Problem Set 2

林小斌

2018年11月21日

1 Logistic Regression

$$
\begin{aligned}
H_{i,j} &= \frac{\partial^2 J(\theta)}{\partial\theta_i \partial\theta_j} \\
&= \frac{\partial}{\partial\theta_i}\left(-\frac{1}{m}\sum_{t=1}^{m}\frac{1}{h_\theta(y^{(t)}x^{(t)})}\frac{\partial}{\partial\theta_j}h_\theta(y^{(t)}x^{(t)})\right) \\
&= \frac{\partial}{\partial\theta_i}\left(-\frac{1}{m}\sum_{t=1}^{m}\frac{1}{h_\theta(y^{(t)}x^{(t)})}h_\theta(y^{(t)}x^{(t)})(1-h_\theta(y^{(t)}x^{(t)}))\frac{\partial}{\partial\theta_j}y^{(t)}\theta^T x^{(t)}\right) \\
&= \frac{\partial}{\partial\theta_i}\left(-\frac{1}{m}\sum_{t=1}^{m}(1-h_\theta(y^{(t)}x^{(t)}))y^{(t)}x_j^{(t)}\right) \\
&= \frac{1}{m}\sum_{t=1}^{m}\frac{\partial}{\partial\theta_i}\left(h_\theta(y^{(t)}x^{(t)}\right)y^{(t)}x_j^{(t)} \\
&= \frac{1}{m}\sum_{t=1}^{m}y^{(t)}x_j^{(t)}h_\theta\left(y^{(t)}x^{(t)}\right)\left(1-h_\theta(y^{(t)}x^{(t)})\right)\frac{\partial}{\partial\theta_i}y^{(t)}\theta^T x^{(t)} \\
&= \frac{1}{m}\sum_{t=1}^{m}\left(y^{(t)}\right)^2 x_i^{(t)}x_j^{(t)}h_\theta\left(y^{(t)}x^{(t)}\right)\left(1-h_\theta(y^{(t)}x^{(t)})\right) \\
&= \frac{1}{m}\sum_{t=1}^{m}x_i^{(t)}x_j^{(t)}h_\theta\left(y^{(t)}x^{(t)}\right)\left(1-h_\theta(y^{(t)}x^{(t)})\right)
\end{aligned}
$$

thus

$$H = \frac{1}{m}\sum_{t=1}^{m}\left[h_\theta\left(y^{(t)}x^{(t)}\right)\left(1-h_\theta\left(y^{(t)}x^{(t)}\right)\right)x^{(t)}\left(x^{(t)}\right)^T\right]$$

where $x^{(t)} \in R^{n+1}$ and $x^{(t)}\left(x^{(t)}\right)^T \in R^{n+1\times n+1}$.

Consider $z \in R^{n+1}$, we get the following formula:

$$z^T H z = \frac{1}{m} \sum_{t=1}^{m} \left[h_\theta \left(y^{(t)} x^{(t)} \right) \left(1 - h_\theta \left(y^{(t)} x^{(t)} \right) \right) z^T x^{(t)} \left(x^{(t)} \right)^T z \right]$$

(1)Consider $g(z)$ is sigmod function, then

$$h_\theta \left(y^{(t)} x^{(t)} \right) \left(1 - h_\theta \left(y^{(t)} x^{(t)} \right) \right) > 0$$

(2)Calculate $z^T x^{(t)} \left(x^{(t)} \right)^T z$

$$\begin{aligned} z^T x^{(t)} \left(x^{(t)} \right)^T z &= \left(z^T x^{(t)} \right) \left(\left(x^{(t)} \right)^T z \right) \\ &= \left(\begin{bmatrix} z_1 & \dots & z_{n+1} \end{bmatrix} \begin{bmatrix} x_1^{(t)} \\ \vdots \\ x_{n+1}^{(t)} \end{bmatrix} \right) \left(\begin{bmatrix} x_1^{(t)} & \dots & x_{n+1}^{(t)} \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_{n+1} \end{bmatrix} \right) \\ &= \left(z^T x^{(t)} \right)^2 \geq 0 \end{aligned}$$

In summary, $z^T H z \geq 0$

2 Regularized Normal Equation for Linear Regression

Consider the cost function

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

The normal equation is to find the parameters that minimize the cost function by solving the following equations.

$$\frac{\partial}{\partial \theta_j} J(\theta) = 0$$

Assuming that there are m training examples, each instance has n characteristics, the training example set is

$$X = \begin{bmatrix} x_0^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \dots & x_n^{(m)} \end{bmatrix}$$

where $x_j^{(i)}$ represents the j feature of the i instance.
Consider

$$\theta = \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix}^T$$

$$Y = \begin{bmatrix} y^{(1)} & y^{(2)} & \dots & y^{(m)} \end{bmatrix}^T$$

thus

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

where L is $m \times m$ matrix and $L = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$

Derivation is equivalent to the following form

$$\frac{1}{2m}\left(\frac{\partial Y^TY}{\partial\theta} - \frac{\partial Y^TX\theta}{\partial\theta} - \frac{\partial\theta^TX^TY}{\partial\theta} + \frac{\partial\theta^TX^TX^T\theta}{\partial\theta} + \lambda L\frac{\partial\theta^2}{\partial\theta}\right)$$

(1)For the first item

$$\frac{\partial Y^TY}{\partial\theta} = 0$$

(2)For the second item

$$\begin{aligned} Y^TX\theta &= \begin{bmatrix} y^{(1)} & y^{(2)} & \dots & y^{(m)} \end{bmatrix} \begin{bmatrix} x_0^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \dots & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix}^T \\ &= \left(x_0^{(1)}y^{(1)} + \dots x_0^{(m)}y^{(m)}\right)\theta_0 + \dots + \left(x_n^{(1)}y^{(1)} + \dots x_n^{(m)}y^{(m)}\right)\theta_n \end{aligned}$$

thus

$$\frac{\partial Y^TX\theta}{\partial\theta} = \begin{bmatrix} \frac{\partial Y^TX\theta}{\partial\theta_0} \\ \frac{\partial Y^TX\theta}{\partial\theta_1} \\ \vdots \\ \frac{\partial Y^TX\theta}{\partial\theta_n} \end{bmatrix} = X^TY$$

(3)For the third item

$$\theta^T X^T Y = \begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix} \begin{bmatrix} x_0^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \dots & x_n^{(m)} \end{bmatrix}^T \begin{bmatrix} y^{(1)} & y^{(2)} & \dots & y^{(m)} \end{bmatrix}^T$$
$$= \left(x_0^{(1)} theta_0 + \dots x_0^{(m)} \theta_n\right) y^{(1)} + \dots + \left(x_n^{(1)} \theta_0 + \dots x_n^{(m)} \theta_n\right) y^{(n)}$$

thus

$$\frac{\partial \theta^T X^T Y}{\partial \theta} = \begin{bmatrix} \frac{\partial \theta^T X^T Y}{\partial \theta_0} \\ \frac{\partial \theta^T X^T Y}{\partial \theta_1} \\ \vdots \\ \frac{\partial \theta^T X^T Y}{\partial \theta_n} \end{bmatrix} = X^T Y$$

(4)For the fourth item

$$\theta^T X^T X \theta = X^T X \left(\theta_0^2 + \theta_1^2 + \dots + \theta_n^2\right)$$

thus

$$\frac{\partial \theta^T X^T X \theta}{\partial \theta} = \begin{bmatrix} \frac{\partial \theta^T X^T X \theta}{\partial \theta_0} \\ \frac{\partial \theta^T X^T X \theta}{\partial \theta_1} \\ \vdots \\ \frac{\partial \theta^T X^T X \theta}{\partial \theta_n} \end{bmatrix} = 2\left(X^T X\right) \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} = 2X^T X \theta$$

(5)For the fifth item

$$\lambda L \frac{\partial \theta^2}{\theta} = 2\lambda L \theta$$

In summary, the normal equation is:

$$\frac{1}{2m}\left(-2X^T Y + 2X^T X \theta + 2\lambda L \theta\right) = 0$$

thus

$$\theta = \left(X^T X + \lambda L\right)^{-1} X^T Y$$

3 Gaussian Discriminant Analysis Model

According to the subject

$$
\begin{aligned}
l(\psi,\mu_0,\mu_1,\Sigma) &= log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\psi,\mu_0,\mu_1,\Sigma) \\
&= log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\psi,\mu_0,\mu_1,\Sigma)p(y^{(i)};\psi) \\
&= \sum_{i=1}^{m} logp(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \sum_{i=1}^{m} logp(y^{(i)};\psi) \\
&= \sum_{i=1}^{m} logp(x^{(i)}|y^{(i)}=0)^{1-y^{(i)}} \cdot p(x^{(i)}|y^{(i)}=1)^{y^{(i)}} + \sum_{i=1}^{m} logp(y^{(i)}) \\
&= \sum_{i=1}^{m}(1-y^{(i)})logp(x^{(i)}|y^{(i)}=0) + \sum_{i=1}^{m} y^{(i)}logp(x^{(i)}|y^{(i)}=1) + \sum_{i=1}^{m} logp(y^{(i)})
\end{aligned}
$$

(1)Finding partial derivatives for ψ

$$
\begin{aligned}
\frac{\partial l(\psi,\mu_0,\mu_1,\Sigma)}{\partial\psi} &= \frac{\sum_{i=1}^{m} logp(y^{(i)})}{\partial\psi} \\
&= \frac{\partial \sum_{i=1}^{m} \psi^{y^{(i)}}(1-\psi)^{1-y^{(i)}}}{\partial\psi} \\
&= \frac{\partial \sum_{i=1}^{m} \left(y^{(i)}log\psi + (1-y^{(i)})log(1-\psi)\right)}{\partial\psi} \\
&= \sum_{i=1}^{m}\left(y^{(i)}\frac{1}{\psi} - (1-y^{(i)})\frac{1}{1-\psi}\right) \\
&= \sum_{i=1}^{m}\left(I(y^{(i)}=1)\frac{1}{\psi} - I(y^{(i)}=0)\frac{1}{1-\psi}\right)
\end{aligned}
$$

where I is Indicator function, let the formula be zero, we get the final ψ:

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

(2)Finding partial derivatives for μ_0 and μ_1

$$
\begin{aligned}
\frac{\partial l(\psi,\mu_0,\mu_1,\Sigma)}{\partial \mu_0} &= \frac{\partial (1-y^{(i)})logp(x^{(i)}|y^{(i)}=0)}{\partial \mu_0} \\
&= \frac{\sum_{i=1}^{m}(1-y^{(i)})\left(log\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)}{\partial \mu_0} \\
&= \sum_{i=1}^{m}(1-y^{(i)})(\Sigma^{-1}(x^{(i)}-\mu_0)) \\
&= I(y^{(i)}=0)\Sigma^{-1}(x^{(i)}-\mu_0)
\end{aligned}
$$

Let the formula be zero, we get the final μ_0:

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

According to symmetry

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

(3)Finding partial derivatives for Σ The following is a partial derivative of Σ. Since only the first two parts of the likelihood function are related to Σ, the first two parts are rewritten as follows

$$
\begin{aligned}
&\sum_{i=1}^{m}(1-y^{(i)})logp(x^{(i)}|y^{(i)}=0) + \sum_{i=1}^{m} y^{(i)}logp(x^{(i)}|y^{(i)}=1) \\
&= \sum_{i=1}^{m}(1-y^{(i)})\left(log\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right) \\
&\quad + \sum_{i=1}^{m} y^{(i)}\left(log\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} - \frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right) \\
&= \sum_{i=1}^{m}\left(log\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} - \frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})\right) \\
&= \sum_{i=1}^{m}\left(-\frac{n}{2}log(2\pi) - \frac{1}{2}log(|\Sigma|)\right) - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})
\end{aligned}
$$

thus

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

The following formula is used for derivation.

$$\frac{\partial|\Sigma|}{\partial\Sigma} = |\Sigma|\Sigma^{-1}, \frac{\partial\Sigma^{-1}}{\partial\Sigma} = -\Sigma^{-2}$$

Let the formula be zero, we get the final Σ:

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

4 MLE for Naive Bayes

(i) According to the question,

$$p^* = arg\ max_{p\in P_y}\sum_{y\in Y} c_y log p_y$$

limited to $\sum_{y\in Y} p_y = 1$. Using Lagrange multiplier method, we have

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

The derivation of $p_1, p_2, \ldots, p_y$ is 0 respectively.

$$\begin{aligned}
&\frac{c_1}{p_1} + \lambda = 0\\
&\frac{c_2}{p_2} + \lambda = 0\\
&\vdots\\
&\frac{c_y}{p_y} + \lambda = 0\\
&\sum_{y\in Y} p_y = 1\\
&\sum_{y\in Y} c_y = N
\end{aligned}$$

Above all, we have

$$p_y^* = \frac{c_y}{N}$$

(ii) Maximum-likelihood Estimates for Naive Bayes

$$
\begin{aligned}
l(\Omega) &= \sum_{i=1}^{m} log\ p(x^{(i)}, y^{i}) \\
&= \sum_{i=1}^{m} log \left(p(y^{(i)}) \prod_{j=1}^{n} p_j(x_j^{(i)} | y^{(i)}) \right) \\
&= \sum_{i=1}^{m} log\ p(y^{(i)}) + \sum_{i=1}^{m} \sum_{j=1}^{n} log p_j(x_j^{(i)} | y^{(i)}) \\
&= \sum_{y=1}^{k} count(y) log\ p(y) + \sum_{j=1}^{n} \sum_{y=1}^{k} \sum_{x \in 0,1} count_j(x|y) log\ p_j(x_j^{(i)} | y^{(i)})
\end{aligned}
$$

where

$$
\begin{aligned}
count(y) &= \sum_{i=1}^{m} I(y^{(i)} = y), \forall y \in 1, 2, \ldots, k \\
count(x|y) &= \sum_{i=1}^{m} I(y^{(i)} = y, x_j^{(i)} = x), \forall y \in 1, 2, \ldots, k, \forall x \in 0, 1
\end{aligned}
$$

The $\frac{\partial l(\Omega)}{\partial p(y)}$ is not related to the second one. Using Lagrange multiplier method, we have

$$
L_1(\Omega, \lambda_1) = \sum_{y=1}^{k} count(y) log\ p(y) + \lambda_1 (\sum_{y=1}^{k}\ p(y) - 1)
$$

where $\sum_{y=1}^{k}\ p(y) = 1$. Thus

$$
\begin{aligned}
\frac{\partial L_1(\Omega, \lambda_1)}{\partial p(y)} &= \frac{count(y)}{p(y)} + \lambda_1 = 0 \\
p(y) &= \frac{-count(y)}{\lambda_1} \\
\sum_{y=1}^{k}\ p(y) &= -\frac{1}{\lambda_1} \sum_{y=1}^{k} count(y) = 1
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\lambda_1 &= -\sum_{y=1}^{k} count(y) = -m \\
p(y) &= \frac{count(y)}{m} = \frac{\sum_{i=1}^{m} I(y^{(i)} = y)}{m} \qquad (1)
\end{aligned}
$$

Similarity, we have

$$L_2(\Omega, \lambda_2) = \sum_{j=1}^{n} \sum_{y=1}^{k} \sum_{x \in 0,1} count_j(x|y) log\ p_j(x|y) + \lambda_2 (\sum_{x \in 0,1} \ p(x|y) - 1)$$

where $\sum_{x \in 0,1} \ p(x|y)$. Thus

$$\begin{aligned}
&\frac{\partial L_2(\Omega, \lambda_2)}{\partial p_j(x|y)} = \frac{count_j(x|y)}{p_j(x|y)} + \lambda_2 = 0 \\
&p_j(x|y) = \frac{-count_j(x|y)}{\lambda_2} \\
&\sum_{x \in 0,1} \ p(x|y) = -\frac{1}{\lambda_2} \sum_{x \in 0,1} count_j(x|y) = 1 \\
&\lambda_2 = -\sum_{x \in 0,1} count_j(x|y) = -\sum_{i=1}^{m} I(y^{(i)} = y)
\end{aligned}$$

Then we have

$$p_j(x|y) = \frac{-count_j(x|y)}{-\sum_{i=1}^{m} I(y^{(i)} = y)} = \frac{\sum_{i=1}^{m} I(y^{(i)} = y, x_j^{(i)} = x)}{\sum_{i=1}^{m} I(y^{(i)} = y)} \tag{2}$$