K-means - Experimentation Notes

Blackbird65536

February 2024

§1 Introduction of the Problem

The main idea of k means is used in classification. Let's assume that we have some dataset D with 10^6 entries with the following conditions.

1. Each item $d \in D$ is a list of three numbers (like (2,3,5)), consisting of solely numerical data for each index.
2. Each $d \in D$ falls in one of eight different categories. For the sake of this example, we assume that each d represents the different cases for the signs of the entries. For example, the cases (2,3,5) and (-2,-3,-5) are on different cases. (This can also be expressed as the eight different octets of $\mathbb{R}^3$.)

So: the goal is the goal for the program to parse all of the data, and draw the lines that will classify the points into clusters, as well as classify how likely these points are in these clusters. The one that is most likely is placed in that cluster.

Example 1.1 (Voting)
Consider a population P with the size of the population ($|P|$) equalling 10^6. Then, we ask voters their stance on certain issues and score them based on three different metrics (Social Issues, Economic Issues, Foreign Issues). We say a negative score is one that leans more to the political left, and a positive score leans to the political right. We wish to classify the population P so that we can target advertisements that promote one side.

(a) We first have to figure out the 'center' of the people that are on the left side of the political spectrum, and similarly for the right.

(b) If we find the center, we can find a sufficient 'distance' away from the center of the right/left-leaning population sample so that we are sure as to whether or not the person is left-leaning, moderate, or right-leaning.

(c) In this case, distance is radial. Therefore, we must try to find a function p_{left} and p_{right}, so that it satisfies the following conditions:

1. As many of the right-leaning has a function p_{right} as close to 1 as possible, similarly for the left-leaning, and the moderate.
2. As few of the right-leaning have a high p_{left}, and the other way around.

I am going to attempt to make a concept of this, and try to put it in the github repository (which is turned into a blog.)