Course › Unit 4 Unsupervised Learning (2 weeks) › Lecture 14. Clustering 2 › 2. Limitations of the K Means Algorithm

Audit Access Expires May 11, 2020
You lose all access to this course, including your progress, on May 11, 2020.

2. Limitations of the K Means Algorithm

Limitations of the K Means Algorithm

0:00 / 0:00 1.25x

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Limitations of the K-Means Algorithm I

1/1 point (graded)

Remember that the K-Means Algorithm is given as below:

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

$$z_j = \operatorname{argmin}_z \sum_{i \in C_j} \left\| x^{(i)} - z \right\|^2 = \frac{\sum_{i \in C_j} x^{(i)}}{|C_j|}$$

where $|C_j|$ is the number of points in C_j.

Which of the following are **false** about K-Means Algorithm? Please choose all those apply.

- [ ] $C_1, \ldots, C_K$ found by the algorithm is always a partition of $\{x_1, \ldots, x_n\}$
- [x] It is always guaranteed that the K representatives $z_1, \ldots, z_K \in \{x_1, \ldots, x_n\}$
- [ ] The algorithm may output different $C_1, \ldots, C_K$ and $z_1, \ldots, z_K$ depending on the initialization of line 1
- [ ] Line 2.2 of the algorithm(Given $C_1, \ldots, C_K$ find the best representatives $z_1, \ldots, z_K$ ...) finds the cost-minimizing representatives $z_1, \ldots z_K$.

Solution:

It is not guaranteed that $z_1, \ldots, z_K \in \{x_1, \ldots, x_n\}$, because as in line 2.2 of the algorithm above, $z_1, \ldots, z_K$ are given by

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

There is no guarantee that the centroid of all $x^{(i)}$ in a cluster will itself belong to $\{x_1, \ldots, x_n\}$. Depending on the application context, such as when clustering Google News articles, it can be problematic that a representative of a clustering is not an actual datapoint.

The other 3 choices are true:

- Clustering always outputs $C_1, \ldots, C_K$ that is a partition of $\{x_1, \ldots, x_n\}$
- The result of clustering depends on the initialization of $z_1, \ldots, z_K$.
- As we saw in the last lecture, line 2.2 of the algorithm

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

minimizes the cost

$$\text{Cost}\,(C_1, \ldots C_K) = \min_{j=z_1,\ldots,z_K} \sum_{j=1}^{k} \sum_{i \in C_j} \text{dist}\,(x^{(i)}, z_j)$$

where the distance function $\text{dist}\,(x^{(i)}, z_j)$ is the squared euclidean distance function $\|x^{(i)} - z_j\|^2$.

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

Limitations of the K-Means Algorithm II

2/2 points (graded)

Suppose we have a 1D dataset drawn from 2 different Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$ where $\mu_1 \neq \mu_2$. The dataset contains n data points from each of the two distributions for some large number n.

Define **optimal clustering** to be the assignment of each point to the more likely Gaussian distribution given the knowledge of the generating distribution.

Consider the case where $\sigma_1^2 = \sigma_2^2$, would you expect a 2-means algorithm to approximate the optimal clustering?

- (•) Yes
- ( ) No

✔

Now if $\sigma_1^2 >> \sigma_2^2$, would you expect a 2-means algorithm to approximate the optimal clustering?

- ( ) Yes
- (•) No

✔

Solution:

When $\sigma_1^2 = \sigma_2^2$, the boundary between the 2 optimal clusters is the midpoint between μ_1 and μ_2. The 2 centroids found by the 2-means algorithm will also be approximately equidistant from this boundary (midpoint between μ_1 and μ_2), and therefore the assignment to clusters will be a similar split around the midpoint. When $\sigma_1^2 >> \sigma_2^2$, the boundary between the 2 optimal clusters is closer to one centroid then the other. Since the 2-means algorithm will always have an equidistant split between the two centroids, this behavior cannot be reproduced and thus k-means clustering will erroneoously assign more points to the cluster with a smaller variance.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Lecture 14. Clustering 2 / 2. Limitations of the K Means Algorithm

Add a Post

Show all posts by recent activity

- The Cs are a partition, but of what
  I recall that Cs are a partition of indexes. Am I wrong? — 4
- Limitations of the K-Means Algorithm II: Relationship between mu and sigma
  Given > Define optimal clustering to be the assignment of each point to the more likely Gaus... — 4
- [To Staff] Limitations of the K-Means Algorithm II - Ambiguous questions
  Although mu1 is different of mu2, we do not know by how much compared to their sigmas s... — 2