CIS 519 Homework 5

Nikhil Jamdade

jnikhil@seas.upenn.edu

1. Purity and Rand Index for K means

	Train		Test	
Number of Clusters	**Purity**	**Rand Index**	**Purity**	**Rand Index**
K = 10	**0.1911**	**0.7566**	**0.2112**	**0.7618**
K=20	**0.2780**	**0.8255**	**0.2635**	**0.8298**
K=100	**0.3368**	**0.8824**	**0.3319**	**0.8804**
K=500	**0.17513**	**0.8986**	**0.3319**	**0.8943**
PCA-reduced data K =10	**0.19518**	**0.74945**	**0.2173**	**0.7405**
PCA-reduced data K =20	**0.27540**	**0.82533**	**0.26156**	**0.82074**

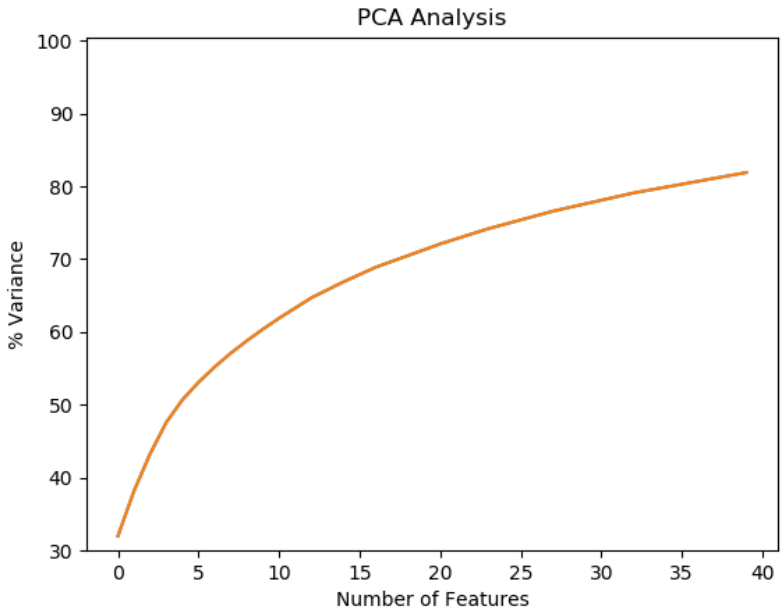
2. Purity and Rand Index for Gaussian Mixture Model

	Train		Test	
Number of Components/ Clusters	**Purity**	**Rand Index**	**Purity**	**Rand Index**
n = 10	**0.2045**	**0.77620**	**0.22736**	**0.7936**
n=20	**0.2352**	**0.81784**	**0.27162**	**0.8233**
n=100	**0.3302**	**0.8815**	**0.3239**	**0.8739**
n=500	**0.1724**	**0.8988**	**0.2152**	**0.68970**
PCA-reduced data n=10	**0.19786**	**0.77645**	**0.1971**	**0.7867**
PCA-reduced data N =20	**0.2633**	**0.8377**	**0.2494**	**0.82959**

Clustering or grouping of samples is done by minimizing the distance between sample and the centroid. i.e. Assign the centroid and optimize the centroid based on the distances from the points to it. A Gaussian mixture model attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset.

The Variance Matrix :

Variance Captured [32. 38.2 43.3 47.6 50.7 53.1 55.2 57.1 58.8 60.4 61.9 63.3 64.7 65.8
66.9 67.9 68.9 69.7 70.5 71.3 72.1 72.8 73.5 74.2 74.8 75.4 76. 76.6
77.1 77.6 78.1 78.6 79.1 79.5 79.9 80.3 80.7 81.1 81.5 81.9]



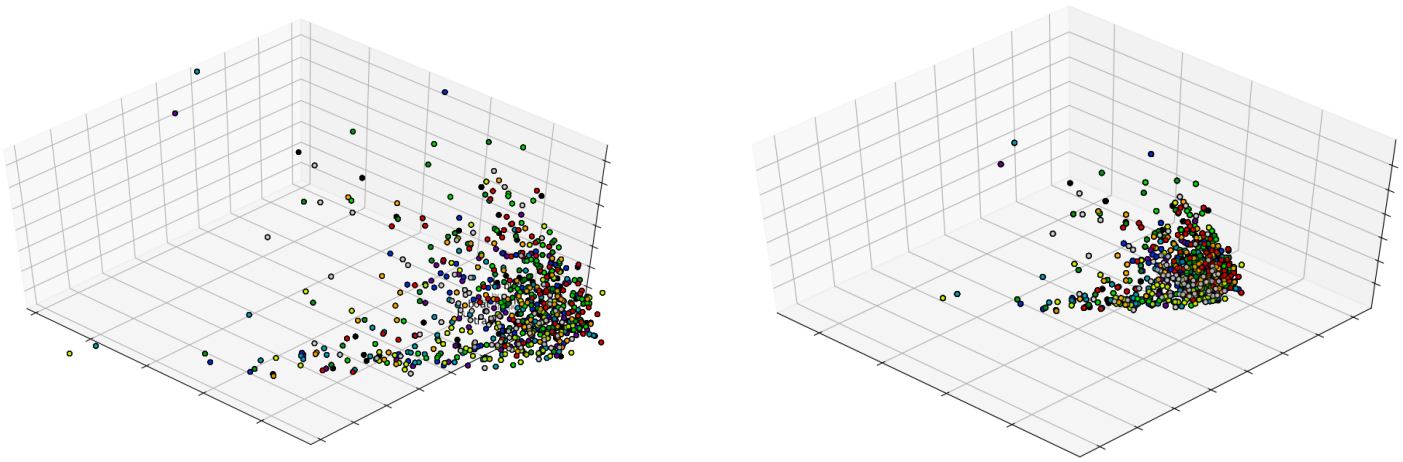
In the above array and figure we can say that the first feature explains roughly 32% of the variance within our data set while the first two explain 38.3 and 40th feature variance is 81.9 so on. If we employ 40 features we capture 82% of the variance within the dataset, thus we gain very little by implementing an additional feature

Efficiency: 0.009765

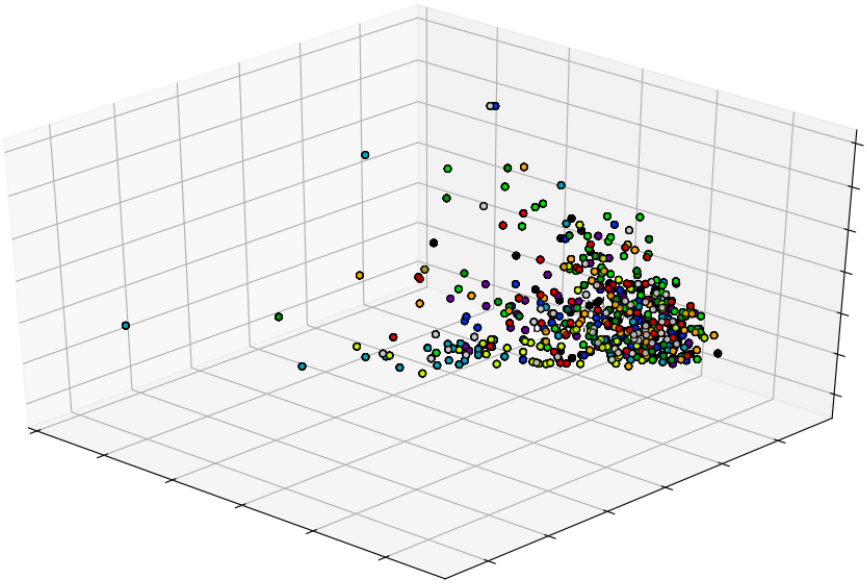
PCA can provide good compression ratios for the cost of implementation.

PCA n=3 3D Plot

Plot for train set



Plot for test set

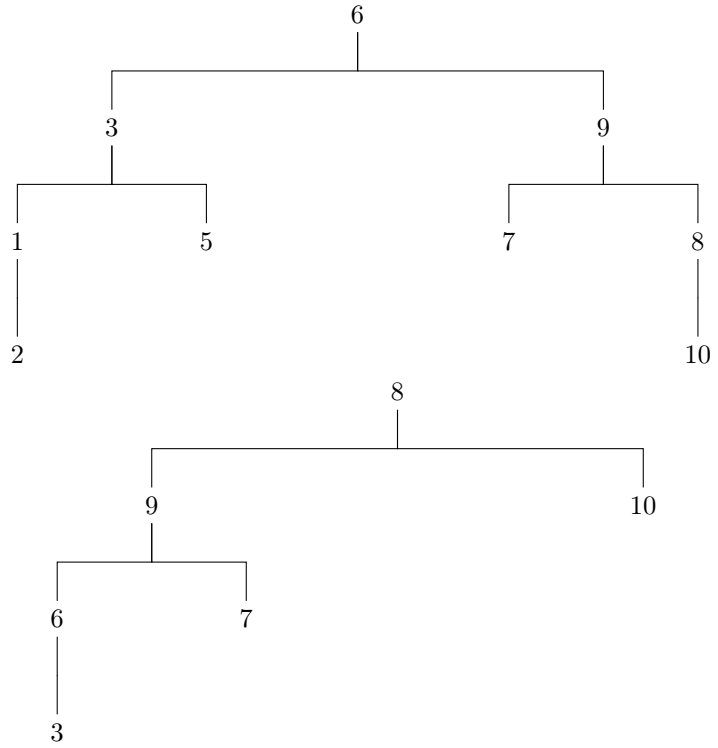


From the data we can say that, *k*-means is not flexible enough and tries to force-fit the data into clusters. This results in a mixing of cluster assignments where the resulting circles overlap. *k*-means—its lack of flexibility in cluster shape and lack of probabilistic cluster assignment. From the observation, we can say that Gaussian Mixture Clustering works better for given data set

1 Tree Dependent Distributionss

(1)
If we construct any two directed trees from the set of nodes, those are equivalent if only their joint probability distributions are equal.

i.e If we select one node from set of node forming a tree and generate a undirected tree then to be equivalent to any undirected tree constructed using any other node from same set of node, their joint probabilities must be equal



$$P_T6(x_1, x_2, ...x_n) = P_T8(x_1, x_2, ..., x_n)$$

For Gernaral Case

$$P_{T_0}(x_1, x_2, x_3...., x_n) = P_{T_1}(x_1, x_2, x_3...., x_n)$$

Where T0 and T1 are two trees formed by set of node

$$(x_1, x_2, x_3...., x_n)$$

(2)
Given: $P(x_i)$ iε 1.....m

$$Pr(x_j|x_i)j\varepsilon 1, 2, ...m$$

$$i\varepsilon 1, 2, ...mi! = j$$

$$i - - - - - j$$

Consider T_i and T_j as two trees constructed keeping x_i and x_j where i!=j from nodes x=$(x_1, x_2, x_3...., x_n)$

As per mentioned above $P_{T_i}(x) = P_{T_j}(x)$

πx : parent node

There is unique path between x_i and x_j As trees are directed, there have different direction between x_i and x_j

$$P_{T_i}(x_1, x_2, x_3...., x_n) = P(x_i) * \prod_{k=1,k!=i}^{n} Px_k/\pi_{x_k}$$

$$P_{T_i}(x_1, x_2, x_3...., x_n) = P(x_i) * P(x_j/x_i) \prod_{k=1,x_k \notin path}^{n} Px_k/\pi_{x_k}$$

Using the Bayes Theorem, the joint probability:

$$= P(x_i, x_j) \prod_{k=1,x_k \notin path}^{n} P(x_k/\pi_{x_k})$$

$$= P(x_j) * P(x_i|x_j) \prod_{k=1,x_k \notin path}^{n} P(x_k/\pi_{x_k})$$

$$= P(x_j) * \prod_{k=1,k!=j}^{n} P(x_k/\pi_{x_k})$$

$$P_{T_i}(x_1, x_2, x_3...., x_n) = P_{T_j}(x_1, x_2, x_3...., x_n)$$

Path is link between x_i and x_j and we assume that other edges which are not path have no directionality in both the trees. From the above, we can say Two trees have same joint distribution

Moreover the given node probability depends only on parent node and other nodes we can write Bayes rule for joint distribution of node independent of non descendant nodes