CSE 575: Statistical Machine Learning Assignment #3
Sushant Trivedi (1213366971)

1. KMEANS

a. Prove that the above procedure will converge in finite steps?

For N data points into k clusters, the total no of clusters possible are finite (equal to k^N). In K-Means, in each iteration new clusters are based only on the old clustering resulting in following scenarios:

1. New cluster is same as the old cluster. Thus, same cost.
2. The new cluster is different from old cluster. The new cluster has lower cost.

In order to reach convergence, it must enter 1st situation in a loop. If its in 2nd situation, it will lead to a cluster with minimum cost as there are finite no of possible clusters. Once it enters the cluster combination with min cost, then it will enter the 1st situation. Thus, the K-Means procedure will converge in finite steps.

b. Suppose we run Kmeans on the following dataset with six data points (i.e., the six black dots) to find two clusters. If we set the initial cluster centers are μ_1 = (0, 1) and μ_2 = (3, 3).

How many iterations does the algorithm take until convergence (3 points)?

It takes 3 steps to converge.

If we only run Kmeans for one iteration, what is the cluster assignment for each data point after Step-a (4 points)?

μ_1 = (1/2, 3/2) -> (0,0), (0,2), (2,0), (0,4)
μ_2 = (2, 3) -> (2,4), (2,2)

What are the updated cluster centers after Step-b (3 points)?

μ_1 = (1/2, 3/2)
μ_2 = (2, 3)

2. MIXTURE MODEL FOR DOCUMENT CLUSTERING

What is P (d=d1)? (You can use the following fact: $0^0 = 1$, $1^0 = 1$ and $0^1 = 0$).

$$P(d = d_1) = P(\theta = \theta_1)P\left(\frac{w_1}{\theta_1}\right)^1 P\left(\frac{w_2}{\theta_1}\right)^1 P\left(\frac{w_3}{\theta_1}\right)^1 P\left(\frac{w_4}{\theta_1}\right)^0 P\left(\frac{w_5}{\theta_1}\right)^0 P\left(\frac{w_6}{\theta_1}\right)^0 + P(\theta = \theta_2)P\left(\frac{w_1}{\theta_2}\right)^1 P\left(\frac{w_2}{\theta_2}\right)^1 P\left(\frac{w_3}{\theta_2}\right)^1 P\left(\frac{w_4}{\theta_2}\right)^0 P\left(\frac{w_5}{\theta_2}\right)^0 P\left(\frac{w_6}{\theta_2}\right)^0$$

$$= 0.5 * 0.2 * 0.3 * 0.3$$

$$= 0.009$$

We run EM algorithm one iteration. What is the updated $E[Z_{ij}]$ (i=1 to 7 ; j = 1,2) after the E-step?

Please note that the calculations are enclosed at the end.

	θ_1	θ_2
D1	1	0
D2	1	0
D3	1	0
D4	1	0
D5	0	1
D6	0	1
D7	0	1

What is the updated $P(\theta_1)$, $P(\theta_2)$, θ_1, and θ_2, respectively, after the M-step?

$P(\theta_1) = 4/7$
$P(\theta_2) = 3/7$
$\theta_1 = [2/7, 2/7, 2/7, 0, 0, 1/7]$
$\theta_2 = [0, 0, 0, 3/8, 3/8, 1/4]$

If we run EM algorithm for another iteration, What is the updated $E[Z_{ij}]$ (i=1 to 7 ; j = 1,2) after the E-step?

	θ_1	θ_2
D1	1	0
D2	1	0
D3	1	0
D4	1	0
D5	0	1
D6	0	1
D7	0	1

3. OVERFITTING

a. Bias-Variance Tradeoff

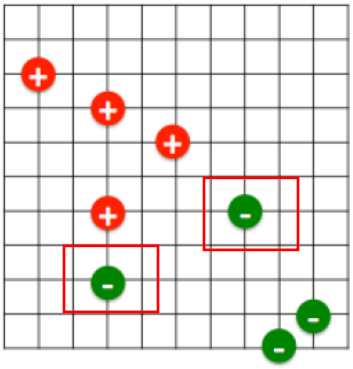
Assume, we are modeling a number of features using a machine learning model. Recall the relationship among bias, variance, and the complexity of hypothesis class, choose the correct answers in the following table.

	High Model Complexity		Low Model Complexity	
Bias	High	Low	High	Low
Variance	High	Low	High	Low

b. Leave-One-Out-Cross Validation

Suppose we use a linear SVM (i.e., no kernel), with some large C value. What is LOOCV for your SVM? Justify your answer.

Ans. LOOCV = 25 %.

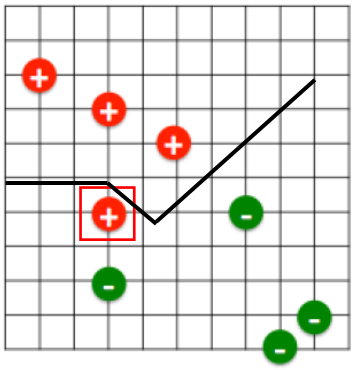


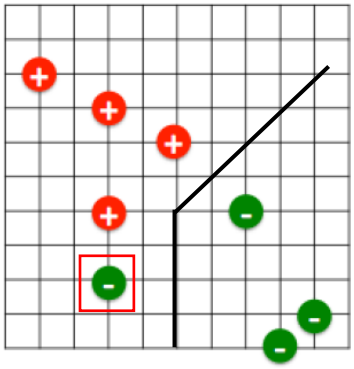
The marked out points when left out affect the decision boundary such that they are incorrectly labelled in the cross validation process.

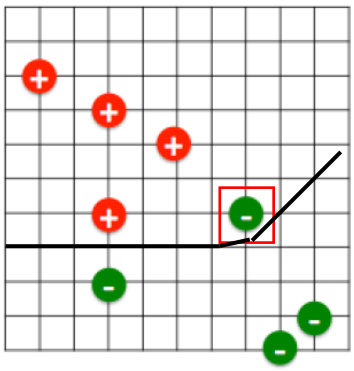
The large C value means that the SVM makes a hard margin.

Suppose We want to train 1NN, using L_2 distance. What is the LOOCV for your 1NN classifier? Justify your answers.

Ans. LOOCV = 3/8







4. PCA

Given First Component: $u = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$

Suppose we add one more data point at (4, 4), how would that affect the first principle component?
There would be no change as its in the same vector angle as the first component.

Suppose we add one more data point at (2, 0), how would that affect the first principle component?
It will skew(rotate) the first component towards the x-axis.

Suppose we add an infinite number of data points at (0, 5), how would that affect the first principle component?

With infinite data points at (0,5) the first component would be drastically rotated towards y-axis. Thus, the new first component would be $u = (0, 1)$

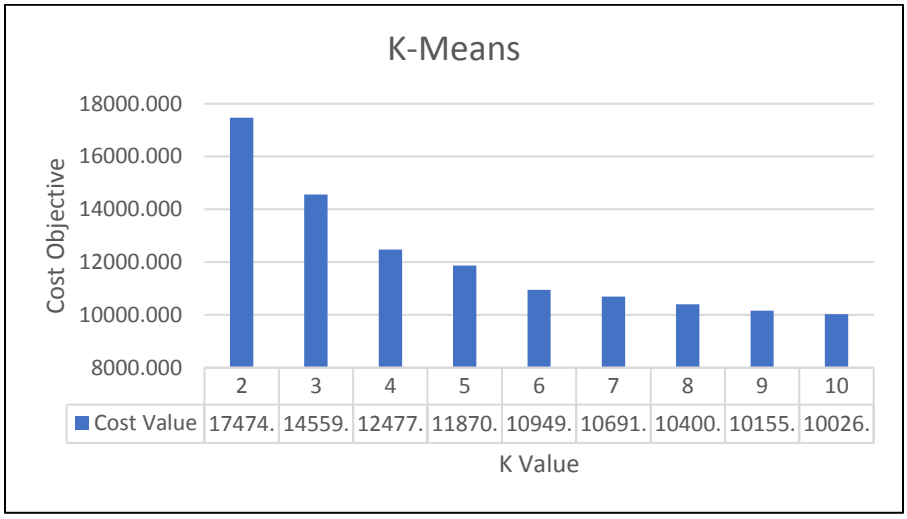
5. IMPLEMENTING KMEANS AND GAUSSIAN MIXTURE MODEL (GMM)

PART-1

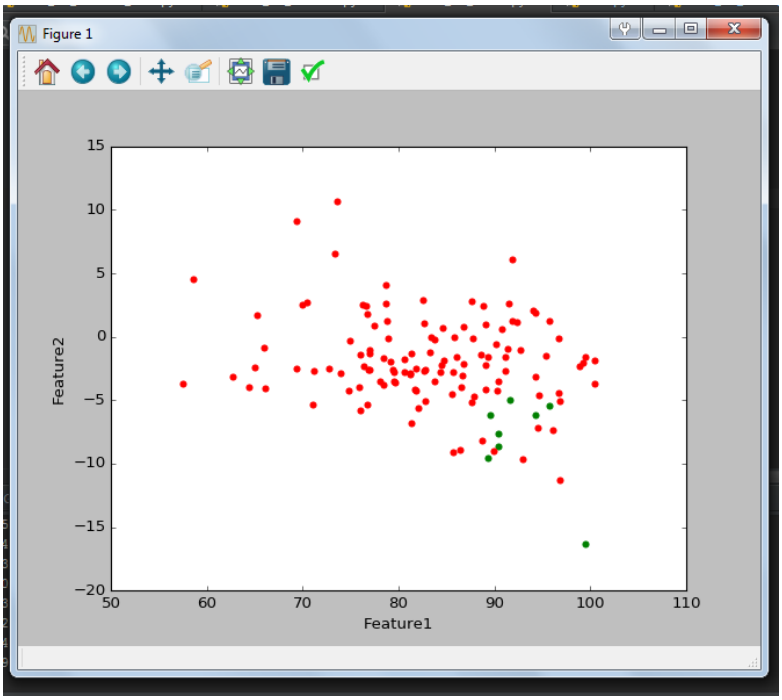
a) K-Means

We successfully implemented k-means algorithm and the cost function vs k values are as follows.

In order to deal with empty clusters convergence, we restart every time an empty cluster is present in the collection.



b) **GMM**



PART-2

a) K-Means

i. We first Normalize the data by using the following formula:

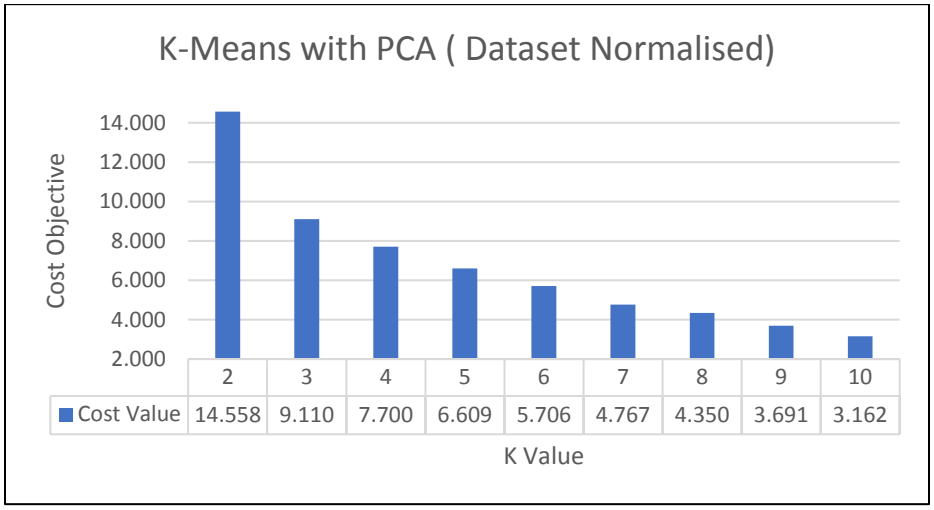
$$X\ (Feature\ Value) = \frac{X - F_{\min}}{Fmax - Fmin}$$ (Code Line 30)

ii. Following that we implement PCA Algorithm

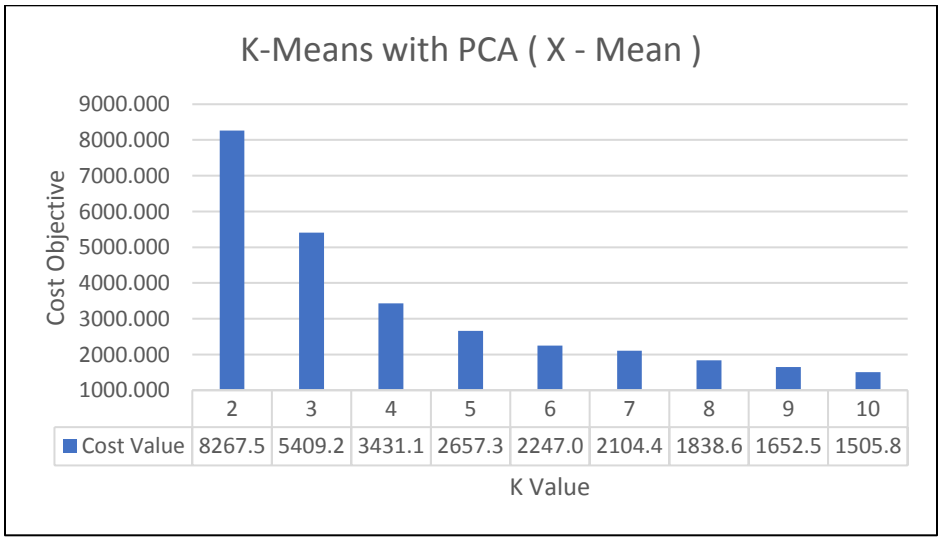
iii. Next, we apply K-Means Algorithm

NOTE: Please see that in code line 28 we had added np.random.seed(10). But that would be ineffective as we restart the iteration to compensate for empty clusters

RESULTS



Replacing the Step (i) with Feature Value – Mean Feature Value



b) GMM

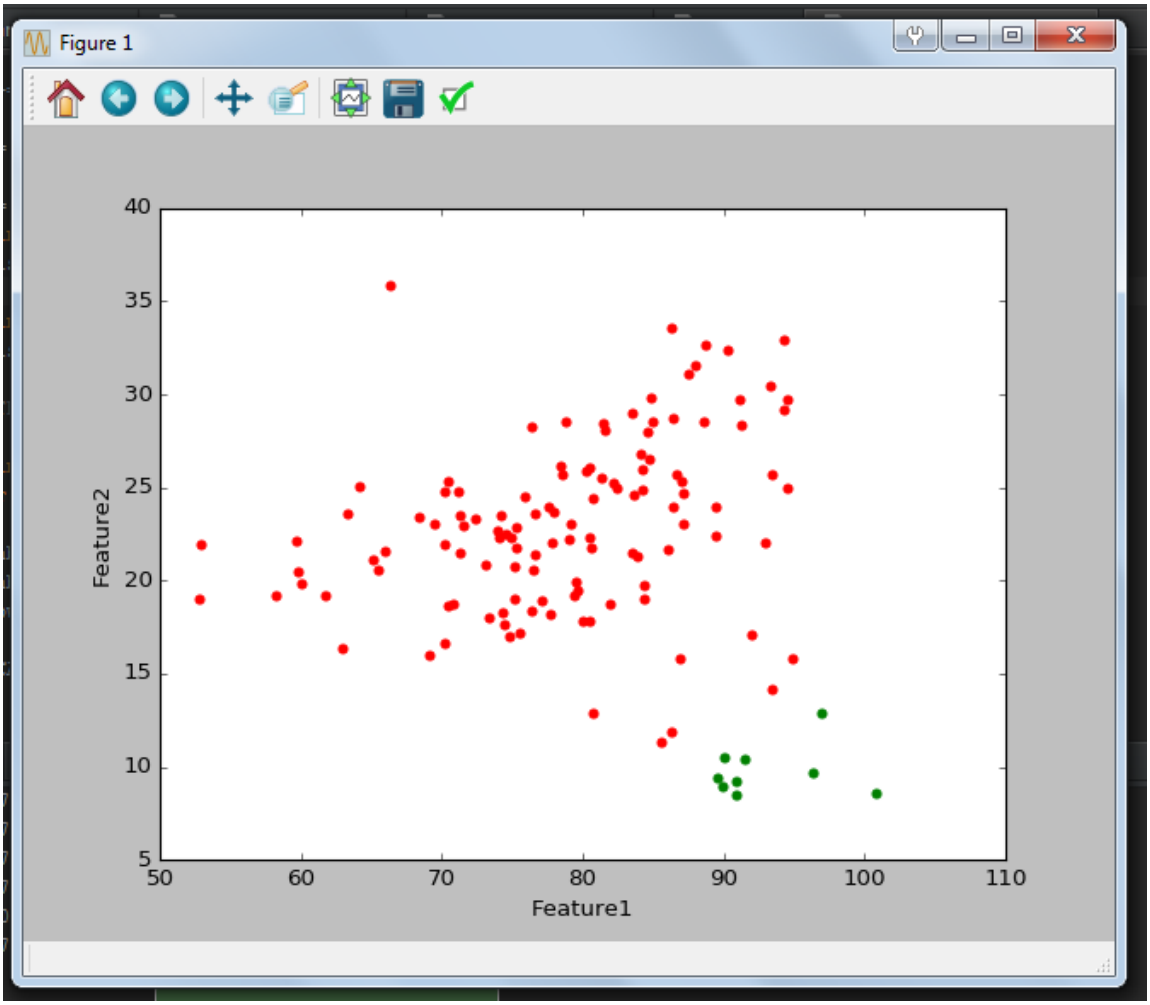


Figure 1
40
35
30
25
20
15
10
5
Feature2
50
60
70
80
90
100
110
Feature1