University of Roehampton London

CMP020L012A	Data Analytics	2022/23

Week-8
K-Nearest Neighbor
Solution
Answer the following Multiple-Choice Questions

1. Assume that you are given 100 samples. What is the correct "K" value that you will consider for classifying a new datapoint in this existing dataset?
    a) 10
    b) 11
    c) 9
    d) Either b or c
2. Multiclass Classification means the label attribute has exactly two class/label values in the dataset.
    a) True
    b) False
3. 5-NN means-
    a) You have 5 samples to calculate Euclidean distance
    b) You have 5 nearest/closest samples to consider for comparing their labels
    c) You have at most 5 samples to calculate Euclidean distance
    d) None of the above
4. Which of the following Machine Learning Algorithm can be used on both Quantitative (Numeric) and Qualitative (Non-numeric) Data?
    a) K-NN
    b) Linear Regression
    c) Both the above
    d) None of the above
5. Which of the following is true for the Euclidean Distance?
    a) It can be directly applied on Categorical/Qualitative Data
    b) It can be directly applied on Continuous/Quantitative Data
    c) Both a and b
    d) None of the above
6. Which of the following distance measure is used in case of Categorical/Qualitative attribute in K-NN?
    a) Euclidean
    b) Hamming
    c) Manhattan
    d) Noe of the above

7. Which of the following will be the Euclidean Distance between Datapoint (1,3) and a new Datapoint (2,3)?
    a) 1
    b) 2
    c) 4
    d) 8
8. Which of the following statements is true about K-NN?
    a) We can choose an Optimal Value of K based on the number of samples
    b) Euclidean Distance treats each datapoint equally
    c) Both a and b
    d) None of the above
9. The statement 'K-NN does not require an explicit training step' is,
    a) True
    b) False
10. Which of the following is true for the K-NN classifier?
    a) The Classification Accuracy is better with larger "K" value
    b) The Classification Accuracy is better with smaller "K" value
    c) Classification is better with not too large and not to small but optimal "K" value
    d) None of the above
11. The Euclidean distance between two numerical attributes used to determine the ____ between them:
    a) Validation Data
    b) Error Rate
    c) Closeness
    d) None of the above
12. Which of the following Machine Learning Algorithms can be used for resolving the Missing Values problem for both continuous and categorical attributes?
    a) Linear Regression
    b) K-Nearest Neighbor
    c) Logistic Regression
    d) None of the above
13. K-NN typically requires more time for Classification as compared to other classifiers
    a) True
    b) False
14. The Classification for any new datapoint may differ when you either increase or, decrease the value of "K" in K-NN.
    a) True
    b) False
15. Taking K=1 in K-NN can cause Mislabeling for the new datapoint.
    a) True
    b) False

University of Roehampton London

Answer the following Question

Can you Estimate the Classification for an unseen Datapoint where the first and second attribute values are 9.1 and 11.0 respectively?

Attribute 1	Attribute 2	Class
0.8	6.3	−
1.4	8.1	−
2.1	7.4	−
2.6	14.3	+
6.8	12.6	−
8.8	9.8	+
9.2	11.6	−
10.8	9.6	+
11.8	9.9	+
12.4	6.5	+
12.8	1.1	−
14.0	19.9	−
14.2	18.5	−
15.6	17.4	−
15.8	12.2	−
16.6	6.7	+
17.4	4.5	+
18.2	6.9	+
19.0	3.4	−
19.6	11.1	+

Answer:

Step:1- Calculating the distance of the new datapoint (9.1, 11.0) with all the datapoint(s) exist in the dataset.

For in instance, the distance of the new datapoint with the first datapoint in the dataset will be,

Sqrt $[(9.1 - 0.8)^2 + (11.0 - 6.3)^2] = 22.73$

Similarly, measure the distance for all the datapoints in the dataset.

Step:2- Determine the value of "K"

Sqrt [total sample] = Sqrt [20] = 4.47213 (Approximately) = 5

So, K=5

Step:3- Identify 5 nearest/closest datapoints from the distance that you have calculated. Also find their corresponding classes. Because you chose K=5 that means you will get a particular class value having higher weight, such as,

5 +s, 0 –s, or,

5 –s, 0 +s, or,

4 –s, 1 +s, or,

1 –s, 4 +s, or,

3 +s, 2 -s, or,

2 +s, 3 -s, etc.

You then choose the "Class" value (+ / -) for the new datapoint according to one with the higher weight.