Final Project

Guanting Wei

Oct.12.2018

Blackground

People in different area will have different preferance. It is one of the most imprtant points that merchant will consider in mercantilism. In order to know preferance of different customers, We need to analyse data so that we can formulate a better business strategy.

Data set

I use a data set called "Wholesale customers data". This data set includes 440 samples, 8 variables(2 are type variable; 6 are numerical variables) and no missing values.

Choosing variables

Due to two type variables included in the data set, I think they will have great impact on the result. As a result, I choose 6 numerical variables only in this project.

Numerical treatment

Due to the difference between each variable, and in order to calculate the distance more precisely, I use standardization for each variable

purpose

My goal is to use one(or more?) clustering techniques to segment customers. Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Thus, there is no outcome to be predicted, and the algorithm just tries to find patterns in the data.

Method

k-means clustering

Given an initial set of k means $m_1^{(1)}$,. . . ,$m_k(1)$, the algorithm proceeds by alternating between two steps:
1.**Assignment step:**Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.

$$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2 \ \forall j, 1 \leq j \leq k\}$$

where x_p is assigned to exactly one $S^{(t)}$, even if it can assigned to two or more.
2.**Update step:** Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

The algorithm has converged when the assignments no longer change.