COMP4434 Big Data Analytics

Assignment 1

PolyU, Hong Kong

Instructor: HUANG Xiao
Logistics: You should submit your solutions through Learn@PolyU (Blackboard). The deadline is 16 Feb Sunday, 11:55 PM. I will not accept submission from any other channels except Learn@PolyU. These are the best exercises that can help you be well prepared for the exam. Therefore, please work independently.

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:	$(1, 1.5)$,	2
instance 2:	$(-0.5, -1)$,	0
instance 3:	$(2, 0.5)$,	1

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

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

Assume that we have initialized

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

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 $\alpha = 0.6$. 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.

Problem 2 (1 points)
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.

Problem 3 (0.5 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.

Problem 4 (3.5 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:	$(1, 1.5)$,	1
instance 2:	$(-0.5, -1)$,	0
instance 3:	$(2, 0.5)$,	1

The logistic regression classifier is defined as follows.

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

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

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

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 $\alpha = 0.6$. 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.