UNIVERSITI MALAYA

WQD7006

WQD7006 MACHINE LEARNING FOR DATA SCIENCE

SEMESTER 1, 2023/2024

FINAL EXAM PART 1 (GROUP 1)

Name	**Matric Number**
Kar Hong Sam	S2191926

LECTURER'S NAME : DR. MUHAMMAD SHAHREEZA SAFIRUZ BIN KASSIM

DATE OF SUBMISSION : 13 JANUARY 2024

MARKS :

Question 1

1 a) $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

b) Example 1 : $x_1 = 3$, $x_2 = 5$, $[\theta_0 = 1, \theta_1 = 3, \theta_2 = 2]$

$h_\theta(x) = 1 + 3(3) + 2(5)$

$= 20$ *

Example 2 : $x_1 = 7$, $x_2 = 2$

$h_\theta(x) = 1 + 3(7) + 2(2)$

$= 26$ *

Example 3 : $x_1 = 4$, $x_2 = 1$

$h_\theta(x) = 1 + 3(4) + 2(1)$

$= 15$ *

c) MSE . training $y = [18, 25]$, training $\hat{y} = [20, 26]$

training MSE $= \frac{1}{2}[(18-20)^2 + (25-26)^2]$

$= \frac{1}{2}[4 + 4]$

$= 4$ *

testing $y = [18]$, testing $\hat{y} = 15$

testing MSE $= (18-15)^2$

$= 9$ *

d) i) Logistics Regression (LR). Because Logistics Regression is a binary classifier that predict values between 1 and 0. Linear Regression only for numerical values prediction.

ii) Logistics Regression (LR) form a linear separable surface. Since the data points in figure above is linear separable hence it can be perfectly classified.

iii) No. Since the data point is not linearly separable, hence no logistics regression classifier can perfectly classify it.

Question 2:

a. Which classifier has the best generalization performance, i.e., most likely would perform the best when applied to unseen data? Why?

Classifier C, because the testing error rate is the lowest compared to other classifiers indicates well performance in unseen data.

b. Which classifier is underfitting the most? Why?

Classifier A, because it has high training error rate and high testing error compared to other classifiers.

c. Which classifier is overfitting the most? Why?

Classifier B, because it has low training error rate but high testing error rate compared to other classifiers.

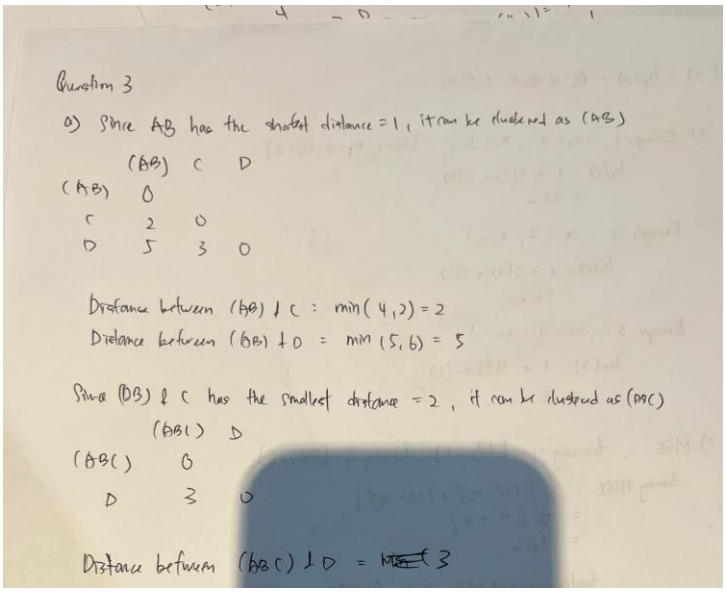
d. Given a dataset, assume we are using linear regression to make a prediction. The dataset is split randomly into training and testing. We increase the training set size gradually. What happens to the mean training and mean testing errors, that is, increase or decrease?

Due to the law of large sample size, the mean for training and testing error will tend to decrease. With more samples to learn the features of data, the model has a better chance to generalize the pattern and identify the relationships in the data. Hence, it will result in decreasing the mean of training and testing errors.

Question 3:

Please this colab link to access the dendogram for both clustering:
https://colab.research.google.com/drive/17ROC3SJLITpkvE2YYhU8HSTRJc8rO5JC?usp=sharing

a. Use Single Linkage



b. Use Complete Linkage