CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

Worksheet 10

Name (Print):________________________________ Section:______________

1 Review

1. True/False:
   A model that performs well on the training data but poorly on test data is likely overfitting

2. Which of the following strategies can reduce overfitting?

   (a) Increase the complexity of the model
   (b) Add more training data
   (c) Remove regularization

3. Which metric is used to evaluate the accuracy of a regression model?

   (a) F1-Score
   (b) Mean Squared Error (MSE)
   (c) Cosine Similarity
   (d) Confusion Matrix

4. ________________ regression uses L1-regularization, which can shrink some coefficients to zero, effectively performing feature selection.

   ________________ regression uses L2-regularization, which shrinks coefficients toward zero but does not eliminate them entirely.

5. True/False:
   Decision trees can handle both categorical and numerical features.

6. What is a potential disadvantage of decision trees?

   (a) They are hard to interpret
   (b) They require feature scaling
   (c) They are prone to overfitting without pruning

7. Which metric is commonly used to determine the optimal clustering number?

   (a) Silhouette or Elbow Method
   (b) Accuracy
   (c) Precision

(d) MSE

8. ________________ does not require pre-specifying the number of clusters, while ________________ requires specifying the number of clusters upfront.

9. ________________ rescales feature values to a range of [0, 1], while ________________ centers data around zero and scales it to have a standard deviation of 1.

10. The lower the MSE value, the ________________ the model's predictions are to the actual values.

11. Circle Manhattan formula:

$$d_{distance} = \sum_{i=1}^{n} |x_i - y_i|$$

$$d_{distance} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2 Naive Bayes Classifier

12. The Naive Bayes classifier is based on ________________ theorem and assumes that the features are ________________ of each other given the class label.

13. Which of the following is not a common distribution assumption made by Naive Bayes classifiers?

(a) Gaussian Distribution

(b) Multinomial Distribution

(c) Bernoulli Distribution

(d) Poisson Distribution

14. True/False:
The Naive Bayes classifier assumes that all features are equally important and contribute independently to the final classification.

15. What Type? The ________________ Naive Bayes classifier is used for text classification tasks, where the features represent word counts or term frequencies.