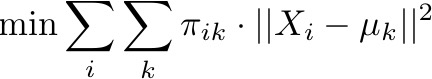
**DS 5230: Unsupervised Machine Learning and Data Mining**

Sarthak Kothari

*Homework 2*  *Date: 9th February, 2018*

**Problem 1:**

Given Kmeans Objective discussed in class with Euclidian distance



*Part A:* Prove that E step update on membership () achieves the minimum objective given the current centroids ()

Solution: Proof by contradiction.

Let us have 2 centroids. We have an array of N data-points (.

Let us assume that for a point :-

………………… **[1]**

We also know that,

We also assume that, at some point has been assigned to cluster with centroid which achieves the minimum objective during that step. Therefore, with this assumption, if was assigned to cluster with centroid , the objective function would comparatively greater.

Therefore,

+ + +

And Since we assumed that is assigned to ,

+ +

If was assigned to the equation would look like: -

+ +

Now we also assumed that assigned to obtains minimum objective,

+ + + +

Since –

Which is a contradiction to our initial claim [1].

Therefore, assigning to cluster with centroid with minimum difference from centroid would obtain the minimum objective during that step.

*Part B:* Prove that M step update on centroids () achieves the minimum objective given the current memberships ().

Solution: Proof by contradiction.

Let us consider, we are working with K=2 and after the E step we compute Let us assume that there exists a such that the difference of between all the data-points belonging to is higher than their difference from .

That is,

We also know that,

We assume that going ahead with helps us obtain the minimum objective function.

That is our minimum objective function with K=2 would look like –

+

However, if we take the centroid of k=1 as , our equation would look like –

+

Equating the two equations,

+ +

Which is a contradiction to our original claim [2].

Thus, updating the to the proper centroids achieves the global minimum.

*Part C:* Explain why KMeans has to stop (converge), but not necessarily to the global minimum objective value.

Solution:

KMeans depends heavily on the **Gradient descent**, that is after every iteration it tries to minimize the function to achieve the global minimum. However, if you don’t initialize the centroids correctly, it is possible that when the algorithm converges we will achieve local minimum and not global minimum.

That is because KMeans is an EM algorithm and if not correctly initialized at the end of every E-step we could assign the data-points to a particular cluster which wouldn’t yield the proper result which would in-turn result the M-step wherein we take the mean of the data-point belonging to the same cluster and recalculate the E step.