Problem Set 2

1 Regularized Normal Equation for Linear Regression

> Given a data set $\{x^{(i)}, y^{(i)}\}_{i=1,...m}$ with $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \mathbb{R}$, the general form of regularized linear regression is as follows
>
> $$\min_\theta \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$
>
> Derive the normal equation.

对于正则化线性回归的代价函数：

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

代入假设函数：$h_\theta(x^{(i)}) = x^{(i)}\theta$，转换为矩阵形式，得到：

$$\begin{aligned} J(\theta) &= \frac{1}{2}\left[(X\theta - Y)^T(X\theta - Y) + \lambda A^T A\right] \\ &= \frac{1}{2}\left[(\theta^T X^T - Y^T)(X\theta - Y) + \lambda A^T A\right] \\ &= \frac{1}{2}(\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y + \lambda A^T A) \end{aligned}$$

其中：

$$L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}, \quad A = L\theta = \begin{bmatrix} 0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

对$J(\theta)$求关于θ的偏导：

$$\begin{aligned} \frac{\partial}{\partial\theta}J(\theta) &= \frac{1}{2}\left[2X^T X\theta - X^T Y - (Y^T X)^T + 0 + 2\lambda A\right] \\ &= \frac{1}{2}(2X^T X\theta - 2X^T Y + 2\lambda A) \\ &= X^T X\theta - X^T Y + \lambda A \end{aligned}$$

令$\frac{\partial}{\partial\theta}J(\theta) = 0$，得：

$$X^T X\theta + \lambda A = X^T Y$$

即：

$$X^T X\theta + \lambda L\theta = X^T Y$$

$$(X^T X + \lambda L)\theta = X^T Y$$

等号两侧左乘$(X^T X + \lambda L)^{-1}$，得：

$$\theta = (X^TX + \lambda L)^{-1}X^TY$$

其中:

$$L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

2 Gaussian Discriminant Analysis Model

> Given m training data $\{x^{(i)}, y^{(i)}\}_{i=1,...m}$, assume that $y \sim Bernoulli(\psi)$, $x|y=0 \sim \mathcal{N}(\mu_0, \Sigma)$, $x|y=1 \sim \mathcal{N}(\mu_1, \Sigma)$. Hence, we have
>
> - $p(y) = \psi^y(1-\psi)^{1-y}$
> - $p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))$
> - $p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))$
>
> The log-likelihood function is
>
> $$\begin{aligned} \ell(\psi, \mu_0, \mu_1, \Sigma) &= log\prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) \\ &= log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)}; \psi, \mu_0, \mu_1, \Sigma)p(y^{(i)}; \psi) \end{aligned}$$
>
> Solve ψ, μ_0, μ_1 and Σ by maximizing $\ell(\psi, \mu_0, \mu_1, \Sigma)$.
>
> Hint: $\nabla_X tr(AX^{-1}B) = -(X^{-1}BAX^{-1})^T$, $\nabla_A|A| = |A|(A^{-1})^T$

由题意知:

$$\begin{aligned} \ell(\psi, \mu_0, \mu_1, \Sigma) &= log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)}; \psi, \mu_0, \mu_1, \Sigma)p(y^{(i)}; \psi) \\ &= \sum_{i=1}^{m}(log\, p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) + log\, p(y^{(i)}; \psi)) \\ &= \sum_{i=1}^{m}\big[log\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} + (-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})) \\ &\quad + y^{(i)}log\, \psi + (1-y^{(i)})log(1-\psi)\big] \\ &= \sum_{i=1}^{m}\big[-\frac{n}{2}log\, 2\pi - \frac{1}{2}log|\Sigma| - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}}) \\ &\quad + y^{(i)}log\, \psi + (1-y^{(i)})log(1-\psi)\big] \end{aligned}$$

对$\ell(\psi, \mu_0, \mu_1, \Sigma)$关于$\psi$求偏导:

$$\frac{\partial \ell(\psi,\mu_0,\mu_1,\Sigma)}{\partial \psi} = \sum_{i=1}^{m}(\frac{y^{(i)}}{\psi} - \frac{1-y^{(i)}}{1-\psi})$$

令该式等于0，有：

$$\begin{aligned}
&\sum_{i=1}^{m}(\frac{y^{(i)}}{\psi} - \frac{1-y^{(i)}}{1-\psi}) = 0 \\
\Rightarrow \quad &\sum_{i=1}^{m}[y^{(i)}(1-\psi) - (1-y^{(i)})\psi] = 0 \\
\Rightarrow \quad &\sum_{i=1}^{m}(y^{(i)} - \psi) = 0 \\
\Rightarrow \quad &m\psi = \sum_{i=1}^{m} y^{(i)}
\end{aligned}$$

得：

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

令$x \in \mathbb{R}^{n\times 1}$，$A \in \mathbb{R}^{n\times n}$，对于$\frac{\partial x^TAx}{\partial x}$，有：

$$\begin{aligned}
\frac{\partial x^TAx}{\partial x} &= \begin{bmatrix} \frac{\partial x^TAx}{\partial x_1} \\ \frac{\partial x^TAx}{\partial x_2} \\ \vdots \\ \frac{\partial x^TAx}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_iA_{ij}x_j}{\partial x_1} \\ \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_iA_{ij}x_j}{\partial x_2} \\ \vdots \\ \frac{\partial \sum_{i=1}^{n}\sum_{j=1}^{n} x_iA_{ij}x_j}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n} A_{i1}x_i + \sum_{j=1}^{n} A_{1j}x_j \\ \sum_{i=1}^{n} A_{i2}x_i + \sum_{j=1}^{n} A_{2j}x_j \\ \vdots \\ \sum_{i=1}^{n} A_{in}x_i + \sum_{j=1}^{n} A_{nj}x_j \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n} A_{i1}x_i \\ \sum_{i=1}^{n} A_{i2}x_i \\ \vdots \\ \sum_{i=1}^{n} A_{in}x_i \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{n} A_{1j}x_j \\ \sum_{j=1}^{n} A_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} A_{nj}x_j \end{bmatrix} \\
&= Ax + A^Tx
\end{aligned}$$

当A为对称矩阵时，有：$A = A^T$，因此：

$$\frac{\partial x^TAx}{\partial x} = Ax + A^Tx = 2Ax$$

而对于对称的协方差矩阵Σ，显然满足该条件，在上式的基础上，对$\ell(\psi,\mu_0,\mu_1,\Sigma)$关于$\mu_0$求偏导：

$$\begin{aligned}\frac{\partial \ell(\psi,\mu_0,\mu_1,\Sigma)}{\partial \mu_0} &= \sum_{i=1}^{m}(\frac{1}{2}\cdot 2\cdot \Sigma^{-1}(x^{(i)}-\mu_0)\cdot 1\{y^{(i)}=0\}) \\ &= \sum_{i=1}^{m}\Sigma^{-1}(x^{(i)}-\mu_0)\cdot 1\{y^{(i)}=0\}\end{aligned}$$

令上式为0：

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

Σ为协方差矩阵，故Σ^{-1}不为0，可约去，有：

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

得：

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

同理可得：

$$\mu_1=\frac{\sum_{i=1}^{m}x^{(i)}\cdot 1\{y^{(i)}=1\}}{\sum_{i=1}^{m}\cdot 1\{y^{(i)}=1\}}$$

由题目中提供的公式$\nabla_X tr(AX^{-1}B)=-(X^{-1}BAX^{-1})^T$, $\nabla_A|A|=|A|(A^{-1})^T$，

对$\ell(\psi,\mu_0,\mu_1,\Sigma)$关于$\Sigma$求偏导：

$$\begin{aligned}\frac{\partial \ell(\psi,\mu_0,\mu_1,\Sigma)}{\partial \Sigma} &= \sum_{i=1}^{m}\big[-\frac{1}{2}\frac{1}{|\Sigma|}\cdot|\Sigma|\cdot(\Sigma^{-1})^T+\frac{1}{2}(\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1})^T\big] \\ &= \sum_{i=1}^{m}\big[-\frac{1}{2}(\Sigma^{-1})^T+\frac{1}{2}(\Sigma^{-1})^T(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T(\Sigma^{-1})^T\big]\end{aligned}$$

由于Σ是对称的，有：

$$\begin{aligned}\frac{\partial \ell(\psi,\mu_0,\mu_1,\Sigma)}{\partial \Sigma} &= \sum_{i=1}^{m}\big[-\frac{1}{2}(\Sigma^T)^{-1}+\frac{1}{2}(\Sigma^T)^{-1}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T(\Sigma^T)^{-1}\big] \\ &= \sum_{i=1}^{m}\big[-\frac{1}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}\big]\end{aligned}$$

令该式等于0，有：

$$
\begin{aligned}
& \sum_{i=1}^{m}\Big[-\frac{1}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}\Big]=0 \\
\Rightarrow\quad & \sum_{i=1}^{m}\big[1-\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\big]=0 \\
\Rightarrow\quad & m=\sum_{i=1}^{m}\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T \\
\Rightarrow\quad & m\Sigma=\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
\end{aligned}
$$

得：

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

3 MLE for Naive Bayes

> Consider the following definition of **MLE problem for multinomials**. The input to the problem is a finite set $\mathcal{Y}$, and a weight $c_y \geq 0$ for each $y \in \mathcal{Y}$.
>
> The output from the problem is the distribution p^* that solves the following maximization problem.
>
> $$p^* = arg\ \underset{p\in\mathcal{P}_{\mathcal{Y}}}{max}\sum_{y\in\mathcal{Y}} c_y log\left(p_y\right)$$

> (i) Prove that, the vector p^* has components
>
> $$p^*_y=\frac{c_y}{N}$$
>
> for $\forall y \in \mathcal{Y}$, where $N=\sum_{y\in\mathcal{Y}} c_y$.
>
> Hint: Use the theory of Lagrange multiplier.

要求解该问题：

$$
\underset{p\in\mathcal{P}_y}{max}\sum_{y\in\mathcal{Y}} c_y log\, p_y
$$

即求解：

$$
\begin{aligned}
& \underset{p\in\mathcal{P}_y}{min}(-\sum_{y\in\mathcal{Y}} c_y log\, p_y) \\
s.t.\quad & \sum_{y\in\mathcal{Y}} p_y=1,\quad p_y\geq 0
\end{aligned}
$$

由上述条件构建拉格朗日问题：

$$
L(c_y,p_y)=-\sum_{y\in\mathcal{Y}} c_y log\, p_y+\lambda(\sum_{y\in\mathcal{Y}} p_y-1)-\sum_{y\in\mathcal{Y}}\mu_y p_y
$$

其中$\mu_y \geq 0$。

关于$\forall y \in \mathcal{Y}$求偏导，均有：

$$\frac{\partial L(c_y, p_y)}{\partial p_y} = -\frac{c_y}{p_y} + \lambda - \mu_y$$

令该式等于0，得：

$$c_y = \lambda p_y - \mu_y p_y$$

由拉格朗日问题性质，有：

$$\mu_y p_y = 0, \quad \forall y \in \mathcal{Y}$$

代入上式，有：

$$p_y = \frac{c_y}{\lambda}$$

由：

$$\begin{aligned} & \sum_{y \in \mathcal{Y}} p_y = 1 \\ \Rightarrow \quad & \sum_{y \in \mathcal{Y}} \frac{c_y}{\lambda} = 1 \\ \Rightarrow \quad & \lambda = \sum_{y \in \mathcal{Y}} c_y \end{aligned}$$

代入上式，得证：

$$p_y = \frac{c_y}{\sum_{y \in \mathcal{Y}} c_y}$$

> (ii) Using the above consequence, prove that, the maximum-likelihood estimates for Naive Bayes model are as follows
>
> $$p(y) = \frac{\sum_{i=1}^m 1(y^{(i)} = y)}{m}$$
>
> and
>
> $$p_j(x|y) = \frac{\sum_{i=1}^m 1(y^{(i)} = y \wedge x_j^{(i)} = x)}{\sum_{i=1}^m 1(y^{(i)} = y)}$$

对于对数似然函数：

$$
\begin{aligned}
\ell(\Omega) &= log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{m} log\, p(x^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{m} log\, (p(y^{(i)}) \prod_{j=1}^{n} p_j(x_j^{(i)} | y^{(i)})) \\
&= \sum_{i=1}^{m} log\, p(y^{(i)}) + \sum_{i=1}^{m} \sum_{j=1}^{n} log\, p_j(x_j^{(i)} | y^{(i)}) \\
&= \sum_{y \in \mathcal{Y}} count(y) log\, p_y + \sum_{j=1}^{n} \sum_{y \in \mathcal{Y}} \sum_{x \in \{-1,+1\}} count_j(x|y) log\, p_j(x|y)
\end{aligned}
$$

其中：

$$
\begin{aligned}
count(y) &= \sum_{i=1}^{m} 1\{y^{(i)} = y\} \\
count_j(x|y) &= \sum_{i=1}^{m} 1\{y^{(i)} = y \wedge x_j^{(i)} = x\}
\end{aligned}
$$

要最大化上式，可分别最大化上式中的两部分。

对于对数似然函数中的第一部分：

$$
\begin{aligned}
& max \sum_{y \in \mathcal{Y}} count(y) log\, p_y \\
s.t. \quad & \sum_{y \in \mathcal{Y}} p_y = 1, \quad p_y \geq 0
\end{aligned}
$$

等价于(i)中的问题，可得：

$$
p_y = \frac{count(y)}{\sum_{y \in \mathcal{Y}} count(y)} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = y\}}{m}
$$

同理，对于第二部分：

$$
\begin{aligned}
& max \sum_{j=1}^{n} \sum_{y \in \mathcal{Y}} \sum_{x \in \{-1,+1\}} count_j(x|y) log\, p_j(x|y) \\
s.t. \quad & \sum_{j=1}^{n} \sum_{y \in \mathcal{Y}} \sum_{x \in \{-1,+1\}} p_j(x|y) = 1, \quad p_j(x|y) \geq 0
\end{aligned}
$$

可得：

$$
p_j(x|y) = \frac{count_j(x|y)}{\sum_{j=1}^{n} \sum_{y \in \mathcal{Y}} \sum_{x \in \{-1,+1\}} count_j(x|y)} = \frac{\sum_{i=1}^{m} 1(y^{(i)} = y \wedge x_j^{(i)} = x)}{\sum_{i=1}^{m} 1(y^{(i)} = y)}
$$