Foundations of Data Analytics (CS910/ CS430) COURSEWORK 3

Akshara Kannan (2045986) MSc Computer Science

December 15 2020

1)

a)

Consider the data points:

$$x = (5, 3, 7, 9); y = (7, 3, 6, 7) \tag{1}$$

i) Hamming Distance

It is possible to compute the Hamming distance for these vectors because x and y are of the same size. The Hamming distance is defined as the number of bit differences in the two vectors. Let x and y of size 4, be represented as (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4).

$$x_1 \neq y_1$$
$$x_2 = y_2$$
$$x_3 \neq y_3$$
$$x_4 \neq y_4$$

Therefore,

$$HammingDistance = 3 \tag{2}$$

ii) Manhattan Distance

For a 2-dimensional Euclidean space, and for two points (x_1, y_1) and (x_2, y_2)

$$ManhattanDistance = |(x_1 - x_2)| + |(y_1 - y_2)| \tag{3}$$

$$ManhattanDistance = |5 - 7| + |3 - 3| + |6 - 7| + |7 - 9|$$
$$L1 = | - 2| + |0| + | - 1| + |2|$$
$$L1 = 2 + 0 + 1 + 2 = 5$$

$$ManhattanDistance = L1 = 5 \tag{4}$$

iii) Euclidean Distance

For a 2-dimensional Euclidean space, and for two points (x_1, y_1) and (x_2, y_2)

$$EuclideanDistance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{5}$$

$$EuclideanDistance = \sqrt{(5 - 7)^2 + (3 - 3)^2 + (7 - 6)^2 + (9 - 7)^2}$$
$$L2 = \sqrt{(-2)^2 + (0)^2 + (-1)^2 + (2)^2}$$
$$L2 = \sqrt{4 + 0 + 1 + 4}$$
$$L2 = \sqrt{9} = 3$$

$$EuclideanDistance = L2 = 3 \tag{6}$$

b)

In this example, length of eyelashes and height of a person are the two attributes chosen to perform clustering on, both measured in millimeters. The average length of eyelashes in people is around 10 mm while the average height of a short person is as high as 1500 - 1600 mm. Such a vast difference between the values of the two attributes will influence the way the clusters are formed, because the effect of eyelash length will be negligible when compared to the height. The use of Euclidean distance measure will only make it worse, because the difference in value is amplified due to the squaring of numbers.

One way to overcome this problem is to use different units for the two attributes, metres for height and millimeters for eyelash length. Then the attributes will be comparable in value, to draw conclusive results.

Another way to overcome this problem is to standardize the attributes using standardisation measures such as Z-scores. This will enable us to compare data with different ranges of values.

c)

Given: $k = 3$ and coordinates of points A-G as $A(2, 10)$, $B(2, 5)$, $C(8, 4)$, $D(5, 8)$, $E(7, 5)$, $F(6, 4)$, $G(1, 2)$.

i)

It is given that points A, D and G are chosen as initial cluster centers. We need to decide how to allocate the remaining points to these clusters.

- To do that, we calculate the distances of the remaining points from the 3 cluster centers. Each cluster is assigned points that are closest to it.
- Then the new cluster centers are found by computing the average of the points assigned to that cluster.

We use the Euclidean distance measure to calculate the distance between two points. For two points in 2-dimensional space, $X(x_1, y_1)$ and $Y(x_2, y_2)$, the euclidean distance is computed using the following Eq. (5):

$$EuclideanDistance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Distance Calculations:-

$$AD = \sqrt{(2-5)^2 + (10-8)^2}$$
$$AD = \sqrt{(-3)^2 + (2)^2}$$
$$AD = \sqrt{9+4}$$
$$AD = \sqrt{13}$$

$$AG = \sqrt{(2-1)^2 + (10-2)^2}$$
$$AG = \sqrt{(1)^2 + (8)^2}$$
$$AG = \sqrt{1+64}$$
$$AG = \sqrt{65}$$

$$AB = \sqrt{(2-2)^2 + (10-5)^2}$$
$$AB = \sqrt{(0)^2 + (5)^2}$$
$$AB = \sqrt{0+25}$$
$$AB = \sqrt{25}$$
$$AB = 5$$

$$DB = \sqrt{(5-2)^2 + (8-5)^2}$$
$$DB = \sqrt{(3)^2 + (-3)^2}$$
$$DB = \sqrt{9+9}$$
$$DB = \sqrt{18}$$

$$DB = 3\sqrt{2}$$

$$GB = \sqrt{(1-2)^2 + (2-5)^2}$$
$$GB = \sqrt{(-1)^2 + (-3)^2}$$
$$GB = \sqrt{1+9}$$
$$GB = \sqrt{10}$$

$$AC = \sqrt{(8-2)^2 + (4-10)^2}$$
$$AC = \sqrt{(6)^2 + (-6)^2}$$
$$AC = \sqrt{36+36}$$
$$AC = \sqrt{72}$$
$$AC = 6\sqrt{2}$$

$$DC = \sqrt{(8-5)^2 + (4-8)^2}$$
$$DC = \sqrt{(3)^2 + (-4)^2}$$
$$DC = \sqrt{9+16}$$
$$DC = \sqrt{25}$$
$$DC = 5$$

$$GC = \sqrt{(8-1)^2 + (4-2)^2}$$
$$GC = \sqrt{(7)^2 + (2)^2}$$
$$GC = \sqrt{49+4}$$
$$GC = \sqrt{53}$$

Similarly all distances are calculated using the formula above and filled in the table below.

Points	A	D	G
A	0	$\sqrt{13}$	$\sqrt{65}$
B	5	$3\sqrt{2}$	$\sqrt{10}$
C	$6\sqrt{2}$	5	$\sqrt{53}$
D	$\sqrt{13}$	0	$2\sqrt{13}$
E	$\sqrt{2}$	$\sqrt{13}$	$3\sqrt{5}$
F	$2\sqrt{13}$	$\sqrt{17}$	$\sqrt{29}$
G	$\sqrt{65}$	$2\sqrt{13}$	0
H	$\sqrt{5}$	$\sqrt{2}$	$\sqrt{58}$

ii)

Now for each point (row in above table), we compare the values of the distances of that point from the three cluster centers. The point belongs to that cluster whose center is closest to it. Thus, the points are clustered into the three clusters as below.

- No points are assigned to Cluster with center A. So it's center remains the same.
- Cluster with center B has C,D,E,F,H points
- Cluster with center G has B,G points

The new cluster centers (Centroids) are computed by averaging the coordinates of points belonging to a cluster in each dimension. In this example, all points were clustered into the cluster with center D, so the new cluster center for this cluster, D^* and G^* is calculated as follows:-

$$D^* = (\frac{(8+5+7+6+4)}{6}, \frac{(4+8+5+4+9)}{6})$$
$$D^* = (\frac{30}{5}, \frac{30}{5})$$
$$D^* = (6, 6)$$

$$G^* = (\frac{(1+2)}{2}, \frac{(5+2)}{2})$$
$$G^* = (\frac{3}{2}, \frac{7}{2})$$
$$G^* = (1.5, 3.5)$$

So, the three clusters for the next iteration are $A(2, 10), D(6, 6)$ and $G(1.5, 3.5)$.

2) Clustering on Breast Cancer Dataset

a) Hierarchical Clustering

Accuracy for Hierarchical Agglomerative Clustering :-

Distance Metric	Single Linkage	Complete Linkage	Average Linkage
Euclidean	70.6294%	55.5944%	69.9301%
Manhattan	70.6294%	55.5944%	69.9301%

For both distance measures, the same accuracies are obtained for correctly clustered instances. This behaviour can be attributed to the type of attributes of the Breast Cancer Dataset. Since the attributes are non numeric in nature, the distance between them is characterised in the same manner.

We can see from the tables above, that single linkage type produces clusters with highest accuracy = 70.6294% for both of the distance measures.

b) Farthest-First Clustering

Range of incorrectly clustered accuracies = 24.4755% to 48.951%.

Seed number	Incorrectly Clustered Instances
1	34.2657%
2	31.1189%
3	37.4126%
4	30.4196%
5	24.4755%
6	29.3706%
7	48.951%
8	37.0629%
9	31.4685%
10	24.8252%

In this clustering method, the arbitrary nature in which initial center is chosen, affects the way clusters are formed. When different centers are chosen at random, the points that form the set of centers vary, thus giving a different accuracy each time, following no pattern. For this dataset, we have only considered range of random seeds as 1 to 10 for which we get the incorrectly clustered accuracies with as much as 100% difference in value.

There is no guarantee we will obtain the best performance on this dataset for this algorithm, proving to be only optimal in the chosen range of seed values, at best.

c) Density Based Clustering

DBSCAN stands for Density Based Spatial Clustering of Applications with Noise. DBSCAN Algorithm clusters points which belong to dense regions together. It overcomes the disadvantage of K-means clustering not being able to work in cases of non-globular clusters.

The two parameters it requires are **epsilon** (how close points should be to each other for the algorithm to cluster them) and **minPoints** (minimum number of points to form a dense region).

For different values of epsilon and minPoints, different models were obtained.

- For some values of parameters, the model is inconclusive because the algorithm marks every point as NOISE.
- For other values, the model produces zero or two clusters.
- When epsilon is in the range of 1.5 - 1.7 and minPoints takes the values of 5 and 6, the algorithm outputs a model with 2 clusters. But even in these cases, the clusters are not of similar size.

Therefore, we can conclude that the DBSCAN is not a suitable algorithm for this dataset.

Also, DBSCAN is not the best method of clustering data because it is not a definitive one. Every round of execution depends on the order of processing of points. Choosing the best value of epsilon is also tough when the dataset contains a large number of attributes.

d) Classification vs Clustering

Classifiers on Breast-Cancer Dataset

Classifier	Accuracy
ZeroR	65.97%
Naive Bayes	71.134%
KNN (k = 5,6,7,8)	**73.1959**%
Decision Tree(J48)	68.0412%
Support Vector Machine(SMO)	70.1031%

Clustering Algorithms on Breast-Cancer Dataset

Algorithm	Accuracy
Hierarchical Clustering	70.6294%
Farthest First Clustering	75.5245%
DBSCAN	Not suitable

The tables above show the accuracies for the different classification and clustering algorithms on the breast-cancer dataset. This dataset contains the class values- recurrence and no-recurrence, so for clustering, we check if we obtain high performance for 2 clusters.

We can see that Farthest-first clustering technique produces the highest accuracy of 75.5254% when compared to the best classifier- KNN with accuracy 73.1959%. Both KNN and Farthest -First algorithms are non-deterministic in nature; the value of accuracy is heavily dependent on the value of random seed(s) or value of number of nearest neighbours(k) considered. So, we can never know the best performing parameters for the algorithm, until we exhaust the values of s/k.

But, due to the fact that this dataset gives us the class value for each data point, we could use supervised learning methods instead of unsupervised.

3)

a)

The settings of eps = 1.1 and minPoints = 10, produced two clusters. The clusters were then visualised for each attribute, clusters on the X axis attributes on the Y axis.

1. Instance_number - The range of values for this attribute (Yaxis) is 0 to 99. Points belonging to both clusters take the entire range of values of the attribute, thus **Instance_number is not separable**.
2. Dealership, M5 - Points that have a value 1 for Dealership fall under cluster 0, whereas points that have a value 0 fall under cluster 1. Very few values of cluster1 take the value 1 for this attribute. Nevertheless, **Dealership and M5 show a separable behaviour**.

3. **ComputerSearch, Z4, Financing, Purchasing -** All of these attributes are **not separable**, but show a similar behaviour with respect to each other. All the data points that take values 0 and 1, are split into both clusters.

4. Showroom - approximately all points in the dataset take a value of 1 for this attribute, and fall under both clusters with equal sizes. **Showroom is not separable** by nature.

5. 3Series - Data points that take a value of 1 for this attribute are in both Cluster0 and Cluster1, whereas points that take a value 0 all fall under cluster0. So, **3Series is not separable**

b)

EM is trained on the dataset for three seed values = 1, 2, 3 for 5 clusters. The following results were found.

Seed = 1

- Cluster 0 - M5 (1), Purchase(0.0058)
- Cluster 1 - Dealership(0), Showroom(1), ComputerSearch(0), M5(0), 3Series(1), Z4(1)
- Cluster 2 - Showroom(1), M5(0), 3Series(0.9995)
- Cluster 3 - Dealership(1), 3Series(0.0803)
- Cluster 4 - M5(1), Financing(1), Purchase (1)

Seed = 2

- Cluster 0 - ComputerSearch(0.0128), M5(1), Financing(0.9021)
- Cluster 1 - Showroom(0.9999), 3Series(1), Purchase(0.0326)
- Cluster 2 - ComputerSearch(0.9133), Financing(0.9984)
- Cluster 3 - Dealership(0.0001), Showroom(1), M5(0), 3Series(1), Financing(1), Purchase(1)
- Cluster 4 - Dealership(0.9992), Financing(0.0616), Purchase (0)

Seed = 3

- Cluster 0 - Dealership(0.0914), Showroom(0.9978), M5(0.0045), 3Series(0.9925), Purchase(0.0163)
- Cluster 1 - Dealership(0.001), Showroom(0.9828), 3Series(0.9437), Financing(0.9316)
- Cluster 2 - M5(0.9446), Z4(0.0543) Financing(0.9619)
- Cluster 3 - Dealership(0.919), Financing(0.0532), Purchase(0.0091)
- Cluster 4 - Dealership(0.9686), Financing(1)

Observations:-

- SEED 2- In the clusters produced when seed value = 2, we can observe that visitors who use a Computer Search, enquire about financing a car but do not take the decision to buy it, whereas visitors who visit a Dealership or a showroom end up making up a purchase. Clusters 0, 2 and 4 correspond to the former scenario, where the Financing takes a very high value, but a purchase does not happen. On the other hand, Clusters 1 and 3 support the latter, where visiting a dealership/showroom leads to a purchase being made, possibly due to the visual appeal of cars presented in the display.
- SEED 3- From the clustering results we obtain for seed = 3, it is observed that across clusters, the tendency to make a purchase is very low. Cluster 0 and 3 illustrate the few instances where visitors make a trip to the dealership/ showroom and enquire about the financing a car and also buy it. That too, in Cluster 3, very few visitors have made a purchase without even looking at car models. For the other Clusters, visitors do not make a purchase even when they visit the dealership and showroom, even after enquiring about the financial aspect.

- SEED 1- The clusters obtained from running the model when seed = 1 are strikingly different as compared to the above observations. There seems to be hardly any confirmed purchasing happening, even with visitors having a high tendency to visit showrooms. This could be attributed to the fact that the visitors, after looking at car models (predominantly the 3Series model) do not even progress to the stage of enquiring its cost with strong confidence. We can also observe that the M5 model is the only one enjoying a purchase, but without a strong emphasis in the attributes related to the place of checking out cars (computerSearch, dealership, showroom fall in mid-range of means).
- Common inter-seed trends- Vistors who visit the showroom are more interested in buying entry level or mid-range car models such as 3series and Z4, but not the M5.There seems to be very less visitors to the dealership; among the few that do visit, rarely do they buy the cars they see. Lastly, people interested in a high-end car model such as M5, always end up making a purchase.

c)

K-means is trained on the dataset for three seed values = 1, 2, 3 for 5 clusters. The following results were found.

Seed = 1

- Cluster 0 - Dealership(0), Showroom(1), 3Series(1)
- Cluster 1 - Dealership(0.9167), Financing(0.0417) Purchase(0)
- Cluster 2 - Dealership(1), showroom(1), Computersearch(0), M5(0.9), Financing(0.9), Purchase(0.1)
- Cluster 3 - ComputerSearch(0.95), 3Series(0.1), Financing(1)
- Cluster 4 - M5(0.9412), Z4(0), Financing(1), Purchase(0.9412)

Seed = 2

- Cluster 0 - Showroom(1), M5(0.1) , Purchase(0)
- Cluster 1 - Dealership(0.0476), Showroom(0.9048), ComputerSearch(0.0476), 3Series(1)
- Cluster 2 - Showroom(0.9118), M5(0.9706)
- Cluster 3 - Showroom(0.0769), Financing(0), Purchase(0)
- Cluster 4 - Dealership(0.9167), Showroom(0.0833), ComputerSearch(0.9167), Financing(1)

Seed = 3

- Cluster 0 - 3Series(0)
- Cluster 1 - Dealership(0), Showroom(0.9091), 3Series(1), Purchase(0)
- Cluster 2 - Financing(1), Purchase(0.9474)
- Cluster 3 - Dealership(1), ComputerSearch(1), Purchase(0)
- Cluster 4 - Z4(1), Financing(0.9333)

Observations:-

- SEED 1- Cluster 4 corresponds to the case where people have enquired about financing a car of their choice, have made the purchase (seen in Cluster 4) whereas in Cluster 2, purchase does not happen when customers enquire about the cost of the car. It is seen that potential customers to a dealership fail to make a purchase (Cluster 1 & 2)
- SEED 2- Very less purchases happen in most clusters of this iteration. Cluster 0, 1 & 2 correspond to high showroom visitation, Cluster 4 corresponds to large number of dealership visits and computer searches. There is also no discernible focus on any one model per cluster (except in Cluster 1 & 2), making it look as though customers are checking out every car model equally.
- SEED 3- Cluster 2 corresponds to a strong case of purchases irrespective of the car model. On the other hand, no purchases are observed even when customers visit the showroom (Cluster 1) and dealership (Cluster 3) or browse on the computer(cluster 3) for car models.
- Common inter-seed trends- The computer search option is always coupled with enquiring the costs of the models browsed. Secondly, we can see that the 3series and M5 models are popular showroom pieces.