CSE 5523

# HOMEWORK V – Report

Pragya Arora

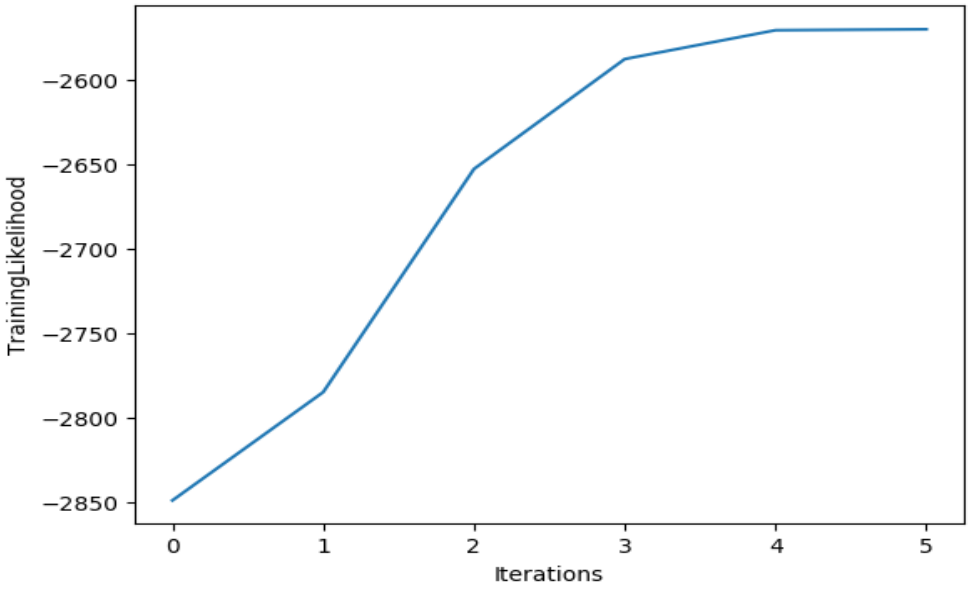
arora.170

# 1. Likelihood vs Iteration

The EM converges at 7.9 iterations approximately for 3 clusters. The train and test likelihood vs iterations graphs for 3 clusters are as follows:

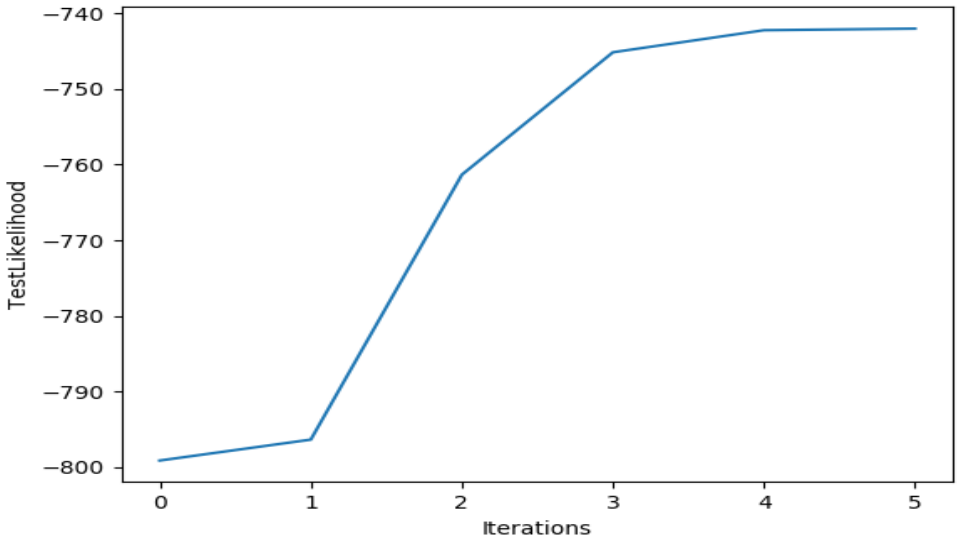
## 1.1. Train Dataset

The train dataset converges at -2570.255 (approx) for 3 clusters.




## 1.2. Test Dataset

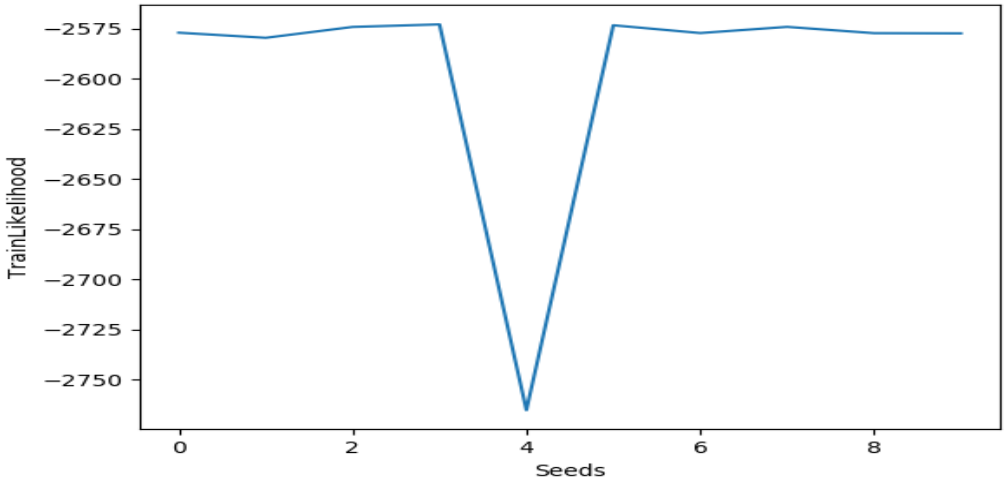
The test dataset converges at -742.328 (approx) for 3 clusters.



# 2. Likelihood vs Random Seeds

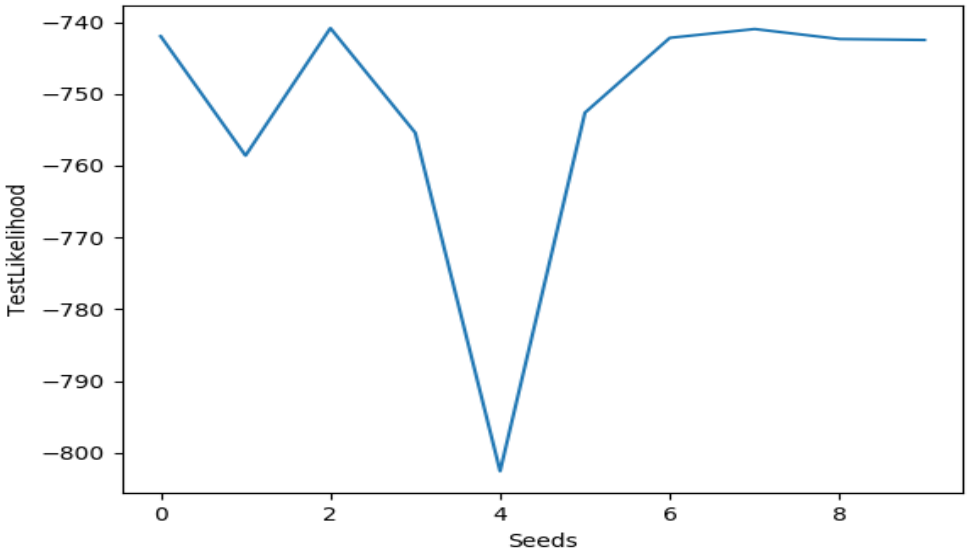
## 2.1. Train Dataset

The log likelihood converges at approximately same value (-2575) in most of the runs for cluster 3 but decreases sometimes when bad parameters are assigned.




## 2.2. Test Dataset

For the test dataset the log likelihood for cluster 3 varies on runs and drops highly when bad parameters are initialized.

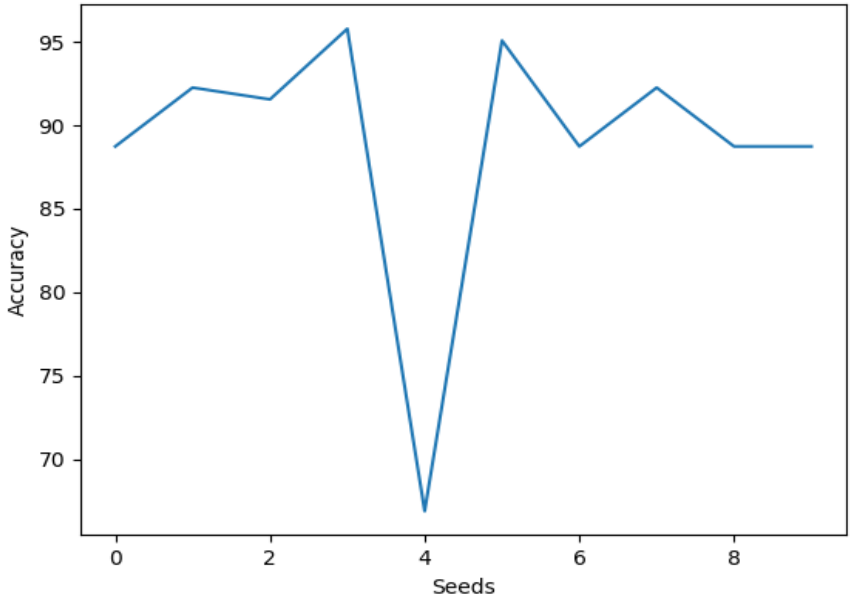


# 3. Accuracy

The most likely cluster is inferred by the following approach.

- For each datapoint probability is calculated for all the clusters and the cluster with the maximum probability is assigned to the datapoint.
- All the datapoints are grouped based on clusters.
- Then true labels of all the datapoints in the cluster is calculated.
- The true label with the maximum datapoints in the cluster is assigned as the predicted label to the datapoints in the cluster.

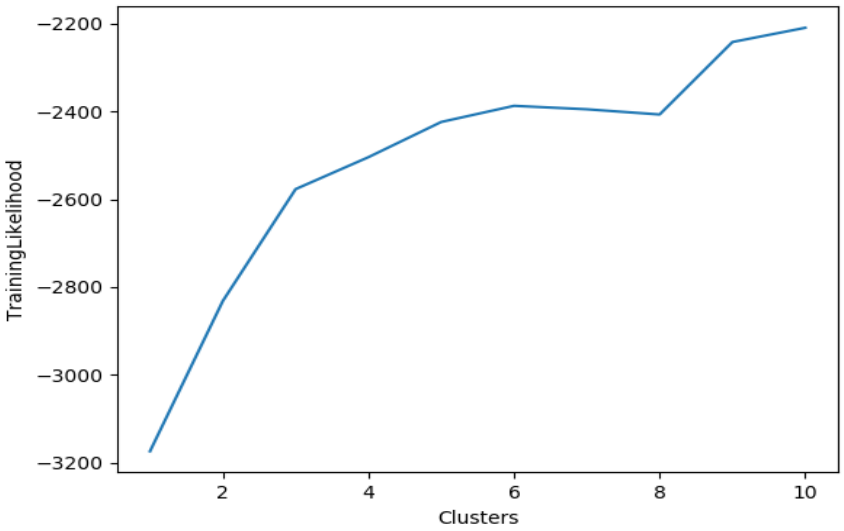
The maximum accuracy of **97.18309**% for 3 clusters is attained by the model predicting 138 out of 142 correctly.



# 4. Likelihood vs Clusters

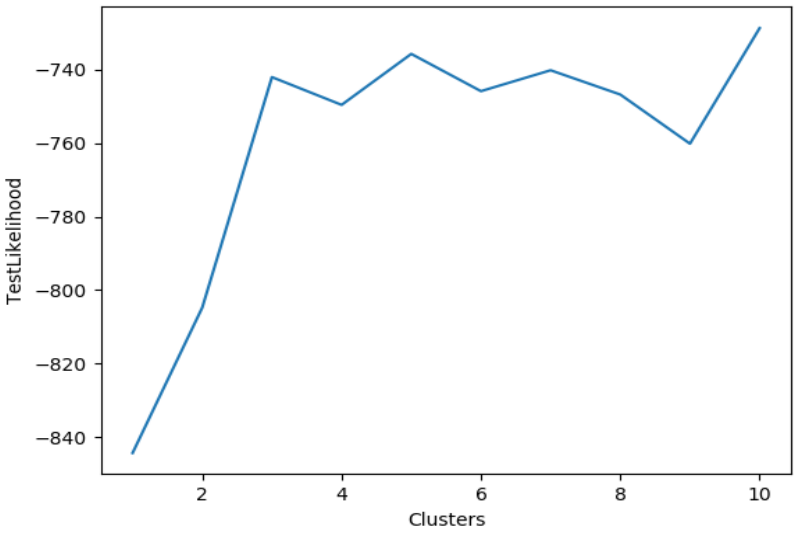
## 4.1. Train Dataset

The training likelihood generally increases with the number of clusters (fig 1). We can see that there is a sharp increase in likelihood when cluster changes from 1 to 3 and it increases slowly with further increase in clusters. The likelihood follows the same trend as accuracy (Section 4.3), the peaks and valleys in the likelihood are like the one in accuracy.



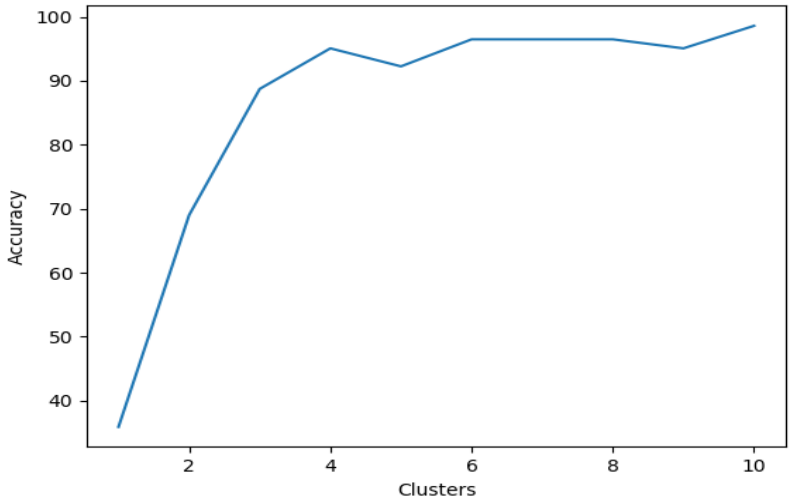

## 4.2. Test Dataset

The test set log likelihood increases sharply when clusters varies from 1 to 3 but with the further increase in clusters it flattens out and remains in the range of -740. The test likelihood follows the same trend as train likelihood.

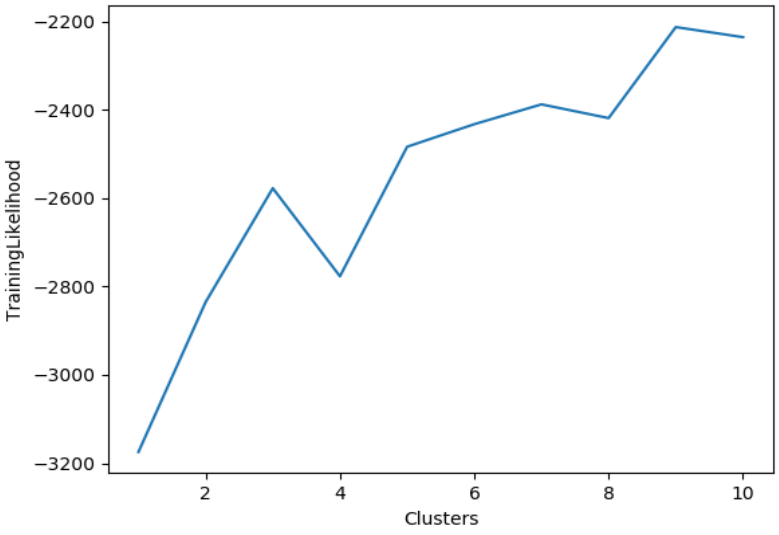


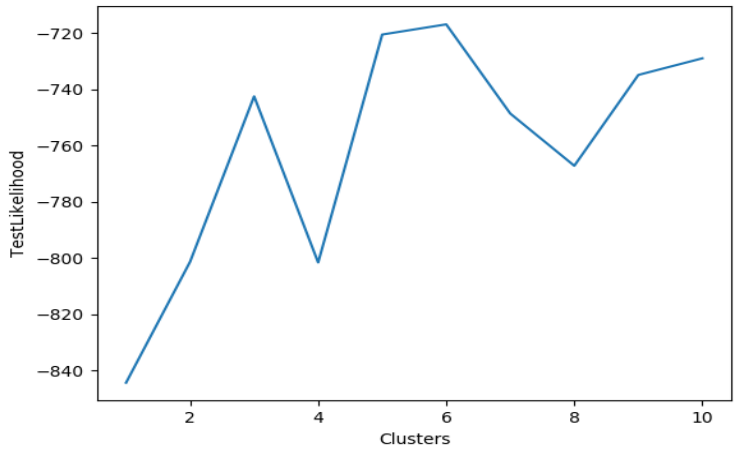
## 4.3. Dataset performance with true number of clusters

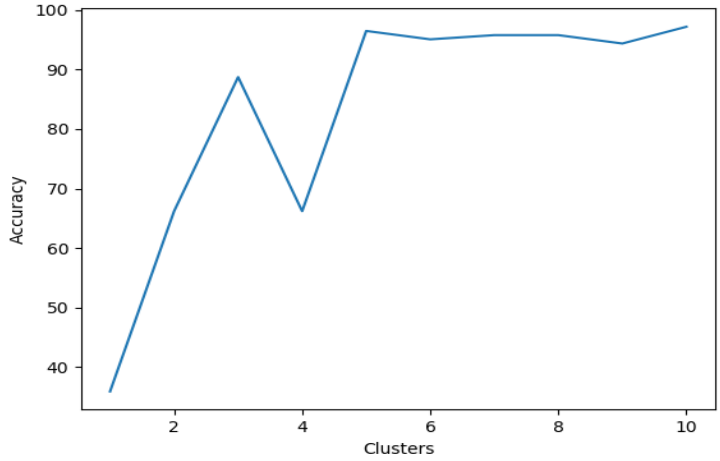
The training data accuracy increases sharply when cluster varies from 1 to 3 and then remain in the same range with the increase in clusters. At cluster 3 the maximum accuracy attained is 97.18% and at cluster 10 the maximum accuracy attained is 98.59%. The optimum value of cluster is attained at 3 that's why the accuracy did not change much.




The trend of train and test likelihood with the accuracy can be also seen in the following graphs:




−720
−740
−760
−780
−800
−820
−840
TestLikelihood
2
4
6
8
10
Clusters


100
90
80
70
60
50
40
Accuracy
2
4
6
8
10
Clusters