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

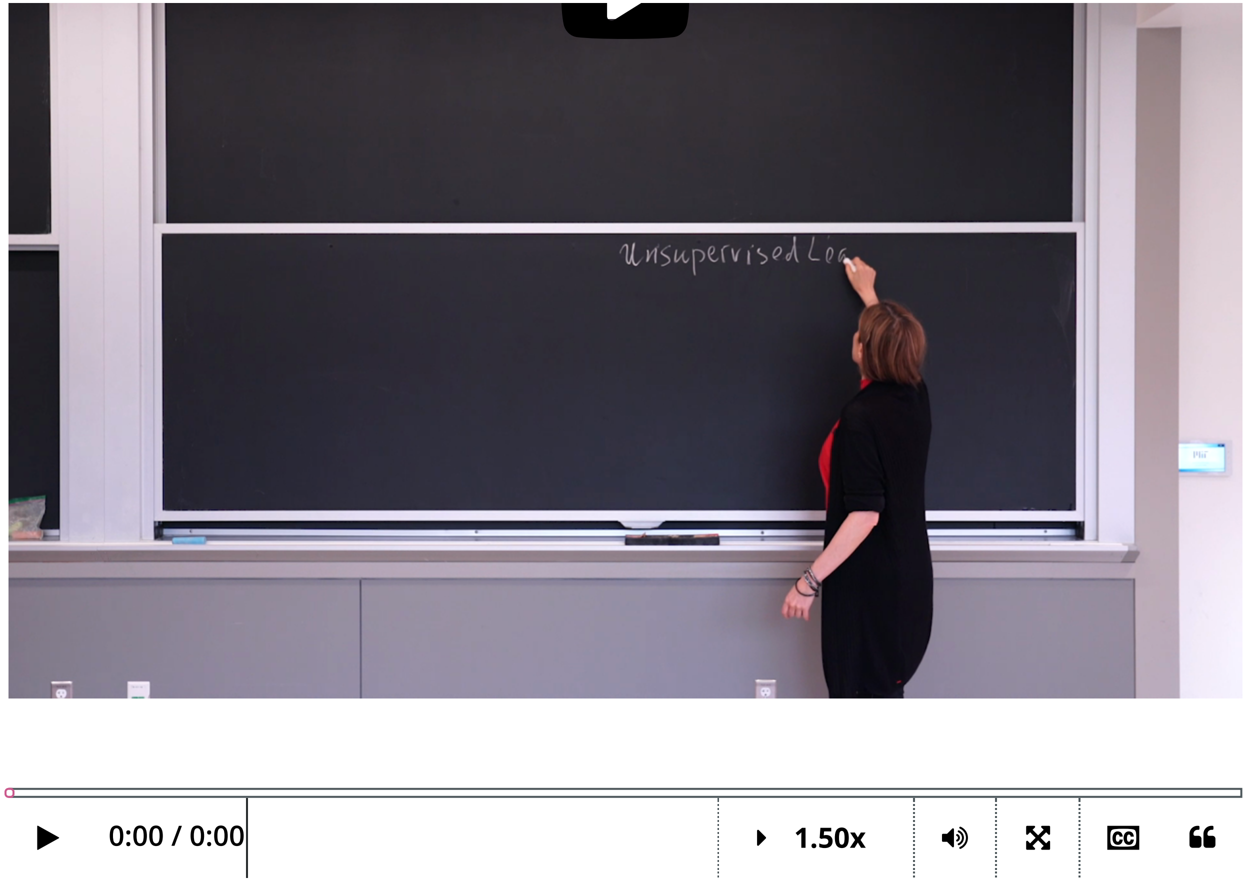
2. Limitations of the K Means Algorithm

Limitations of the K Means Algorithm

Start of transcript. Skip to the end.

So today, we will continue the conversation

about unsupervised learning, unsupervised learning.



And we will continue talking about clustering.

And I will start my lecture by briefly summarizing

what we've seen last time, which covered the discussion

about K-means algorithm

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Limitations of the K-Means Algorithm

1/1 point (graded)

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

1. Randomly select $z_1, \dots, z_K$
2. Iterate
    1. Given $z_1, \dots, z_K$, assign each $x^{(i)}$ to the closest z_j. i.e., assign each $x^{(i)}$.
    2. Given $C_1, \dots, C_K$ find the best representatives $z_1, \dots, z_K$ such that

$$\operatorname{argmin}_{z_1,\dots,z_K} \sum_{j=1}^{k} \sum_{i \in C_j} \left\| x^{(i)} - z_j \right\|^2$$

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

☐ $C_1, \dots, C_K$ found by the algorithm is always a partition of $\{x_1, \dots, x_n\}$

☑ It is always guaranteed that the K representatives $z_1, \dots, z_K \in \{x_1, \dots, x_n\}$ ✓

☐ The algorithm may output different $C_1, \dots, C_K$ and $z_1, \dots, z_K$ depending on the initialization of line 1

☑ Line 2b of the algorithm(Given $C_1, \dots, C_K$ find the best representatives $z_1, \dots, z_K$ such that ...) finds the cost-minimizing representatives $z_1, \dots z_K$ for all cost functions ✓

✓

Solution:

It is not guaranteed that $z_1, \ldots, z_K \in \left\{x_1, \ldots, x_n\right\}$ because as in line 2b 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 $\left\{x_1, \ldots, x_n\right\}$. 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.

Also, as we saw in the last lecture, line 2b of the algorithm

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

is a simplification(or special case) of

$$\text{Cost}\left(C_1, \ldots C_K\right) = \min_{j = z_1, \ldots, z_K} \sum_{j=1}^{k} \sum_{i \in C_j} \left\|x^{(i)} - z_j\right\|^2$$

when the cost function is the euclidean distance function($\left\|x^{(i)} - z_j\right\|^2$).

These two points are the **limitations** of the K-Means algorithm. We saw in the last lecture that clustering always outputs $C_1, \ldots, C_K$ that is a partition of $\{x_1, \ldots, x_n\}$, and that the result of clustering depends on the initialization of $z_1, \ldots, z_K$.

Submit You have used 1 of 3 attempts

ℹ Answers are displayed within the problem

Limitations of the K-Means Algorithm 2

2/2 points (graded)

Suppose we have a 1D dataset drawn from 2 different Gaussian distribution $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)$. The dataset contains n data points from each of the two distributions for some large number n. If we define the optimal clustering is to assign each point to the most 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 equidistant from this boundary and therefore the assignment to clusters will be a similar split around the midpoint.
When $\sigma_1^2 >> \sigma_2^2$, the boundary betwwen 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 2 of 2 attempts

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

❮ All Posts

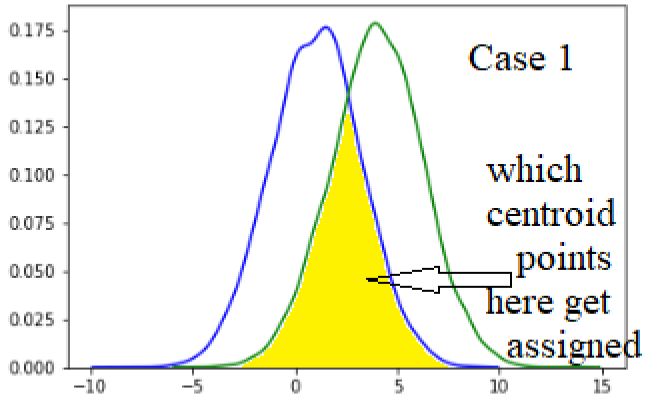
Limitations of kmeans algorithm 2

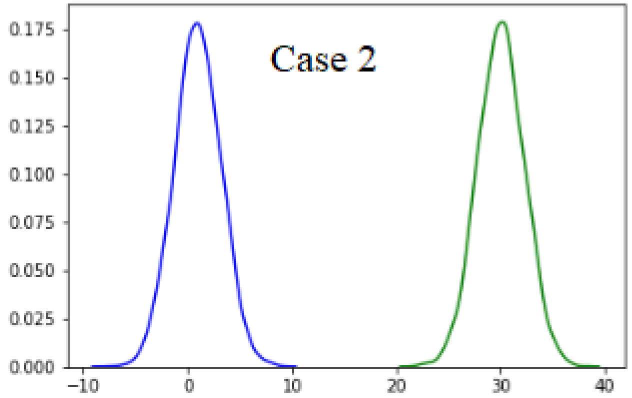
discussion posted about 15 hours ago by **sandipan_dey**

Should not the answer for the first case ($\sigma_1 = \sigma_2$) depend on the difference between μ_1 , μ_2 and how they are related to σ_1, σ_2? For example, let's consider the following 2 cases:

1. $\mu_2 \in [\mu_1 - \sigma_1, \mu_1 + \sigma_1]$
2. $\mu_2 \notin [\mu_1 - 3\sigma_1, \mu_1 + 3\sigma_1]$

Will the chance that the k-means will find the optimal partition be same in these two cases, assuming all other variabilities (e.g., initialization etc.) fixed?





This post is visible to everyone.

Add a Response

2 responses

nr7116

about 11 hours ago

I had the same doubt. But since it was not given, I made simplifying assumption that means are well separated with no overlapping point - that gives an answer that is accepted. I have realized that though this was s a lecture on a unsupervised learning, question are mostly straight out of lecture with little googling read:-) making it supervised learning for me Previous units had enough cases of unsupervised googling read to get answers!

Add a comment

BrendanWood

about 4 hours ago

It should not matter what the means of the underlying Gaussian distributions are in this case. Note that we are not concerned whether the representatives of the clusters are equal to the means, we are just concerned that the two clusters (that the K-means algorithm determines) happen to coincide with the optimal clustering as defined in the question.

(Actually, I guess it matters that $\mu_1 \neq \mu_2$, or else the problem doesn't really make any sense, since there's nothing to separate)

Add a comment

Showing all responses

Add a response:

Preview

Submit