COMP4434 Big Data Analytics

Assignment 1 solution

PolyU, Hong Kong

Instructor: HUANG Xiao

Problem 1 (3 points)
Assume that we are training a multivariate linear regression model using three instances as follows.

instance id:	(x_1, x_2) aka $\mathbf{x}$,	target value y
instance 1:	$(1, -0.5)$,	1
instance 2:	$(0.5, 1.5)$,	0
instance 3:	$(2, -0.5)$,	1

The cost function is based on mean squared error as follows.

$$\min_{\theta_0,\theta_1,\theta_2} \frac{1}{3}\sum_{i=1}^{3}(h_\theta(\mathbf{x}^{(i)}) - y^{(i)})^2.$$

Assume that we have initialized

$$\mathbf{w} = [\theta_0, \theta_1, \theta_2] = [0.5, 0, 1].$$

You apply gradient descent algorithm to compute a $\mathbf{w}$ that minimizes the cost function w.r.t. the three given instances. Assume that in the first iteration, we set learning rate $\alpha = 0.5$. In the second iteration, we set $\alpha = 0.4$. Please concisely show how $\mathbf{w}$ is updated in the first and second iterations.

Answer:
As there are two features, the linear regression model would be defined as follows,

$$h_\theta\left(x_1, x_2\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2.$$

The cost function would be defined as follows,

$$J\left(\theta_0, \theta_1, \theta_2\right) = \min_{\theta_0,\theta_1,\theta_2} \frac{1}{3}\sum_{i=1}^{3}\left(h_\theta\left(x_1^{(i)}, x_2^{(i)}\right) - y^{(i)}\right)^2.$$

The gradient descent algorithm repeats the following calculations,

$$\theta_0 = \theta_0 - \alpha \cdot \frac{\partial J\left(\theta_0, \theta_1, \theta_2\right)}{\partial \theta_0}.$$
$$\theta_1 = \theta_1 - \alpha \cdot \frac{\partial J\left(\theta_0, \theta_1, \theta_2\right)}{\partial \theta_1}.$$
$$\theta_2 = \theta_2 - \alpha \cdot \frac{\partial J\left(\theta_0, \theta_1, \theta_2\right)}{\partial \theta_2}.$$

By using chain rule, we can simplify $\frac{\partial J(\theta_0,\theta_1,\theta_2)}{\partial\theta_0}$ as,

$$\begin{aligned}\frac{\partial J\left(\theta_0,\theta_1,\theta_2\right)}{\partial\theta_0} &= \frac{\partial}{\partial\theta_0}\left(\frac{1}{3}\sum_{i=1}^{3}\left(h_\theta\left(x_1^{(i)},x_2^{(i)}\right)-y^{(i)}\right)^2\right)\\ &= \frac{\partial}{\partial\theta_0}\left(\frac{1}{3}\sum_{i=1}^{3}\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}-y^{(i)}\right)^2\right)\\ &= \frac{1}{3}\sum_{i=1}^{3}\frac{\partial}{\partial\theta_0}\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}-y^{(i)}\right)^2\\ &= \frac{2}{3}\sum_{i=1}^{3}\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}-y^{(i)}\right).\end{aligned}$$

Similarly, we have,

$$\begin{aligned}\frac{\partial J\left(\theta_0,\theta_1,\theta_2\right)}{\partial\theta_1} &= \frac{\partial}{\partial\theta_1}\left(\frac{1}{3}\sum_{i=1}^{3}\left(h_\theta\left(x_1^{(i)},x_2^{(i)}\right)-y^{(i)}\right)^2\right)\\ &= \frac{2}{3}\sum_{i=1}^{3}\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}-y^{(i)}\right)\cdot x_1^{(i)}.\end{aligned}$$

$$\begin{aligned}\frac{\partial J\left(\theta_0,\theta_1,\theta_2\right)}{\partial\theta_2} &= \frac{\partial}{\partial\theta_2}\left(\frac{1}{3}\sum_{i=1}^{3}\left(h_\theta\left(x_1^{(i)},x_2^{(i)}\right)-y^{(i)}\right)^2\right)\\ &= \frac{2}{3}\sum_{i=1}^{3}\left(\theta_0+\theta_1x_1^{(i)}+\theta_2x_2^{(i)}-y^{(i)}\right)\cdot x_2^{(i)}.\end{aligned}$$

Given the three training instances, in the first iteration, the gradient for θ_0 is
$(2/3)*[(0.5+0*1-0.5*1-1)+(0.5+0*0.5+1.5*1-0)+(0.5+0*2-0.5*1-1)] = 0.$
The gradient for θ_1 is
$(2/3)*[(0.5+0*1-0.5*1-1)+(0.5+0*0.5+1.5*1-0)*0.5+(0.5+0*2-0.5*1-1)*2] = -1.33333.$
The gradient for θ_2 is
$(2/3)*[-0.5*(0.5+0*1-0.5*1-1)+(0.5+0*0.5+1.5*1-0)*1.5-0.5*(0.5+0*2-0.5*1-1)] = 2.66667.$
$\theta_0 = 0.5-0.5*0 = 0.5.$
$\theta_1 = 0+0.5*1.33333 = 0.66665.$
$\theta_2 = 1-0.5*2.66667 = -0.333333.$
In the second iteration, the gradient for θ_0 is
$(2/3)*[(0.5+0.66665*1+0.5*0.33333-1)+(0.5+0.66665*0.5-1.5*0.33333-0)+(0.5+0.66665*2+0.5*0.33333-1)] = 1.11107.$
The gradient for θ_1 is
$(2/3)*[(0.5+0.66665*1+0.5*0.33333-1)*1+(0.5+0.66665*0.5-1.5*0.33333-0)*0.5+(0.5+0.66665*2+0.5*0.33333-1)*2] = 1.66661.$
The gradient for θ_2 is
$(2/3)*[-(0.5+0.66665*1+0.5*0.33333-1)*0.5+(0.5+0.66665*0.5-1.5*0.33333-0)*1.5-(0.5+0.66665*2+0.5*0.33333-1)*0.5] = -0.11110.$
$\theta_0 = 0.5-0.4*1.11107 = 0.05557.$
$\theta_1 = 0.66665-0.4*1.66661 = 0.000006.$
$\theta_2 = -0.333333+0.4*0.11110 = -0.28889.$

Problem 2 (1 points)
Assume that there are 1,200 documents in total. Among them, 800 documents are related to big data analysis. You build a model to identify documents related to big data analysis. As a result, your model returns 700 documents, but only 500 of them are relevant to big data analysis. What is the recall of your model? What is the F1 score of your model? Briefly justify your answer.

Answer:
TP = 500. FP = 200. FN = 300. TN = 200. Precision = 500/700. Recall = 500/(500+300).
F1 = $(2 * 500/700 * 500/800)/(500/700 + 500/800) = 0.66667$.

Problem 3 (1 points)
Explain why, in the context of cross-validation, test data and training data are randomly selected from the same dataset, even though it is generally stated that test data should be independent of training data. Provide a detailed explanation that reconciles this apparent contradiction.

Answer:
In k-fold cross-validation, the dataset is partitioned into k equal parts. In each iteration, one fold is designated as the test set, while the remaining k-1 folds are used for training. This process is repeated k times, with each fold serving as the test set once. Although the test and training sets are derived from the same dataset, they remain independent in each iteration. This independence is crucial for assessing the model's performance on unseen data. Cross-validation is particularly advantageous when data is limited, as it maximizes the use of available data for both training and testing, providing a comprehensive evaluation of the model's generalization capabilities.

Problem 4 (3 points)
Assume that we are training a logistic regression classifier using three instances as follows.

instance id:	(x_1, x_2) aka $\mathbf{x}$,	label y
instance 1:	$(1, -0.5)$,	1
instance 2:	$(0.5, 1.5)$,	0
instance 3:	$(2, -0.5)$,	1

The logistic regression classifier is defined as follows.

$$h_\theta(\mathbf{x}) = h_\theta(x_1, x_2) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}.$$

The cost function is based on **logistic loss** as follows.

$$\min_{\theta_0, \theta_1, \theta_2} -\frac{1}{3} \sum_{i=1}^{3} \left[y^{(i)} \log(h_\theta(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(\mathbf{x}^{(i)})) \right].$$

Assume that we have initialized

$$\mathbf{w} = [\theta_0, \theta_1, \theta_2] = [0.5, 0, 1].$$

You apply gradient descent algorithm to compute a $\mathbf{w}$ that minimizes the cost function w.r.t. the three given instances. Assume that in the first iteration, we set learning rate $\alpha = 0.5$. In the second iteration, we set learning rate $\alpha = 0.4$. Please concisely show how $\mathbf{w}$ is updated in the first and second iterations.

Answer:
Based on the chain rule, the equations for gradient descent algorithm are:

$$\theta_0 = \theta_0 - \alpha \cdot \frac{1}{3}\sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)}),$$

$$\theta_1 = \theta_1 - \alpha \cdot \frac{1}{3}\sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)}) \cdot x_1^{(i)},$$

$$\theta_2 = \theta_2 - \alpha \cdot \frac{1}{3}\sum_{i=1}^{3}(h_\theta(x^{(i)}) - y^{(i)}) \cdot x_2^{(i)}.$$

In the first iteration, we have,

$$[h_\theta(\mathbf{x}^{(i)})] = [\frac{1}{1+e^0}, \frac{1}{1+e^{-2}}, \frac{1}{1+e^0}] = [0.5, 0.8808, 0.5].$$

The corresponding labels are $[y^{(i)}] = [1, 0, 1]$.

$$\theta_0 = 0.5 - 0.5 \cdot \frac{1}{3}[(0.5-1) + (0.8808-0) + (0.5-1)] = 0.51986,$$
$$\theta_1 = 0 - 0.5 \cdot \frac{1}{3}[(0.5-1)\times 1 + (0.8808-0)\times 0.5 + (0.5-1)\times 2] = -0.5/3 * [-1.0596] = 0.1766,$$
$$\theta_2 = 1 - 0.5 \cdot \frac{1}{3}[(0.5-1)\times(-0.5) + (0.8808-0)\times 1.5 + (0.5-1)\times(-0.5)] = 0.69646.$$

Thus we update the value of $\mathbf{w}$ to $[0.520, 0.177, 0.696]$.
In the second iteration, we have

$$\begin{aligned}[h_\theta(\mathbf{x}^{(i)})] =& [\frac{1}{1+e^{-(0.51986+0.1766*1+0.69646*(-0.5))}}, \frac{1}{1+e^{-(0.51986+0.1766*0.5+0.69646*1.5)}}, \\ & \frac{1}{1+e^{-(0.51986+0.1766*2+0.69646*(-0.5))}}], \\ =& [0.586188, 0.83927, 0.62827].\end{aligned}$$

$$\theta_0 = 0.51986 - 0.4 \cdot \frac{1}{3}[(0.586188-1) + (0.83927-0) + (0.62827-1)] = 0.51269626.$$
$$\theta_1 = 0.1766 - 0.4 \cdot \frac{1}{3}[(0.586188-1)\times 1 + (0.83927-0)\times 0.5 + (0.62827-1)\times 2] = 0.274516.$$
$$\begin{aligned}\theta_2 =& 0.69646 - 0.4 \cdot \frac{1}{3}[(0.586188-1)\times(-0.5) + (0.83927-0)\times 1.5 + (0.62827-1)\times(-0.5)] \\ &= 0.476236.\end{aligned}$$

Then we update the value of $\mathbf{w}$ to $[0.513, 0.275, 0.476]$.