# Data Mining Assignment #5

**TEAM MEMBERS**

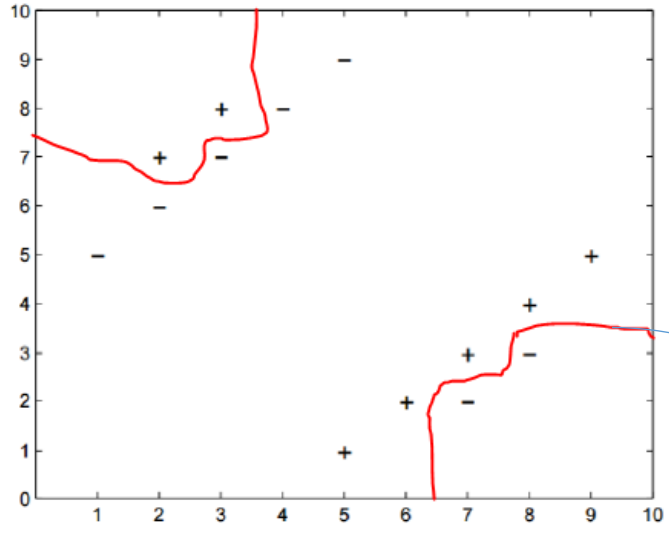
***1. Nithyadharshni Sampathkumar***
***2. Saai Krishnan Udayakumar***
***3. Sibi Senthur Muthusamy***

## Problem 1

**a)** We are given that each point can be its own neighbor. So, effectively "**k=1**" (as each point themselves are their first nearest neighbors) minimizes the training error for the data set. The net error in such a scenario would be effectively zero.

**b)** Generally, in k nearest neighbor's → **large values of k leads to under fitting and small values of k results in over fitting.** Large values of k results in a smoother boundary, less noise but may miss the local structure. Based on empirical results, k=n (size of the entire data set) is the same as a naïve rule which classifies all records according to the majority class and basically under fits as the decision surfaces becomes simpler. On the other hand, smaller values of k (k=1, 2, 3 …) capture local structure in data but also has less noise leading to overfitting – as it encompasses a decision boundary vulnerable to the training set. Let's understand the classification accuracy for larger values of k, **for k=13,** all the points in the data are misclassified using leave one-out cross validation.

Smaller k values lead to overfitting as could be inferred from the below decision boundary graph. For K=1 or for other smaller values we have a decision boundary that is fit exactly to classify points according to the nearest neighbor. We can infer how strict the decision boundary is for smaller values of k (K=1 in our case) from the decision boundary marked in red from the following graph. Also, to choose the best value of K, we use cross validation and chose k values which gives the lowest misclassification error on the validation set.




Stringent Decision Boundary vulnerable to overfitting for k=1, if there is a slight shift in the data points, the classification boundary is susceptible for a change.

**c)** We start with k=1 to predict the misclassification rate. We, use leave-out cross validation to predict the misclassification error. We can infer that **K=5 and K=7** have the lowest leave-one out cross validation error the data set with an error rate **(4/14).** For, even-values of K, the misclassification error would majorly depend on how we decide to break the tie if there are (for Example) three instances (neighbor) in the positive class and three instances in the negative class. Hence, our analysis is mainly focused on the odd valued K-neighbors, although we have accounted for even number of instances as well. **Based on complete analysis, "K=5 or K=7" would have the least misclassification error rate of 4/14 (~ 28.57 %).** The following column provides a comprehensive results of the misclassification rate for various values of k (k=1, 2, 3…13) from which we have based our results upon.

| K | Misclassification Rate | Points Classified Correctly |
|---|---|---|
| 1 | 10/14 | **(5,1), (9,5), (1,5), (5,9)** |
| 2 | Depends on how the tie is broken to determine the majority class | → Either all points are misclassified or none depending on how the tie is broken between nearby points. |
| 3 | 6/14 | **All points except (7,2), (7,3), (3,7) (8,3), (2,7) and (3,8)** |
| 4 | Depends on how the tie is broken to determine the majority class | |
| **5** | **4/14** | **All points except (7,2), (8,3), (2,7) and (3,8)** |
| 6 | Depends on how the tie is broken to determine the majority class | Depends on how the point (5, 9) is classified as the distance from the point to (1, 5) and (9, 5) are the same which seem to have opposite classes. |
| **7** | **4/14** | **All points except (7,2), (8,3), (2,7) and (3,8)** |
| 8 | Depends on how the tie is broken to determine the majority class | |
| 9 | **>> 4/14 < E <= 1** | **'E' implies the misclassification rate here.** |
| 10 | Depends on how the tie is broken to determine the majority class | |
| 11 | **>> 4/14 < E <1** | **'E' implies the misclassification rate here.** |
| 12 | Depends on how the tie is broken to determine the majority class | |
| 13 | **14/14 (~100 %)** | **All the points are being misclassified** |

**Answer: "K=5 or K=7" would have the least misclassification error rate of 4/14 (~ 28.57 %)**

## Problem 2

Let the two clusters be terms as C1 and C2, we perform k-means clustering from the given data points.

Given that the initial centroids are **c1 (1, 1) and c2 (8, 8),** Distance of the points from the two centroids is given from the table below **(Iteration #1)**

| (X,Y) | A (8,4) | B (3,3) | C (4,5) | D (0,1) | E (10,2) | F (3,7) | G (0,9) | H (8,1) | I (4,3) | J (9,4) |
|---|---|---|---|---|---|---|---|---|---|---|
| D(c1) | 10 | 4 | 7 | 1 | 10 | 8 | 9 | 7 | 5 | 11 |
| D(c2) | 4 | 10 | 7 | 15 | 8 | 6 | 9 | 7 | 9 | 5 |

Points will be assigned to clusters based on their minimum distance to centroid

| A (8,4) | B (3,3) | C (4,5) | D (0,1) | E (10,2) | F (3,7) | G (0,9) | H (8,1) | I (4,3) | J (9,4) |
|---|---|---|---|---|---|---|---|---|---|
| C2 | C1 | C1 | C1 | C2 | C2 | C1 | C1 | C1 | C2 |

**Cluster 1 are the points (3, 3), (4,5), (0,1), (0,9), (8,1) and (4,3)**

**Cluster 2 are the points (8, 4), (10, 2), (3, 7) and (9, 4)**

**Computing the Centroid again based on the Clusters and computing the distance again, (Iteration #2)**

**Centroid 1 (19/6, 22/6) → (3.166, 3.666) (Average of Points in Cluster #1)**
**Centroid 2 (30/4, 17/4) → (7.5, 4.25) (Average of Points in Cluster #2)**

**Computing the Distance of the Points to the Centroid(s),**

| (X,Y) | A (8,4) | B (3,3) | C (4,5) | D (0,1) | E (10,2) | F (3,7) | G (0,9) | H (8,1) | I (4,3) | J (9,4) |
|---|---|---|---|---|---|---|---|---|---|---|
| D(c1) | 5.17 | 0.832 | 2.168 | 5.832 | 8.5 | 3.5 | 8.5 | 7.5 | 1.5 | 6.168 |
| D(c2) | 0.75 | 5.75 | 4.25 | 10.75 | 4.75 | 7.25 | 12.25 | 3.75 | 4.75 | 1.75 |

Points will be assigned to clusters based on their minimum distance to centroid

| A (8,4) | B (3,3) | C (4,5) | D (0,1) | E (10,2) | F (3,7) | G (0,9) | H (8,1) | I (4,3) | J (9,4) |
|---|---|---|---|---|---|---|---|---|---|
| C2 | C1 | C1 | C1 | C2 | C1 | C1 | C2 | C1 | C2 |

**Cluster 1 are the points (3, 3), (4,5), (0,1), (3,7), (0,9), (4,3)**

**Cluster 2 are the points (8, 4), (10, 2), (8, 1), (9, 4)**

**Computing the Centroid again based on the Clusters and computing the distance again, (Iteration #3)**

**Centroid 1 (14/6, 28/6) → (2.333, 4.666) (Average of Points in Cluster #1)**
**Centroid 2 (35/4, 11/4) → (8.75, 2.75) (Average of Points in Cluster #2)**

**Computing the Distance of the Points to the Centroid(s),**

| (X,Y) | A (8,4) | B (3,3) | C (4,5) | D (0,1) | E (10,2) | F (3,7) | G (0,9) | H (8,1) | I (4,3) | J (9,4) |
|---|---|---|---|---|---|---|---|---|---|---|
| D(c1) | 6.333 | 2.333 | 2 | 6 | 10.333 | 3 | 6.667 | 9.333 | 3.333 | 7.333 |
| D(c2) | 2 | 6 | 7 | 10.5 | 4.75 | 10 | 15 | 2.5 | 5 | 1.5 |

Points will be assigned to clusters based on their minimum distance to centroid

| A (8,4) | B (3,3) | C (4,5) | D (0,1) | E (10,2) | F (3,7) | G (0,9) | H (8,1) | I (4,3) | J (9,4) |
|---|---|---|---|---|---|---|---|---|---|
| C2 | C1 | C1 | C1 | C2 | C1 | C1 | C2 | C1 | C2 |

**Cluster 1 are the points (3, 3), (4,5), (0,1), (3,7), (0,9), (4,3)**

**Cluster 2 are the points (8, 4), (10, 2), (8, 1), (9, 4)**

**Hence, we could infer that at the end of third iteration our Clusters start to Converge and we will be having the same clusters and centroid value.**

**b)** The point given is (5, 3) and we calculate the Manhattan distance of the point (5, 3) from each of the points in two dimensional space.

| Points | Manhattan Distance | Class |
|---|---|---|
| (8,4) | 4 | Yes |
| **(3,3)** | **2** | **Yes** |
| **(4,5)** | **3** | **No** |
| (0,1) | 7 | No |
| (10,2) | 6 | Yes |
| (3,7) | 6 | Yes |
| (0,9) | 11 | No |
| (8,1) | 5 | No |
| **(4,3)** | **1** | **Yes** |
| (9,4) | 5 | Yes |

From the table we can infer that the three closest points in the data set are (4, 3), (3, 3) and (4, 5) as they have the lowest Manhattan distance (1,2 and 3 units respectively) from (5, 3) among the ten points given in the two dimensional space.

The predicted class for (5, 3) would be the majority class of the points (4, 3), (4, 5), (3, 3) which is the majority outcome of classes {yes, No, yes} which is yes.

**Hence, the predicted class for the point (5, 3) would be YES**