Assignment 5

12/01/2020

Name: Yusuf Elnady

Problem 1. K-means Clustering

P1.1

After the first iterations we will have Clusters A = {0.1}, B={0.2,0.4}, C={0.5,0.6,0.8,0.9}, then we calculate and update the centroids to be $c_A = 0.1, C_B = 0.3, C_C = 0.7$

After the second iteration we will still have the same clustering and the same centroids **A = (0.1), B=(0.2,0.4), C=(0.5,0.6,0.8,0.9), then we update the centroids to be** $\mathbf{c}_A = 0.1, C_B = 0.3$,$\mathbf{C}_C$ **= 0.7**

and so the same also for the Third Iteration. So, that's conclude the answer.

P1.2

SSE = 0.01+0.01+0.04+0.01+0.01+0.04 = 0.12

P1.3

First, we start by having two clusters C_1 and C_2 with centroids A(0.1) and B(0.9), so C_1 will get these points : (0.1,0.2,0.4) and C_2 will get (0.5,0.6,0.8,0.9), so SSE of C_1 is $0 + 0.01 + 0.09 = 0.1$, and SSE of C_2 is $0.16 + 0.09 + 0.01 + 0 = 0.26$

So, we will split the cluster that has the largest SSE into C_2 and C_3 with the centroids A(0.5) and B(0.9) respectively. C_2 will get the points (0.5, 0.6) and C_3 will get the points (0.8,0.9).

Finally, the total SSE for all clusters is is $0.1 + (0 + 0.01) + (0.01 + 0) = 0.12$

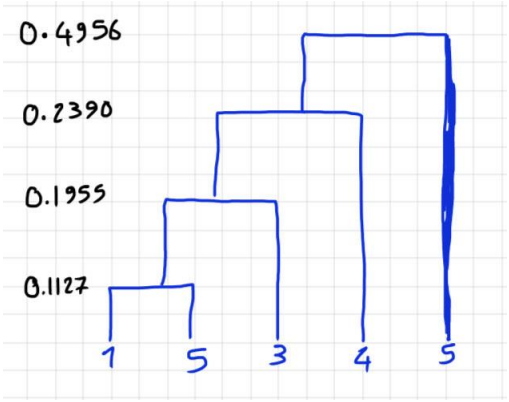
Iter	0.1	0.2	0.4	0.5	0.6	0.8	0.9	A	B
0	—	—	—	—	—	—	—	0.1	0.9
1	0.1	0.1	0.1	0.9	0.9	0.9	0.9	0.5	0.9
2	0.1	0.1	0.1	0.5	0.5	0.9	0.9	0.1	0.4

P1.4

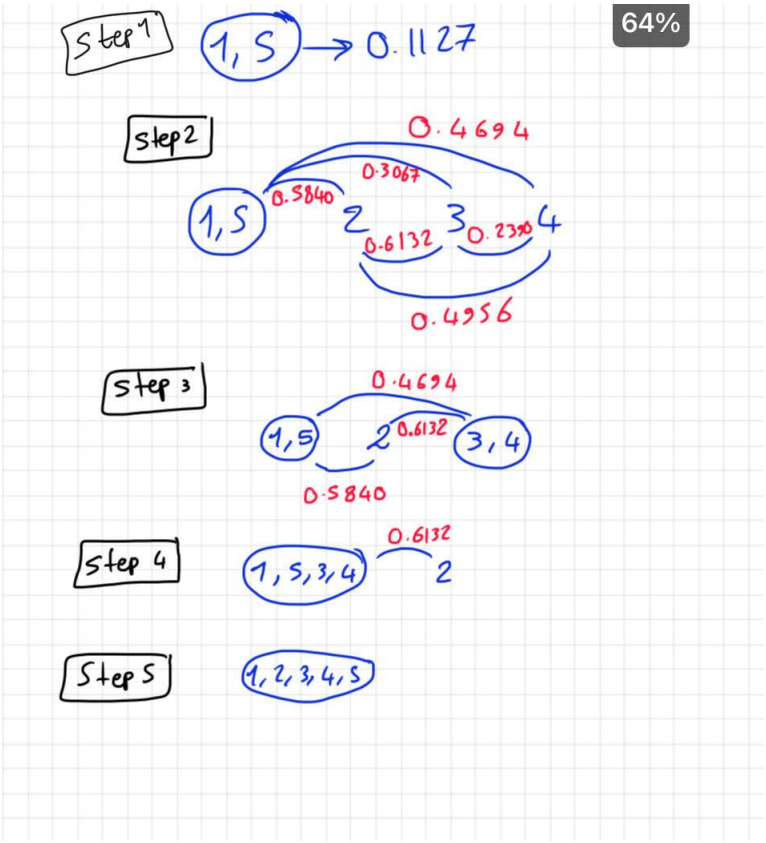
As both K-means and Bisecting K-means gave the same SSE, so for this dataset, they are all equal in terms of their performance.

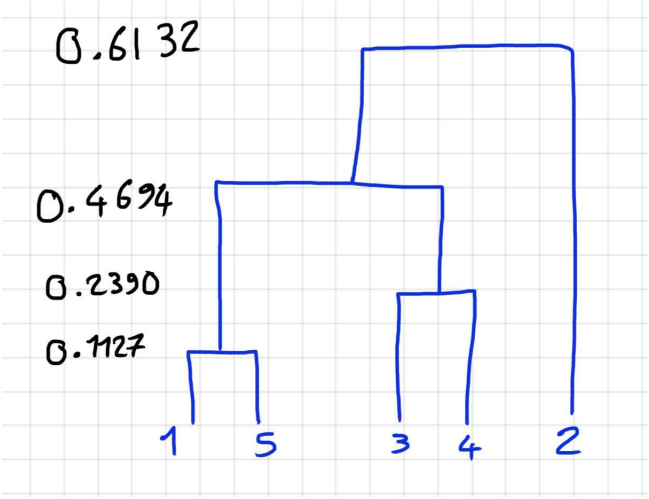
Problem 2. Dendograms

Single Linkage: We start by looking for the smallest numbers which is 0.1127, so we merge 1 and 5 into (1,5). Next, we have 0.1955 so we merge (3) into (1,5) to get the cluster (1,3,5).
Next, we will merge 4 into (1,3,5) as it has the minimum distance with 0.2390, so we get the cluster (1,4,3,5). Finally, we just have to group all into one cluster with having (1,2,3,4,5), as the remaining distance is 0.4956.



Complete Linkage: Please see the steps in the following pictures:





Problem 3. DBSCAN Clustering

P3.1

The core points are: **a,b,c,d,e,f,g,h,i,j,k,l,q,r,s,t,x**

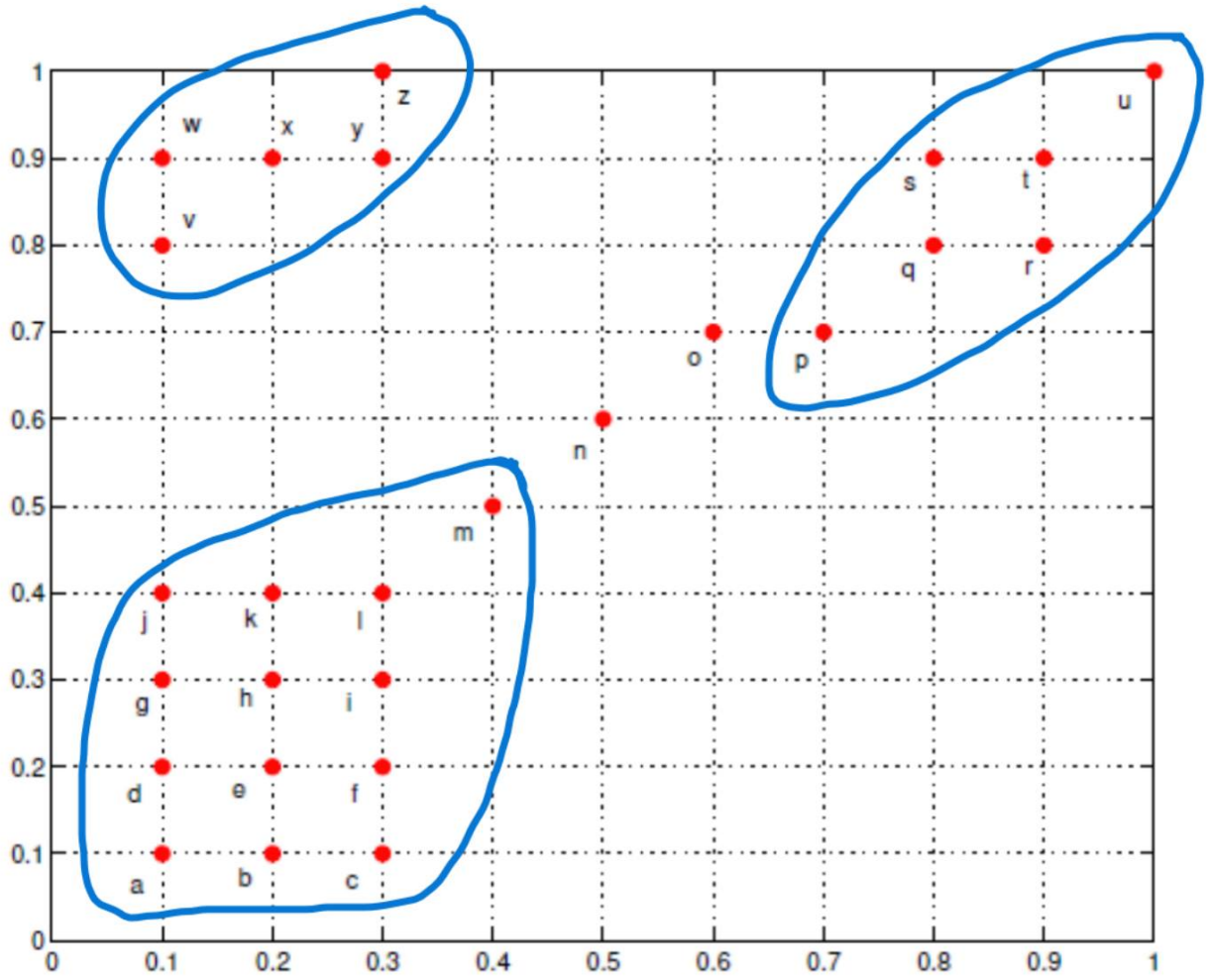
P3.2

The border points are: **m,p,u,v,w,y and z**

P3.3

The noise points are: **n, o**

P3.4



Problem 4. Confusion Matrices

P4.1

For the pure clusters, that mean both of the clusters have exactly 50 data points assigned to the correct class. Calculating the entropy we get the following: $Entropy(Cluster1) = \frac{50}{100} * (\frac{50}{50} \log \frac{50}{50}) = 0$, and also $Entropy(Cluster2) = \frac{50}{100} * (\frac{50}{50} \log \frac{50}{50}) = 0$, so the **Total Entropy = 0+0 = 0**.

For the purity we have the following: $Purity(Cluster1) = \frac{50}{100} * \frac{50}{50} = \frac{1}{2}$, and also $Purity(Cluster2) = \frac{50}{100} * \frac{50}{50} = \frac{1}{2}$, so the **Total Purity = 1.**

For the normalized mutual information, we can start by calculating $H_1 = -(\frac{50}{100}\log\frac{50}{100}+\frac{50}{100}\log\frac{50}{100}) = 1$, and the same for $H_2 = -(\frac{50}{100}\log\frac{50}{100}+\frac{50}{100}\log\frac{50}{100}) = 1$, Finally $\mathbf{NMI} = = 2[(\frac{50}{100}\log\frac{50*100}{50*50}+\frac{50}{100}\log\frac{50*100}{50*50}]/(H_1+H_2) = 2/2$**=1.**

P4.2

For the first clustering, when calculating the Entropy:

$$Entropy(Cluster1) = \frac{60}{100} * (-1)(\frac{40}{60}\log\frac{40}{60}+\frac{20}{60}\log\frac{20}{60}) = 0.551$$

$$Entropy(Cluster2) = \frac{40}{100} * (-1)(\frac{10}{40}\log\frac{10}{40}+\frac{30}{40}\log\frac{30}{40}) = 0.325$$

Entropy For First Clustering = 0.876

For the Second clustering, when calculating the Entropy:

$$Entropy(Cluster1) = \frac{50}{100} * (-1)(\frac{35}{50}\log\frac{35}{50}+\frac{15}{50}\log\frac{15}{50}) = 0.441$$

$$Entropy(Cluster2) = \frac{10}{100} * (-1)(\frac{5}{10}\log\frac{5}{10}+\frac{5}{10}\log\frac{5}{10}) = 0.1$$

$$Entropy(Cluster3) = \frac{40}{100} * (-1)(\frac{10}{40}\log\frac{10}{40}+\frac{30}{40}\log\frac{30}{40}) = 0.325$$

Entropy For First Clustering = 0.866.

The second solution is the best as it has the lower entropy.

P4.3

For the first clustering, when calculating the Purity:

$$Purity(Cluster1) = \frac{60}{100} * \frac{40}{60} = 0.4$$

$$Purity(Cluster2) = \frac{40}{100} * \frac{30}{40} = 0.3$$

The Purity for the first clustering = 0.7

For the Second clustering, when calculating the Purity:

$$Purity(Cluster1) = \frac{50}{100} * \frac{35}{50} = 0.35$$

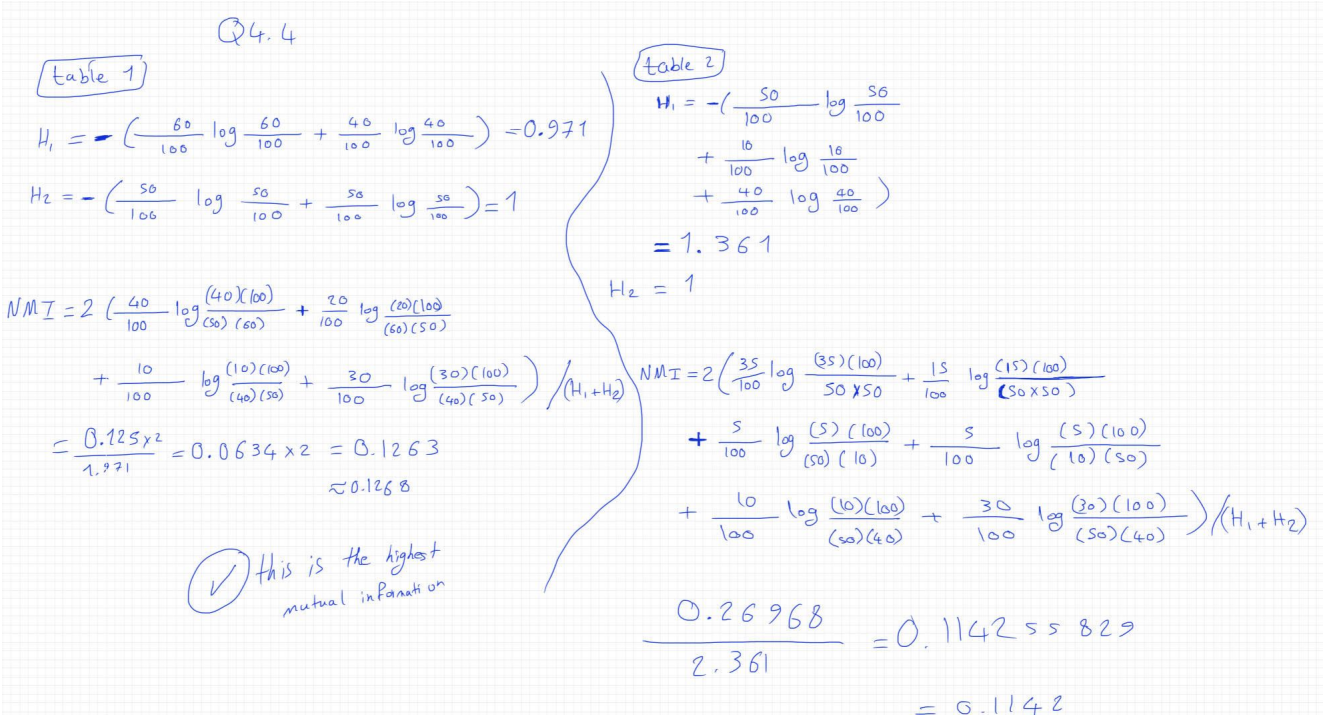
$$Purity(Cluster2) = \frac{10}{100} * \frac{5}{10} = 0.05$$

$$Purity(Cluster3) = \frac{40}{100} * \frac{30}{40} = 0.3$$

The Purity for the second clustering = 0.7

Both Solutions has the same Purity.

P4.4



P4.5

NMI is better, the details are in the quiz.

Problem 5. K-means Clustering Code

Please see the attached Jupyter Notebook or see the attached pictures in the quiz.

Problem 6. Regression Analysis Code

Please see the attached Jupyter Notebook or see the attached pictures in the quiz.