Problem Set 2

1.设

$$X=\begin{bmatrix}(x^{(1)})^T\\(x^{(2)})^T\\\cdots\\(x^{(m)})^T\end{bmatrix}, \quad Y=\begin{bmatrix}y^{(1)}\\y^{(2)}\\\cdots\\y^{(m)}\end{bmatrix},$$

因此，$X\theta-Y=\begin{bmatrix}(x^{(1)})^T\theta\\(x^{(2)})^T\theta\\\cdots\\(x^{(m)})^T\theta\end{bmatrix}-\begin{bmatrix}y^{(1)}\\y^{(2)}\\\cdots\\y^{(m)}\end{bmatrix}=\begin{bmatrix}h_\theta(x^{(1)})-y^{(1)}\\h_\theta(x^{(2)})-y^{(2)}\\\cdots\\h_\theta(x^{(m)})-y^{(m)}\end{bmatrix}$,

损失函数可以表达为$J(\theta)=\frac{1}{2m}[(X\theta-Y)^T(X\theta-Y)+\lambda\theta^T\theta]$,

$\nabla_\theta J(\theta)=\nabla_\theta\frac{1}{2m}[(X\theta-Y)^T(X\theta-Y)+\lambda\theta^T\theta]$
$=\frac{1}{2m}[\nabla_\theta(X\theta-Y)^T(X\theta-Y)+\nabla_\theta\lambda\theta^T\theta]$

$\nabla_\theta\lambda\theta^T\theta=\lambda\nabla_\theta\theta^T\theta=\lambda\nabla_\theta tr(\theta\theta^T)=\lambda L\theta$

因此，$\nabla_\theta J(\theta)=\frac{1}{2m}(X^TX\theta-X^TY+\lambda L\theta)$

令$\nabla_\theta J(\theta)=0$，当$X$矩阵各列向量线性独立时，$X^TX$矩阵可逆，存在唯一解$\theta=(X^TX+\lambda L)^{-1}X^TY$.

2.将概率分布代入对数似然函数,

$l(\psi,\mu_0,\mu_1,\sum)=\sum_{i=1}^m logp_{X|Y}(x^{(i)}|y^{(i)};\mu_0,\mu_1,\sum)+\sum_{i=1}^m logp_Y(y^{(i)};\psi)$

$=\sum_{i=1}^m(1-y^{(i)})log\frac{1}{(2\pi)^{n/2}|\sum|^{1/2}}exp(\frac{1}{2}(x^{(i)}-\mu_0)^T\sum^{-1}(x^{(i)}-\mu_0))$

$+\sum_{i=1}^m y^{(i)}log\frac{1}{(2\pi)^{n/2}|\sum|^{1/2}}exp(\frac{1}{2}(x^{(i)}-\mu_1)^T\sum^{-1}(x^{(i)}-\mu_1))$

$+\sum_{i=1}^m log\psi^{y^{(i)}}(1-\psi)^{1-y^{(i)}}$

求取$l(\psi,\mu_0,\mu_1,\sum)$的最大值，令

$\frac{\partial}{\partial\psi}l(\psi,\mu_0,\mu_1,\sum)=0$ (1)

$\nabla_{\mu_0}l(\psi,\mu_0,\mu_1,\sum)=0$ (2)

$\nabla_{\mu_1}l(\psi,\mu_0,\mu_1,\sum)=0$ (3)

$\nabla_{\sum}l(\psi,\mu_0,\mu_1,\sum)=0$ (4)

对于（1）式:

$\frac{\partial}{\partial\psi}\sum_{i=1}^m y^{(i)}log\psi+(1-y^{(i)})log(1-\psi)=0$

$\sum_{i=1}^m\frac{y^{(i)}}{\psi}+\frac{1-y^{(i)}}{1-\psi}=0$

$\sum_{i=1}^m y^{(i)}(1-\psi)+(1-y^{(i)})\psi=0$

$\sum_{i=1}^m y^{(i)}=m\psi$

$\psi=\frac{\sum_{i=1}^m 1\{y^{(i)}=1\}}{m}$

对于（2）式:

$\nabla_{\mu_0}\sum_{i=1}^m(1-y^{(i)})(x^{(i)}-\mu_0)^T\sum^{-1}(x^{(i)}-\mu_0)=0$

$\sum_{i=1}^m(1-y^{(i)})(x^{(i)}-\mu_0)^T\sum^{-1}(x^{(i)}-\mu_0)=0$

$\sum_{i=1}^{m}(1-y^{(i)})[\sum^{-1}(x^{(i)}-\mu_0)d(x^{(i)}-\mu_0)^T+(x^{(i)}-\mu_0)^T\sum^{-1}d(x^{(i)}-\mu_0)]=0$

$\sum_{i=1}^{m}(1-y^{(i)})\sum^{-1}(x^{(i)}-\mu_0)=0$

$\sum_{i=1}^{m}(1-y^{(i)})(x^{(i)}-\mu_0)=0$

$\sum_{i=1}^{m}(1-y^{(i)})x^{(i)}=\sum_{i=1}^{m}(1-y^{(i)})\mu_0$

$\mu_0=\sum_{i=1}^{m}1\{y^{(i)}=0\}x^{(i)}/\sum_{i=1}^{m}1\{y^{(i)}=0\}$

对于（3）式，类同（2）式：

$\mu_0=\sum_{i=1}^{m}1\{y^{(i)}=1\}x^{(i)}/\sum_{i=1}^{m}1\{y^{(i)}=1\}$

对于（4）式：

$\nabla_{\sum}(-\frac{m}{2}log|\sum|)-\frac{1}{2}\sum_{i=1}^{m}(1-y^{(i)})(x^{(i)}-\mu_0)^T\sum^{-1}(x^{(i)}-\mu_0)-\frac{1}{2}\sum_{i=1}^{m}y^{(i)}(x^{(i)}-\mu_1)^T\sum^{-1}(x^{(i)}-\mu_1)=0$

$\nabla_{\sum}(mlog|\sum|)+\nabla_{\sum}\sum_{i=1}^{m}(1-y^{(i)})(x^{(i)}-\mu_0)^T\sum^{-1}(x^{(i)}-\mu_0)+\nabla_{\sum}\sum_{i=1}^{m}y^{(i)}(x^{(i)}-\mu_1)^T\sum^{-1}(x^{(i)}-\mu_1)=0$

已知协方差矩阵$S_i=\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_i)(x^{(i)}-\mu_i)^T$，将通过$S_i$简化表达上式

$\nabla_{\sum}\sum_{i=1}^{m}(x^{(i)}-\mu_i)^T\sum^{-1}(x^{(i)}-\mu_i)$

$=\nabla_{\sum}tr(\sum_{i=1}^{m}(x^{(i)}-\mu_i)^T\sum^{-1}(x^{(i)}-\mu_i))$

$=\nabla_{\sum}tr(\sum_{i=1}^{m}(x^{(i)}-\mu_i)(x^{(i)}-\mu_i)^T\sum^{-1})$

$=\nabla_{\sum}tr(m_iS_i\sum^{-1})$

其中$m_i=\sum_{k=1}^{m}1\{y^{(k)}=i\}$，

$\nabla_{\sum}tr(m_iS_i\sum^{-1})=-m_iS_i^T\sum^{-2}$，

而$\nabla_{\sum}(mlog|\sum|)=m\frac{1}{|\sum|}|\sum|\sum^{-1}=m\sum^{-1}$，

因此，（4）式可简化为

$m\sum^{-1}-\sum_{i}^{2}m_iS_i^T\sum^{-2}=0$

$\sum=\frac{1}{m}\sum_{i}^{2}m_iS_i^T$

$\sum=\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T(x^{(i)}-\mu_{y^{(i)}})$

3.(i)设拉格朗日函数为$L(\Omega,\alpha)=\sum_{y\in Y}c_ylogp_y-\alpha(\sum_{y\in Y}p_y-1)$，其中$\alpha$为拉格朗日乘子，

对p_y求偏导，令$\frac{\partial}{\partial p_y}L(\Omega,\alpha)=0$，

求得$p_y^*=\frac{c_y}{\alpha}$，代入$\sum_{y\in Y}p_y^*=1$得$\frac{\sum_{y\in Y}c_y}{\alpha}=1$，

而$N=\sum_{y\in Y}c_y$，因此$\alpha=N$，进而$p_y^*=\frac{c_y}{N}$

(ii)贝叶斯的最大似然模型的目标函数为

$max\sum_{i=1}^{m}logp(y^{(i)})+\sum_{i=1}^{m}\sum_{j=1}^{n}logp_j(x_j^{(i)}|y^{(i)})$

设标签种类数为k，则$p(y)$满足约束$\sum_{i=1}^{k}p(y)=1$，以及$p(x_j|y)$满足约束$\sum_{j=1}^{n}p(x_j|y)=1$，且所有概率均是非负的。

注意到加号两边可以分开独立进行优化，对于加号左边考虑优化模型：

$max\sum_{i=1}^{m}logp(y^{(i)})$

$s.t.\sum_{i=1}^{k}p(y)=1$

将标签y在训练集中的出现次数$cnt(y)$视为权重c_y，其中$cnt(y)=\sum_{i=1}^{m}1(y^{(i)}=y)$，因此

$max\sum_{i=1}^{m}logp(y^{(i)})=max\sum_{i=1}^{k}cnt(y)logp(y)$，根据第一问的结论有$p^*(y)=\frac{cnt(y)}{m}=\frac{\sum_{i=1}^{m}1(y^{(i)}=y)}{m}$.

同理，将特征x_j在训练集标签为y的样本中的出现次数$cnt(x_j|y)$视为权重c_y，其中
$cnt(x_j|y)=\sum_{i=1}^{m}1(y^{(i)}=y\wedge x_j^{(i)}=x)$，因此

$$
\begin{aligned}
&\quad max \sum_{i=1}^{m}\sum_{j=1}^{n} logp_j(x_j^{(i)}|y^{(i)}) \\
&= max \sum_{j=1}^{n}\sum_{i=1}^{m} logp_j(x_j^{(i)}|y^{(i)}) \\
&= max \sum_{j=1}^{n} cnt(x_j|y)logp_j(x_j|y)
\end{aligned}
$$

根据第一问的结论有$p_j^*(x_j|y) = \frac{cnt(x_j|y)}{cnt(y)} = \frac{\sum_{i=1}^{m} 1(y^{(i)}=y \wedge x_j^{(i)}=x)}{\sum_{i=1}^{m} 1(y^{(i)}=y)}$，证毕。