867002105_HW2

Charles Dotson

September 13, 2022

Contents

1 Let the training examples be $x^{(1)} = [1, 0]^T$, $x^{(2)} = [1, 1]^T$, $y^{(1)} = 10$, $y^{(2)} = -10$. Let the parameters of the linear regression model be $\theta = [-1, 3]^T$. Calculate the likelihood of the training data under the model two steps: 1) write down the general likelihood equation for linear regression (assuming $\sigma = 1$); 2) plug in the data and parameters to calculate the likelihood (no need to get the value but express your results using exp without the x, y, and θ symbols).

The genral likelihood equation is:

$$L(\theta; \{x^{(i)}, y^{(i)}\}_{i=1}^m) = \prod_{i=1}^m Pr(y^{(i)}|x^{(i)} : \theta)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2\right\}$$

Plugging in our training data and assumptions we arrive at:

$$L(\theta; \{x^{(i)}, y^{(i)}\}_{i=1}^m) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left\{-\left[\left(10 - [-1, 3]\begin{bmatrix}1\\0\end{bmatrix}\right)^2 + \left(-10 - [-1, 3]\begin{bmatrix}1\\1\end{bmatrix}\right)^2\right]\right\}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\{-265\}$$

2 The likelihood function of linear regression in the lecture note assumes that the variance σ^2 is the same for each training example. Now assume that the i -th training example has a specific variance σ_i^2, where the variances for two different examples can be different. Prove that the likelihood of the i training example goes to 0 as $\sigma_i \to \infty$.

We will start by restating the likelihood formulation:

$$L(\theta; \{x^{(i)}, y^{(i)}\}_{i=1}^{m}) = \prod_{i=1}^{m} Pr(y^{(i)}|x^{(i)} : \theta)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{m} \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2\right\}$$

Where the product is due to the independence of training examples. Furthere more, adding Gaussian noise to the error terms in the linear approximation where the following is true,

$$(1) \qquad \epsilon^{(i)} \sim N(0, \sigma^2)$$

$$(2) \qquad y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

$$(3) \qquad Pr(\epsilon^{(i)}; 0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right\}$$

$$(4) \qquad Pr(y^{(i)}|x^{(i)} : \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma^2}\right\}$$

Making one small adjustment to equation four for the purposes of σ^2 being different for all training examples:

$$Pr(y^{(i)}|x^{(i)} : \theta) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma_i^2}\right\}$$

Thus we can see the following:

$$L(\theta; \{x^{(i)}, y^{(i)}\}_{i=1}^{m}) = \prod_{i=1}^{m} Pr(y^{(i)}|x^{(i)} : \theta)$$

$$= (Pr(y^{(1)}|x^{(1)} : \theta))(Pr(y^{(2)}|x^{(2)} : \theta)) \; ... \; (Pr(y^{(m)}|x^{(m)} : \theta))$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(y^{(1)} - \theta^T x^{(1)})^2}{\sigma_1^2}\right\}\right) \left(\frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(y^{(2)} - \theta^T x^{(2)})^2}{\sigma_2^2}\right\}\right) \; ... \; \left(\frac{1}{\sqrt{2\pi}\sigma_m} \exp\left\{-\frac{(y^{(m)} - \theta^T x^{(m)})^2}{\sigma_m^2}\right\}\right)$$

Thus for one single likelihood, if $\sigma_i \to \infty$, we can see that $\frac{1}{\sqrt{2\pi}\sigma_i} \to \frac{1}{\infty} \to 0$ and thus the likelihood of the trainging example $\to 0$ due to multiplicative law of 0. From the above formulation we can also see that the general likelihood will also approach 0 if any $\sigma_i \to \infty$ $\blacksquare$

3 Prove that $\frac{\partial}{\partial z}\log(\sigma(-z)) = -\sigma(z)$

We state that

$$\frac{\partial}{\partial z}\log(\sigma(-z)) = -\sigma(z)$$

We state the following truth: $1-\sigma(z) = \sigma(-z)$ thus, $-\sigma(z) = \sigma(-z) - 1$

Following this, we derive the following:

$$\frac{\partial}{\partial z}\log(\sigma(-z)) = -\sigma(z)$$

$$\frac{1}{\sigma(-z)}\left(\frac{\partial}{\partial z}\sigma(-z)\right) = \sigma(-z) - 1 \qquad \text{Derivative of } \log(g(x))$$

$$e^z + 1\left(\frac{\partial}{\partial z}\frac{1}{e^z+1}\right) = \frac{1}{e^z+1} - 1$$

$$e^z + 1\left(-\frac{\frac{\partial}{\partial z}[e^z+1]}{(e^z+1)^2}\right) = \frac{1}{e^z+1} - \frac{e^z+1}{e^z+1} \qquad \text{Reciprocal Rule}$$

$$e^z + 1\left(-\frac{e^z}{(e^z+1)^2}\right) = -\frac{e^z}{e^z+1} \qquad \text{Derivative of } e^x$$

$$-\frac{e^z}{e^z+1} = -\frac{e^z}{e^z+1} \qquad \blacksquare$$

4 Let the training examples be $x^{(1)} = [1,0]^T$, $x^{(2)} = [1,1]^T$, $y^{(1)} = 1$, $y^{(2)} = 0$. Evaluate the log-likelihood of a logistic regression with parameter $\theta = [-1,3]^T$

$$\log L(\theta) = \sum_{i=1}^{m}\{y^{(i)}\log\sigma(z^{(i)}) + (1-y^{(i)})\log\sigma(-z^{(i)})\}$$

Where $\sigma(z) = \frac{1}{1+e^{-z}}$ and $z^{(i)} = \theta^T x^{(i)}$. Thus,

$$z^1 = [-1,3]\begin{bmatrix}1\\0\end{bmatrix} = -1$$

$$z^2 = [-1,3]\begin{bmatrix}1\\1\end{bmatrix} = 2$$

Thus,

$$\sigma(z^{(1)}) = \frac{1}{1+e^{-z^{(1)}}} = \frac{1}{1+e^{1}}$$

$$\sigma(-z^{(1)}) = \frac{1}{1+e^{z^{(1)}}} = \frac{1}{1+e^{-1}}$$

$$\sigma(z^{(2)}) = \frac{1}{1+e^{-z^{(2)}}} = \frac{1}{1+e^{-2}}$$

$$\sigma(-z^{(2)}) = \frac{1}{1+e^{z^{(2)}}} = \frac{1}{1+e^{2}}$$

Inputing these values,

$$\log L(\theta) = \sum_{i=1}^{m} \{y^{(i)} \log \sigma(z^{(i)}) + (1-y^{(i)}) \log \sigma(-z^{(i)})\}$$

$$= \left[y^{(1)} \log \sigma(z^{(1)}) + (1-y^{(1)}) \log \sigma(-z^{(1)})\right] + \left[(y^{(2)} \log \sigma(z^{(2)}) + (1-y^{(2)}) \log \sigma(-z^{(2)})\right]$$

$$= \log \frac{1}{1+e^{1}} + \log \frac{1}{1+e^{2}}$$

$$= \log \left[\left(\frac{1}{1+e^{1}}\right)\left(\frac{1}{1+e^{2}}\right)\right]$$

$$= \log \frac{1}{1+e^{3}+e^{2}+e^{1}}$$

$$= -3.4402$$

5 **(Graduate only) Newton method for multi-class logistic regression requires the Hessian matrix that contains second-order derivatives. Let $z_j = \theta_j^T x$. Derive the second-order partial derivative of the log of the softmax output $\phi_j(x) = \frac{\exp(z_j)}{\sum_{l=1}^k \exp(z_l)}$ for class j. Formally, prove that $\frac{\partial^2 log\phi_j(x)}{\partial\theta_j \partial\theta_i} = -\phi_j(\delta_{ij} - \phi_i)XX^T \in \mathbb{R}^{n\mathbf{x}n}$, where n is the dimension of x, and $\delta_{ij} = \mathbb{1}[i = j] = 1$ if $i = j$ and 0 otherwise. (Hints: start from the gradient of $\log \phi_j$ w.r.t θ_j.)**

$$\frac{\partial}{\partial\theta_j}\log(\phi_j(x)) = \frac{\partial}{\partial\theta_j}\log\left[\frac{\exp(\theta_j^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}\right]$$

$$= \frac{\partial}{\partial\theta_j}\left[\log(\exp(\theta_j^T X)) - \log(\sum_{l=1}^k \exp(\theta_l^T X))\right]$$

$$= \frac{\partial}{\partial\theta_j}\log(\exp(\theta_j^T X)) - \frac{\partial}{\partial\theta_j}\log(\sum_{l=1}^k \exp(\theta_l^T X))$$

$$= \frac{\partial}{\partial\theta_j}\theta_j^T X - \frac{1}{\sum_{l=1}^k \exp(\theta_l^T X)}\frac{\partial}{\partial\theta_j}\sum_{l=1}^k \exp(\theta_l^T X)$$

$$= \mathbb{1}X - \frac{\exp(\theta_j^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}\frac{\partial}{\partial\theta_j}\theta_j^T X$$

$$= [\mathbb{1} - \phi_j]X$$

$$\frac{\partial^2}{\partial\theta_j\partial\theta_i}\log(\phi_j(x)) = \frac{\partial}{\partial\theta_i}[\mathbb{1}-\phi_j]X$$

$$= \frac{\partial}{\partial\theta_i}\left[\mathbb{1}-\frac{\exp(\theta_j^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}\right]X$$

$$= \left[\frac{\partial}{\partial\theta_i}\mathbb{1}-\frac{\partial}{\partial\theta_i}\frac{\exp(\theta_j^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}\right]X$$

$$= \left[-\left(\frac{\frac{\partial}{\partial\theta_i}\exp(\theta_j^T X)\sum_{l=1}^k \exp(\theta_l^T X)-\frac{\partial}{\partial\theta_i}\sum_{l=1}^k \exp(\theta_l^T X)\exp(\theta_j^T X)}{\left(\sum_{l=1}^k \exp(\theta_l^T X)\right)^2}\right)\right]X$$

$$= \left[-\left(\frac{\delta_{ij}X\exp(\theta_j^T X)\sum_{l=1}^k \exp(\theta_l^T X)-\exp(\theta_i^T X)X\exp(\theta_j^T X)}{\left(\sum_{l=1}^k \exp(\theta_l^T X)\right)^2}\right)\right]X$$

$$= \left[-\left(\delta_{ij}X\frac{\exp(\theta_j^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}\frac{\sum_{l=1}^k \exp(\theta_l^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}-X\frac{\exp(\theta_j^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}\frac{\exp(\theta_i^T X)}{\sum_{l=1}^k \exp(\theta_l^T X)}\right)\right]X$$

$$= (-\delta_{ij}\phi_j X-\phi_j\phi_i X)X$$

$$= -\phi_j(\delta_{ij}-\phi_i)XX^T\blacksquare$$