Homework Assignment 1

Monday, 29 January 2024 3:07 AM

Ans 1

K - means clustering

cost function for k-means

$$\text{Cost}(T) = \sum_{z \in T} \sum_{x \in C_z} \|x - z\|^2$$

it can also be written as

$$\text{cost}(C_1, \dots, C_K ; z_1, \dots, z_K) = \sum_{j=1}^{K} \sum_{x \in C_j} \|x - z_j\|^2$$

$\Rightarrow$ For any specific $j = 1, \dots k$

$$\text{Cost}(C_j ; z_j) = \sum_{x \in C_j} \|x - z_j\|^2 \quad — (1)$$

This represents the sum of squared distances from every point in 'C' to the centroid 'z'

we want to proove

$$\text{Cost}(C_j ; z_j) = \text{Cost}(C_j ; \text{mean}(C_j)) + |C_j| \|z_j - \text{mean}(C_j)\|^2 \quad — (2)$$

where $\text{mean}(C_j)$ is the mean (centroid) of points in set C_j

$$\Rightarrow \quad \text{mean}(C_j) = \frac{1}{|C_j|} \sum_{x \in C_j} x \quad — (3)$$

where $|C_j|$ is the cardinality of C_j

Expanding eq (1),

$$\text{Cost}(C_j ; z_j) = \sum_{x \in C_j} \|x - z_j\|^2$$

$$= \sum_{x \in C_j} \left(\|x\|^2 - 2x \cdot z_j + \|z_j\|^2 \right)$$

$$= \sum_{x \in C_j} \|x\|^2 - 2z_j \sum_{x \in C_j} x + \|z_j\|^2 \sum_{x \in C_j} 1$$

$$= \sum_{x \in C_j} \|x\|^2 - 2z_j |C_j| \,\text{mean}(C_j) + |C_j| \|z_j\|^2 \quad — (4) \quad \Big(\text{Using } (3)$$

Expanding Cost $(C_j ; \text{mean}(C_j)$,

$$\text{Cost}(C_j ; \text{mean}(C_j)) = \sum_{x \in C_j} \| x - \text{mean}(C_j) \|^2 \qquad (\text{From } 1)$$

$$= \sum_{x \in C_j} \| x \|^2 - 2\,\text{mean}(C_j) \sum_{x \in C_j} x + \| \text{mean}(C_j) \|^2 \sum_{x \in C_j} 1$$

$$= \sum_{x \in C_j} \| x \|^2 - 2 |C_j| \| \text{mean}(C_j) \|^2 + |C_j| \| \text{mean}(C_j) \|^2 \quad (\text{Using } 3)$$

$$= \sum_{x \in C_j} \| x \|^2 - |C_j| \| \text{mean}(C_j) \|^2 \quad \text{— } 5$$

Subtracting 4 & 5 (4 - 5)

$$\text{cost}(C_j ; z_j) - \text{cost}(C_j ; \text{mean}(C_j))$$

$$= \left(\sum_{x \in C_j} \| x \|^2 - 2 z_j |C_j| \,\text{mean}(C_j) + |C_j| \| z_j \|^2 \right) - \left(\sum_{x \in C_j} \| x \|^2 - |C_j| \| \text{mean}(C_j) \|^2 \right)$$

$$= |C_j| \| z_j \|^2 - 2 z_j |C_j| \,\text{mean}(C_j) - |C_j| \| \text{mean}(C_j) \|^2$$

$$= |C_j| \left(\| z_j \|^2 - 2 z_j \,\text{mean}(C_j) + \| \text{mean}(C_j) \|^2 \right)$$

$$= |C_j| \| z_j - \text{mean}(C_j) \|^2$$

$\Rightarrow$ The equation can be re-written as

$$\text{cost}(C_j ; z_j) = \text{Cost}(C_j ; \text{mean}(C_j)) + |C_j| \| z_j - \text{mean}(C_j) \|^2$$

is same as eq 2

Hence Proved

Ans 2

k-means algorithm is defined as :

For

initial centers $z_1, \dots, z_k \in \mathbb{R}^d$ & clusters $C_1, \dots, C_k$

repeat until there is no change in cost

for each j : $C_j \leftarrow \{x \in S$ whose closest center is $z_j\}$

for each j : $z_j \leftarrow \text{mean}(C_j)$

We have to prove that the cost of K-means algorithm is monotonically decreasing.

Proof :

The k-means algorithm has 2 main steps

1. Assignment Step
   where each data point is assigned to it's nearest centroid

2. Updation step
   The centroids for each cluster are calculated again as the mean of all the points in the cluster

Let $C_j^{(t)}$ be the set of points in cluster C_j at iteration 't'

with $z_j^{(t)}$ as it's centroid.

The cost function becomes

$$\text{Cost}^{(t)} = \sum_{j=1}^{k} \sum_{x \in C_j^{(t)}} \| x - z_j^{(t)} \|^2$$

For the function to be monotonically decreasing

$$\text{Cost}^{(t+1)} \leq \text{Cost}^{(t)}$$

Step 1 : Assignment Step

At each iteration, each point x is assigned to the nearest centroid.

$\Rightarrow$ distance of points at iteration 't' $= \| x - z_j^{(t)} \|^2$

& iteration 't+1' $= \| x - z_j^{(t+1)} \|^2$

Since points are reassigned to a closer centroid in iteration 't+1' by definition

$$\Rightarrow \quad \| x - z_j^{(t+1)} \|^2 \leq \| x - z_j^{(t)} \|^2$$

Therefore, the cost associated with each point cannot increase i.e. we see that summing effect cannot increase total cost. which is the distance

It can also be thought as

$$\text{cost}\left(C_1^{(t+1)}, \dots, C_k^{(t+1)}; z_1^{(t)}, \dots, z_k^{(t)}\right) \leq \text{cost}\left(C_1^{(t)}, \dots C_k^{(t)}; z_1^{(t)}, \dots z_k^{(t)}\right)$$

Step 2 : Updation Step

We calculate new centroids $z_j^{(t+1)}$ as the mean of all the points in cluster $C_j^{(t+1)}$

By definition of mean, it minimizes the sum of squared distances from the points in the cluster to its centroid (or mean)

Therefore, on calculating the centroids of the clusters at iteration '$t+1$' cannot increase the cost.

$$\Rightarrow \sum_{x \in C_j^{(t+1)}} \| x - z_j^{(t+1)} \|^2 \leq \sum_{x \in C_j^{(t)}} \| x - z_j^{(t)} \|^2$$

It can also be thought as

$$\text{cost}\left(C_1^{(t+1)}, \dots C_k^{(t+1)}; z_1^{(t+1)}, \dots z_k^{(t+1)}\right) \leq \text{cost}\left(C_1^{(t+1)}, \dots, C_k^{(t+1)}; z_1^{(t)}, \dots, z_k^{(t)}\right)$$

On combing the 2 steps, we get

$$\text{cost}^{(t+1)} = \sum_{j=1}^{k} \sum_{x \in C_j^{(t+1)}} \| x - z_j^{(t+1)} \|^2$$

$$\leq \sum_{j=1}^{k} \sum_{x \in C_j^{(t+1)}} \| x - z_j^{(t)} \|^2$$

$$\leq \sum_{j=1} \sum_{x \in C_j^{(t)}} \| x - z_j^{(t)} \|^2$$

$$\leq \text{cost}^{(t)}$$

Hence Proved