，再利用$|a_ib_i|\leq\frac{|a_i|^p}{p}+\frac{|b_i|^q}{q}$，两边求和即可知$||ab||_1<1$，从而得证。

最后说明原结论：由绝对值三角不等式直接展开可知

$$||a+b||_p^p\leq\sum_{i=1}^n|a_i||a_i+b_i|^{p-1}+\sum_{i=1}^n|b_i||a_i+b_i|^{p-1}$$

利用赫尔德不等式与$p=qp-q$进一步得到

$$\sum_{i=1}^n|a_i||a_i+b_i|^{p-1}+\sum_{i=1}^n|b_i||a_i+b_i|^{p-1}\leq||a||_p||a+b||_p^{p/q}+||b||_p||a+b||_p^{p/q}$$

由此有$||a+b||_p^p\leq||a||_p||a+b||_p^{p/q}+||b||_p||a+b||_p^{p/q}$，两边同除以$||a+b||_p^{1/q}$由 p，q 关系知结论。

5.

1 阶推 0 阶：

若某个x_0,y_0,t_0使零阶条件不成立，记$p=t_0x_0+(1-t_0)y_0$，下证不存在列向量α使$f(y_0)\geq f(p)+\alpha(y_0-p)$、$f(x_0)\geq f(p)+\alpha(x_0-p)$同时成立。只需要考虑$x_0,y_0,p$所在

的直线上的分量，也即记$g(t)=f(tx_0+(1-t)y_0)$，只需说明$g(1)\geq g(t_0)+a(1-t_0)$与$g(0)\geq g(t_0)-at_0$不会同时成立。

从第一个式子可知$a\leq\frac{g(1)-g(t_0)}{1-t_0}$，第二个式子可知$a\geq\frac{g(t_0)-g(0)}{t_0}$，但是零阶条件不成立可化为$\frac{g(1)-g(t_0)}{1-t_0}<\frac{g(t_0)-g(0)}{t_0}$，从而不等式无解，由此得证。

0 阶推 1 阶:

零阶条件可化为$f(y)\geq f(x)+\frac{f(x+t(y-x))-f(x)}{t}$，而右侧令$t\to 0$即为$\nabla f(x)^T(y-x)$，从而得证。

第二章：

1.

交叉验证：由于分层划分要求，给出的将为正负例个数相同的训练集，因此模型将随机预测，正确率 50%。

留一法：由于训练集中个数较多的与留出的样本不同，正确率 0%。

2.

$$TPR=\frac{TP}{TP+FN},FPR=\frac{FP}{FP+TN},P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN}$$

从数值上来说，真正例率与查全率是相同的，都代表将正例中的多少找了出来，但与它们对比的东西不同：假正例率是将反例识别为正例的概率，而查准率是将识别的正例中真实的概率，前者将真实正反分开，方便考虑代价，而后者将准确与全面两个属性刻画出，方便计算性能。

3.

教材上给的近似曲线绘制方法只有在每个样本概率值不同时才能成立，而此时l_{rank}中相等项根本不成立。为了考虑到一般的情况，我们考虑在某步同时将 m 个正例 n 个反例视为正例(也即对这些计算出的概率值相同)时的情况。

由于此曲线单调递增，所谓的“曲线之上的面积”其实用“曲线左侧到坐标轴”的面积看待更为准确。

若 m=0，此时只考虑了反例，不产生面积，否则，利用直角梯形面积公式可知曲线左侧的面积会是

$$\frac{1}{2}(FPR_a+FPR_b)(TPR_a-TPR_b)$$

其中 a 代表 after，即这步之后的情况，b 表示 before。接下来说明，这段的面积正好是添入的 m 个正例所产生的“罚分”。

$$FPR_b=\frac{FP_b}{m^-},FPR_a=\frac{FP_b+n}{m^-},TPR_b=\frac{TP_b}{m^+},TPR_a=\frac{TP_b+m}{m^+}$$

代入计算可知面积为$\frac{1}{m^-m^+}\left(FP_bm+\frac{1}{2}mn\right)$，而$FP_b$恰好是这步之前(也就是在这 m 个正例之前)的反例，n 恰好是同时的反例，从而可知成立。

4.

在提供两学习器分类差别联列表时，若两学习器接近独立，则可以认为不存在显著差别，否则认为存在显著差别。

为了寻找卡方检验方式，我们提出原假设与预备假设。

原假设H_0：两学习器无显著差别。

备择假设H_1：两学习器有显著差别。

H_0成立的情况下：对于随机抽取的 n 个样本，根据数据，有$\frac{e_{00}+e_{10}}{n}$个被 A 认为正确的样本，有$\frac{e_{00}+e_{01}}{n}$个被 B 认为正确的样本。若存在显著差别，则这两件事相互独立，于是被两算法都认为正确的样本的理论频数w_{00}应为$n\frac{e_{00}+e_{10}}{n}\frac{e_{00}+e_{01}}{n}=\frac{(e_{00}+e_{10})(e_{00}+e_{01})}{n}$。类似可知，$w_{01}=\frac{(e_{01}+e_{00})(e_{01}+e_{11})}{n}, w_{10}=\frac{(e_{10}+e_{00})(e_{10}+e_{11})}{n}, w_{11}=\frac{(e_{11}+e_{01})(e_{11}+e_{10})}{n}$。

利用正态分布可知，$\sum_{i,j}\frac{(w_{ij}-e_{ij})^2}{w_{ij}}$在两学习器独立时应具有卡方分布，由此可以计算此值小于某个数的概率，将 n 替换为条件中的总数$e_{00}+e_{01}+e_{10}+e_{11}$可将公式化简为$\frac{(e_{00}e_{11}-e_{01}e_{10})^2(e_{00}+e_{01}+e_{10}+e_{11})}{(e_{00}+e_{10})(e_{00}+e_{01})(e_{11}+e_{10})(e_{11}+e_{01})}$。

由于此统计量呈卡方分布，可以根据统计量的结果与积分理论结果α比对，若在范围内，则无法拒绝假设，认为无显著差别，否则能认为存在显著差别。

(对更一般的检验，会有不同的呈卡方分布的统计量，对结果进行检验即可。)

Homework 2

PB20000296 郑滕飞

第三章：

1.

3.18 非凸：

只需举出一个例子。考虑$\omega=(\omega_1,0,\dots,0)$，这时式子化为$y=\frac{1}{1+e^{-\omega_1 x_1+b}}$，其对$\omega_1$的二阶导为$-\frac{e^{b+\omega_1 x_1}(-e^b+e^{\omega_1 x_1})x_1^2}{(e^b+e^{\omega_1 x_1})^3}$，只要$x_1$不为 0，可取到$\omega_1$使得$-e^b+e^{\omega_1 x_1}>0$，从而二阶导小于 0，若原函数为凸函数，其 Hesse 阵应正定，利用正定阵性质可知对任何分量的二阶导大于 0，矛盾。

(由此，只要x不为零向量，3.18 就不可能为凸函数)

3.27 为凸：

由于正定阵的前若干行若干列构成的矩阵正定，只需要说明 3.27 对β是凸的，不考虑 b 这个分量时其对ω必然也是凸的。

利用书中的式 3.31 可知 3.27 的二阶导数可以写为XDX^T，其中 X 的第 i 列为x_i，而 D 是对角元素为$p_1(\hat{x_i};\beta)\big(1-p_1(\hat{x_i};\beta)\big)$的对角阵。由定义，D 的对角元素全部为正，对任何向量α，可发现$\alpha^T\nabla^2 l(\beta)\alpha=\left|\sqrt{D}(X^T\alpha)\right|^2$，因此$\nabla^2 l(\beta)$为半正定阵，$l(\beta)$为凸函数。

2.

	f1	f2	f3	f4	f5	f6	f7	f8	f9
C1	+1	+1	+1	+1	+1	+1	+1	+1	+1
C2	+1	+1	+1	-1	-1	-1	-1	-1	-1
C3	-1	-1	-1	+1	+1	+1	-1	-1	-1
C4	-1	-1	-1	-1	-1	-1	+1	+1	+1

根据书上的原理，需要使 4 个类之间两两产生的 6 个海明距离的最小值最大(注：此处采取最小值是因为，若用求和来刻画，会导致九个分类器全部是 C1C2/C3C4 时仍能取到最大值，但这是不符合分开任意两类的要求的)。对 2v2 分类器，产生的海明距离的和共为 4，而 1v3 分类器的和为 3，计算可知最小值最多为$4\times 9\div 6=6$。而上方的构成方法中达到了最小值为 6，因此为最优。

3.

先证明：AA^T的秩与A相同。

考虑奇异值分解$A=P\Sigma Q$，则$AA^T=P\Sigma\Sigma^T P^T$，由于 P、Q 为正交阵，左右乘并不改变秩，而对于对角元为正的对角阵Σ，其与其转置乘积相当于对每个分量作平方，不会改变非零分量的个数，因此秩不变。

再证明：$A+B$的秩不超过A,B的秩之和。

将秩看作是线性无关的列向量的个数，由于$A+B$的列向量一定可以被 A 的列向量与 B 的列向量线性表出，而 A 的列向量一定可以被线性无关的列向量表出，结论成立。
由于μ可用所有μ_i线性表出，于是μ_N可被其他μ_i与μ表出，将μ_N拆分计算发现

$$\sum_{i=1}^{N} m_i(\mu_i-\mu)(\mu_i-\mu)^T=\sum_{i=1}^{N-1} m_i'(\mu_i-\mu)(\mu_i-\mu)^T$$

也即事实上只存在 N-1 个可能线性无关的分量。利用前面证明的两个结论，S_b的秩不超过 N-1。

4.

可将其看作$\min_W -tr(W^TS_bW)$使得$tr(W^TS_wW)=1$。

在A对称时，由于$tr(W^TAW)=\sum_k\sum_{s,t} w_{sk}\,a_{st}w_{tk}$，计算其对$w_{ij}$求导：由于第二个下标是$j$，左侧的求和只有$k=j$一项有意义，只需计算$\sum_{s,t} w_{sj}\,a_{st}w_{tj}$对$w_{ij}$求导，而这可以分为$s=i,t=i,s=t=i$讨论得到结果为$2\sum_{k=1}^{d} a_{ik}\,w_{kj}$。

于是，$\frac{\partial tr(W^TAW)}{\partial W}=2AW$，由拉格朗日乘子法，令乘子为λ即知需要求解$2S_bW=2\lambda S_wW$，两边同除以 2 即得原式。

5.
投影矩阵定义为$A^2=A$，直接计算

$$(X(X^TX)^{-1}X^T)^2=X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T=X(X^TX)^{-1}X^T$$

从而得证。
将行向量x_i估算得到的y排成一列，可发现结果恰好为$X(X^TX)^{-1}X^Ty$，也即满秩情况下，事实上将y到线性空间$\omega^T\mathrm{X}+\mathrm{b}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix},\omega\in\mathbb{R}^n,\mathrm{b}\in\mathbb{R}$作了投影，这样得到的估计即能使得距离的损失最小。

第四章：

1.
对属性数量归纳证明。
在只有一个属性时，由于无矛盾，直接区分即可构造出一层的决策树，下面假设 n-1 个属性时结论成立。
假设训练集有 n 个属性，m 个样例。假设所有样例对第 1 个属性总共有 a 种不同取值(若属性为连续，a 个值可以被分到 a 个区间种)，有$a\le m$。
构造决策树，对第一个属性产生 a 个分支，将所有样例按取值分成 a 类。若某类中存在矛盾，又由于它们第一个属性相同，可知原特征向量存在矛盾，不符合条件。因此，对每个分支中的数据，忽略已经分类的第一个属性，都可以应用归纳假设得到一棵子树。由此即构造出了 n 个属性时的生成树。

2.

利用与式 4.12 相同的符号，记

$$Gini\_index(D,a) = \rho \sum_{v=1}^{V} \tilde{r}_v Gini(\tilde{D}^v)$$

其中$\text{Gini}(\tilde{D}) = 1 - \sum_{k=1}^{|y|} \tilde{p}_k^2$。

3.

由于约束要求所有p_k和为 1，即要对$\sum_k(\lambda p_k - p_k \log_2 p_k) - \lambda$计算极值。

其对p_i的偏导是$\lambda - \log_2 p_i - \frac{1}{p_i \ln 2}$，计算发现此式对$p_i$单调减，因此对于每个$p_i$的零点必须相同，必须取到均匀分布。将均匀分布与$p_1 = 1$，其他为 0 的分布比较可知此时熵确实是最大值(而不是最小值)。

4.

(a) 包含正负两个类，各占 0.5，熵为$-0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$。

(b) A 划分后，T 中 3+1-，F 中 2+4-，熵分别为 0.811，0.918，增益为 0.125。
B 划分后，T 中 2+3-，F 中 3+2-，熵分别为 0.971，0.971，增益为 0.029。

(c) <=1.5 中 1+0-，>1.5 中 4+5-，熵分别为 0，0.991，增益为 0.108。
<=2.5 中 2+0-，>2.5 中 3+5-，熵分别为 0，0.954，增益为 0.237。
<=3.5 中 2+1-，>3.5 中 3+4-，熵分别为 0.918，0.985，增益为 0.035.
<=4.5 中 3+1-，>4.5 中 2+4-，熵分别为 0.811，0.918，增益为 0.125。
<=5.5 中 3+3-，>5.5 中 2+2-，熵分别为 1，1，增益为 0。
<=6.5 中 4+3-，>6.5 中 1+2-，熵分别为 0.985，0.918，增益为 0.035。
<=7.5 中 5+4-，>7.5 中 0+1-，熵分别为 0.991，0，增益为 0.108。

(d) A 划分后 TF 的纯度分别为 0.375，0.444，基尼指数 0.417。
B 划分后 TF 的纯度分别为 0.48，0.48，基尼指数 0.48。
以 A 为凭据划分。

(e)

```
root
----C <= 2.5 正
----C > 2.5
--------A = T
------------C <= 5.5 反
------------C > 5.5 正
--------A = F
------------C <= 4.5
----------------B = T 反
----------------B = F 正
------------C > 4.5 反
```

Homework 3

PB20000296 郑滕飞

第五章：

1.

第一，用线性函数作为激活函数，事实上计算可发现线性复合后只是改变了权重 w 与阈值 theta 中，并没有增添新的信息。

第二，理想的神经元需要区分是否激活，从而为阶跃函数。模仿阶跃函数构造的激活函数必须在 0 点处有从 0 到 1 的大幅改变，而线性函数并不满足。

第三，线性函数无法将大范围的输入控制到小范围的输出当中，从而误差可能累计扩大，不利于学习。

2.

如果直接计算，在$\sum_{j=1}^{C}\exp(x_j)$的部分都可能出现上溢。为避免溢出，记$m=\max_{i=1}^{c}x_i$，

$x_i'=x_i-m$，则只需计算$\frac{\exp(x_i')}{\sum_{j=1}^{C}\exp(x_j')}$与$\log\sum_{j=1}^{C}\exp(x_j')+m$即可。由于最大数对应的 exp 此时为 1，而其他为小于 1 的数，避免了溢出(过小而被计算为 0 不会影响结果正确性，只是精度问题)。

3.

以下记$M=\sum_{j=1}^{C}\exp(x_j)$。

左：对x_i求导为$\frac{M\,exp(x_i)-exp(2x_i)}{M^2}$，对其他$x_k$求导为$-\frac{\exp(x_i+x_k)}{M^2}$。

右：对x_i求导为$1-\frac{\exp(x_i)}{M}$，对其他x_k求导为$-\frac{\exp(x_k)}{M}$。

4.

记 A、B 到 1、2 的权重a_1,a_2,b_1,b_2，1、2 到 3 的权重x_1,x_2，有

$$Out=\max(x_1\max(a_1A+b_1B,0)+x_2\max(a_2A+b_2B,0),0)$$

代入可知初始估计值为 0.274，误差约 0.056。

由于当前参数均为正，对应位置 ReLU 函数的导数均为 1。于是：

$$\frac{\partial Out}{\partial a_1}=\mathrm{A}x_1,\frac{\partial Out}{\partial b_1}=\mathrm{B}x_1,\frac{\partial Out}{\partial a_2}=\mathrm{A}x_2,\frac{\partial Out}{\partial b_2}=\mathrm{B}x_2$$

$$\frac{\partial Out}{\partial x_1}=a_1\mathrm{A}+b_1\mathrm{B},\frac{\partial Out}{\partial x_2}=a_2\mathrm{A}+b_2\mathrm{B}$$

而$\frac{\partial E}{\partial x}=\frac{\partial E}{\partial Out}\frac{\partial Out}{\partial x}=(Out-y)\frac{\partial Out}{\partial x}$，这里$Out-y=-0.226$。

于是以误差偏导的相反数更新，即

$$\Delta a_1=0.02260,\Delta a_2=0.03616,\Delta b_1=0.03390,\Delta b_2=0.05424$$

$$\Delta x_1=0.04068,\Delta x_2=0.05198$$

$a_1 = 0.6226, a_2 = 0.13616, b_1 = 0.2339, b_2 = 0.75424, x_1 = 0.54068, x_2 = 0.85198$
更新后估计值约为 0.321，误差约为 0.016，有所减小。

第六章：

1.
在样本线性不可分时，LDM 最终的判别一定为线性，但核支持 SVM 求出的特征空间中的划分退回原空间后不可能为线性(否则样本线性可分)，于是两者基本无关；然而，哪怕在线性可分时，依然无法确定等价，这是由于 SVM 只考虑了支持向量。
一个例子是，当在平面上(1,0)为正例，(1,2)、(2,999)、(3,1000)为负例，这时的 SVM 得到结果是 y=1，只与(1,0)、(1,2)有关，但由于 LDA 判定方式是类中心的距离与每类的协方差之和，计算发现得到的方向是$y = -\frac{14}{97}x$，和 SVM 得到的结果并无太大关系。
从这个例子中可以看出，哪怕是十分简单的情况，LDA 和 SVM 也会得到不同的结果，因此基本不可能从样本分布中得到较强的等价结论，于是，只好转而研究二者得到的结果。"等价"的意义是二者的决策边界重合。SVM 得到的结果即为决策边界，而 LDA 得到的决策边界与投影到的直线垂直，且经过两类分别均值的中点投影。当 SVM 得到的平面与 LDA 得到的决策边界垂直且经过两类中心的中点时，两者等价。
*此处有一个较为简单的例子：当所有样本都是支持向量且两类样本分别的均值连线与 SVM 得到的平面垂直时，两者等价。这是由于所有样本都是支持向量时，两类样本的平均值也都与平面等距，从而其中点在平面上。此外，在垂直方向的投影使得同类样本距离全部为 0，两类样本均值的连线线段长度即为投影后的距离，于是垂直方向必然是最佳方向。

2.
先考虑硬间隔的情况。由于 SVM 希望找出与支持向量有最大距离的超平面，近处的噪声可能直接导致支持向量判断错误，甚至一个噪点就会破坏原空间或特征空间中线性可分性，对噪声的容许程度极低。
对软间隔 SVM，对噪声的容许程度有一定提升，但由于本质原理不变，少数的靠近分界面的噪点依然会对结果造成大幅度的影响。

3.
由于截距仍可通过增广 X 出一列 1 来构造，此处不考虑截距。这时，考虑映射到特征空间的函数ϕ，可发现损失函数变为

$$\mathrm{l}(\omega) = \sum_i -y_i \omega^T \phi(x_i) + \ln\left(1 + \exp\left(\omega^T \phi(x_i)\right)\right)$$

若直接求解，只能确定$\left(\omega^T \phi(x_i)\right)^2 = \kappa(x_i, x_i)\omega^T\omega$，而无法确定值的正负。

由此，利用引理：最优解ω一定能被所有$\phi(x_i)$线性表出。
引理证明：由于由于回归是寻找最优的$||y - X\omega||_2$，可将ω分解为 X 生成空间中的分量ω_1与其正交补中的分量ω_2。计算可发现$||y - X\omega||_2^2 = ||y - X\omega_1||_2^2$，于是$\omega_1$亦为最优解。
*加入 L2 正则项时上述结论仍成立，利用勾股定理知ω_1模长不超过ω。
于是，设$\omega = \sum_i \beta_i \phi(x_i)$，损失函数即为

$$l(\beta)=\sum_i\left(-y_i\sum_j\beta_j\,\kappa(x_j,x_i)+\ln\left(1+\exp\left(\sum_j\beta_j\,\kappa(x_j,x_i)\right)\right)\right)$$

再对β最小化即可。

4.

如下构造：

$$\alpha'=(\alpha,\hat{\alpha}),u_i=\begin{cases}1 & i\le m\\ -1 & i>m\end{cases},v_i=\begin{cases}-y_i-\epsilon & i\le m\\ y_{i-m}-\epsilon & i>m\end{cases},K_{ij}=\begin{cases}\kappa(x_i,x_j) & i\le m,j\le m\\ -\kappa(x_{i-m},x_j) & i>m,j\le m\\ -\kappa(x_i,x_{j-m}) & i\le m,j>m\\ \kappa(x_{i-m},x_{j-m}) & i>m,j>m\end{cases}$$

则问题即化为$\max_{\alpha'} g(\alpha')=\alpha'^T v-\frac{1}{2}\alpha'^T K\alpha'$，使得$C\geqslant\alpha'\geqslant 0$且$\alpha'^T u=0$。

5.

假设x为n维，由于$\left(x_i^T x_j\right)^2=\left(\sum_k x_{ik}\,x_{jk}\right)^2=\sum_{1\le k\le n,1\le l\le n}x_{ik}\,x_{jk}x_{il}x_{jl}$，记$M_{k,l}=x_{ik}x_{il}$，其中$M$为$n\times n$的矩阵，将$M$按照行先列后展开成向量即为$\phi(x_i)$。

Homework 4

PB20000296 郑滕飞

第八章：

1.

$$\frac{\partial \ell(H|D)}{\partial H} = -\ell'\big(-H(x)\big)P(f(x)=1|x) + \ell'\big(H(x)\big)P(f(x)=-1|x)$$

下记$P_1 = P(f(x)=1|x), P_{-1} = P(f(x)=-1|x)$。由于$\ell\big(-f(x)H(x)\big)$对$fH$在给定区间单调减，$\ell\big(H(x)\big)$必然在$[-\delta,\infty]$单调增，于是对$H$导数大于 0。

当$H(x)=0$时，此式变为$\ell'(0)(P_{-1}-P_1)$，符号与$P_{-1}-P_1$相同。当$P_1 > P_{-1}$，由此式为负，即这点处损失函数局部单调减，其右侧必有极小值，即$H(x)>0$，$sign(H(x))=1$。反之，当$P_{-1} < P_1$时，此式为正，这点处损失函数局部单增，极小值在左侧，即$H(x)<0$，$sign\big(H(x)\big)=-1$，从而得证。

[由于损失函数一般具有凸性，极小值点可以作为最小值点；这里假设损失函数可以被极小化，不考虑不收敛的情况。]

2.

Multi Boosting：优点为参与投票的是训练结果已较强的 boosting 结果，单个的准确率都得到提升，但缺点是 boosting 过程中越往后的学习器具有的独立性越差。
Iterative Bagging：优点为通过后续迭代进一步增强了 bagging 投票的结果，缺点为 bagging 的投票结果已经具有了某种意义上“均匀”的分布，迭代对结果的改进可能并不明显。

第九章：

1.

由于度量建立在长度相同的向量上，可不妨设为单位球上，所有向量模长为 1。

对余弦距离，不妨考虑三维，令$x=(1,0,0), y=\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2},0\right), z=(0,1,0)$，则

$$1-x^Ty+1-y^Tz = 2-\sqrt{2} < 1 = 1-x^Tz$$

于是余弦距离不为度量。
为了证明余弦夹角为度量，只需要在两边同取 cos 时说明成立，即证明

$$\cos\arccos x^Tz \geq \cos(\arccos x^Ty + \arccos y^Tz)$$

其等价于$x^Tz \geq x^Tyy^Tz - \sqrt{(1-(x^Ty)^2)(1-(y^Tz)^2)}$，由于旋转不影响结果可不妨设 y 只有第一个分量为 1，其他为 0，则原式移项后同平方可化简为$(1-x_1^2)(1-z_1^2) \geq (x_2z_2+\cdots+x_nz_n)^2$，注意到 x、z 的模为一，此即柯西不等式，原命题得证。

2.

首先，m 个样本 k 聚类的总可能数m^k，只有有限种结果，因此只要迭代有极限，必然在

有限步之后等于最终值且不再改变。
构造能量函数$E(T,\mu)=||x-T\mu||_F^2$，下面说明其在迭代过程中单调减：
将某个 x 划分到离它最近的类时，由于$||x_i-\mu_{new}||<||x_i-\mu_{old}||$，而其他行不变，于是$E(T_{new},\mu)<E(T_{old},\mu)$。
更新某类均值向量时，求导可知$\sum_i||x_i-\alpha||^2\leq\sum_i||x_i-\bar{x}||^2$，其中$\bar{x}$代表均值，而其他类不变，于是$E(T,\mu_{\mathrm{new}})\leq E(T,\mu_{\mathrm{old}})$。
由于其单调减且聚类可能性有限，总会在某一步后其不再变化。注意到，算法中只要更改某个 x 的所在类，能量一定严格减小，因此不再变化蕴含着聚类结果不变，即算法收敛。

3.
这时的能量函数为$E(T,\mu)=\sum_{i=1}^{m}||x_i-t_i^T\mu||$，此处的范数表示采用的度量，可以根据计算需求取次方(如欧氏距离取了平方)。在T固定时，优化μ即相当于对每个簇的x_i寻找最小值，即$\mu=\arg\min_{\alpha}\sum_{x\in C_c}||x-\alpha||$，假设距离函数具有一定光滑性，可以令其对α偏导数为 0 进行求解。
(例如，当采用曼哈顿距离时，得到的μ的每个分量为此类中的x对应分量的中位数。)

Homework 5

PB20000296 郑滕飞

第十章：

1.

样本无限时$P(c|z)=P(c|x)$，设$|\mathcal{Y}|=n$，$P(c|x)$分别为$p_1,\dots,p_n$，不妨设$p_1\geq p_2\dots\geq p_n$：

第一个不等号即$\sum_i p_i^2\leq p_1$，由于$2(a^2+b^2)=(a+b)^2+(a-b)^2$，当和一定时相差越大平方和越大，左侧在尽量多分量取到p_1时取最大值，于是当$p_1\in\left(\frac{1}{k},\frac{1}{k-1}\right]$时，左减右$\leq(k-1)p_1^2+(1-(k-1)p_1)^2-p_1=(kp_1-1)\big((k-1)p_1-1\big)\leq 0$，从而得证。

第二个不等号计算化简得$(n-1)\sum_{i>1}p_i^2\geq 1-2p_1+p_1^2$，利用柯西不等式可知固定$p_1$时左侧$\geq(1-p_1)^2=1-2p_1+p_1^2$，从而成立。

2.

利用奇异值分解计算可知XX^T的特征值是X对应奇异值的平方，而通过 QR 方法等办法直接计算奇异值分解可以规避大矩阵的乘法，加快效率。

3.

*δ_i^j在$i=j$时为 1，否则为 0，记w_i为W的第i个列向量。

记XX^T的第 i 行第 j 列为z_{ij}，则直接计算可知$\mathrm{tr}(W^TXX^TW)=\mathrm{tr}(XX^TWW^T)=\sum_{i=1}^{d}\sum_{j=1}^{d}z_{ji}\sum_{l=1}^{d'}w_{il}\,w_{jl}$，从而利用$z_{ij}=z_{ji}$计算得其对$w_{st}$求偏导的结果为$2\sum_{i=1}^{d}z_{si}\,w_{it}$，即$\frac{\partial tr(W^TXX^TW)}{\partial W}=2XX^TW$。而右侧的等式约束事实上是$d'^2$个约束$w_i^Tw_j=\delta_i^j$，假设对每个的拉格朗日乘子为$\lambda_{ij}$（注意到由内积对称性$\lambda_{ij}=\lambda_{ji}$），由于$w_i^Tw_j$对$w_{st}$求导后为$\delta_i^sw_{tj}+\delta_j^sw_{ti}$，乘子部分对$w_{st}$导数为$\sum_{i=1}^{d'}\sum_{j=1}^{d'}\lambda_{ij}\big(\delta_i^sw_{tj}+\delta_j^sw_{ti}\big)=2\sum_{i=1}^{d'}\lambda_{is}\,w_{ti}$，于是乘子部分对$W$的导数为$2W\Lambda$，其中$\Lambda=(\lambda_{ij})$。利用乘子法极值条件可知，$W$在符合约束且存在对称矩阵$\Lambda$使得$XX^TW=W\Lambda$时取到极值。

由于实对称方阵可以正交相似对角化，若存在这样的Λ，设其对角化为QDQ^T，计算可得$XX^TWQ=WQD$，由于相似不影响 tr 的值，取$W'=WQ$计算得仍然取到极值且符合约束条件，因此，可直接寻找使Λ为对角阵的解W，而这意味着W的每一列都是XX^T的特征值。

代入在原式中代入$XX^Tw_i=\lambda_iw_i$，计算可知$tr(W^TXX^TW)=\sum_{i=1}^{d'}\lambda_i\,||w_i||_2^2$，利用约束条

件可知$||w_i||_2^2 = 1$，于是 tr 变为$\sum_{i=1}^{d'} \lambda_i$，由于实对称方阵特征值是非负实数，其不超过最大的$d'$个特征值(重复特征值算多个)，下面证明其可以取到。
利用实对称方阵可以正交相似对角化，假设对角化为$Q^T X X^T Q = D_w$，通过置换阵相似可不妨设D_w的特征值按从大到小排列。取Q的前d'列为W，直接计算可知$WW^T = I$，且此时即取到最大值，故得证(可发现此时W每列恰好是对应特征值的一个特征向量)。

4.
不是。只要举出例子即可，当右侧只有一个$x_i - x_j$时，条件可化为$||Px||_2 \geq 1$，其中x即为对应的差值。不妨令二维情况$x = (1,0)$，则P需要满足$P_{11}^2 + P_{12}^2 \geq 1$，即四维空间中的某种"圆柱面外部"，当$P$取$\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, \begin{pmatrix}-1 & 0\\ 0 & 0\end{pmatrix}$时均符合，但$\begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix}$不符合，从而限定区域不凸，不是凸优化。
(对于更一般的情况，分析容易发现大部分情况下由于区域非凸，不构成凸优化。)

第十一章：

1.
当最小情况下平方误差项等值线恰好与 L1 范数等值面相切时，产生的解不稀疏。
计算可以发现，由于平方误差项等值线是凸的，在形状不同时相切的概率在特定维度下不会超过给定值(如二维不会超过 1/2)，但 L2 范数正则化以 1 的概率产生不稀疏的解，因此 L1 更容易产生稀疏解。

2.
零范数光滑性过差(不连续)，不存在有效的梯度，于是所有利用梯度递降等的数值求解思路都无法使用，剩下的求解方法开销较大。
(此外零范数非凸也会导致求解更加困难。)

3.
[以下考虑损失函数时都在 X 中增加一列以将 b 并入 w]

线性回归损失函数$(y - Xw)^T(y - Xw)$，梯度为$2X^T(Xw - y)$，计算发现
$$||2X^T(Xw_1 - y) - 2X^T(Xw_2 - y)|| = ||2X^T X(w_1 - w_2)|| \leq ||2X^T X||||w_1 - w_2||$$
其中矩阵外的范数符号代表矩阵二范数，由此可知其满足 L-Lipschitz 条件，对应的 L 为$||2X^T X||$。

对率回归损失函数$\sum_i -y_i w^T x_i + \ln(1 + \exp(w^T x_i))$，梯度为$\sum_i \left(-y_i + \frac{1}{1+\exp(-z_i)}\right) x_i$，其中$z_i = w^T x_i$。首先，乘正交阵可不妨设$x_i$只有第一个分量非零，从而估计知
$$||(w_1^T - w_2^T)x_i|| \leq ||w_1 - w_2||||x_i||$$
另一方面，$\frac{1}{1+\exp(-z)}$对 z 求导得$-\frac{e^z}{(1+e^z)^2}$，进一步求导知其不超过 1/4，因此$\left|\frac{1}{1+exp(-z_1)} - \frac{1}{1+exp(-z_2)}\right| \leq \frac{1}{4}|z_1 - z_2|$。综合以上可知

$$||\Delta f(w_1) - \Delta f(w_2)|| \leq \sum_i \left| \frac{1}{1+exp(-z_{1i})} - \frac{1}{1+exp(-z_{2i})} \right| ||x_i||$$

$$\leq \frac{1}{4} \sum_i |z_{1i} - z_{2i}| ||x_i|| \leq ||w_1 - w_2|| \frac{1}{4} \sum_i ||x_i||^2 = \frac{1}{4} ||X||_F^2 ||w_1 - w_2||$$

由此可知其满足 L-Lipschitz 条件，对应的 L 可以取$\frac{1}{4}||X||_F^2$。

Homework 6

PB20000296 郑滕飞

第七章：

1.

由于只需要比较大小，又由 ln 单调性，比较

$$\ln P\left(c\right)\prod_{i=1}^{d}P(x_i|c)=\ln P\left(c\right)+\sum_{i=1}^{d}\ln P\left(x_i|c\right)$$

即可。

2.

由两样本同分布，LDA 中$\Sigma_0=\Sigma_1=\Sigma$，于是$\omega=\Sigma^{-1}\frac{\mu_0-\mu_1}{2}$，分类边界也即考虑

$$min\{\left|x^T\omega-\mu_0^T\omega\right|,\left|x^T\omega-\mu_1^T\omega\right|\}$$

于是分类边界为$x^T\omega=\frac{\mu_0+\mu_1}{2}^T\Sigma^{-1}\frac{\mu_0-\mu_1}{2}$，即$2x\Sigma^{-1}(\mu_0-\mu_1)=(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)$，直接展开得到

$$2x\Sigma^{-1}\mu_0-\mu_0^T\Sigma\mu_0=2x\Sigma^{-1}\mu_1-\mu_1^T\Sigma\mu_1$$

同减$x^T\Sigma^{-1}x$即得到

$$(x-\mu_0)^T\Sigma(x-\mu_0)=(x-\mu_1)^T\Sigma(x-\mu_1)$$

贝叶斯分类器中，分类边界也即$P(+)P(x|+)=P(-)P(x|-)$，计算即

$$P(+)c\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)=P(-)c\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

其中由于协方差矩阵相同，c 为同样的系数。
由于同先验，消去得到

$$(x-\mu_0)^T\Sigma(x-\mu_0)=(x-\mu_1)^T\Sigma(x-\mu_1)$$

与 LDA 相同。此外，显然μ_0和μ_1分别分到+与-中，因此对边界两侧的判定相同，从而可知结果一致。

3.

只需说明需要最大化的边际似然$ln\,P\left(X\middle|\Theta\right)$在算法中单调上升，又由有界性即知收敛，由于此函数与估计 Z 的过程无关，只需要证明 M 步更新估计时上升即可。

根据贝叶斯公式，有$\ln P\left(X\middle|\Theta^{\mathrm{t}}\right)=\ln P\left(X,Z\middle|\Theta^{\mathrm{t}}\right)-\ln P\left(Z\middle|X,\Theta^{\mathrm{t}}\right)$，对$\mathrm{P}\left(Z|\mathrm{X},\Theta^t\right)$求和可得(左侧与 Z 无关因此直接提取消去)

$$\ln P\left(X\middle|\Theta^t\right)=\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(X,Z\middle|\Theta^{\mathrm{t}}\right)-\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(Z\middle|X,\Theta^t\right)$$

同理

$$\ln P\left(X\middle|\Theta^{t+1}\right)=\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(X,Z\middle|\Theta^{t+1}\right)-\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(Z\middle|X,\Theta^{t+1}\right)$$

由于Θ^{t+1}为使得$\sum_z P\left(Z\middle|X,\Theta^{\mathrm{t}}\right)\ln P\left(X,Z\middle|\Theta\right)$最大的$\Theta$，必有

$$\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(X,Z\middle|\Theta^{t+1}\right)\geq\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(X,Z\middle|\Theta^{\mathrm{t}}\right)$$

于是只需证明

$$\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(Z\middle|X,\Theta^{t+1}\right)\leq\sum_z P\left(Z\middle|X,\Theta^t\right)\ln P\left(Z\middle|X,\Theta^t\right)$$

即

$$\sum_z P\left(Z\middle|X,\Theta^t\right)\ln\frac{P\left(Z\middle|X,\Theta^{t+1}\right)}{P\left(Z\middle|X,\Theta^t\right)}\leq 0$$

下面证明对任何概率分布函数f,g有$\mathrm{E_f}\left[\ln\frac{\mathrm{g(x)}}{\mathrm{f(x)}}\right]=\sum_x f(x)\ln\frac{g(x)}{f(x)}\leq 0$。

根据琴生不等式，由 f 是离散分布，ln 为凸函数，有$\sum_x f(x)\ln\frac{g(x)}{f(x)}\leq\ln\left(\sum_x f(x)\frac{g(x)}{f(x)}\right)$，

而右侧即为$\sum_x g(x)=1$，于是取 ln 为 0，得证。

综上即得结论。

4.

由条件概率定义，原式即$\frac{P(x_1,\ldots,x_{n+1})}{P(x_1,\ldots,x_n)}$，由此只需要计算$P(x_1,\ldots,x_n)$。

根据马尔可夫链的定义，$P(x_1,y_1,\ldots,x_n,y_n)=P(y_1)\prod_{t=2}^n P(y_t|y_{t-1})\prod_{t=1}^n P(x_t|y_t)$，于是$P(x_1,\ldots,x_n)=\sum_{y_1,\ldots,y_n}P(y_1)\prod_{t=2}^n P(y_t|y_{t-1})\prod_{t=1}^n P(x_t|y_t)$。

而根据上式化简有

$$P(x_1,\ldots,x_{n+1})=\sum_{y_1,\ldots,y_n}P(y_1)\prod_{t=2}^{n}P(y_t|y_{t-1})\prod_{t=1}^{n}P(x_t|y_t)\sum_{y_{n+1}}P(y_{n+1}|y_n)P(x_{n+1}|y_{n+1})$$

由于$\frac{P(y_1)\prod_{t=2}^n P(y_t|y_{t-1})\prod_{t=1}^n P(x_t|y_t)}{\sum_{y_1,\ldots,y_n}P(y_1)\prod_{t=2}^n P(y_t|y_{t-1})\prod_{t=1}^n P(x_t|y_t)}=P(y_1,\ldots,y_n|x_1,\ldots,x_n)$，最终可写出

$$P(x_{n+1}|x_1,\ldots,x_n)=\sum_{y_{n+1}}P(y_{n+1}|y_n)P(x_{n+1}|y_{n+1})\,P(y_1,\ldots,y_n|x_1,\ldots,x_n)$$

在转移概率 A，观测概率 B，初始状态概率π的情况下，可以写为

$$P(x_{n+1}|x_1,\ldots,x_n)=\sum_{i_{n+1}}a_{i_{n+1},i_n}\,b_{x_{n+1},i_{n+1}}P(y_1,\ldots,y_n|x_1,\ldots,x_n)$$

其中$P(y_1,\ldots,y_n|x_1,\ldots,x_n)=\frac{\pi_{i_1}\prod_{t=2}^n a_{i_t,i_{t-1}}\prod_{t=1}^n b_{x_t,i_t}}{\sum_{i_1,\ldots,i_n}\pi_{i_1}\prod_{t=2}^n a_{i_t,i_{t-1}}\prod_{t=1}^n b_{x_t,i_t}}$，$i_t$代表$y_t$对应的编号。

第十四章：

1.

(1)

$\mathrm{p}(D,\mu,\lambda)=p(D|\mu,\lambda)p(\mu,\lambda)$，直接计算得为

$$\frac{(2\pi)^{-(m+1)/2}b_0^{a_0}\sqrt{\kappa_0}}{\Gamma(a_0)}\lambda^{(m-1)/2+a_0}\exp\left(-\frac{\lambda}{2}\left(\sum_i(x_i-\mu)^2+\kappa_0(\mu-\mu_0)^2\right)-b_0\lambda\right)$$

(2)

这里$X=(x_1,\dots,x_n),Z=(\lambda,\mu),\Theta=(a_0,b_0,\mu_0,\kappa_0)$，记欲确定的参数为$(a,b,u,\kappa)$(μ记号重复)，证据下界为$\int_{\mu,\lambda}Q(\mu,\lambda)\ln\frac{P(\mu,\lambda,D)}{Q(\mu,\lambda)}$，其中

$$Q(\mu,\lambda)=\sqrt{\frac{\kappa}{2\pi}}\frac{b^a}{\Gamma(a)}\lambda^{-1/2+a}\exp\left(-\frac{\lambda}{2}\kappa(\mu-u)^2-b\lambda\right)$$

利用凸函数的积分琴生不等式，由于 Q 非负且积分为 1 可知

$$\int_{\mu,\lambda}Q(\mu,\lambda)\ln\frac{P(\mu,\lambda,D)}{Q(\mu,\lambda)}\le\int_{\mu,\lambda}ln\frac{P(\mu,\lambda,D)Q(\mu,\lambda)}{Q(\mu,\lambda)}=\int_{\mu,\lambda}ln\,P(\mu,\lambda,D)=ln\,P(D)$$

即说明了它是下界。

代入两式可知结果为

$$\int_{\lambda=0}^{\infty}\int_{\mu=-\infty}^{\infty}\sqrt{\frac{\kappa}{2\pi}}\frac{b^a}{\Gamma(a)}\lambda^{-1/2+a}\exp\left(-\frac{\lambda}{2}\kappa(\mu-u)^2-b\lambda\right)\cdot(A+B)$$

其中

$$A=ln\left(\frac{(2\pi)^{-\mathrm{m}/2}b_0^{a_0}\sqrt{\kappa_0}\Gamma(\mathrm{a})}{\mathrm{b}^{\mathrm{a}}\sqrt{\kappa}\Gamma(a_0)}\lambda^{m/2+a_0-a}\right)+(b-b_0)\lambda$$

$$B=-\frac{\lambda}{2}\left(\sum_i(x_i-\mu)^2+\kappa_0(\mu-\mu_0)^2-\kappa(\mu-u)^2\right)$$

先对μ积分，记$G=\frac{b^a\lambda^{a-1}}{\Gamma(a)}exp(-b\lambda)$可得

$$\int_{\lambda=0}^{\infty}G(A+B')$$

其中

$$B'=-\frac{\lambda}{2}\left(\frac{m+\kappa_0-\kappa}{\lambda\kappa}+\sum_i(x_i-u)^2+\kappa_0(\mu_0-u)^2\right)$$

而由于$\int_0^{\infty}\frac{b^a\lambda^{a-1}}{\Gamma(a)}exp(-b\lambda)=1$，$\int_0^{\infty}\frac{b^a\lambda^a}{\Gamma(a)}exp(-b\lambda)=\frac{\mathrm{a}}{\mathrm{b}}$，$\int_0^{\infty}\frac{b^a\lambda^{a-1}}{\Gamma(a)}exp(-b\lambda)\ln\lambda=-ln\,b+f(a)$可算得积分为

$$\ln\left(\frac{(2\pi)^{-\mathrm{m}/2}b_0^{a_0}\sqrt{\kappa_0}\Gamma(\mathrm{a})}{\mathrm{b}^{\mathrm{a}}\sqrt{\kappa}\Gamma(a_0)}\right)+\frac{a(b-b_0)}{b}-\frac{m+\kappa_0-\kappa}{2\kappa}$$

$$-\frac{a}{2b}\left(\sum_i(x_i-u)^2+\kappa_0(\mu_0-u)^2\right)+\left(\frac{m}{2}+a_0-a\right)\left(-ln\,b+f(a)\right)$$

其中$f(a)=\frac{d}{da}\ln\Gamma(a)$。

(3)

证据下界对 a 求导为(消去共同部分$f(a)-\ln b$)

$$\frac{b-b_0}{b}-\frac{1}{2b}\left(\sum_i(x_i-u)^2+\kappa_0(\mu_0-u)^2\right)+f'(a)\left(\frac{m}{2}+a_0-a\right)$$

对 b 求导为

$$-\frac{a}{b}-\frac{ab_0}{b^2}+\frac{a}{2b^2}\left(\sum_i(x_i-u)^2+\kappa_0(\mu_0-u)^2\right)-\frac{m+2a_0-2a}{2b}$$

对 u 求导为

$$-\frac{a}{b}\left(\sum_i(u-x_i)+\kappa_0(u-\mu_0)\right)$$

对κ求导为

$$-\frac{1}{2\kappa}+\frac{m+\kappa_0}{2\kappa^2}$$

于是

$$u=\frac{\kappa_0\mu_0+\sum_i x_i}{m+\kappa_0},\kappa=m+\kappa_0$$

代入 b 中部分消去可得对 a 求导为

$$-\left(\frac{m}{2}+a_0-a\right)+f'(a)\left(\frac{m}{2}+a_0-a\right)$$

其为 0 也即

$$a=a_0+\frac{m}{2}$$

带回原式化简可得 b 需满足

$$-(b-b_0)+\frac{1}{2}\left(\sum_i(x_i-u)^2+\kappa_0(\mu_0-u)^2\right)=0$$

即$b=b_0+\frac{1}{2}(\sum_i(x_i-u)^2+\kappa_0(\mu_0-u)^2)$，代入 u 化简知

$$b=b_0+\frac{1}{2}\left(\sum_i x_i^2+\kappa_0\mu_0^2-\frac{(\kappa_0\mu_0+\sum_i x_i)^2}{m+\kappa_0}\right)$$

最终结果:

$$u=\frac{\kappa_0\mu_0+\sum_i x_i}{m+\kappa_0},\kappa=m+\kappa_0$$

$$a=a_0+\frac{m}{2},b=b_0+\frac{1}{2}\left(\sum_i x_i^2+\kappa_0\mu_0^2-\frac{(\kappa_0\mu_0+\sum_i x_i)^2}{m+\kappa_0}\right)$$

2.

[Viterbi 算法]

由预测问题的性质，我们希望找到观测序列 x 时使得 P(y|x)最大的 y，而根据条件有

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=2}^{n}\lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_{l=1}^{K_2}\sum_{i=1}^{n}\mu_l s_l(y_i, \boldsymbol{x}, i)\right)$$

如 ppt 中，记

$$f_k(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} t_k(y_{i-1}, y_i, \boldsymbol{x}, i), & k = 1, \dots, K_1 \\ s_l(y_i, \boldsymbol{x}, i), & l = K_1 + l; l = 1, \dots, K_2 \end{cases}$$

$$F_t(y_{t-1}, y_t, \boldsymbol{x}) = \big(f_1(y_{i-1}, y_i, \boldsymbol{x}, i), f_2(y_{i-1}, y_i, \boldsymbol{x}, i), \dots, f_K(y_{i-1}, y_i, \boldsymbol{x}, i)\big), K = K_1 + K_2$$

$$\boldsymbol{w} = \big(\lambda_1, \dots, \lambda_{K_1}, \mu_1, \dots, \mu_{K_2}\big)$$

则

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{t=1}^{n}\boldsymbol{w}^{\top}F_t(y_{t-1}, y_t, \boldsymbol{x})\right)$$

注意到，这里F_t只与 x、y 的前 t 个分量有关，于是假设最优的 y 为y^*，必然满足条件：

$$y^*_{1,\dots,t-1} = \mathrm{argmax}_{y_1,\cdots,y_{t-1}}\mathrm{P}(y_1, \dots, y_{t-1}|y_t, x_1, \dots, x_t)$$

否则，改变$y_1, \dots, y_{t-1}$可以使$\sum_{i=1}^{t} w^T F_i(y_{i-1}, y_i, x)$更大，而其后不受影响，总结果更大。

根据这个最优子结构性质，我们构造如下的算法(假设$n > 1$，不妨设y_i取值 1 到 m)：

第一步：对每个y_2，直接计算使得$P(y_1|y_2, x_1, \dots, x_t)$最大的$y_1$，并记为$y_1^{2,t}$，t 为 1 到 m。

第二步，从$y_i^{l,s}$计算$y_i^{l+1,t}$：

注意到，只要$y_l = s$确定，之前的路径y_1到y_{l-1}随之确定，所以为了计算$y_{l+1} = t$时最优的$y_1, \dots, y_l$，只需要在 x 确定时取$y_l = \mathrm{argmin}_{y_l}\{P\big(y_{l+1} = t, y_l = s, y_i = y_i^{l,s}\big)\}$。

由此，计算出$y_i^{n,s}$，并比较所有 s 中使概率最大的，取这条路径即是最终结果。

由于每步迭代需要进行$O(m^2)$次比较(对$y^{l,t}$的每个 t 需要取 m 个中的最大值)，共进行了$O(n)$次，比较次数$O(m^2 n)$。

由于共需要记录 m 条路径，每条路径长$O(n)$，空间复杂度$O(mn)$。