HW7 and HW14 Reference

（作业证明题证明思路不唯一，表明思路与关键步骤即可）

1 HW7

1.1 实践中使用式(7.15) 决定分类类别时，若数据的维数非常高，则概率连乘$\prod_i^d P(x_i|c)$的结果通常会非常接近于0从而导致下溢.试述防止下溢的可能方案.

（言之有理即可）

解：通常采用取对数的方法将连乘变为连加：$\Pi_{i=1}^d P(x_i|c) \rightarrow log[\Pi_{i=1}^d P(x_i|c)] = \Sigma_{i=1}^d logP(x_i|c)$

1.2 试证明:二分类任务中两类数据满足高斯分布且方差相同时，线性判别分析产生贝叶斯最优分类器.

（思路与关键步骤正确即可）

解：假设数据满足高斯分布：$P(x|c) \sim N(\mu_c, \Sigma)$，模型中需要确定的参数有均值$\mu_1$和$\mu_0$，以及共同的方差$\Sigma$，$\Sigma$为对称正定矩阵。数据集的对数似然为：

$$\begin{aligned} LL(\mu_1, \mu_0, \Sigma) &= \sum_i logP(x_i|c_i) \\ &= \sum_i log\{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(x_i - \mu_{c_i})^T \Sigma^{-1}(x_i - \mu_{c_i})]\} \end{aligned}$$

通过最大化对数似然可以求得参数估计：

$$\nabla_{\mu_c} LL = 0 \Rightarrow \mu_c = \frac{1}{|D_c|} \sum_{x_i \in D_c} x_i$$

$$\nabla_{\Sigma^{-1}} LL = 0 \Rightarrow \Sigma = \frac{1}{m} \sum_i (x_i - \mu_{c_i})(x_i - \mu_{c_i})^T$$

上式中求取Σ^{-1}梯度时应用了关系$\nabla_A |A| = |A|(A^{-1})^T$。那么，该贝叶斯分类器的决策函数为：

$$h_{Bayes}(x) = arg \max_c P(c)P(x|c)$$

对于二分类任务，这等价于：

$$
\begin{aligned}
h_{Bayes}(x) &= sign[P(1)P(x|1) - P(0)P(x|0)] \\
&= sign\{exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)] - exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)]\} \\
&= sign[\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1 + 2(\mu_1-\mu_0)^T\Sigma^{-1}x]
\end{aligned}
$$

上式中第二行采取了同先验假设，亦即P(0)=P(1)=1/2. 在3.4节线性判别分析(LDA)中，关于μ_1, μ_0的定义与上面求得的μ_c完全相同，而根据(3.33)式可知，$S_w = m\Sigma$。在3.4节中求得最优投影直线方向为$w = S_w^{-1}(\mu_0 - \mu_1)$， LDA对于新数据的分类是根据投影点距离两个投影中心的距离远近决定的，可以将其表达为:

$$
\begin{aligned}
h_{LDA}(x) &= sign[(w^Tx - w^T\mu_0)^2 - (w^Tx - w^T\mu_1)^2] \\
&= sign\{[2w^Tx - w^T(\mu_1+\mu_0)][w^T(\mu_1-\mu_0)]\}
\end{aligned}
$$

注意到上式第二行右边项$w^T(\mu_1-\mu_0) = -(\mu_1-\mu_0)^TS_w^{-1}(\mu_1-\mu_0)$，而$S_w$为对称正定矩阵，因此该项恒为负，因此，上式可以进一步化简为：

$$
\begin{aligned}
h_{LDA}(x) &= sign[w^T(\mu_1+\mu_0) - 2w^Tx] \\
&= sign[(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_1+\mu_0-2x)] \\
&= sign[\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1 + 2(\mu_1-\mu_0)^T\Sigma^{-1}x]
\end{aligned}
$$

对比可知，决策函数$h_{Bayes}(x)$和$h_{LDA}(x)$完全相同，因此可以说LDA产生了最优Bayes分类。

1.3 证明EM算法的收敛性

（证明单调性即可，也可使用Jensen不等式）

解：证明EM算法的收敛性，即证明EM算法每次迭代得到的Θ^t满足：

$$
P(X|\Theta^{t+1}) \geq P(X|\Theta^t)
$$

因为$lnP(X|\Theta) = lnP(X,Z|\Theta) - lnP(Z|X,\Theta)$，两边取关于$Z|X,\Theta^t$的期望有：

$$
\mathbb{E}_{Z|X,\Theta^t} lnP(X|\Theta) = \mathbb{E}_{Z|X,\Theta^t} lnP(X,Z|\Theta) - \mathbb{E}_{Z|X,\Theta^t} lnP(Z|X,\Theta)
$$

因为$lnP(X|\Theta)$与Z无关，所以：

$$
\mathbb{E}_{Z|X,\Theta^t} lnP(X|\Theta) = \int_Z P(Z|X,\Theta^t) lnP(X|\Theta) dZ = lnP(X|\Theta)
$$

同时有$\mathbb{E}_{Z|X,\Theta^t} lnP(X,Z|\Theta) = Q(\Theta,\Theta^t)$，记$H(\Theta,\Theta^t) = \mathbb{E}_{Z|X,\Theta^t} lnP(Z|X,\Theta)$

1)

因为$\Theta^{t+1} = \underset{\Theta}{arg\ max}\, Q(\Theta,\Theta^t)$,

所以$Q(\Theta^{t+1},\Theta^t) \geq Q(\Theta,\Theta^t)$,

令$\Theta = \Theta^t$，则 $Q(\Theta^{t+1},\Theta^t) \geq Q(\Theta^t,\Theta^t)$,

2)

$H(\Theta^{t+1},\Theta^t) - H(\Theta^t,\Theta^t) = -D_{KL}[P(Z|X,\Theta^t)||P(Z|X,\Theta^{t+1})]$

根据KL散度的性质可知，上式小于等于0。

综上可知，$P(X|\Theta^n)$单调递增，上界为1，所以EM算法是收敛的。

1.4 在HMM中，求解概率$P(x_{n+1}|x_1, x_2, \cdots, x_n)$

（注意题目要求的是关于观测序列的条件概率，不是联合概率$P(x_1, y_1, \cdots, x_n, y_n)$）
解：在前向算法中，

$$P(x_1, x_2, \cdots, x_n|\lambda) = \sum_{i=1}^{N} \alpha_n(i)$$

所以

$$\begin{aligned} P(x_{n+1}|x_1, x_2, \cdots, x_n) &= \frac{P(x_1, x_2, \cdots, x_{n+1}|\lambda)}{P(x_1, x_2, \cdots, x_n|\lambda)} \\ &= \frac{\sum_{i=1}^{N} \alpha_{n+1}(i)}{\sum_{i=1}^{N} \alpha_n(i)} \\ &= \frac{\sum_{i=1}^{N} \sum j = 1^{N} \alpha_n(j) a_{j,i} b_{i,x_n}}{\sum_{i=1}^{N} \alpha_n(i)} \end{aligned}$$

求解概率$P(x_{n+1}|x_1, x_2, \cdots, x_n)$的步骤为：
（1）初值：$\alpha_1(i) = \pi_i b_{i,x_1}$
（2）递推：$\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j) a_{j,i} b_{i,x_t}$
（3）终止：$P(x_{n+1}|x_1, x_2, \cdots, x_n) = \frac{\sum_{i=1}^{N} \sum j=1^{N} \alpha_n(j) a_{j,i} b_{i,x_n}}{\sum_{i=1}^{N} \alpha_n(i)}$

2 HW14

2.1 假设数据集$D = x_1, x_2, \cdots, x_m$，任意$x_i$是从均值为$\mu$、方差$\lambda^{-1}$的正态分布$N(\mu, \lambda^{-1})$中独立采样而得到。假设$\mu$和$\lambda$的先验分布为$p(\mu, \lambda) = N(\mu|\mu_0, (\kappa 0\lambda)^{-1}) Gam(\lambda|a_0, b_0)$，其中$Gam(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0{}^{a_0} \lambda^{a_0 - 1} exp(-b_0\lambda)$

（1）请写出联合概率分布$p(D, \mu, \lambda)$

（2）请写出证据下界（即变分推断的优化目标），并证明其为观测数据边际似然$\sum_{i=1}^{m} logp(x_i)$的下界

（3）请用变分推断法近似推断后验概率$p(\mu, \lambda|D)$

（第一问注意采样，第二问主要依据KL与证据下界关联和KL性质证明，第三问通过求导来计算分布）
解：（1）
因为x_i从正态分布$N(\mu, \lambda^{-1})$中采样而来，所以有$p(x_i|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} exp(-\frac{\lambda(x_i - \mu)^2}{2})$。
所以

$$\begin{aligned} p(D, \mu, \lambda) &= p(D|\mu, \lambda) p(\mu, \lambda) \\ &= \Pi_{i=1}^{m} p(x_i|\mu, \lambda) p(\mu, \lambda) \\ &= \Pi_{i=1}^{m} \sqrt{\frac{\lambda}{2\pi}} exp(-\frac{\lambda(x_i - \mu)^2}{2}) \sqrt{\frac{\kappa_0}{2\pi}} \frac{1}{\Gamma(a_0)} b_0{}^{a_0} \lambda^{a_0 - \frac{1}{2}} exp(-\frac{\kappa_0 \lambda(\mu - \mu_0)^2}{2} - b_0\lambda) \\ &= (\frac{1}{2\pi})^{\frac{m+1}{2}} \frac{\sqrt{\kappa_0}}{\Gamma(a_0)} b_0{}^{a_0} \lambda^{a_0 + \frac{m-1}{2}} exp(-\sum_{i=1}^{m} \frac{\lambda(x_i - \mu)^2}{2} - \frac{\kappa_0 \lambda(\mu - \mu_0)^2}{2} - b_0\lambda) \end{aligned}$$

（2）证据下界为：

$$L = E_q[log\ p(x,z)] - E_q[log\ q(z)] = E_q[p(x|\mu,\lambda)] + E_q[log\ p(\lambda)] - E_q[q(\mu)] - E_q[log\ q(\mu)]$$

变分目标为找到

$$q^*(z) = \underset{q(z)}{arg\ min}\, KL(q(z)||p(z|x))$$

即需要找到$q^*(z) \approx p(z|x)$来近似得到$p(z|x)$，又：

$$KL(q(z)||p(z|x)) = E_q[log\ q(z)] - E_q[log\ p(x,z)] + log\ p(x) \geq 0$$

所以:

$$\sum_{i=1}^{m} log\ p(x_i) = log\ p(x) \geq E_q[log\ p(x,z)] - E_q[log\ q(z)] = L$$

可知，证据下界即为$\sum_{i=1}^{m} logp(x_i)$的下界，得证

（3）

通过最大化L来最小化$KL(q(z)||p(z|x))$

令

$$\frac{\partial L}{\partial q_\lambda(\mu)} = E_\lambda(log\ p(\mu|\lambda)) + E_\lambda(log\ p(D|\mu,\lambda)) - log\ q(\mu) = 0$$

有

$$\begin{aligned} log\ q^*(\mu) &= -\frac{1}{2}E(\lambda\kappa_0)(\mu-\mu_0)^2 - \frac{1}{2}E(\lambda)\sum_{i=1}^{m}(x_i-\mu)^2 \\ &= -\frac{1}{2}E(\lambda)[(\kappa_0+m)\mu^2 + \sum_{i=1}^{m}{x_i}^2 - 2\mu(\kappa_0\mu_0+m\bar{x})] \\ &= -\frac{1}{2}E(\lambda)[(\kappa_0+m)(\mu - \frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m})^2 + \sum_{i=1}^{m}{x_i}^2 - \frac{(\kappa_0\mu_0+m\bar{x})^2}{\kappa_0+m}] \\ &\sim N(\mu|\frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}, [(\kappa_0+m)E(\lambda)^{-1}]) \end{aligned}$$

令

$$\frac{\partial L}{\partial q_\mu(\lambda)} = E_\mu(log\ p(D|\mu|,\lambda)) + E_\mu(log\ p(\lambda)) - log\ q(\lambda) = 0$$

有

$$\begin{aligned} log\ q^*(\lambda) &= -\frac{1}{2}\lambda E_\mu(\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2) + (a_0-1)log\ \lambda - b_0\lambda + \frac{m+1}{2}log\ \lambda \\ &= (a_0 + \frac{m-1}{2})log\ \lambda - \lambda[b_0 + \frac{1}{2}E_\mu(\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2)] \\ &\sim Gam(\lambda|a_0+\frac{m-1}{2}, b_0 + \frac{1}{2}E_\mu(\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{m}(x_i-\mu)^2)) \end{aligned}$$

所以

$$p(\mu,\lambda|D) \sim N(\mu|\frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}, [(\kappa_0+m)E(\lambda)^{-1}])Gam(\lambda|a_0+\frac{m-1}{2}, b_0+\frac{1}{2}E_\mu(\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^{m}(x_i-\mu)^2))$$

2.2 给出CRF的预测问题的解法

（CRF预测问题可通过维特比算法求解）
解：预测问题，即寻找序列 $y=(y_1,y_2,\cdots,y_n)$, 使得 $P(y|x)$ 最大

$$\begin{aligned} y^* &= arg\ \underset{y}{max}\, P_w(y|x) \\ &= arg\ \underset{y}{max}\, \frac{expw\cdot F(y,x)}{Z_w(x)} \\ &= arg\ \underset{y}{max}\, expw\cdot F(y,x) \\ &= arg\ \underset{y}{max}\, w\cdot F(y,x) \end{aligned}$$

可以看出，CRF的预测问题变为求非规范路径概率最大化的最优路径问题

$$arg\ \underset{y}{max}\, w\cdot F(y,x)$$

此时只需计算非规范化概率，不必计算概率。为了求解最优路径，将优化目标写成如下形式：

$$\underset{y}{max}\sum_{i=1}^{n} w\cdot F_i(y_{i-1},y_i,x)$$

其中，

$$F_i(y_{i-1},y_i,x)=f_1(y_{i-1},y_i,x),f_2(y_{i-1},y_i,x),\cdots,f_K(y_{i-1},y_i,x)^T$$

为局部特征向量。上述最优路径问题可使用维特比算法求解，算法如下：
输入：模型特征向量$F(y,x)$，和权值向量w，观测序列$x=(x_1,x_2,\cdots,x_n)$；
输出：最优路径$y^*=(y_1^*,y_2^*,\cdots,y_n^*)$.
（1）初始化

$$\delta_i(j)=w\cdot F_1(y_0=start,y_1=j,x),j=1,2,\cdots,m$$

（2）递推。对$i=1,2,\cdots,n$

$$\delta_i(l)=\underset{1\le j\le m}{max}\,\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x),l=1,2,\cdots,m$$

$$\Psi_i(l)=arg\ \underset{1\le j\le m}{max}\,\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x),l=1,2,\cdots,m$$

（3）终止

$$\underset{y}{max}\, w\cdot F(y,x)=arg\ \underset{1\le j\le m}{max}\,\delta_n(j)$$

$$y_n^*=arg\ \underset{1\le j\le m}{max}\,\delta_n(j)$$

（4）返回路径

$$y_i^*=\Psi_{i+1}(y_{i+1}^*),i=n-1,n-2,\cdots,1$$

求得最优路径$y^*=(y_1^*,y_2^*,\cdots,y_n^*)$。