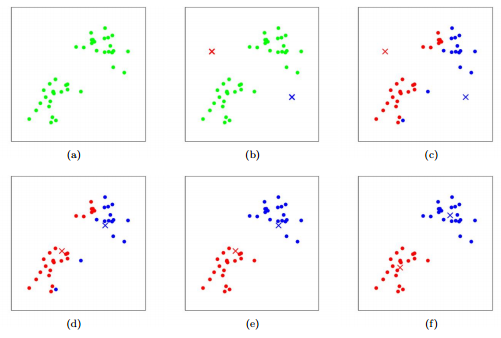
**k-means**

1. **introduction**

K-means, as an unsupervised learning, is largely used in different analysis regarding separating data into different groups. It regards each observation as a point and measure the distance among them. This method requires us to set the number of clusters at the beginning and it will provide a clustering result with the same number of clusters using the algorithm below.

1. **Algorithm**

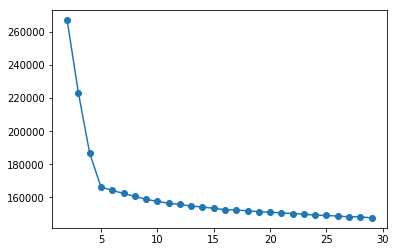
The following graphs describe the algorithm of K-means clustering.



http://stanford.edu/~cpiech/cs221/handouts/kmeans.html

1. Is the original dataset.
2. Assume we choose the number of clusters equals to two. This method will then first randomly choose two points.
3. Find out all the points that are closer to red cross than the blue cross and label them as red. The same for all the blue points.
4. Calculate the centroid (center) of each group of the points and mark them as new crosses.
5. Recalculate the distance of each point to both of the crosses. If the point is closer to blue, then label it as blue, otherwise red.
6. Repeat above procedures several times until no points will change colors (converge).
7. **How to choose the number of clusters.**

The method we are using is called “elbow method”. The idea of this method is to try different number of clusters and calculate how “well” each number of clusters performs. “well” is defined as: after clustering, how close the points are to their cluster’s center (mean value of all the points in the same cluster). The closer each point is to its cluster’s center, the better. We can put this information into a line graph as shown below. We use sum of squares (SSE, will be described more in details in latter section) to describe how close the points are to their center. The graph is drawn by using “Retail Community\_FOS Concepts” data. The y-axis is a measure of how close each point is to their cluster’s center (SSE).

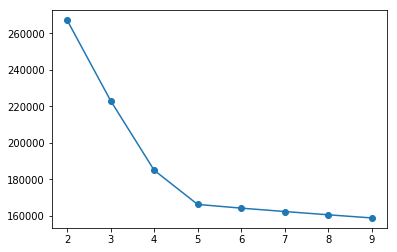


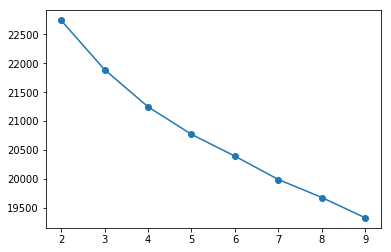
“Elbow method” tries to find out the elbow point in the above graph which could be 5 or 6. A smaller number of clusters results in a larger distance within each cluster which means the number we chose cannot describe the true segmentation well or the variation in each cluster is too large. A larger number of clusters contributes little to reduce such distance which indicates similarities among some of the clusters and we need to combine them together (we need to be aware that as we choose infinite number of clusters, the distance within each cluster will goes to zero).

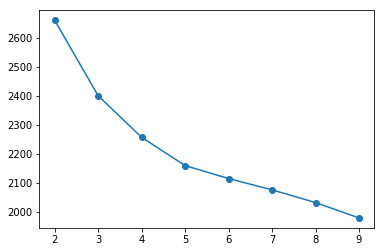
In the code, we cannot decide the number of clusters by looking at the graph. So, we choose the number of clusters by comparing this distance measure through some formula that will be described in the next section.

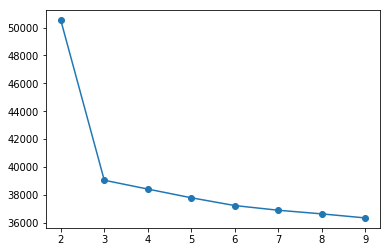
1. **The results on 5 datasets**

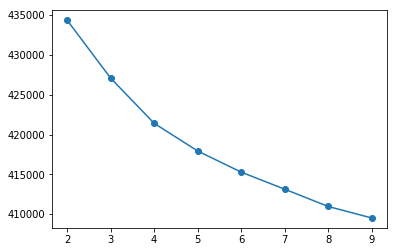
|  |  |
| --- | --- |
| Dataset | Number of clusters |
| Retail Community\_FOS Concepts | 5 |
| Toys Community\_Parents and Brand Websites | 4 |
| Healthcare Community\_Employer Wellness Portal | 4 |
| Retail Panel\_Brand | 3 |
| Media Panel\_Moviegoing | 4 |

 (1)

 (2)

 (3)

 (4)

 (5)

1. **Theoretical explanation**
2. How to calculate the distance among points.

In mathematics, people use the term “norm” to describe the way we calculate the distance.

Assume we have two points:and

Distance between two points in 2-norm.

Distance between two points in 1-norm.

2-norm is also called Euclidean distance which is the method we use every day. The distance measured in 2-norm in 3-dimension space is the length of line segment between 2 points, which is our common sense of “distance”.

There are infinitely many different norms we can use in measuring distance. But it’s rarely to use norm other than 1-norm and 2-norm.

The reasons we use 2-norm in K-means are as follows:

1. The functions in 2-norm are usually differentiable. Taking the derivatives of a function makes it possible to use gradient method to do optimization which is a usual part in most statistical methods.
2. There is little difference on clustering results by using these 2 norms.
3. An optimization problem with a 1-norm objective function and 2-norm objective function have similar results (prove in appendix). The reason we care about this is because the process of finding the proxy for centroid is also an optimization problem (will discuss in next session).

Appendix:

Proof:

Proof:

With these two statements, we have

Thus, to some extent, 2-norm is bounded by 1-norm on both sides and they are close to some extent.

1. sum of squares (SSE)

means the i-th cluster and is the mean value of all the points in i-th cluster.

Thus, the SSE is the sum of the variance in each cluster. And if the number of clusters is well-set, then SSE is supposed to be small.

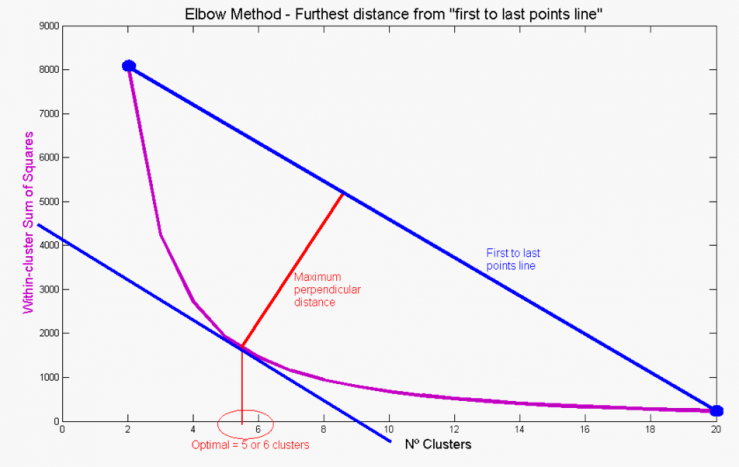
1. How to find out the centroid.

Then centroid means the “center” of a cluster. In the introduction, we use the mean. The way we get it is to minimize the SSE over and we get is the mean value of all the points in each cluster.

And get

Notice that if we use 1-norm instead of 2-norm, the centroid is the median rather than mean value.

1. How to find out the elbow



<https://quantdare.com/clustering-two-company-three-crowd/>

To calculate the distance between each point on the purple line to the blue line above, we use the process below.

Assume the points on the purple lines are and each of them has the attributes .

1. Get vector between first and last point - this is the line

The blue line is determined by the first and the last point on the purple line which can be expressed in vector form:

Then find out the unit vector with the same direction:

1. Vector between all points and first point

We also derive the vector through the first point and every point by using:

1. Distance to the line

To calculate the distance to the line, we split into two components, one that is parallel to the line and one that is perpendicular (vertical) . Then, we take the norm of the part that is perpendicular to the line () which is the distance.

We find the vector parallel to the line by projecting onto the line.

The perpendicular vector is

.

The distance of each point to the line is

1. Find out the elbow

After calculating all the distances, we choose the point with the largest distance as the elbow point.