# CSE 575 - ASSIGNMENT 3

**TEJA KUNDERU**

**1211384158**

**Question 1:**

**1.**

After each iteration of k means, new clustering will either have a lower cost or have the same cost as compared to the old clustering.

If cost remains same, this implies k means has converged.

There are N data points and K clusters.
So, there are (K)^N ways we can divide these points into K clusters.
In the worst-case scenario, the number of iterations k means requires to classify is finite.

Therefore, we can say that K means converges in finite number of iterations.

**2.**

Considering spherical Gaussian and hard assignments

$$\prod_{j=1}^{m}\sum_{i=1}^{k} p(x_j|y=i)\ \alpha \prod_{j=1}^{m} exp\left(-\frac{1}{2^{\sigma^2}}\left\|x_j-\mu_{c(j)}\right\|^2\right)$$

Taking log MLE of the above term
Max() = $\sum_{j=1}^{m}(-\frac{1}{2^{\sigma^2}}\left\|x_j-\mu_{(c(j))}\right\|^2)$
The minimization term can also be written as,
Min() = $\sum_{j=1}^{m}(\left\|x_j-\mu_{(c(j))}\right\|^2)$
Therefore argmin $\sum_{j=1}^{m}(\left\|x_j-\mu_{(c(j))}\right\|^2)$ represents $r_{n,k}$ in k means

$r_{n,k}$ = 1 if k = argmin $\sum_{j=1}^{m}(\left\|x_j-\mu_{(c(j))}\right\|^2)$
= 0 otherwise

Hence, proved that $p(z_k = 1 \mid x_n) = r_{n,k}$

**Question 2:**

**1.**

F must be of n x k size and G must be of k x d size.
Here, k is the number of clusters.
Additional constraints are that elements in the matrix F should be 0 or 1 and each row should sum up to 1. Also, each row of G should be a cluster center.

**2.**

The step (a) corresponds to the step of k-means algorithm where we adjust the cluster centers. Step (b) corresponds to starting with a random choice of points for the cluster centers in the k-means algorithm.

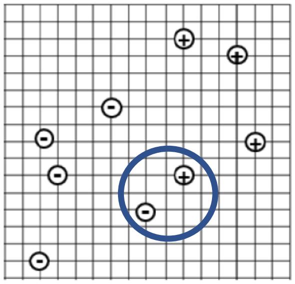
**Question 3:**

(a) LOOCV is 100%. Because, if we remove '-' point and try to put it back, it will be classified as '+' since nearest point is '+'.
Even if we take '+' point first, it will result in wrong classification.

(b) LOOCV is 100%. Because, if we remove any of the given data points and try to put the point back, the result will be a wrong classification.
This is because the nearest data point to the point under consideration is always of the opposite sign.

(c) LOOCV is 100%. Because, if we remove any of the given data points and try to put the point back, the result will be a wrong classification.
This is the result of the nearest data point to the point under consideration always having the opposite sign.

(d) LOOCV is (2/9) x 100 = 22.22%.
When compared to the original sign, the results will be of opposite sign for only two points in the following figure.
For other points, if we remove and put the data points back to their position, the nearest data point will be of the same sign.



**Question 4:**

**1.**

Classification will always be correct if points are not on the support vectors.
Only the points on the support vectors need to be considered.
If the '+' (2,2) point is removed, the support vector will shift to x1 = 1 from x1 = 2.
The hyperplane will be at x1 = 2.5 and therefore, if we put the point back, it will be correctly classified as positive.

If any of the '-' points (4,3) or (4,1) were to be removed, the support vector will remain unchanged and hence, there will be no classification error.

LOOCV for SVM is zero (since there are no errors).

**2.**

Upper bound on LOOCV is given by the equation -

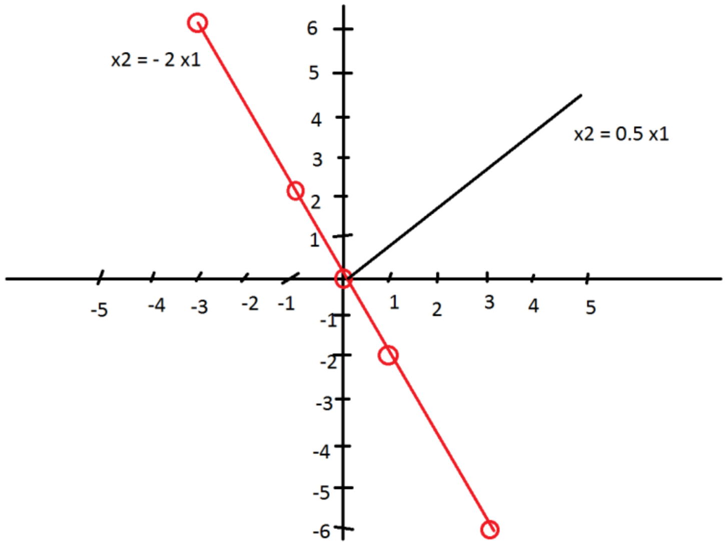
LOOCV error $\leq \frac{Number\ of\ support\ vectors}{n}$

Where, Number of support vectors = 2,
data points n = 9

Therefore, Upper bound of LOOCV of SVM is $\frac{2}{9}$ or (22.22 %)

**Question 5:**

**1.**



**2.**

By observation, all points together represent a line given by the equation,
PC1 => $x_2 = - 2 * x_1$

This line can be chosen as the first principle component as all the data points will have maximum variance on it.

**3.**

By the property of the principle component, all principle components are orthonormal to each other.

Hence, slope of PC1 x slope of PC2 = -1.

Therefore, PC2 => $x_2 = 0.5 * x_1$

**Question 6:**

We have 3 hidden states – location, person name, background
and 4 distinct obsercations say w1, w2, w3, w4.

**1.**

Corresponding state transtion matrix has a size of 3 x 3 = 9.

**2.**

The size of state observation probability matrix is 3 x 4 = 12.

**3.**

We have a paragraph containing 100 segments and each segment can have only 4 possible observations.

No. of obervations = 100.

The length of the path of state = No. of observations = 100.

**4.**

As w1 → c is fixed, we have $3^{99}$ total possible state paths for 100 observations.

**Question 7:**

**1. Pseudo Code**

- The data from the txt file is first converted to a .mat file called ‘data.mat’

- Data is stored such that the first row is the class and the rest are features.

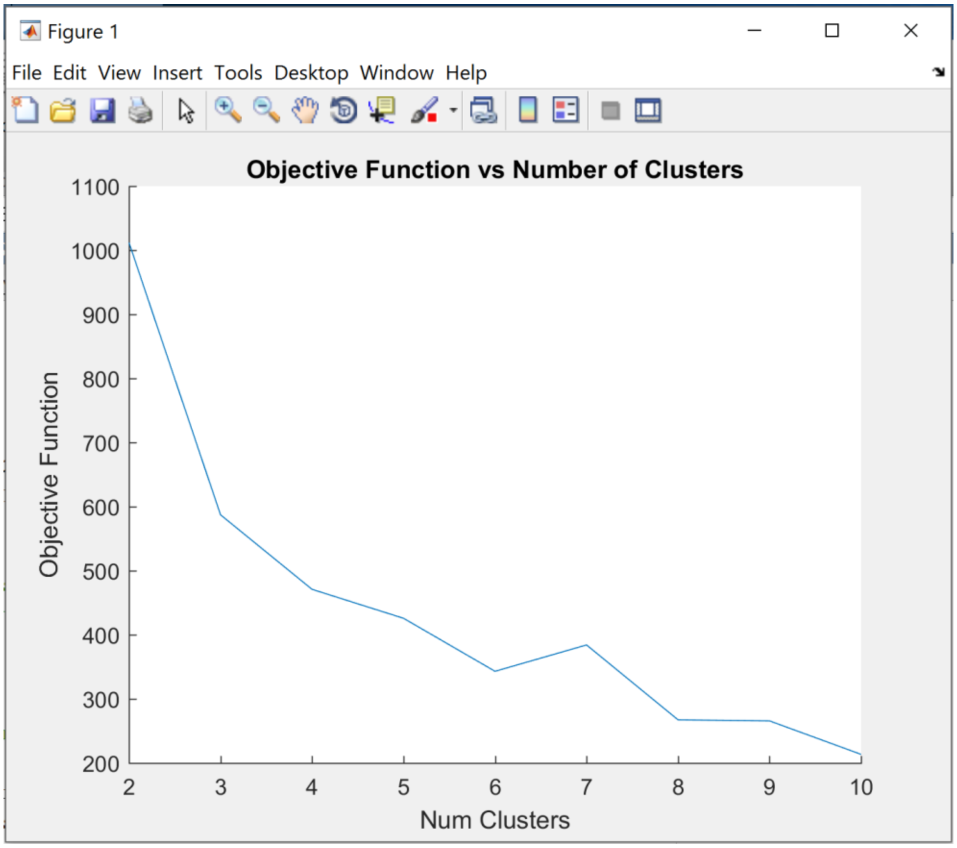
- A threshold a 0.0001 is chosen and the clusters are updated with new centers until the difference between the new and old centers is less than the threshold.

- Centers are randomly chosen at the beginning. All data points are allotted the closes center and the centers are updated by calculating the mean.

- The objective function is calculated and the process is repeated for all cluster numbers from 2 to 10.

- The objective function is then plotted against the number of cluster.

**2.**



We can notice from this plot that the objective function is minimum when the number of clusters is 10 and maximum when the number of clusters is 2.