Dimensionality reduction & clustering

Lecture 18

Changho Suh

October 5, 2021

Outline

1. Revist the role of clustering.

2. Study the most popluar clustering method and its simple variant:

   K-means algorithm

   K-medoids algorithm

3. Explore another popluar method:

   Hierarchical clustering (agglomerative clustering)

Role of clustering

Suppose: Data distribution is pretty wide

We need lots of data examples to ensure good generalization performance.

In case # of examples is not so big, one may want to classify the examples such that the distribution of the classified examples is concentrated.

Clustering is often employed for such classification.

K-means in words

K indicates the number of resultant clusters.

“mean” serves as a representative of each cluster.

How *K*-means works

1. Choose *K* points randomly.

2. *(Assignment step):* Map each data point to either one of the *K* points depending on its distance.

3. *(Update step):* Compute the means of such *K* clusters.

4. Repeat 2 & 3 until assignment is not changed further.

A variant of *K*-means

K-medoids

Very similar to *K*-means.

The only distinction is that we take "median (medoid)" instead of "mean".

Note: Robust to outliers.

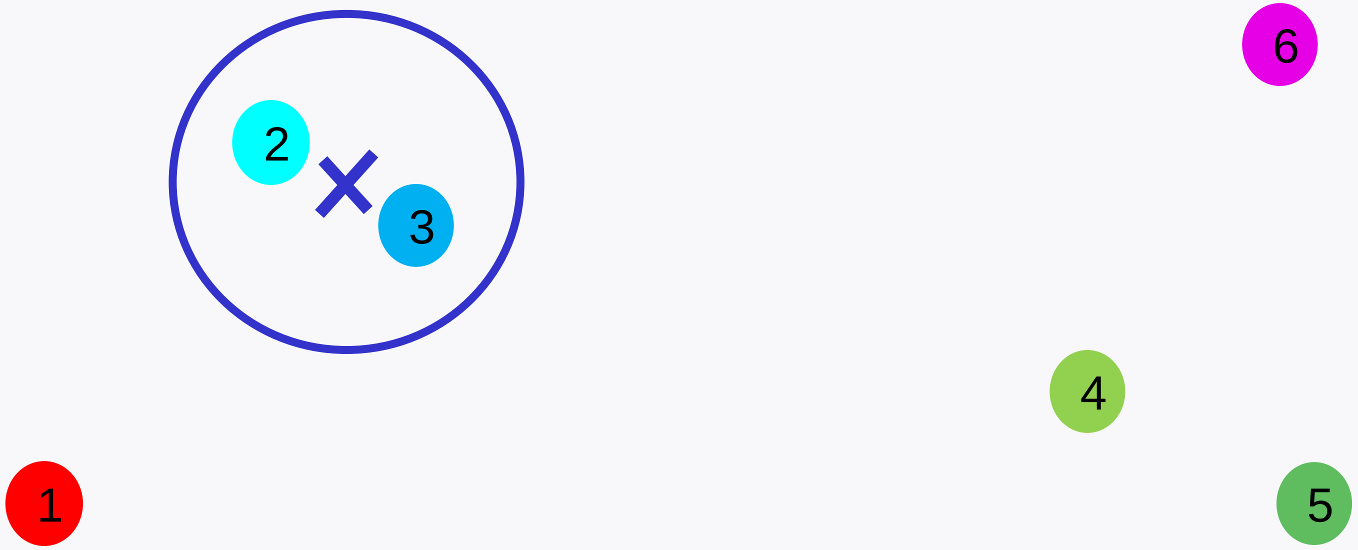
Hierarchical clustering in words

Do clustering in a **hierarchical** manner.

1. Start with the largest number of clusters (same as data points).
2. Merge clusters according a certain rule.
3. Repeat such merging until we reach down to K clusters.

How hierarchical clustering works

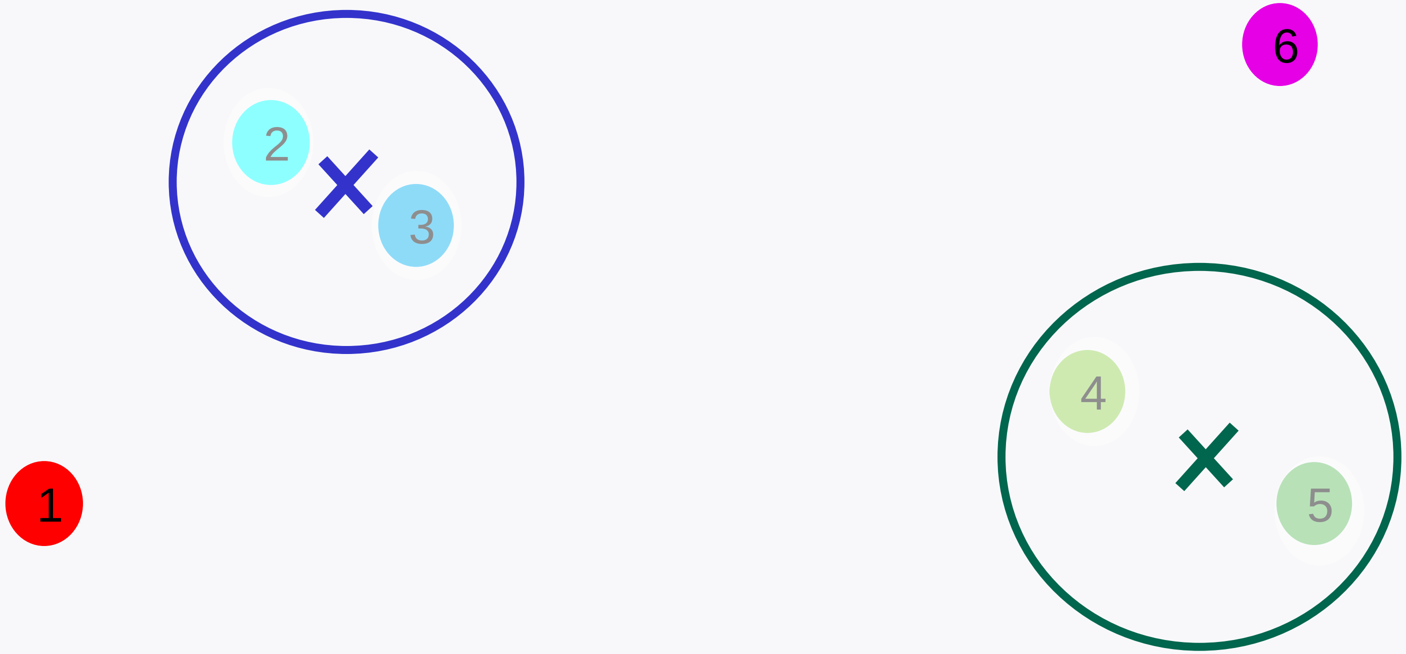
Start with 6 clusters (same # of data points)



1. Choose a pair of cluster centroids with minimum distance

2. Compute the centroid of the pair.

How hierarchical clustering works



3. Repeat Step 1 & 2: Choosing a pair of *updated* centroids with minimum distance, and then update centroids.

4. Repeat until *K* clusters are formed.

Dimensionality reduction + clustering?

How to combine?

Often: Do clustering **after** dimensionality reduction.

What is next?

From lots of project experiences with Hyundai Motor, found that many people are interested in:

1. Anamoly detection

2. Fusion learning

Important techniques for the problems

From lots of project experiences with Hyundai Motor, found that many people are interested in:

1. Anamoly detection

⇨ **autoencoder**

2. Fusion learning

⇨ **matrix completion**

Look ahead

Will study the two techniques:

1. **autoencoder**

2. **matrix completion**