Course › Unit 4 Unsupervised Learning (2 weeks) › Lecture 13. Clustering 1 › 8. The K-Means Algorithm: the Specifics

8. The K-Means Algorithm: the Specifics

The K-Means Algorithm: the Specifics

seen the k-means algorithm, we've

seen the conversion properties of this algorithm,

and we also realize an important drawback of this algorithm,

that it's very sensitive to initialization.

And we understood how the properties of initialization

will be impacting the final results.

So with that, we completed the discussion

on k-means algorithm.

End of transcript. Skip to the start.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Finding the Representative Z

3/3 points (graded)

Find a simplified form of the following expression:

$$\frac{\partial}{\partial z_j} \sum_{i \in \mathbb{C}_j} \left\| x^{(i)} - z_j \right\|^2$$

● $\sum_{i \in \mathbb{C}_j} -2\left(x^{(i)} - z_j\right)$ ✓

○ $-2\left(z_j - \sum_{i \in \mathbb{C}_j} x^{(i)}\right)$

○ $\sum_{i \in \mathbb{C}_j} -\left(x^{(i)} - z_j\right)$

○ $\sum_{i \in \mathbb{C}_j} x^{(i)}$

Now, what is the value of z_j that minimizes the sum?

● $\frac{\sum_{i \in C_j} x^{(i)}}{|C_j|}$ ✓

○ $\sum_{i \in C_j} x^{(i)}$

Regarding update of z_j, which of the following statements is true (select all that apply)?

□ The value of z_j is affected by points $x_i \notin C_j$

☑ The value of z_j is only affected by points $x_i \in C_j$ ✓

☑ The obtained z_j is the centroid (center of mass assuming each $x^{(i)}$ has equal mass) of the jth cluster ✓

✓

Solution:

Note that

$$z_j = \frac{\sum_{i \in C_j} x^{(i)}}{|C_j|}$$

is the center of mass, or centroid, of the jth cluster.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Impact of Intialization

1/1 point (graded)

Remember that the K-Means algorithm is given by

1. Randomly select $z_1, \ldots, z_K$
2. Iterate
    1. Given $z_1, \ldots, z_K$, assign each data point $x^{(i)}$ to the closest z_j, so that

    $$\text{Cost}\left(z_1, \ldots z_K\right) = \sum_{i=1}^{n} \min_{j=1,\ldots,k} \left\| x^{(i)} - z_j \right\|^2.$$

    2. Given $C_1, \ldots, C_K$ find the best representatives $z_1, \ldots, z_K$, i.e. find $z_1, \ldots, z_K$ such that

    $$z_j = \text{argmin}_z \sum_{i \in C_j} \|x^{(i)} - z\|^2.$$

Which of the following is true about the initialization and output of the K-Means algorithm? Select all those apply.

☑ Step 2.1 decreases or does not change the cost of clustering output ✓

☑ Step 2.2 decreases or does not change the cost of clustering output ✓

☑ The clustering output that the K-Means algorithm converges to depends on the intialization ✓

Solution:

While steps 2.1 and 2.2 of the algorithm always decreases the cost or keeps it the same at least, the output of the algorithm largely dependes on the intialization of step 1. Thus, in practice, it is wise to make sure that $z_1, \ldots z_K$ are intialized so that they are well spread out. Another alternative is to try multiple initializations and choose the clustering output that appears the most commonly.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

What if K is 1?

1/1 point (graded)
Now, assume that we are given with $K = 1$ as the number of clusters. Now, does initialization matter at all?

● No, because cluster assignment does not change in step 2.1 ✓

○ Yes, because representative selection changes in step 2.2

Solution:

Because if $K = 1$ cluster assignment can never change, initialization does not matter. Also note that the algorithm will converge (have same assignment and same representative from there on) after just 1 iteration.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Lecture 13. Clustering 1 / 8. The K-Means Algorithm: the Specifics