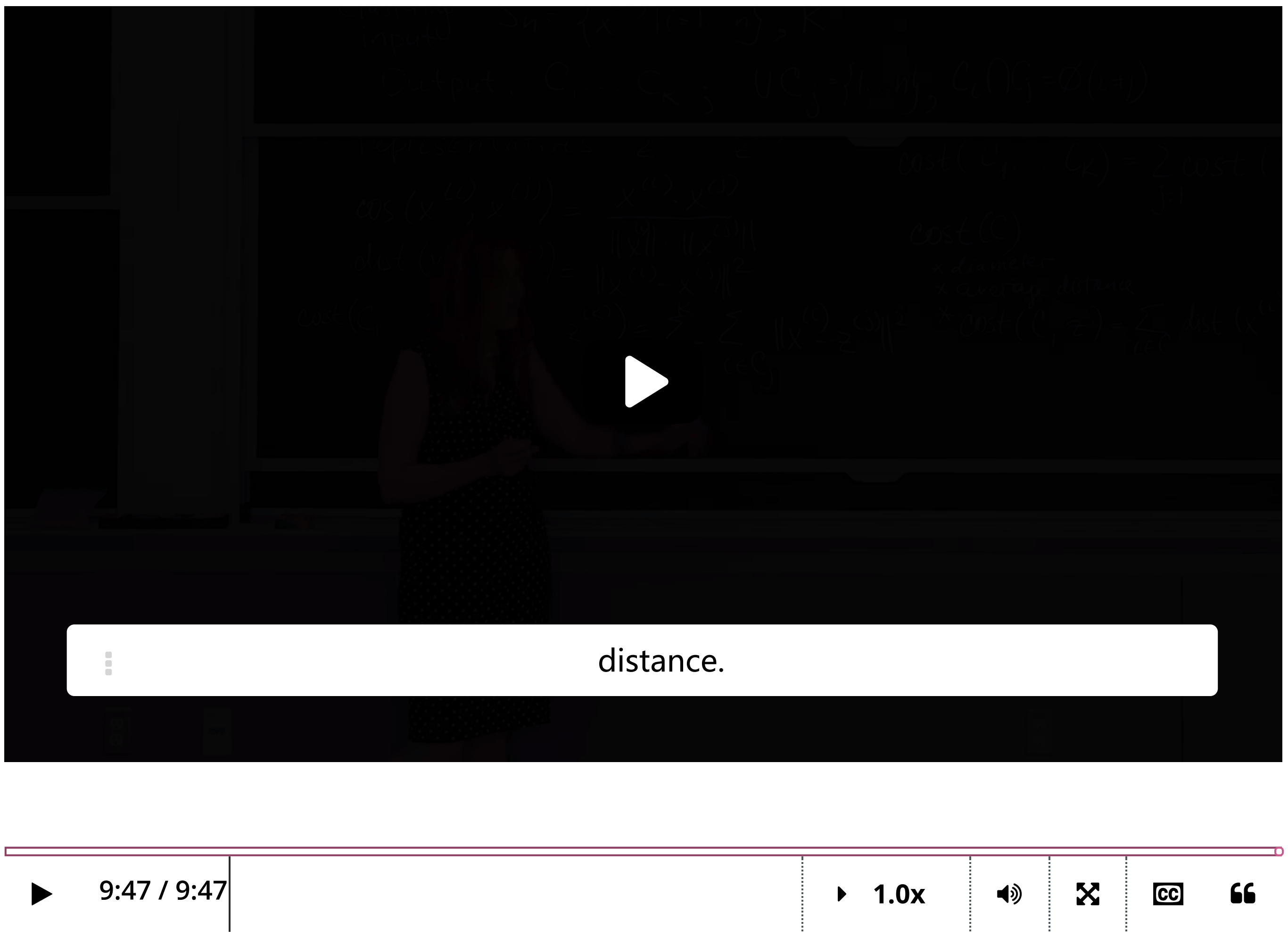
6. Similarity Measures-Cost functions

Similarity Measures-Cost functions



And as we will continue our lecture,

we will see how to unify them together.

But at this point, I would want you

to think about this cost, which we will be optimizing.

I would assume that we are given the clusters,

we are given the representatives,

and this is our way to compute the cost with squared Euclidean

distance.

End of transcript. Skip to the start.

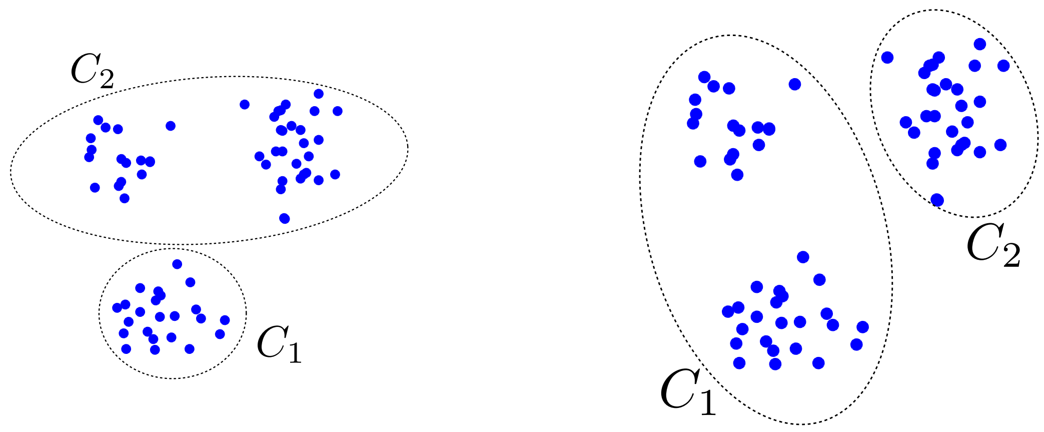
Video
Download video file

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

The Need to Define Costs

1/1 point (graded)

Note that it is possible to have multiple clustering results given the same set of feature vectors. For example, in the following picture, we can have two scenarios of clustering outputs given the same set of feature vectors.



Output 1 Output 2

What is a good method for deciding which clustering output is more preferable?

○ Randomly select a scenario because all clustering outputs are possible

● Define a measure of homogeneity inside cluster assignments and compare the measure of each scenario ✓

○ Always use the average distance of points in the cluster from its center

Solution:

A clustering output is preferable if and only if the data assigned inside each cluster are homogeneous to each other enough. Thus we define a measure of homogeneity inside cluster assignments and compare the measure of each scenario.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Choosing the Right Similarity Measure

1/1 point (graded)

Now, let's think about the Google News example the professor has mentioned in the beginning of the lecture. We want to measure the similarity between two Google News articles.

In the feature space, each article is represented as with the bag-of-words approach. For example, if "I", "love", "you", "more", "than", "Kevin" are the list of all unique vocabulary mentioned in all articles, the article "I love you" is represented as a vector $[1, 1, 1, 0, 0, 0]$ while another article "you love Kevin more than I" is represented as a vector $[1, 1, 1, 1, 1, 1]$. Note that each entry of vector is a binary indicator whether given word exists in an article or not.

You assume that the length of an article does not tell any useful information about the article, and hence choose a similarity measure that does not depend on the length of the article.

Which of the following similarity measure could be the one you chose?

○ Euclidean distance

● Cosine distance ✓

Solution:

It can be thought that longer articles will have larger norms, since they are more likely to contain unique words. Because it is assumed that the length of the article does not contain any important information, it is not ideal to use the Euclidean distance.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Possible Ways to Define Costs

1/1 point (graded)

Remember from the lecture above that the total cost of clustering output is defined as the sum of the cost inside each cluster. In other words,

$$\mathrm{Cost}\left(C_1, \ldots, C_K\right) = \sum_{j=1}^{K} \mathrm{Cost}\left(C_j\right)$$

Note that the cost $\mathrm{Cost}\left(C_j\right)$ is supposed to measure "how homogeneous" the assigned data are inside the jth cluster C_j. Which of the following are valid ways to define $Cost$? Select all those apply.

☑ The diameter of a cluster ✓

☑ The average distance between points inside a cluster ✓

☑ The sum of distance between the representative and all points inside a cluster ✓

✓

Solution:

As mentioned in the lecture, all three choices are possible. Note that different cost measures take different characteristics into considerations. For example,the diameter of a cluster will be decided by the outlier of a cluster. On the other hand, the average distance between points will have all points equally contribute to the cost.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Calculating Costs

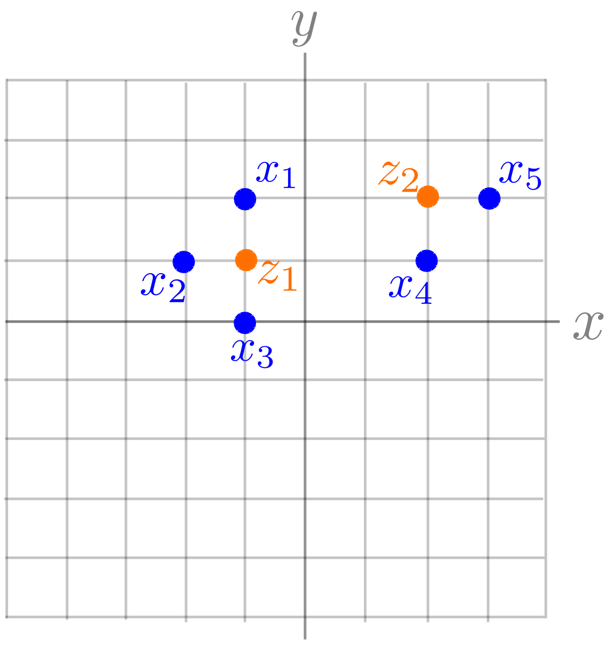
3/3 points (graded)

As in the picture below, the set of feature vectors is given by

$$S_n = \{x_1, \dots, x_5\}$$

and the number of clusters $K = 2$. S_n is clustered such that

$$\begin{aligned} 1,2,3 &\in C_1 \quad \text{whose representative is } z_1 \\ 4,5 &\in C_2 \quad \text{whose representative is } z_2 \end{aligned}$$



If the coordinates of points are given by

$$\begin{aligned} &x_1 = (-1,2)\,, x_2 = (-2,1)\,, x_3 = (-1,0)\,, z_1 = (-1,1) \\ &x_4 = (2,1)\,, x_5 = (3,2)\,, z_2 = (2,2) \end{aligned}$$

The cost of a clustering output is given by the sum of the squared euclidean distance of all points in a cluster with the representative for each of its clusters, i.e.

$$\text{Cost}\,(C_1, \dots, C_K) = \sum_{j=1}^{K} \text{Cost}\,(C_j) = \sum_{j=1}^{K} \sum_{i \in C_j} \|x_i - z_j\|^2$$

What is $\text{Cost}\,(C_1)$?

$\text{Cost}\,(C_1) =$

3

✓ Answer: 3

Now, What is $\text{Cost}\,(C_2)$?

$\text{Cost}\,(C_2) =$

2

✓ Answer: 2

Finally, what is the cost of this clustering output?

$\text{Cost}\,(C_1, C_2) =$

5

✓ Answer: 5

Solution:

Because $x_1, x_2, x_3 \in C_1$ and $x_4, x_5 \in C_2$, the cost of the clustering output is given by

$$\begin{aligned} \text{Cost}\,(C_1, C_2) &= \|x_1 - z_1\|^2 + \|x_2 - z_1\|^2 + \|x_3 - z_1\|^2 \\ +\|x_4 - z_2\|^2 + \|x_5 - z_2\|^2 & \\ &= 1 + 1 + 1 + 1 + 1 \\ &= 5 \end{aligned}$$

Submit

You have used 1 of 3 attempts

Answers are displayed within the problem

Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Lecture 13. Clustering 1 / 6. Similarity Measures-Cost functions

Show Discussion