CSCI 183
Winter 2024
Quiz 2
2/23/2024
Time Limit: 40 Minutes

Name: Arman Niri
SCU ID: _____

This exam contains 4 pages (including this cover page) and 7 questions.
Total of points is 15.

Grade Table (for teacher use only)

Question	Points	Score
1	4	
2	1	
3	2	
4	2	
5	2	
6	2	
7	2	
Total:	15	

1. (4 points) For the following statements, state whether they are True/False. **If False, justify your answer**

(a) MAE is more sensitive to outliers than MSE

$y_i - \hat{y} = 4$

False, MSE gives a larger penalization to big prediction error by squaring it while MAE treats all errors the same

MSE = 16
MAE = 4

(b) A high F1 score always indicates that both precision and recall are high.

True

$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ weighted AVG of both

(c) KNN is computationally expensive during the training phase compared to the testing phase.

False, expensive at test time as it needs to sort through all traing data to find the nearest neighbors. Distance computations with N traing points (D features each)

Store data from training in memory so it can be true?

(d) Having missing values in an instance in a dataset does not impact the KNN algorithm.

~~False it calculates the distance of the points it there is no second point~~ True, the algo should predict missing values

(e) Overfitting is generally associated with lower performance on the training dataset compared to the testing dataset.

→ this is underfitting

False, means learned the data too well and picked up noise memorization not learning thus overperform on not on traing

2. (1 point) When looking at true and predicted labels, it is observed that count(1,1) = 25, count(0,0) = 10, count(1,0) = 10 and count(0,1) = 5, where count(true,pred) indicates the count of the different true-pred label combination.

What will be precision and recall of this system?

Pred \ Actual	T	F
T	25	5
F	10	10

$\text{Precision} = \frac{TP}{TP+FP} = \frac{25}{30}$ $\text{Recall} = \frac{TP}{TP+FN} = \frac{25}{25+10} = \frac{25}{35}$

3. (2 points) Chandler is developing a fraud detection system (don't ask why) for online transactions. During training he gets a 5% error, but during testing he gets a 35% error. What strategy should Chandler employ to improve his model's ability to generalize to new fraud patterns, and why would this help?

~~He should use cross validation to be able to see more data in the train phase and hopefully better his testing data as well.~~

The other option is a new model as he is underfitting. this is since the training data isn't getting a solid pattern thus the low accuracy

4. (2 points) Joey and Chandler are working on evaluating a (same) regression model using the same test set. Joey uses MAE and Chandler uses MSE. Joey gets $MAE > MSE$. Is this possible? If so why, if not why not? Explain your answer.

yes this is possible. this only happens if the data for when the error is a decimal. this is the effect of squaring in MSE which will cause the error to shrink

Ex $y_i - \hat{y} = .5$ $MSE = .25$ $\rightarrow$ Proven
$MAE = .5$

5. (2 points) Ross (who believes he is that kind of 'Dr.'), develops an algorithm to diagnose a rare disease, with the disease present in only 1% of the dataset. His algorithm achieves 95% precision. Is precision the best metric to evaluate Ross's model? Should he consider other metrics? Explain your answer.

No the best in this case would be an F1 score as we care about Recall and precision due to the nature of the algorithm. Since it is health you don't want FP or FN Precision is only FP. thus we care on recall for FN as well. Fortunately F1 is the weighted average of these two thus, I believe it is the best metric

6. (2 points) Consider the following dataset containing height (in centimeters) and weight (in kilograms) of individuals along with their classified health status:

	Height (cm)	Weight (kg)	Health Status
1	170	65	Healthy
2	160	70	Unhealthy
3	175	80	Unhealthy
4	180	75	Healthy

euclidean distance

$\sqrt{(p_1-p_2)^2+(q_1-q_2)^2}$ given (p_1,q_1) (p_2,q_2)

A new individual with a height of 168 cm and a weight of 65 kg needs to be classified using the 3-NN algorithm. What would be the predicted output? Show the steps.

$d(1,5) = \sqrt{(170-168)^2+(65-65)^2} = 2$ (healthy point)

$d(2,5) = \sqrt{(160-168)^2+(70-65)^2} = 40$ (unhealthy point)

$d(3,5) = \sqrt{(175-168)^2+(80-65)^2} = 105$ (unhealthy point)

$d(4,5) = \sqrt{(180-168)^2+(75-65)^2} = 13.8924$ (healthy point)

the point will be given as healthy as it is closer to the two healthy points compared to unhealthy

7. (2 points) Joey and Chandler are performing K-NN using Euclidean distance on the same dataset (same #features & #instances) to detect Yes(1) or No(0). They want to predict the label for a new (same) test point X. Joey uses 3-NN and Chandler uses 7-NN. They both get very different computational times. Is this possible? Explain your answer.

Yes ~~No~~, since their dataset is the same the # of folds should'nt change the computational time. If the data sizes were different it would since KNN searchs all train data to find the nearest neighbors. Folds effect the split of train and testing however so it could make sense since there will be many more train points to compute.