DS 5230

GAUTAM AJEY KHANAPURI

HW - 2

PROBLEM - 1

K-Means Objective:

$$\min \sum_{i=1}^{N} \sum_{K=1}^{K} \pi_{iK} \cdot \|X_i - \mu_K\|^2$$

$\Rightarrow$ We must minimize the variance within each cluster.

π_{iK}

no. of clusters/regions

$$\begin{array}{c|c} & \begin{array}{ccccc} 1 & 2 & 3 & \cdots & K \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ N \end{array} & \begin{bmatrix} 0 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ & & & & \end{bmatrix} \end{array}$$

all points

represents membership in K^{th} cluster

(A)

In order to prove that the E-step update on memberships π achieves the minimum objective given the current centroids (μ), we must show that reassigning memberships of each point to the nearest centroid minimizes the K-means objective function.

For each datapoint, the objective function yields a single term

$$\|x_i - \mu\|^2$$

Therefore, minimizing the objective function means reducing the value of $\|x_i - \mu\|^2$ for each datapoint

There is only 1 term because each datapoint can only belong to one cluster.

Therefore, when the centroids, μ, are fixed, we must reassign memberships (π) of each datapoint to the closest centroid.

$$K = \arg\min_k \|x_i - \mu_k\|^2.$$

$$\pi_{ik} = \begin{cases} 1, & \text{if } k = \arg\min_k \|x_i - \mu_k\|^2 \\ 0, & \text{for all other } k \end{cases}$$

When this process is repeated for all datapoints, the total objective function becomes:

$$obj = \sum_{i=1}^{N} \min_k \|x_i - \mu_k\|^2$$

Since each x_i is assigned to the cluster with the closest centroid, the overall objective is minimized for the given centroids μ_k.

(B)

We must show that updating each centroid to the mean of the datapoints in its cluster minimizes the objective function with respect to the centroids.

Focussing on a single cluster k:

$$obj_k = \sum_{x_i \in k} \|x_i - \mu_k\|^2$$

To minimize the objective function, we must find a optimal μ_k. This is because the function is exponential & not linear. Larger distances are penalized more heavily compared to shorter distances.

$$obj_k = \sum_{x_i \in k} \left[(x_i - \mu_k)^T (x_i - \mu_k) \right]$$

$$= \sum \left[\|x_i\|^2 - 2 x_i^T \mu_k + \|\mu_k\|^2 \right]$$

since $\|x_i\|^2$ is independent of μ_k, it is a constant. Thus we can focus on minimizing:

$$obj_k = -2 \sum_{x_i \in k} x_i^T \mu_k + \underbrace{|k|}_{\text{no. of points in cluster } k} \|\mu_k\|^2$$

To minimize obj_k, take its gradient wrt μ_k.

$$\frac{\partial \, obj_k}{\partial \mu_k} = \frac{\partial \left\{ -2 \sum_{x_i \in k} x_i^T \mu_k + |k| \, \|\mu_k\|^2 \right\}}{\partial \mu_k}$$

$$= -2\sum_{x_i \in k} x_i + 2|k|\mu_k$$

set derivative to 0 to find the local minima

$$\frac{\partial \, obj_k}{\partial \mu_k} = 0$$

$$-2\sum_{x_i \in k} x_i + 2|k|\mu_k = 0$$

$$\cancel{2}|k|\mu_k = \cancel{2}\sum_{x_i \in k} x_i$$

$$\mu_k = \frac{\sum_{x_i \in k} x_i}{|k|}$$

$$\mu_k = \frac{\sum_{x_i \in k} x_i}{\sum_{i=1}^{N} \pi_{ik}}, \quad |k| = \sum_{i=1}^{N} \pi_{ik}$$

∴ sum of 1s in the k^{th} col is equal to the number of datapoints in the k^{th} cluster

$$\mu_k = \frac{\sum_{i=1}^{N} \pi_{ik} x_i}{\sum_{i=1}^{N} \pi_{ik}} \Rightarrow$$

The optimal ~~point~~ centroid is the mean of all points in the cluster

(C)

The E-step and M-step alternatively reduce the K Means objective function

1. The E-step assigns each point to the ~~cl~~ nearest center, which either reduces the obj func or keeps it the same.

2. The M-step, updates the centroids, which minimizes the $(\text{distance})^2$ between the points and centroids for each cluster.

Neither step increases ~~$~~ obj.

obj ≥ 0.

and since both steps diminish, obj, they must converge to a finite value.

There are ~~n~~ N datapoints & each of them can belong to 1 cluster only.
Max number of unique cluster combinations is K^N.
Each iteration over E & M step moves to a new configuration that reduces J, so the configuration must eventually reach a local minima.

Why K-Means does not necessarily converge to the global minimum.

→ The global minimum is 0 when the number of centroids is equal to the number of datapoints.

→ Therefore for any ~~all~~ other value of K, obj > 0.

→ Moreover, initial selection of centroids may lead to a poor income.