**Business Data Mining**

**Assignment 5**

**Name: Fahad Khalid Alofani KSU ID: 437104121**

**Name: Faisal Abdullah Almahmood KSU ID: 437104470**

**Problem 1 [40pts]**. Consider the following points in two-dimensional space and their classifications:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | x | y | Class | EU D | MAN D |
| 1 | 8 | 4 | Y | 2.83 | 4 |
| 2 | 3 | 3 | Y | 3.16 | 4 |
| 3 | 4 | 5 | Y | 3.61 | 5 |
| 4 | 0 | 1 | Y | 6.08 | 7 |
| 5 | 10 | 2 | Y | 4 | 4 |
| 6 | 3 | 7 | Y | 5.83 | 8 |
| 7 | 0 | 9 | N | 9.22 | 13 |
| 8 | 8 | 1 | N | 2.24 | 3 |
| 9 | 4 | 3 | Y | 2.24 | 3 |
| 10 | 9 | 4 | N | 3.61 | 5 |
| 11 | 1 | 5 | Y | 5.83 | 6 |
| 12 | 10 | 8 | N | 7.21 | 10 |
| 13 | 6 | 1 | N | 1 | 1 |
| 14 | 8 | 3 | Y | 2.24 | 3 |
| 15 | 10 | 3 | N | 4.12 | 5 |
| 16 | 0 | 5 | N | 6.71 | 9 |
| 17 | 1 | 8 | Y | 7.81 | 11 |
| 18 | 5 | 5 | N | 3.16 | 4 |
| 19 | 5 | 6 | Y | 4.12 | 5 |
| 20 | 9 | 2 | N | 3 | 3 |

For this exercise, we will use Euclidean and Manhattan distance metrics. For example: the distance from (3, 3) to (8, 1) using Euclidean distance metric is , and the distance using Manhattan is |3 – 8| + |3 – 1| = 7.

Suppose we are interested in a binary (Y/N) output. Considering a new point **(6, 2)**:

1. Using Euclidean distance function, what are the five closest points in our data set to (6, 2)?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **X** | **Y** | **CLASS** | **Euclidean Distance** |
| **1** | **8** | **4** | **Y** | **2.83** |
| **8** | **8** | **1** | **N** | **2.24** |
| **9** | **4** | **3** | **Y** | **2.24** |
| **13** | **6** | **1** | **N** | **1** |
| **14** | **8** | **3** | **Y** | **2.24** |

1. Using Manhattan distance function, what are the five closest points in our data set to (6, 2)?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **X** | **Y** | **CLASS** | **Manhattan Distance** |
| **8** | **8** | **1** | **N** | **3** |
| **9** | **4** | **3** | **Y** | **3** |
| **13** | **6** | **1** | **N** | **1** |
| **14** | **8** | **3** | **Y** | **3** |
| **20** | **9** | **2** | **N** | **3** |

1. Using the K-nearest neighbors approach, what would be the predicted output for (6, 2) using *k* = 2 and *k* = 5 neighbors in both distance functions?

Euclidean distance: When k=5 the outcome is Y

When k=2 We select randomly because there are 2 Y and 2 N

Manhattan distance: When k= 5 the outcome is N

When k=2 the outcome is N

**Problem 2 [20pts]**. Consider the following two-dimensional plot:



1. Assume using KNN classifier with Euclidean distance, and that a point can be its own neighbor (k = 0). What value of k minimizes the training set error for this dataset? What is the resulting training error?

k=0 100% accuracy, and 0 training error.

1. When using KNN classifier, why might using too large values of *k* be bad in this dataset? Why might too small values of *k* also be bad?

When k is too large that means it will take other points far from it. That will decrease the accuracy.

When k is too small that mean it will take near points from it. That will increase the accuracy, but it is overfitting.

1. Can decision tree technique have high accuracy for this data? Why or why not?

NO, because when we used gain\_rati , info\_gain we got 50% accuracy and. When we used gini\_index , accuracy we got 64.29% that means it is bad for this dataset, but we have also tried without selecting apply pruning we got 100% but it is overfitting.

1. Can logistic regression technique have high accuracy for this data? Why or why not?

NO, because when we used LR we got 71.43%.

**Problem 3 [40pts]**. Consider the following points in three-dimensional space and their classifications:

(8,4,0) (3,3,3) (9,9,2) (4,5,0) (0,1,0) (8,13,5) (10,2,4) (7,12,5) (3,7,2) (0,9,1)

(8,1,2) (4,3,1) (9,4,3) (6,9,6) (2,5,3) (7,10,5) (5,2,0) (4,8,2) (8,6,1) (2,8,3)

For this exercise, use the Manhattan distance metric.

*For example*: the distance from (3, 3, 1) to (8, 1, 5) is |3 – 8| + |3 – 1| + |1 – 5| = 11

We want to find two clusters. Beginning with centroids at (1, 1, 1) and (8, 8, 4), do THREE iterations of the 2-means clustering algorithm as follows:

* Allocate the points to centroids, then find the new centroids.
* Again, allocate the points to the centroids, and then get the new centroids.

If a point is equidistant between the centroids, assign it to the centroid that starts at (1, 1, 1). What are the resulting centroids and resulting clusters?

To properly present your answers, do each iteration separately, show your results in table format, and state the new centroids after each iteration.

**Iteration1:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | Y | Z | Distance1 | Distance2 | Cluster |
| 8 | 4 | 0 | 11 | 8 | 2 |
| 3 | 3 | 3 | 6 | 11 | 1 |
| 9 | 9 | 2 | 17 | 4 | 2 |
| 4 | 5 | 0 | 8 | 11 | 1 |
| 0 | 1 | 0 | 2 | 19 | 1 |
| 8 | 13 | 5 | 23 | 6 | 2 |
| 10 | 2 | 4 | 13 | 8 | 2 |
| 7 | 12 | 5 | 21 | 6 | 2 |
| 3 | 7 | 2 | 9 | 8 | 2 |
| 0 | 9 | 1 | 9 | 12 | 1 |
| 8 | 1 | 2 | 8 | 9 | 1 |
| 4 | 3 | 1 | 5 | 12 | 1 |
| 9 | 4 | 3 | 13 | 6 | 2 |
| 6 | 9 | 6 | 18 | 5 | 2 |
| 2 | 5 | 3 | 7 | 10 | 1 |
| 7 | 10 | 5 | 19 | 4 | 2 |
| 5 | 2 | 0 | 6 | 13 | 1 |
| 4 | 8 | 2 | 11 | 6 | 2 |
| 8 | 6 | 1 | 12 | 5 | 2 |
| 2 | 8 | 3 | 10 | 7 | 2 |
| Centroid1 | 1 | 1 | 1 |
| Centroid2 | 8 | 8 | 4 |

**Iteration2:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | Y | Z | Distance1 | Distance2 | Cluster |
| 8 | 4 | 0 | 7.04 | 8.09 | 1 |
| 3 | 3 | 3 | 2.2 | 8.59 | 1 |
| 9 | 9 | 2 | 12.54 | 4.75 | 2 |
| 4 | 5 | 0 | 4.04 | 8.59 | 1 |
| 0 | 1 | 0 | 6.46 | 16.59 | 1 |
| 8 | 13 | 5 | 18.54 | 8.41 | 2 |
| 10 | 2 | 4 | 11.2 | 9.75 | 2 |
| 7 | 12 | 5 | 16.54 | 6.41 | 2 |
| 3 | 7 | 2 | 4.54 | 5.59 | 1 |
| 0 | 9 | 1 | 8.8 | 10.25 | 1 |
| 8 | 1 | 2 | 8.2 | 9.09 | 1 |
| 4 | 3 | 1 | 1.7 | 9.59 | 1 |
| 9 | 4 | 3 | 8.54 | 6.09 | 2 |
| 6 | 9 | 6 | 13.54 | 4.91 | 2 |
| 2 | 5 | 3 | 4.3 | 7.59 | 1 |
| 7 | 10 | 5 | 14.54 | 4.41 | 2 |
| 5 | 2 | 0 | 4.7 | 10.59 | 1 |
| 4 | 8 | 2 | 6.54 | 4.25 | 2 |
| 8 | 6 | 1 | 8.04 | 5.09 | 2 |
| 2 | 8 | 3 | 7.3 | 5.25 | 2 |
| Centroid1 | 2.88 | 3.33 | 1.25 |
| Centroid2 | 6.75 | 7.67 | 3.17 |

**Iteration3:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | Y | Z | Distance1 | Distance2 | Cluster |
| 8 | 4 | 0 | 5.5 | 8.9 | 1 |
| 3 | 3 | 3 | 3.5 | 9.5 | 1 |
| 9 | 9 | 2 | 11.1 | 4.7 | 2 |
| 4 | 5 | 0 | 2.5 | 9.5 | 1 |
| 0 | 1 | 0 | 7.9 | 17.5 | 1 |
| 8 | 13 | 5 | 17.1 | 7.5 | 2 |
| 10 | 2 | 4 | 11.1 | 9.7 | 2 |
| 7 | 12 | 5 | 15.1 | 5.5 | 2 |
| 3 | 7 | 2 | 4.5 | 6.5 | 1 |
| 0 | 9 | 1 | 8.9 | 10.3 | 1 |
| 8 | 1 | 2 | 8.1 | 9.9 | 1 |
| 4 | 3 | 1 | 1.5 | 10.5 | 1 |
| 9 | 4 | 3 | 7.1 | 6.9 | 2 |
| 6 | 9 | 6 | 12.1 | 4.1 | 2 |
| 2 | 5 | 3 | 4.5 | 8.5 | 1 |
| 7 | 10 | 5 | 13.1 | 3.5 | 2 |
| 5 | 2 | 0 | 4.5 | 11.5 | 1 |
| 4 | 8 | 2 | 5.1 | 4.5 | 2 |
| 8 | 6 | 1 | 6.5 | 5.9 | 2 |
| 2 | 8 | 3 | 7.5 | 5.5 | 2 |
| Centroid1 | 3.7 | 4 | 1.2 |
| Centroid2 | 6.8 | 8.1 | 3.6 |