Solution for Assignment 1

Chinh La

2018-09-25

---

1 Newton's method for computing least squares

(a) Find the Hessian of cost function $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2$

Answer:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 \tag{1}$$

$$\frac{\partial J(\theta)}{\partial \theta_i} = \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})x^{(i)} \tag{2}$$

$$\frac{\partial^2 J(\theta)}{\partial \theta_i \partial \theta_j} = \sum_{i=1}^{m} x^{(j)} x^{(i)} \tag{3}$$

The **Hessian** matrix of cost function $J(\theta)$ is an $n \times n$ matrix where

$$H_{ij} = \sum_{i=0}^{m} x^i x^j = X^T X_{ji}$$

with $X = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(m)} \end{bmatrix}^T$

(b) Show that the first iteration of Newton's method gives us $\theta^* = (X^T X)^{-1} X^T \vec{y}$

Answer: As the notes 1 show, $\theta \in R^{1xn}$, Newton's method performs the following update:

$$\theta_{i+1} := \theta_i - H^{-1}\nabla_\theta l(\theta_i) \tag{4}$$

In this problem, $\nabla_\theta l(\theta) = X^T(X\theta^T - \vec{y})$, $\vec{y} = \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(m)} \end{bmatrix}^T$, $H^{-1} = (X^T X)^{-1}$.

Let $\theta_0 = \vec{0}^T$, we have

$$\Rightarrow \theta_1 = \theta_0 - (X^T X)^{-1} X^T (X\theta^T - \vec{y}) \tag{5}$$

$$\Rightarrow \theta_1 = \vec{0}^T - (X^T X)^{-1} X^T (X\dot{0} - \vec{y}) \tag{6}$$

$$\Rightarrow \theta_1 = (X^T X)^{-1} X^T \vec{y} \tag{7}$$

So θ_1 is optimal solution of cost function $l(\theta)$ and we only need 1 iteration.

---

2 Locally-weighted logistic regression

See a1q2.py file

---

3 Multivariate least squares

(a) The cost function for this case is

$$J(\Theta) = \frac{1}{2}\sum_{n=1}^{m}\sum_{j=1}^{p}\left(\left(\Theta^T x^{(i)}\right)_j - y_j^{(i)}\right)^2$$

Find the matrix way to write it.

Answer:

Let $X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}, Y = \begin{bmatrix} y_1^{(1)} & y_2^{(2)} & \cdots & y_p^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \cdots & y_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y_p^{(m)} & y_2^{(m)} & \cdots & y_p^{(m)} \end{bmatrix}, \Theta = \begin{bmatrix} \theta_{1(1)} & \theta_{1(2)} & \cdots & \theta_{1(p)} \\ \theta_{2(1)} & \theta_{2(2)} & \cdots & \theta_{2(p)} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{n(1)} & \theta_{n(2)} & \cdots & \theta_{n(p)} \end{bmatrix}$

We have

$$J(\Theta) = tr\left((X\Theta - Y)^T(X\Theta - Y)\right) \tag{8}$$

$$J(\Theta) = \sum_{i=1}^{m}\left((X\Theta - Y)^T(X\Theta - Y)\right)_{ii} \tag{9}$$

$$J(\Theta) = \sum_{i=1}^{m}\left((X\Theta - Y)^T(X\Theta - Y)\right)_{ii} \tag{10}$$

where $\left((X\Theta - Y)^T(X\Theta - Y)\right)_{ii}$ is the ith element on main diagonal of $(\Theta^T X - Y)^T(\Theta^T X - Y)$ and

$$(x\Theta - Y) = \begin{bmatrix}\theta^T x^{(1)} - & \theta^T x^{(2)} & \cdots & \theta^T x^{(m)}\end{bmatrix} \tag{11}$$

$$\left((X\Theta - Y)^T(X\Theta - Y)\right)_{ii} = \sum_{j=1}^{p}\left(\left(\Theta^T x^{(i)}\right)_j - y_j^{(i)}\right)^2 \tag{12}$$

$$(10)(12) \Rightarrow J(\Theta) = \frac{1}{2}\sum_{n=1}^{m}\sum_{j=1}^{p}\left(\left(\Theta^T x^{(i)}\right)_j - y_j^{(i)}\right)^2 \tag{13}$$

(b) Find the closed form solution for Θ which minimizes $J(\Theta)$. This is equivalent to the normal equations for multivariate case.

Answer:

$$\nabla_\Theta J(\Theta) = \frac{1}{2}\nabla_\Theta tr\left((X\Theta - Y)^T(X\Theta - Y)\right) \tag{14}$$

$$\Rightarrow \nabla_\Theta J(\Theta) = \frac{1}{2}\nabla_\Theta tr\left((\Theta^T X^T - Y^T)(X\Theta - Y)\right) \tag{15}$$

$$\Rightarrow \nabla_\Theta J(\Theta) = \frac{1}{2}\nabla_\Theta tr(\Theta^T X^T X\Theta - \Theta^T X^T Y - Y^T X\Theta + Y^T Y) \tag{16}$$

$$\Rightarrow \nabla_\Theta J(\Theta) = \frac{1}{2}\nabla_\Theta tr(\Theta^T X^T X\Theta) - \nabla_\Theta tr(\Theta^T X^T Y) - \nabla_\Theta tr(Y^T X\Theta) \tag{17}$$

$$\Rightarrow \nabla_\Theta J(\Theta) = \frac{1}{2}\left((X^T X)^T\Theta + (X^T X)\Theta - X^T Y - (Y^T X)^T\right) \tag{18}$$

$$\Rightarrow \nabla_\Theta J(\Theta) = (X^T X)\Theta - X^T Y \tag{19}$$

Let $\nabla_\Theta J(\Theta) = 0$ and we have the optimal Θ^*

$$\Theta^* = (X^T X)^{-1} X^T Y \tag{20}$$

(c) Suppose instead of considering the multivariate vetors $y^{(i)}$ all at once, we instead cumpute each variable $y_j^{(i)}$ separately for each $j = 1, \ldots, p$. In this case, we have a p individual linear models, of the form

$$y_j^{(i)} = \theta_j^T x^{(i)}, j = 1, \ldots, p$$

(So here, each $\theta_j \in \Re^n$). How do the parameters from these p independent least squares problems compare to the multivariate solution?

Answer:

As we have

$$\theta_i = (X^T X)^{-1} X^T \vec{y_i} \tag{21}$$

where $\Theta = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_m \end{bmatrix}, Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}$

Then the solution of each independent model become the solution of multivariate case.

---

4 Naive Bayes

In this problem, we look at maximum likelihood parameter estimation using the naive Bayes assumption. Here, the input features $x_j, j = 1, \ldots, n$ to our model are discrete, binary-valued variable, so $x_j \in \{0, 1\}$. We call $x = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}$ to be the input vector. For each training example, our output targets are single binary-value $y \in \{0, 1\}$. Our model is then parameterized by $\phi_{j|y=0} = p(x_j = 1|y_j = 0), \phi_{j|y=1} = p(x_j = 1|y_j = 1)$ and $\phi_y = p(y = 1)$. We model the joint distribution of (x, y) according to

$$\begin{aligned}
p(y) &= (\phi_y)^y (1 - \phi_y)^{1-y} \\
p(x|y = 0) &= \prod_{j=1}^{n} p(x_j|y = 0) \\
&= \prod_{j=1}^{n} (\phi_{j|y=0})^{x_j} (1 - \phi_{j|y=0})^{1-x_j} \\
p(x|y = 1) &= \prod_{j=1}^{n} p(x_j|y = 1) \\
&= \prod_{j=1}^{n} (\phi_{j|y=1})^{x_j} (1 - \phi_{j|y=1})^{1-x_j}
\end{aligned}$$

(a) Find the joint likelihood function $l(\varphi) = \log \left(\prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \varphi) \right)$ in terms of the model parameters given above. Here, φ represents the entire parameters $\{\phi_y, \phi_{y=0}, \phi_{y=1}, j = 1, \ldots, n\}$.

Answers:

$$l(\varphi) = \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \varphi) \tag{22}$$

$$l(\varphi) = \sum_{i=1}^{m} \log \left(p(x^{(i)}, y^{(i)}; \varphi) \right) \tag{23}$$

$$l(\varphi) = \sum_{i=1}^{m} \log \left(p(x^{(i)}|y^{(i)}; \varphi) p(y^{(i);\varphi}) \right) \tag{24}$$

$$l(\varphi) = \sum_{i=1}^{m} \left(\log p(x^{(i)}|y^{(i)}; \varphi) + \log p(y^{(i)}; \varphi) \right) \tag{25}$$

$$l(\varphi) = \sum_{i=1}^{m} \left(\log \prod_{j=1}^{n} p(x_j^{(i)}|y^{(i)}; \varphi) + \log p(y^{(i)}; \varphi) \right) \tag{26}$$

$$l(\varphi) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log p(x_j^{(i)}|y^{(i)}; \varphi) + \log p(y^{(i)}; \varphi) \right) \tag{27}$$

$$l(\varphi) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \left(x_j^{(i)} \log \phi_{j|y} + (1 - x_j^{(i)}) \log(1 - \phi_{j|y}) \right) + y^{(i)} \log \phi_y + (1 - y^{(i)}) \log(1 - \phi_y) \right) \tag{28}$$

(b) Show that the parameters which maximize the likelihood function are the same as those given in the lecture notes: i.e., that

$$\phi_{j|y=0} = \frac{\sum_{i=1}^m 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}}$$
$$\phi_{j|y=1} = \frac{\sum_{i=1}^m 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}}$$
$$\phi_y = \frac{\sum_{i=1}^m 1\{y^{(i)} = 1\}}{m}$$

Answers:

$$\nabla_{\phi_{j|y=0}} l(\varphi) = \sum_{i=1}^m \left(\frac{x_j^{(i)}}{\phi_{j|y=0}} 1\{y^{(i)} = 0\} - \frac{1 - x_j^{(i)}}{1 - \phi_{j|y=0}} 1\{y^{(i)} = 0\} \right) \tag{29}$$

$$\Rightarrow \nabla_{\phi_{j|y=0}} l(\varphi) = 0, \sum_{i=1}^m \left(\frac{x_j^{(i)}}{\phi_{j|y=0}} - \frac{1 - x_j^{(i)}}{1 - \phi_{j|y=0}} \right) 1\{y^{(i)} = 0\} = 0 \tag{30}$$

$$\Rightarrow \sum_{i=1}^m \left(x_j^{(i)}(1 - \phi_{j|y=0}) - (1 - x_j^{(i)})\phi_{j|y=0} \right) 1\{y^{(i)} = 0\} = 0 \tag{31}$$

$$\Rightarrow \sum_{i=1}^m \left(x_j^{(i)} - \phi_{j|y=0} \right) 1\{y^{(i)} = 0\} = 0 \tag{32}$$

$$\Rightarrow \phi_{j|y=0} = \frac{\sum_{i=1}^m x_j^{(i)} 1\{y^{(i)} = 0\}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}} \tag{33}$$

$$\Rightarrow \phi_{j|y=0} = \frac{\sum_{i=1}^m 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}} \tag{34}$$

With $\nabla_{\phi_{j|y=1}} l(\varphi) = 0$ we have the same result

$$\phi_{j|y=1} = \frac{\sum_{i=1}^m 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}} \tag{35}$$

With $\nabla_{\phi_y} l(\varphi) = 0$ we have:

$$\nabla_{\phi_y} l(\varphi) = \sum_{i=1}^m \left(\frac{y^{(i)}}{\phi_y} - \frac{1 - y^{(i)}}{1 - \phi_y} \right) \tag{36}$$

$$\sum_{i=1}^m \left(y^{(i)}(1 - \phi_y) - (1 - y^{(i)})\phi_y \right) = 0 \tag{37}$$

$$\sum_{i=1}^m \left(y^{(i)} - \phi_y \right) = 0 \tag{38}$$

$$\sum_{i=1}^m y^{(i)} - m\phi_y = 0 \tag{39}$$

$$\phi_y = \frac{\sum_{i=1}^m y^{(i)}}{m} = \frac{\sum_{i=1}^m 1\{y^{(i)} = 1\}}{m} \tag{40}$$

(c) Consider making a prediction on some new data point x using the most likely class estimate generated by the naive Bayes algorithm. Show that the hypothesis returnd by naive Bayes is a linear classifier - i.e., if $p(y = 0|x)$ and $p(y = 1|x)$ are the class probabilities return by naive Bayes, show that there exists some $\theta \in \Re^{n+1}$ such that

$$p(y = 1|x) \geq p(y = 0|x) \text{ if and only if } \theta^T \begin{bmatrix} 1 \\ x \end{bmatrix} \geq 0$$

(Assume θ_0 is an intercept terms.)

Answers:

$$
\begin{aligned}
& p(y=1|x) \geq p(y=0|x) && (41)\\
\iff & \frac{p(y=1|x)}{p(y=0|x)} \geq 1 && (42)\\
\iff & \frac{\left(\prod_{j=1}^{n} p(x_j|y=1)\right) p(y=1)}{\left(\prod_{j=1}^{n} p(x_j|y=0)\right) p(y=0)} \geq 1 && (43)\\
\iff & \frac{\left(\prod_{j=1}^{n} (\phi_{j|y=1})^{x_j} (1-\phi_{j|y=1})^{1-x_j}\right) \phi_y}{\left(\prod_{j=0}^{n} (\phi_{j|y=0})^{x_j} (1-\phi_{j|y=0})^{1-x_j}\right) (1-\phi_y)} \geq 1 && (44)\\
\iff & \sum_{j=1}^{n} \left(x_j \log\left(\frac{\phi_{j|y=1}}{\phi_{j|y=0}}\right) + (1-x_j)\log\left(\frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}}\right)\right) + \log\left(\frac{\phi_y}{1-\phi_y}\right) \geq 0 && (45)\\
\iff & \sum_{j=1}^{n} x_j \log\left(\frac{(\phi_{j|y=1})(1-\phi_{j|y=0})}{(\phi_{j|y=0})(1-\phi_{j|y=1})}\right) + \sum_{j=1}^{n} \log\left(\frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}}\right) + \log\left(\frac{\phi_y}{1-\phi_y}\right) \geq 0 && (46)\\
\iff & \theta^T \begin{bmatrix} 1 \\ x \end{bmatrix} \geq 0, && (47)
\end{aligned}
$$

where

$$
\begin{aligned}
\theta_0 &= \sum_{j=1}^{n} \log\left(\frac{1-\phi_{j|y=1}}{1-\phi_{j|y=0}}\right) + \log\left(\frac{\phi_y}{1-\phi_y}\right)\\
\theta_j &= \log\left(\frac{(\phi_{j|y=1})(1-\phi_{j|y=0})}{(\phi_{j|y=0})(1-\phi_{j|y=1})}\right), j = 1, \ldots, n
\end{aligned}
$$