CS 345: Data Mining Names:_____

In-Class Activity 7 _____

K-Nearest Neighbors

1. [True or False] k-NN algorithm does more computation during test time rather than train time.

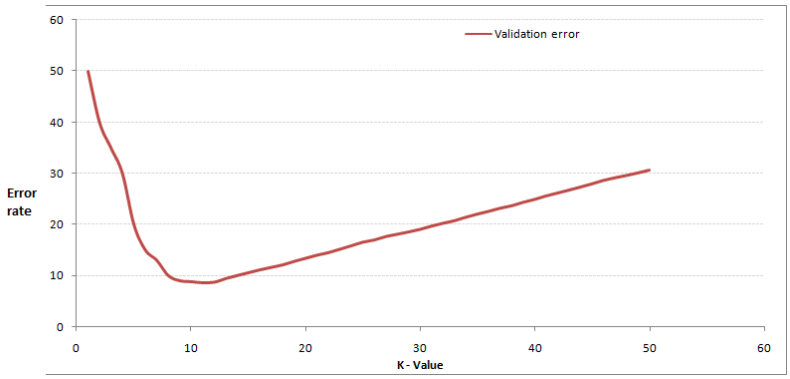
2. In the image below, which would be the best value for k, assuming you are using the k-nearest neighbor algorithm?

   a. 3

   b. 10

   c. 20

   d. 50



3. Which of the following distance metrics cannot be used in k-NN?

   A) Manhattan

   B) Minkowski

   C) Hamming

   D) Supremum

   E) Euclidean

   F) All can be used

4. Which of the following option is true about k-NN algorithm?

   A) It can be used for classification

   B) It can be used for regression

   C) It can be used in both classification and regression

5. Which of the following is true about Manhattan distance?

   A) It can be used for continuous variables

   B) It can be used for categorical variables

   C) It can be used for categorical as well as continuous

   D) None of these

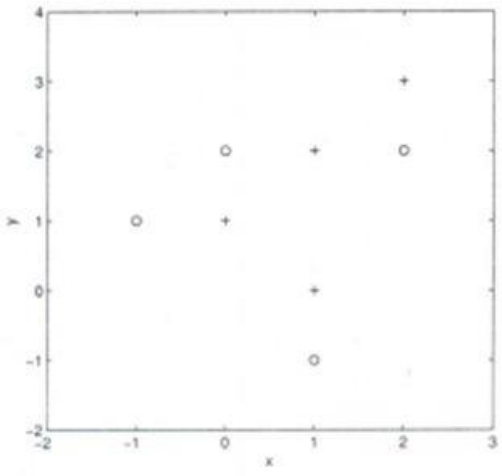
6. What is the **Euclidean** distance between two data points A (1, 3) and B (2, 3)?

7. What is the **Manhattan** distance between two data points A (1, 3) and B (2, 3)?

Use the following information for Questions 8-9.

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable. On the right is a scatter plot of the data in the table in 2D space.

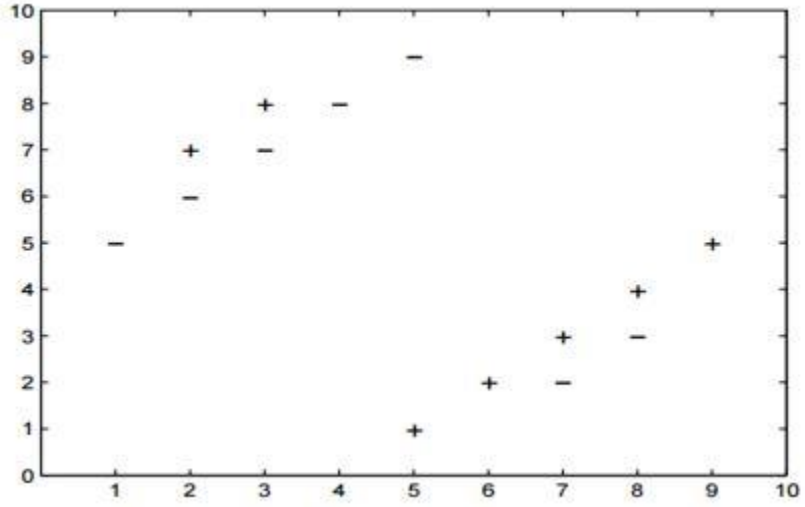
x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$



8. Suppose, you want to predict the class of new data point x=1 and y=1 using Euclidean distance in **3**-NN. In which class this data point belong to?

9. Suppose, you want to predict the class of new data point x=1 and y=1 using Euclidean distance in **7**-NN. In which class this data point belong to?

Use the following to answer questions 10-11. Suppose you have given the following 2-class data where “+” represents a positive class and “-” represents a negative class.



10. Which of the following value of k in k-NN would minimize the leave one out cross validation accuracy?
    a. 3
    b. 5
    c. Both have the same.
    d. None of these.

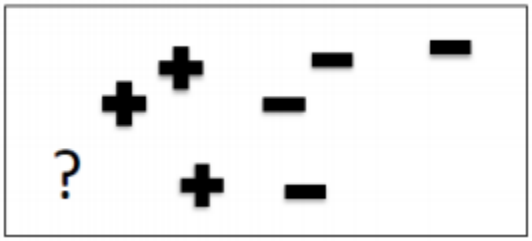
11. Which of the following would be the leave one out cross validation accuracy for k=5?
    a. 4/14
    b. 6/14
    c. 8/14
    d. 10/14

12. In k-NN it is very likely to overfit due to the curse of dimensionality. Which of the following option would you consider to handle such problem?
    a. Dimensionality Reduction
    b. Feature selection
    c. Both
    d. Neither

13. When you find noise in data, which of the following options would you consider in k-NN?
    a. Increase the value of k.
    b. Decrease the value of k.
    c. Noise is not dependent on value of k.
    d. None of these.

14. A company has built a k-NN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong?
    Note: Model has successfully deployed and no technical issues are found at client side except the model performance.
    a. It is probably an overfitted model.
    b. It is probably an underfitted model.
    c. Can't say.
    d. None of these.

15. Given the following labeled data set.



    For what (minimal) value of k will the query point "?" be classified as negative? Ties are broken at random, so try to avoid them.

16. Which of the following statements are true for k-NN?
    a. If k is a very large value, we may include points from other classes into the neighborhood.
    b. If k is a very small value, the algorithm is very sensitive to noise.
    c. Both are true.
    d. Neither is true.