Tutorial 2 for Chapter 3

A Review of K-means Clustering

Reference: 数据挖掘原理与应用

For the course AMA546 Statistical Data Mining
Lecturer: Dr. Catherine Liu
AMA, PolyU, HKSAR

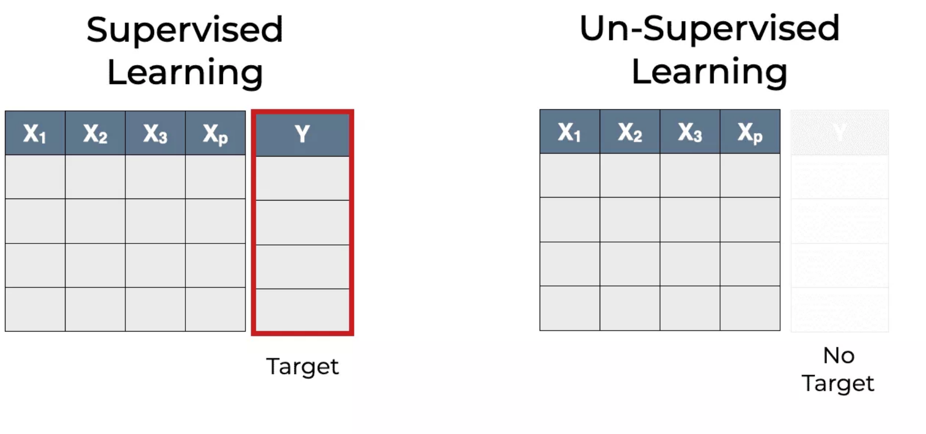
Content:

1 Clustering

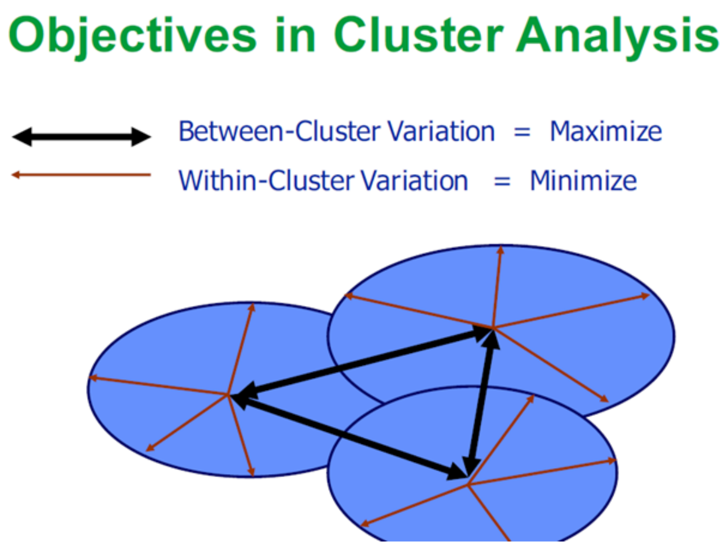
The **characteristic of clustering** task is that the observed data **only has features without labels**, which is called **unsupervised learning**.

Supervised learning and Unsupervised learning:

- **Supervised learning** involves the use of **labeled data** to train a model to make predictions or decisions based on new input data. **The goal of supervised learning** is to **minimize the error between the predicted output and the actual output**, also known as the **ground truth**. Common examples of supervised learning applications include **classification** and **regression**.
- **Unsupervised learning** involves the analysis of **unlabeled data** to identify patterns or structure within the data, so there is **no ground truth** in the unsupervised learning. In unsupervised learning, the algorithm learns to **identify meaningful patterns or relationships** among the input data, such as **clustering** or **dimensionality reduction**.



The clustering model needs to divide the observed samples into different groups according to their features. **The goal of cluster analysis** is to make the samples in the **same group have high similarity (minimize the within-cluster variance)** and the objects in **different groups have great divergence (maximize the between-cluster variance)**.



Normally, clustering algorithms generally uses the **iterative** technique that **involves trial and failure** to find the best group.

2 K-means

K-means is the most commonly used clustering algorithm.

- **Advantages**: **Simple**, easy to **interpreter**, **fast** computation.
- **Disadvantages**: It can only be applied to **continuous data** (the centroid of discrete data is not defined), and **the number of clusters** needs to be specified before clustering.

2.1 The concept of K-means algorithm

The **goal** of the K-means algorithm is to **divide n sample points into k groups**, **each group has a centroid**, and **each point in the group has a shorter distance to the centroid of the group to which it belongs than to the centroid of other groups**. In physics, centroid is the center of gravity of the points, assuming that the weight of each point is equal.

2.1.1 Centroid

Centroid is the core concept of K-means clustering algorithm. Centroid is the **central point** obtained by calculating the **mean value of each coordinate of all samples in a group**. The formal definition is stated in **section 2.1.2**.

For example, assume we have four points A, B, C and D in $\mathbf{R}^2$.

Item	x1	x2
A	7	9
B	3	3
C	4	7
D	3	8

Then the cluster centroids, or the **mean of all the variables within the cluster**, are as follows:

Centroid

Cluster	$\overline{x}_1$	$\overline{x}_2$
(A,B)	$\dfrac{7+3}{2}=5$	$\dfrac{9+3}{2}=6$
(C,D)	$\dfrac{4+3}{2}=3.5$	$\dfrac{1+8}{2}=4.5$

.

.

2.1.2 Mathematical definition of K-means clustering

Given a set of observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where each observation is a d-dimensional real vector, K-means clustering aims to partition the n observations into $k(\leq n)$ sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to **minimize the within-cluster sum of squares (WCSS)**. Formally, the objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{1}$$

where $\boldsymbol{\mu}_i$ is the **centroid** of points in S_i, i.e.

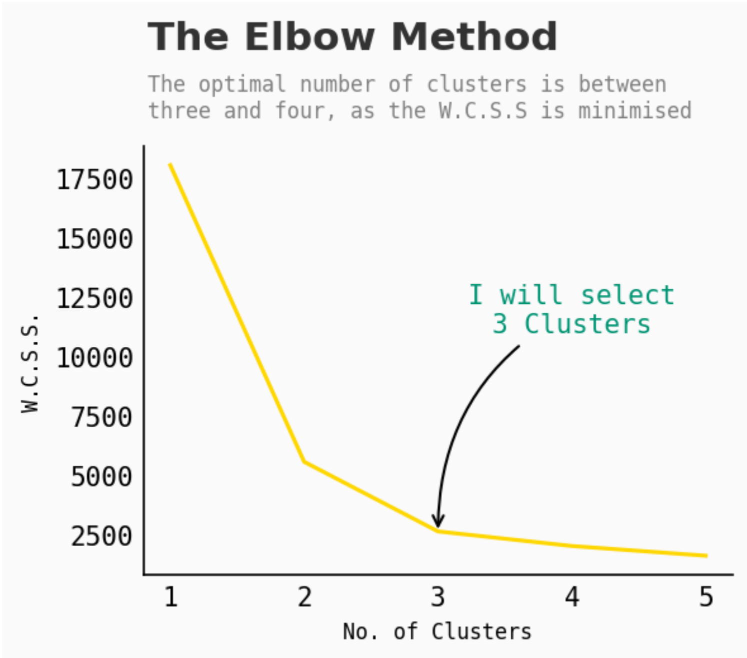
$$\boldsymbol{\mu}_i = \frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{x} \tag{2}$$

$|S_i|$ is the size of S_i, and $\|\cdot\|$ is the usual L^2 norm (Euclidian distance).

One can prove that **minimizing the within-cluster sum of squares (WCSS)** is **equivalent** to **maximizing the between-cluster sum of squared (BCSS)** (see [here (https://en.wikipedia.org/wiki/K-means_clustering#Description)](https://en.wikipedia.org/wiki/K-means_clustering#Description)). Therefore, the **objective of K-means** clustering is exactly **the same as** the **objective in cluster analysis** we introduced in seciton 1.

2.1.3 Method to Find the Best Value of K

The last problem is how to choose the most important parameter: the optimal number of clusters K. Here we will introduce the most common way to **determine the optimal K**: **Elbow Plot Method**.



Recall that the basic idea behind the k-means clustering is to define clusters such that the **within-cluster sum of squares (WCSS) is minimized**. The total WCSS measures the compactness of the clustering, and we want it to be as small as possible. The elbow method runs K-means clustering on the dataset **for a range of values of K** (say 1 to 10). In the elbow method, we **plot WCSS with respect to K** and look for the **elbow point where the rate of decrease shifts**, that is, the decreasing rate of WCSS is fast before elbow point and is slow after elbow point.

In the Elbow plot above, the elbow point appears in K=3.

2.1.4 K-means algorithm and clustering process

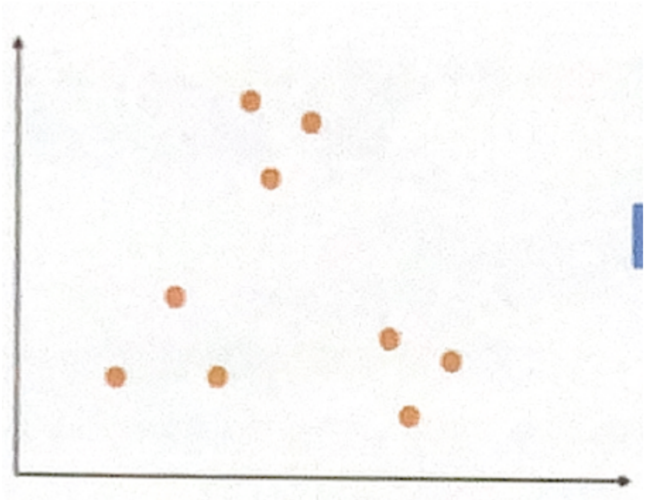
The following algorithm is the **K-means algorithm** we taught in class. We will explain the algorithm with illustrations to understand the process of K-means clustering.

```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

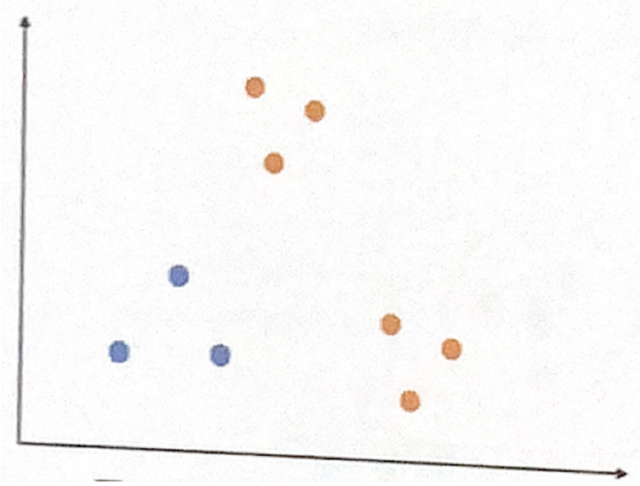
K-means always converges to a solution.
K-means reaches a state in which no points are shifting from one cluster to another

Introduction to Data Mining, 2nd Edition
Tan, Steinbach, Karpatne, Kumar

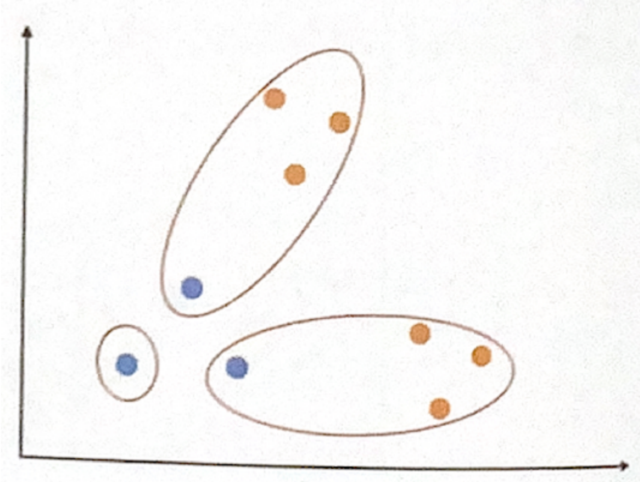
1. We have **9 yellow sample points** in the figure below, and let's say we want to divide it into **3 clusters**, so the K value of the K-means is set to 3:



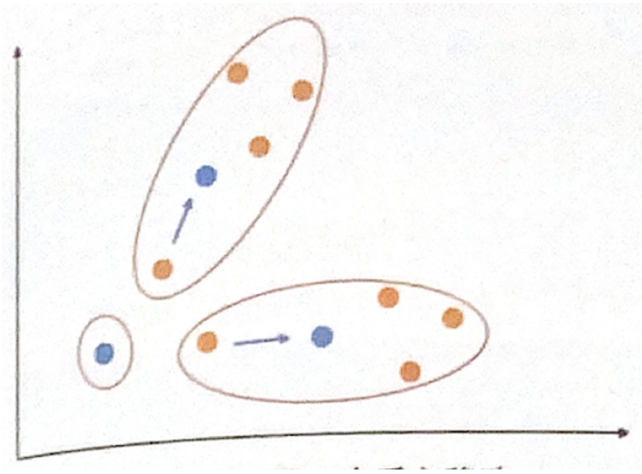
2. Following the first line in the algorithm, we **randomly select three points** (the three in the bottom left corner) and mark them blue: (*you can also select points other than the sample points in the space*)



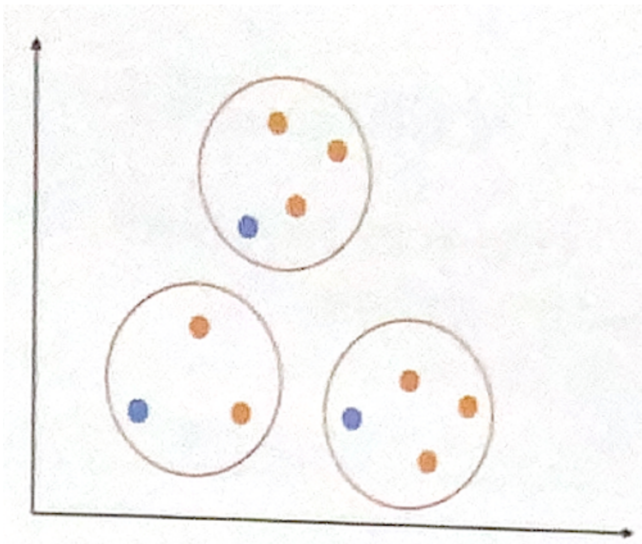
3. And then, we calculate the **distance from each yellow point to each blue point**, determine **which blue point is closest to each yellow point**, and **group** them together (line 3 in the algorithm). Clusters are circled in red:



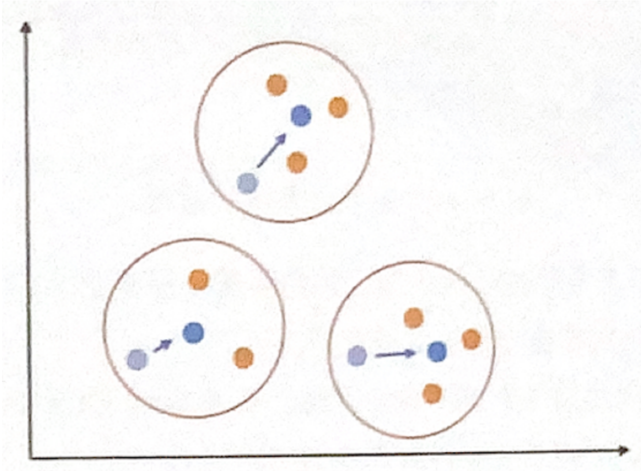
4. After that, we **recompute the centroid of each cluster** (line 4 in the algorithm). The formula of the centroid is stated above. The new centroid are colored in blue and the arrow shows the movement of the centroid:



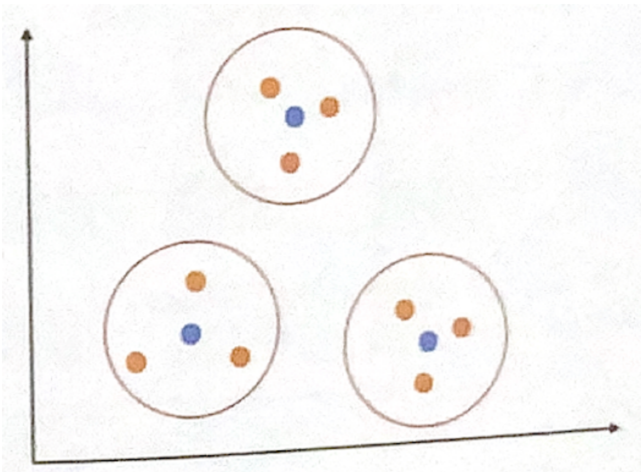
5. The first cycle is over. Since the centroid changed, we **start the second loop** (line 5 in the algorithm). We **assign all points to the cloest centroid** like what we did in step 3 (line 3 in the algorithm):



6. Then we **recompute the centroid of each cluster** (line 4 in the algorithm):



7. The second cycle is over. Since the centroid changed, we **start the third loop** (line 5 in the algorithm). We **assign all points to the cloest centroid**. As you can see, all the points are assigned to the same group, and the **centroid doesn't change any more.** Therefore, K-means reaches a state in which no points are shifting from one cluster to another (line 5 in the algorithm) and **terminated.**



Summary:

This is **K-means clustering**, and the whole idea is to **minimize the distance between the sample point and the center of mass**.

We first **randomly initializing the centroid points**, then we iterated the following two processes:

1. Clustering the sample points with the **shortest distance from the centroid points** into a group;
2. Use the center point of the group as the **new centroid point**

until the centroid **no longer changes**.