Sheet_03_P2

Saturday, 11. November 2023 19:52

Given. set of n points $p_i \in \mathbb{R}^d$, $i \in \{1,2,3,\dots,n\}$
$k \rightarrow$ no. of clusters
K-Means clustering:-
find centers of k clusters c_j, $j \in \{1,2,\dots,k\}$

Loss function to be minimized:—
$$L(c) = \sum_{i=1}^{n} \min_{j \in \{1,2,\dots k\}} \|p_i - c_j\|_2^2$$

New variable introduced for approximation:—
$z_i \in \operatorname{argmin}_{j \in \{1,\dots k\}} \|p_i - c_j\|_2^2$ for each data point p_i.

K-Means Clustering:—
(i) assignment:— updating z_i
(ii) Updating centers / Refitting:—
$$c_j = \frac{1}{|\{i : z_i = j\}|} \sum_{i: z_i = j} p_i$$

To prove:- k-means is guaranteed to converge (to a local optimum)
prove separately for assignment step & refitting step.

Proof:
For Assignment Step:-
Let at iteration t,
loss function is given by $L_t(c)$ and
assignment variable be z_t.

The minimization of loss function is given by:
$$z_i^{(t+1)} = \operatorname{argmin}_{j \in \{1,\dots,k\}} \|p_i - c_{j_t}\|_2^2$$

Let L_t be the contribution of i-th point to the loss at iteration t.
Then,
$$L_t = \sum_{i=1}^{n} \min_{j \in \{1,\dots,k\}} \|p_i - c_j^{(t)}\|_2^2$$

Total loss at iteration t:—
$$L_t(c) = \sum_{i=1}^{n} L_t^{(i)}$$

Then, loss at iteration t+1:—
$$L_{t+1}(c) = \sum_{i=1}^{n} \min_{j \in \{1,\dots k\}} \|p_i - c_j^{(t+1)}\|_2^2$$

By defn of assignment step: $\forall j \in \{1,\dots,k\}$
$$\|p_i - c_{t+1}^{(z_i^{(t+1)})}\|_2^2 \leq \|p_i - c_j^{(t+1)}\|_2^2$$

Thus,
$$L_{t+1} = \sum_{i=1}^{n} \min_{j \in \{1,\dots k\}} \|p_i - c_j^{(t+1)}\|_2^2 \leq \sum_{i=1}^{n} \min_{j \in \{1,\dots,k\}} \|p_i - c_j^t\|_2^2 = L_t$$

Loss for step t+1 $\leq$ loss for step t.
$\Rightarrow$ loss function is decreasing monotonically in each step.

For Refitting Step:-
For the step t+1, the new center in the refitting step is given as:
$$c_j^{(t+1)} = \frac{1}{|\{i : z_i^{(t+1)} = j\}|} \sum_{i: z_i^{(t+1)} = j} p_i$$

Loss function at iteration t+1 is given by
$$L_{t+1}(c) = \sum_{i=1}^{n} \min_{j \in \{1,\dots k\}} \|p_i - c_j^{(t+1)}\|^2$$

After updating the clusters, we get
$$\|p_i - c_j^{(t+1)}\|_2^2 \leq \|p_i - c_j^t\|_2^2 \quad [\because \text{ new cluster center is chosen to minimize the dist.}]$$

Now, Summing over i=1 to n
$$\sum_{i=1}^{n} \|p_i - c_j^{(t+1)}\|_2^2 \leq \sum_{i=1}^{n} \|p_i - c_j^t\|_2^2$$

or,
$$\frac{1}{|\{i : z_i^{(t+1)} = j\}|} \sum_{i=1}^{n} \|p_i - c_j^{(t+1)}\|_2^2 \leq \frac{1}{|\{i : z_i = j\}|} \sum_{i=1}^{n} \|p_i - c_j^t\|^2$$

which is equivalent to saying $L_{t+1}(c) \leq L_t(c) \Rightarrow$ loss function is decreasing monotonically.

$\therefore$ We prove that K-means is guaranteed to converge.

proved