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

Worksheet 09

Name (Print):_____ Section:_____

1. Induction = learning _____ concepts from _____ examples

2. Name 2 biggest reasons for poor model performance:

   (a)

   (b)

3. Identify the type of fit (over or underfitting)

Type	Definition
	the model performs well on the training data but poorly on the testing data
	the model has poor performance both on the training and testing data

4. Provide an example of overfitting and underfitting

   (a)

   (b)

5. Define Classification task:

6. Provide Class labels:

   - Categorizing Email Messages -
   - Identifying tumor cell -

7. There are two main types of classification: _____ classification and _____ classification.

8. Logistic regression outputs a _____ function value to determine class membership

9. True/False
   Decision trees can handle both categorical and numerical data

10. Which classification algorithm finds the hyperplane that best separates classes in a high-dimensional space?

    (a) k-Nearest Neighbors
    (b) Logistic Regression
    (c) Support Vector Machines (SVM)
    (d) Naive Bayes

11. Decision trees use simple decision rules such as _____ to classify data.

12. True/False
    Decision trees are biased when some classes dominate the dataset.

13. What type of tree is used when the dependent variable has continuous values?

    (a)
    (b) Classification Tree
    (c) Regression Tree
    (d) Hyperparameter Tree
    (e) None of the above

14. A _____ vote is used in k-Nearest Neighbors to assign a class label based on the most frequently represented label around a data point.

15. Why do you think that clean, high-quality data important for decision trees?

16. List classifications models that you are not familiar with: