Exercise 5: Clustering

1. (a) Consider the data points: $x : (5, 3, 7, 9)$ and $y : (7, 3, 6, 7)$

   i. The Hamming distance is 3 for there 3 differences

   ii. The Manhattan (L_1) distance is $|5-7| + |3-3| + |7-6| + |9-7| = 5$

   iii. The Euclidean (L_2) distance is $\sqrt{(5-7)^2 + (3-3)^2 + (7-6)^2 + (9-7)^2} = 3$

   (b) Clustering using a distance measure such as Euclidean distance would be problematic on this data because the ranges of values are vastly different, for example the height of a person normally falls between 1000 and 2000 millimetres, however the length of one's eyelash is most likely just tens of millimetres. Using Euclidean distance will only reflect the difference in height, since the difference in eyelash length would be negligible compared to that. In this case we may want to normalise the data first, for example apply statistical scaling (subtract mean and divide by standard deviation) to each attribute before clustering.

   (c) Initially we assign A, D, and G as the centre of each cluster.

   i. $d(B, A) = 5$, $d(B, D) = 4.2426$, $d(B, G) = 3.1623$, B is assigned to cluster G

   $d(C, A) = 8.4853$, $d(C, D) = 5$, $d(C, G) = 7.2801$, C assigned to cluster D

   $d(E, A) = 7.0711$, $d(E, D) = 3.6056$, $d(E, G) = 6.7082$, E assigned to D

   $d(F, A) = 7.2111$, $d(F, D) = 4.1231$, $d(F, G) = 5.831$, F assigned to D

   $d(H, A) = 2.2361$, $d(H, D) = 1.4142$, $d(H, G) = 7.6158$, H assigned to D

   ii. The new cluster centres after this step will be:

   $(2, 10)$, for cluster with centre A still only has one point

   $(6, 6)$, the average of C, D, E, F and H

   $(1.5, 3.5)$, the average of B and G

2. (a) The percentage of incorrectly clustered instances with Hierarchical Clustering:

	single	complete	average
Euclidean distance	29.3706%	44.4056%	30.0699%
Manhattan distance	29.3706%	44.4056%	29.7203%

The combination of both distances with single link would perform best for accuracy of predicting the class value. For the linkage types single and complete, using Euclidean and Manhattan distance yield the same result, because all different values for a nominal attribute are "equally different" (the distance is always either 0 or 1). For any two instances, if they have n attributes with different values, the Euclidean distance would be $\sqrt{n}$ and the Manhattan distance would be n, thus when we measure distance between clusters to find the closest pair, comparing min and max distance will always give us the same choice (which pair of clusters to merge), but for average distance it could be different.

(b) The percentage of incorrectly clustered instances with Furthest Point Clustering:

seed value	1	2	3	4	5
percentage	34.2657%	31.1189%	37.4126%	30.4196%	24.4755%

seed value	6	7	8	9	10
percentage	29.3706%	48.951%	37.0629%	31.4685%	24.8252%

The range of accuracy results is between around 51 and 75.5%, the error rate of the worst case being twice of the best case, so we know the randomness in choose the initial seed point to start with can clearly influence the final clustering result for this dataset.

(c) The percentage of incorrectly clustered instances with DBSCAN (Euclidean distance):

For the default Epsilon = 0.9 and minPoints = 6, all data points are classed as "noise".

Epsilon	0 - 1.0	1.1 - 1.4	1.5 - 1.7	1.8 - 2.0
minPoints = 6	-	15.7343%	20.979%	29.7203%
minPoints = 5	-	12.5874%	22.7273%	29.7203%
minPoints = 4	-	14.3357%	25.5245%	29.7203%
minPoints = 3	0.3497%	14.6853%	26.2238%	29.7203%
minPoints = 2	12.2378%	24.4755%	27.2727%	29.7203%

We see that 0 < Epsilon < 1 (can be anything from, 1.0E-8, for example, to 1.0) and minPoints = 3 gives the best result, only one instance from the entire dataset was incorrectly clustered! So we can say that this clustering method is very suitable for this data.

(d) Recall the accuracy of different classifiers on this dataset were all below 75%, but with DBSCAN clustering it seems easy to reach an accuracy over 85%, even more than 99% if the best parameters are chosen. So the ability of clustering to predict the class value is high.

3. (a) The attributes dealership and M5 are mostly divided by the clustering.

(b) Using Weka 3.8.6, the significant dimensions for each cluster are:

seed		cluster0	cluster1	cluster2	cluster3	cluster4
1	> 0.9	M5	showroom, 3Series, Z4	showroom, 3Series	dealership	M5, *financing*, *purchase*

seed		cluster0	cluster1	cluster2	cluster3	cluster4
	< 0.1	purchase	dealership, computer search, M5	M5	3Series	-
2	> 0.9	M5, financing	showroom, 3Series	computer search, financing	showroom, 3Series, *financing*, *purchase*	dealership
	< 0.1	computer search	purchase	-	dealership, M5	*financing*, *purchase*
3	> 0.9	showroom, 3Series	showroom, 3Series, financing	M5, financing	dealership	dealership, financing
	< 0.1	dealership, M5, purchase	dealership	Z4	*financing*, *purchase*	-

There is some consistency across the different clusterings, for example the pairs "showroom" and "3Series", "financing" and "purchase", "M5" and "financing" often cooccur among the significant dimensions for a cluster (either both are more than 0.9 or both less than 0.1), this is probably due to some correlations, for example one must enter a show room to look at a model, one tend to enquired about financing a car before completing the purchase, (or perhaps the other way around, one is more likely to make the enquiry if they are at least considering purchasing the car,) and one who looked at the M5 model is more likely to be interested in financing (because it is too expensive).

We may also observe that for seed = 1, cluster 1 and 2 are very similar (they have a lot of significant dimensions in common and no conflicting ones), and for seed = 3 there are cluster 0 and 1. Chances are they could be merged and we might not need as many as 5 clusters.

(c) The significant dimensions for each cluster are:

seed		cluster0	cluster1	cluster2	cluster3	cluster4
1	> 0.9	showroom, M5	dealership	dealership, showroom, M5, financing	computer search, financing	M5, *financing*, *purchase*
	< 0.1	dealership, 3Series	*financing*, *purchase*	computer search	-	Z4
2	> 0.9	showroom	showroom, computer search	showroom, M5	-	dealership, computer search, financing
	< 0.1	purchase	dealership, 3Series	-	showroom, *financing*, *purchase*	showroom

seed		cluster0	cluster1	cluster2	cluster3	cluster4
3	> 0.9	-	showroom, 3Series	-	dealership, computer search	3Series, financing
	< 0.1	3Series	dealership, purchase	*financing*, *purchase*	purchase	-

Here going into the showroom also cooccurs with looking at specific models (M5 or 3Series), so do enquiring about finance and making the purchase. Clustering using K Means generally have fewer significant dimensions than using the EM algorithm (especially as the seed value increases), in such case the attributes are less divided by the clustering.

Problem Set 5

1. (a) For 10-fold cross validation, take only one of the 10 point as test data and the rest as train data, $k = 1$, so every test point will be classified according to its closest point, which belongs the opposite class. None of the ten tests would be correct, so the cross validation error is 1.

   (b) We mark the points as $A(1, 3)$, $B(1, 4)$, $C(2, 4)$, $D(2, 5)$, $E(3, 1)$, $F(3, 5)$, $G(4, 1)$, $H(4, 2)$, $I(5, 2)$, $J(5, 3)$. $\{A, C, F, G, I\}$ belong to $\circ$, $\{B, D, E, H, J\}$ belong to $+$.

   i. $k = 3$, A's 3 nearest neighbours are $\{B, C, D\}$, class $+$, wrong;

   B - $\{A, C, D\}$, $\circ$, wrong; C - $\{B, D, A \ or \ F\}$, $+$, wrong;

   D - $\{B, C, F\}$, $\circ$, wrong; F - $\{B, C, D\}$, $+$, wrong;

   The same for $\{E, G, H, I, J\}$, the 10-fold cross validation error is 1.

   ii. $k = 4$, A's 4 nearest neighbours are $\{B, C, D, E \ or \ F\}$, suppose we choose randomly from E and F, and if F is chosen (a tie) we break it randomly, both with 50% probability. Then the chance is 25% to get it right and 75% to get it wrong.

   B - $\{A, C, D, F\}$, $\circ$, wrong; D - $\{A, B, C, F\}$ - $\circ$, wrong; F - similar to A;

   C - $\{A, B, D, F\}$, break the tie randomly, 50% probability to get it wrong.

   The same for $\{E, G, H, I, J\}$, error = $(0.75 + 1 + 0.5 + 1 + 0.75)/5 = 0.8$

   iii. $k = 5$, A's 5 nearest neighbours are $\{B, C, D, E, F\}$, class $+$, wrong;

   B - $\{A, C, D, F, E \ or \ H\}$, $\circ$, wrong; D - the same as B; F - the same as A;

   C - $\{A, B, D, F, H\}$ - $+$, wrong; the same for $\{E, G, H, I, J\}$, error = 1.

   iv. $k = 9$, every test point will be classified according to all other 9 points, 5 of which belong the opposite class. None of the tests would be correct, error is 1.

   $k = 4$ yields the minimum number of 10-fold cross validation errors.

(c) For (x_1, y_1) and (x_2, y_2), define $d(1,2) = \sqrt{(x_1 - 1 - x_2)^2 + (y_1 - 1 - y_2)^2}$

A's nearest neighbour is B, class $+$, wrong; B's is A, class $\circ$, wrong; C's is A, class $\circ$, correct, D's is B, class $+$, correct; E's is G, class $\circ$, wrong; F's is C, class $\circ$, correct; G's is E, class $+$, wrong; H's is E, class $+$, correct; I's is G, class $\circ$, correct; J's is H, class $+$, correct. The 10-fold cross validation error of 1-NN is $4/10$. Basically this is still Euclidean distance, but we move the test point one unit down the x bar and another unit down the y bar before we set out to calculating distances and finding its nearest neighbour.

2. (a) Consider the points $\{1, 2, 3, 4\}$ in the 1-dimensional Euclidean space, the optimal clustering for $K = 2$ is having 1 and 2 in one cluster while 3 and 4 in the other. Starting with one centre from $\{1, 2\}$ and the other from $\{3, 4\}$ is sure to give us the desired result in just one round. Suppose 1 and 2 are selected as the initial centres, after the first iteration we have clusters $\{1\}$ and $\{2, 3, 4\}$, the new centre of the latter will be updated to 3, after the second iteration 2 will move to cluster $\{1\}$ and we get our optimal clustering. Similarly, starting with 3 and 4 as initial centres, we will get $\{1, 2, 3\}$ and $\{4\}$ after one round, the centre of the first cluster will be updated to 2, and $\{1, 2\}$, $\{3, 4\}$ will be ready in another round. Lloyd's K-means algorithm always yields the optimal clustering for $K = 2$.

(b) Suppose we begin with 1 and 2 again. After one iteration we have the clusters $\{1\}$ and $\{2, 3, a\}$, and new centre for the latter would be $\frac{a+5}{3}$. There are three cases:

i. $3 < a < 4$, so that $|2 - \frac{a+5}{3}| < |2 - 1|$, iteration ends here. If this result is not optimal then we assume either $\{1, 2\}$, $\{3, a\}$ or $\{1, 2, 3\}$, $\{a\}$. Consider that $a < 4$, $\frac{1}{2} + \frac{(a-3)^2}{2} < 2$, so it should be $\{1, 2\}$, $\{3, a\}$ rather than the other, we can write:

$\frac{1}{2} + \frac{(a-3)^2}{2} < (2 - \frac{a+5}{3})^2 + (3 - \frac{a+5}{3})^2 + (a - \frac{a+5}{3})^2$

Solving the above inequality would give us $(a-1)^2 > 3$ which $3 < a < 4$ naturally satisfies. So we can simply add a small ε to 3 to satisfy the sub-optimal requirement, say $a = 3.05$, so that $a - 0.1$ would be be smaller than 3 and thus not a valid solution.

ii. $4 < a < 6$, so that $|2 - \frac{a+5}{3}| > |2 - 1|$ and $|3 - \frac{a+3}{2}| < |3 - \frac{3}{2}|$, 2 will be pulled to $\{1\}$ but 3 will remain with a, the clusters we get are $\{1, 2\}$ and $\{3, a\}$.

iii. $a > 6$, so that $|3 - \frac{a+3}{2}| > |3 - \frac{3}{2}|$, 2 will be pulled to $\{1\}$ and 3 will be pulled to $\{1, 2\}$ in the next iterations, so eventually we will get $\{1, 2, 3\}$ and $\{a\}$.

Since we look for a small a there's no need to dive into the other two cases.