Q1

Assume that we are training a multivariate linear regression model using three instances as follows.

instance id (i)	(x1, x2) aka x (j)	target value
instance 1:	(1, 1.5)	2
instance 2:	(−0.5, −1)	0
instance 3:	(2, 0.5)	1

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

Cost Function:

$$J(\theta_0, \theta_1, \theta_2) = \min_{\theta_0,\theta_1,\theta_2} \frac{1}{6}\sum_{i=1}^{3}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Model:

$$h_\theta\left(x^{(i)}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Gradient Descent Algorithm:

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

Error

$$\epsilon = h_\theta\left(x^{(i)}\right) - y^{(i)}$$

Partial Derivative

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(\epsilon_i)$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(\epsilon_i){x_j}^{(i)}$$

Assume that we have initialized w = [θ0, θ1, θ2] = [0, 0.5, 1].

You apply gradient descent algorithm to compute a w that minimizes the cost function w.r.t. the three given instances. Assume that in the first iteration, we set learning rate α = 0.6. In the second iteration, we set learning rate α = 0.4. Please concisely show how w is updated in the first and second iterations.

For each iteration:

1. Substitute in x1 and x2 into(2)for each instance

2. Compute Errors for each instance using $\epsilon \ = \ h_\theta\left(x^{(i)}\right) - y^{(i)}$
3. Compute partial derivative for each step ($x_j{}^{(i)}$ is the j-th feature of the i-th instance
4. Compute weights

1st iteration

Computing Predictions

$$h_\theta\left(x^{(i)}\right) = \ \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h_\theta(x^1) = \ 0 \ + \ 0.5 \ \cdot \ 1 + 1 \cdot 1.5 = 2$$

$$h_\theta(x^2) = \ 0 \ + \ 0.5 \ \cdot \ \text{-}0.5 + 1 \cdot -1 \ = \ -1.25$$

$$h_\theta(x^3) = \ 0 \ + \ 0.5 \ \cdot \ 2 \ + 1 \cdot 0.5 \ = \ 1.5$$

Computing Errors

$$\epsilon_i \ = \ h_\theta\left(x^{(i)}\right) - y^{(i)}$$

$$\epsilon_1 \ = \ 2 \ - \ 2 \ = \ 0$$

$$\epsilon_2 \ = \ -1.25 \ - \ 0 \ = \ -1.25$$

$$\epsilon_3 \ = \ 1.5 \ - \ 1 \ = \ 0.5$$

Computing derivatives

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_0} \ = \ \frac{1}{3} \sum_{i=1}^{3} (\ \epsilon_i)$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_0} \ = \ \frac{1}{3}(0 \ - 1.25 \ + \ 0.5) \ = \ -0.25$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_j} \ = \ \frac{1}{m} \sum_{i=1}^{m} (\ \epsilon_i) x_j{}^{(i)}$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_1} \ = \ \frac{1}{3}(0 \ \cdot \ 1 \ - 1.25 \cdot \ -0.5 \ + \ 0.5 \cdot \ 2) \ = \ 0.625$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_2} \ = \ \frac{1}{3}(0 \ \cdot \ 1.5 \ - 1.25 \cdot \ -1 \ + \ 0.5 \cdot \ 0.5) \ = \ 0.5$$

Compute Weights

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

$$\theta_1 = \ 0 \ - \ 0.6 \ \cdot \ -0.25 \ = \ 0.15$$

$$\theta_2 = \ 0.5 \ - \ 0.6 \ \cdot \ 0.625 \ = \ 0.125$$

$$\theta_3 = 1 - 0.6 \cdot 0.5 = 0.7$$

$$w_1 = [0.15 \quad 0.125 \quad 0.7]$$

2nd iteration

Computing predictions

$$h_\theta\left(x^{(i)}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h_\theta(x^1) = 0.15 + 0.125 \cdot 1 + 0.7 \cdot 1.5 = 1.325$$

$$h_\theta(x^2) = 0.15 + 0.125 \cdot -0.5 + 0.7 \cdot -1 = -0.6625$$

$$h_\theta(x^3) = 0.15 + 0.125 \cdot 2 + 0.7 \cdot 0.5 = 0.825$$

Computing errors

$$\epsilon = h_\theta\left(x^{(i)}\right) - y^{(i)}$$

$$\epsilon_1 = 1.325 - 2 = -0.675$$

$$\epsilon_2 = -0.6625 - 0 = -0.6625$$

$$\epsilon_3 = 0.825 - 1 = -0.175$$

Computing Derivatives

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_0} = \frac{1}{3}\sum_{i=1}^{3}(\epsilon_i)$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_0} = \frac{1}{3}(-0.675 - 0.6625 - 0.175) = -0.5042$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_1} = \frac{1}{3}(-0.675 \cdot 1 - 0.6625 \cdot -0.5 - 0.175 \cdot 2) = -0.1458$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_2} = \frac{1}{3}(-0.675 \cdot 1.5 - 0.6625 \cdot -1 - 0.175 \cdot 0.5) = -0.1458$$

Computing Weights

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

$$\theta_1 = 0.15 - 0.4 \cdot -0.25 = 0.3517$$

$$\theta_2 = 0.125 - 0.4 \cdot 0.625 = 0.1833$$

$$\theta_3 = 0.7 - 0.4 \cdot 0.5 = 0.7583$$

$$w_2 = [0.3517 \quad 0.1833 \quad 0.7583]$$

Q2

Assume that there are 1,500 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 900 documents, but only 600 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.

$$Recall = \frac{\text{Number of relevant documents retrieved}}{\text{Number of total of related documents}} = \frac{600}{800} = 0.75$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$Precision = \frac{\text{Number of relevant documents retrieve}}{Number\ of\ documents\ retrieved} = \frac{2}{3}$$

$$F1 = 2 \cdot \frac{\frac{2}{3} \cdot 0.75}{\frac{2}{3} + 0.75} = 0.70588235294\ldots \approx 0.7059$$

The recall of the model is 0.75 and the F1 score is approximately 0.7069. The recall is the ratio of number of relevant documents retrieved to the number of total related documents, which highlights how well the model captures all relative documents. The F1 score is the harmonic mean between the precision and recall, representing both equally. This score implies that the model is pretty good at identifying documents, but there is room for improvement.

Q3

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.

Cross validation divides available data into multiple sets, and for each iteration, the model uses one of the sets for validation for each iteration and training the model with the remaining sets. Then, the results can be averaged out to provide a single estimation.

This method, while using test data and training data from the same dataset, is reconciled in many ways:

Firstly, during each iteration of training, the test set is excluded from training, so they are independent during the same fold. As such, this prevents data overlapping during scaling. In addition, data points would only appear in the test set once throughout all iterations of training, preventing overfitting. This method is like independent sampling, obtaining different samples from each dataset so different features that the dataset has can be represented well when predicting using new data. Averaging out the result allows the features that represent the dataset fully to affect the model the most, assuming the dataset is representative, which gives it a more accurate performance on unseen data.

In addition, while cross-validation uses the same dataset, we can use a final, separated testing dataset to test the data again, which can evaluate the model.

Q4

Assume that we are training a **logistic regression classifier** using three instances as follows

instance id (i)	(x1, x2) aka x (j)	target value
instance 1:	(1, 1.5)	1
instance 2:	(−0.5, −1)	0
instance 3:	(2, 0.5)	1

The **logistic regression classifier** is defined as follows:

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

The cost function is defined as follows (gradient)

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

Assume that we have initialised

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

You apply gradient descent algorithm to compute a w that minimizes the cost function w.r.t. the three given instances. Assume that in the first iteration, we set learning rate α = 0.6. In the second iteration, we set learning rate α = 0.4. Please concisely show how w is updated in the first and second iterations.

1st Iteration:

Computing Predictions

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

$$h_0(x) = \frac{1}{1 + e^{-(0+0.5\cdot 1+1\cdot 1.5)}} = \frac{1}{1 + e^{-2}} \approx 0.8808$$

$$h_1(x) = \frac{1}{1 + e^{-(0+0.5\cdot -0.5+1\cdot -1)}} = \frac{1}{1 + e^{1.25}} \approx 0.2227$$

$$h_2(x) = \frac{1}{1 + e^{-(0+0.5\cdot 2+1\cdot 0.5)}} = \frac{1}{1 + e^{-1.5}} \approx 0.8176$$

Compute Gradient

$$-\frac{1}{3}\sum_{i=1}^{3}\left[y^{(i)}log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right]$$

$$\epsilon_i = y^{(i)} - h_\theta\left(x^{(i)}\right)$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_j} = -\frac{1}{3}\sum_{i=1}^{3}(\epsilon_i)x_j{}^{(i)}$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_0} = -\frac{1}{3}[(1 - 0.8808) + (0 - 0.2227) + (1 - 0.8176)]$$

$$\approx -\frac{1}{3}(0.1192 - 0.2227 + 0.1824) \approx 0.0265$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_1} = -\frac{1}{3}[(1 - 0.8808)\cdot 1 + (0 - 0.2227)\cdot(-0.5) + (1 - 0.8176)\cdot 2]$$

$$\approx -\frac{1}{3}(0.1192 + 0.1114 + 0.3648) \approx -0.1985$$

$$\frac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_2} = -\frac{1}{3}[(1 - 0.8808)\cdot 1.5 + (0 - 0.2227)\cdot -1 + (1 - 0.8176)\cdot 2]$$

$$\approx -\frac{1}{3}(0.1788 + 0.2227 + 0.0912) \approx -0.1642$$

Update Weights

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

$$\theta_0 = -0 - 0.6 \cdot 0.0265 \approx -0.0159$$

$$\theta_1 = 0.5 - 0.6 \cdot -0.1985 = 0.6191$$

$$\theta_2 = 1 - 0.6 \cdot -0.1642 = 1.0985$$

2nd Iteration

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

$$h_0(x) = \frac{1}{1 + e^{-(-0.0159+0.6191 \cdot 1+1.0985\cdot 1.5)}} = \frac{1}{1 + e^{-1.832}} \approx 0.8621$$

$$h_1(x) = \frac{1}{1 + e^{-(-0.0159+0.6191 \cdot -0.5+1.0985\cdot -1)}} = \frac{1}{1 + e^{1.433}} \approx 0.1925$$

$$h_2(x) = \frac{1}{1 + e^{-(-0.0159+0.6191 \cdot 2+1.0985\cdot 0.5)}} = \frac{1}{1 + e^{-1.433}} \approx 0.8075$$

Compute Gradient

$$-\epsilon_i = y^{(i)} - h_\theta\left(x^{(i)}\right)$$

$$\frac{\partial J(\theta_0,\theta_1,\theta_2)}{\partial \theta_j} = -\frac{1}{3}\sum_{i=1}^{3}(\epsilon_i)x_j{}^{(i)}$$

$$\frac{\partial J(\theta_0,\theta_1,\theta_2)}{\partial \theta_0} = -\frac{1}{3}[(1-0.8621)+(0-0.1925)+(1-0.8075)]$$

$$\approx -\frac{1}{3}(0.1379-0.1925+0.1925) \approx -0.0457$$

$$\frac{\partial J(\theta_0,\theta_1,\theta_2)}{\partial \theta_1} = -\frac{1}{3}[(1-0.8621)\cdot 1+(0-0.1925)\cdot(-0.5)+(1-0.8075)\cdot 2]$$

$$\approx -\frac{1}{3}(0.1379+0.0963+0.385) \approx -0.2064$$

$$\frac{\partial J(\theta_0,\theta_1,\theta_2)}{\partial \theta_2} = -\frac{1}{3}[(1-0.8621)\cdot 1.5+(0-0.1925)\cdot(-1)+(1-0.8075)\cdot 0.5]$$

$$\approx -\frac{1}{3}(0.2069+0.1925+0.0963) \approx -0.1652$$

Update Weights

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

$$\theta_1 = -0.0159 - 0.4\cdot(-0.0457) \approx 0.0024$$

$$\theta_2 = 0.6191 - 0.4\cdot(-0.2064) \approx 0.7017$$

$$\theta_3 = 1.0985 - 0.4\cdot(-0.1652) \approx 1.164$$

$$w \approx [0.0024, 0.7017, 1.1646]$$