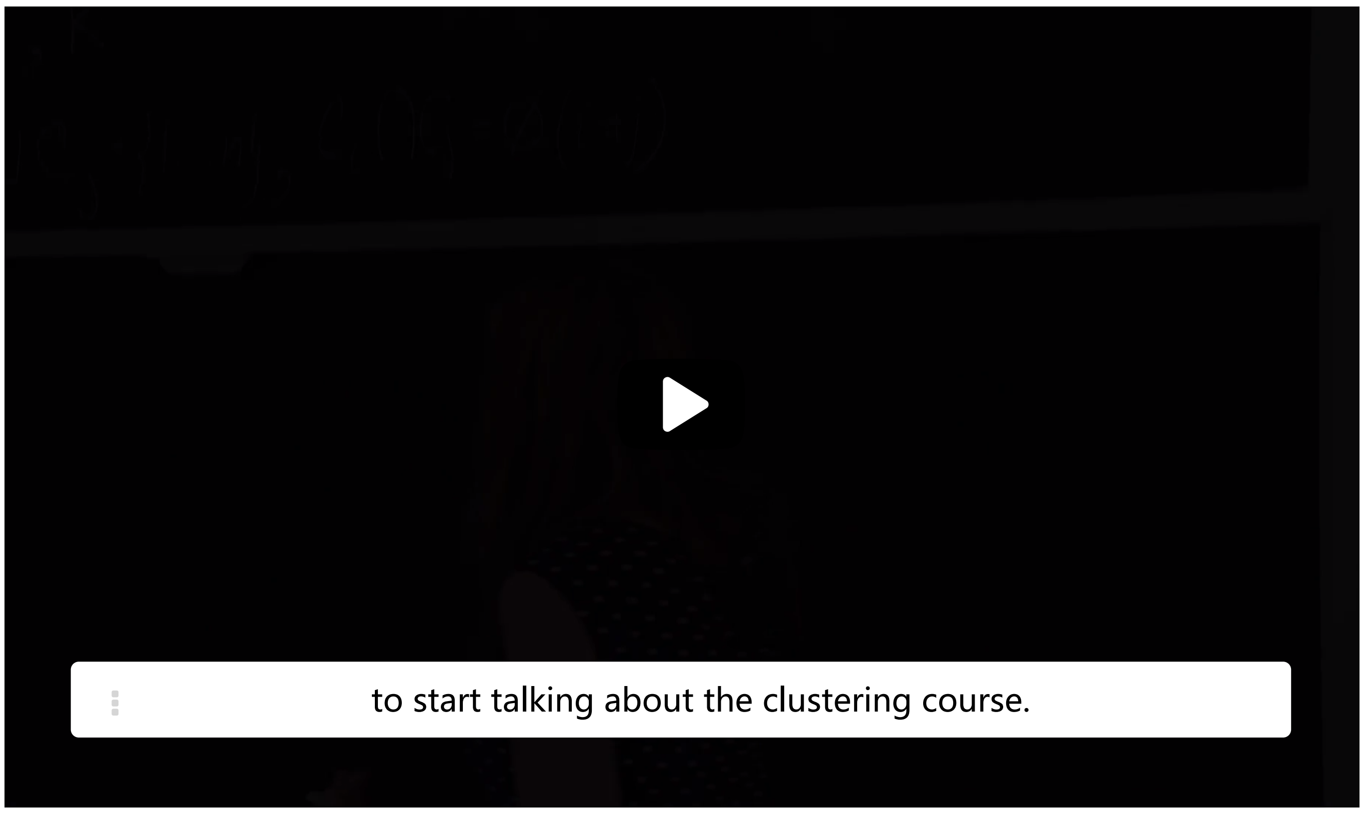
5. Clustering Definition

Clustering Definition



3:57 / 3:57 1.0x

If you know who is a representative,

you may be able to guess who is a constituent,

but you will see how you can actually

unify these two views together.

So at this point, we completed the discussion

about clustering definition, and we've seen some examples.

So the next part for us is actually

to start talking about the clustering course.

End of transcript. Skip to the start.

Video
Download video file

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

Partition Definition

1/1 point (graded)

A **partition** of a set is a grouping of the set's elements into non-empty subsets, in such a way that **every** element is included in one and only one of the subsets. In other words, $C_1, C_2, \ldots, C_K$ is a partition of $\{1, 2, \ldots, n\}$ if and only if

$$C_1 \cup C_2 \cup \ldots \cup C_K = \{1, 2, \ldots, n\}$$

and

$$C_i \cap C_j = \emptyset \quad \text{for any } i \neq j \text{ in } \{1, \ldots, k\}$$

(Union of all C_j's is the original set and the intersection of any C_i and C_j is an empty set.)

For example,

$$\{3\}, \{1\}, \{2\},$$

$$\{2, 1\}, \{3\},$$

$$\{2,3,1\}$$

are all partitions of the set $\{1,2,3\}$.

Now, which of the following is a partition of $\{1,2,\ldots,10\}$? Select all those apply.

- [x] $\{1,2,3\},\{4\},\{5,6,7,8,9,10\}$ ✔
- [ ] $\{1,2\},\{2,3,4,5\},\{6,7,8,9,10\}$
- [ ] $\{1\},\{3,4,5\},\{6,7,8,9,10\}$
- [x] $\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\}$ ✔
- [x] $\{1,2,3,4,5,6,7,8,9,10\}$ ✔

✔

Solution:

The intuitive meaning of partition is **grouping of elements such that each element belongs to exactly one partition**. Thus, choice 2 ("$\{1,2\},\{2,3,4,5\},\{6,7,8,9,10\}$") is not a partition because the element "2" belongs to two sets, not exactly one set. Also, choice 3 ("$\{1\},\{3,4,5\},\{6,7,8,9,10\}$") is not a partition because the element "2" does not belong to any set.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Clustering Definition: the Input

1/1 point (graded)
Remember that classification takes the training set

$$S_n = \{\,(x^{(i)}, y^{(i)})\,|i = 1,\ldots,n\}$$

and the number of classes as input. (where $x^{(i)}$ is the feature vector and $y^{(i)}$ is the label). (In other words, these were **given** so that we can find a classifier that will best classify the test set into the given number of classes.)

Remember in the lecture above that now we are discussing clustering, which has a different setting and a different goal from classification. Which of the following are the inputs (givens) of clustering? Select all those apply.

- [x] Set of feature vectors $S_n = \{x^{(i)}|i = 1,\ldots,n\}$ ✔
- [ ] Set of feature vectors and their labels $S_n = \{\,(x^{(i)}, y^{(i)})\,|i = 1,\ldots,n\}$
- [x] The number of clusters K ✔
- [ ] The representatives of each cluster $z_1,\ldots,z_K$

✔

Solution:

First, it is important to note that clustering is **unsupervised learning**, which means we do not have labels to start from. Now the goal is much less specific and more focus on the big picture. The goal is not to predict to which class each data will fall into, but to **visualize data** that we do not have much sense of. As in the example of partitioning Google News articles, clustering is used in settings where we do not have much information on data, and would like to visually get a sense of how many groupings the data consists of. Thus, clustering takes Set of feature vectors $S_n = \{x^{(i)}|i=1,\dots,n\}$ and the number of clusters K as input.

Finding the optimal number of K is itself another problem; K is a hyperparameter.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Clustering Definition: the Output

1/1 point (graded)
In the last problem, you have figured out the clustering input. Which of the following are outputs of the clustering? Or in other words , which of the following are determined by the clustering algorithm? (More details of the algorithm will be on the next page.)

(Select all those apply.)

- [x] A partition of indices $\{1,\dots,n\}$ into K sets, $C_1,\dots,C_K$ ✔
- [x] "Representatives" in each of the k partition sets, given as $z_1,\dots,z_K$ ✔
- [ ] Number of clusters K
- [ ] Set of feature vectors $S_n = \{x^{(i)}|i=1,\dots,n\}$

✔

Solution:

Clustering outputs (1) partitions of data indices into each of the K clusters and (2) representative in each cluster.

Remark: Deciding a good partition and good representatives are two tasks that are intertwined. In the next sections, we discuss (1) how to measure the "goodness" of certain assignment and representative selection and (2) how to collectively find good assignment and good representatives.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Lecture 13. Clustering 1 / 5. Clustering Definition