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

$$\min \sum_i \sum_k \pi_{ik} \cdot ||X_i - \mu_k||^2$$

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

Solution: Proof by contradiction.

Let us have 2 centroids $\mu_1 \; and \; \mu_2$. We have an array of N data-points $(X_1, \dots, X_n)$.
Let us assume that for a point X_j :-

$$\|X_j - \mu_1\| > \|X_j - \mu_1\| \quad \dots\dots\dots\dots\dots\dots \textbf{[1]}$$

We also know that,

$$\pi_{ik} = 1 \; if \; i \; belongs \; to \; cluster \; K, 0 \; otherwise.$$

We also assume that, at some point X_j has been assigned to cluster with centroid μ_1 which achieves the minimum objective during that step. Therefore, with this assumption, if X_j was assigned to cluster with centroid μ_2, the objective function would comparatively greater.

Therefore,

$\sum_{i=0}^{j-1}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2+\pi_{j1}.\left\|X_j-u_1\right\|^2+\pi_{j2}.\left\|X_j-u_2\right\|^2+\sum_{i=j+1}^{n}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2$

And Since we assumed that X_j is assigned to u_1, $\pi_{j1}=1\ and\ \pi_{j2}=0$.

$\therefore\ \sum_{i=0}^{j-1}\sum_{k=1}^{2}\pi_{ik}\ .\|X_i-\mu_k\|^2+\pi_{j1}.\left\|X_j-\mu_1\right\|^2+\sum_{i=j+1}^{n}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2$

If X_j was assigned to μ_2 the equation would look like: -

$\sum_{i=0}^{j-1}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2+\pi_{j2}.\left\|X_j-\mu_2\right\|^2+\sum_{i=j+1}^{n}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2$

Now we also assumed that X_j assigned to u_1 obtains minimum objective,

$\therefore\ \sum_{i=0}^{j-1}\sum_{k=1}^{2}\pi_{ik}\ .\|X_i-\mu_k\|^2+\pi_{j1}.\left\|X_j-\mu_1\right\|^2+\sum_{i=j+1}^{n}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2<\sum_{i=0}^{j-1}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2+\pi_{j2}.\left\|X_j-\mu_2\right\|^2+\sum_{i=j+1}^{n}\sum_{k=1}^{2}\pi_{ik}\ .\ \|X_i-\mu_k\|^2$

$$\therefore\ \pi_{j1}.\left\|X_j-\mu_1\right\|^2\ <\ \pi_{j2}.\left\|X_j-\mu_1\right\|^2$$

Since $\pi_{j_1}=1\ in\ LHS\ and\ \pi_{j2}=2\ in\ RHS-$

$$\therefore\ \left\|X_j-\mu_1\right\|^2\ <\ \left\|X_j-\mu_1\right\|^2$$

Which is a contradiction to our initial claim [1].
Therefore, assigning X_j to cluster with centroid μ_2 with minimum difference from centroid would obtain the minimum objective during that step.

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

Solution: Proof by contradiction.

Let us consider, we are working with K=2 and after the E step we compute $\mu_1 \; and \; u_2$. Let us assume that there exists a u_1' such that the difference of between all the data-points belonging to u_1 is higher than their difference from μ_1'.

That is,

$\sum_i^n \|X_i - \mu_1\| > \sum_i^n \| \; X_i - \; \mu_1' \| \; \forall \; i \; belonging \; to \; cluster \; with \; centeroid \; u_1$. **[2]**

We also know that,

$$\pi_{ik} = 1 \; if \; i \; belongs \; to \; cluster \; K, 0 \; otherwise.$$

We assume that going ahead with μ_1 helps us obtain the minimum objective function.
That is our minimum objective function with K=2 would look like –

$$\sum_{i=0}^n \pi_{i1} \; . \; \| X_i - \mu_1 \|^2 \; + \sum_{i=0}^n \pi_{i2} \; . \; \| X_i - \mu_2 \|^2$$

However, if we take the centroid of k=1 as μ_1', our equation would look like –

$$\sum_{i=0}^n \pi_{i1} \; . \; \| X_i - \mu_1' \|^2 \; + \sum_{i=0}^n \pi_{i2} \; . \; \| X_i - \mu_2 \|^2$$

Equating the two equations,

$$\sum_{i=0}^n \pi_{i1} \; . \; \| X_i - \mu_1 \|^2 + \sum_{i=0}^n \pi_{i2} \; . \; \| X_i - \mu_2 \|^2 \; < \; \sum_{i=0}^n \pi_{i1} \; . \; \| X_i - \mu_1' \|^2 + \sum_{i=0}^n \pi_{i2} \; . \; \| X_i - \mu_2 \|^2$$

$$\therefore \; \sum_{i=0}^n \pi_{i1} \; . \; \| X_i - \mu_1 \|^2 \; < \; \sum_{i=0}^n \pi_{i1} \; . \; \| X_i - \mu_1' \|^2$$

$$\pi_{i1} = 1 \; for \; only \; where \; X_i \; belongs \; to \; k = 1 \; ; otherwise \; 0.$$

$$\therefore \; \sum_{i=0}^n \; \| X_i - \mu_1 \|^2 \; < \; \sum_{i=0}^n \; \| X_i - \mu_1' \|^2$$

Which is a contradiction to our original claim [2].
Thus, updating the μ 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.