K-Means Overview

Ahmed Alawami

Colorado State University

November 19, 2015
Clustering: The Problem

Some Clustering Methods

Approach

Examples

K-Means

Properties

Minimizing Criterion

Hard Problem

Algorithm

Algorithm Illustrated

Some Issues

Local Optimum Illustration

Number of Clusters?

Other Issues

R

The End
Clustering: The Problem
We want to find subgroups (clusters) in our data set.

Contrast this to classification where we want to assign a class to each observation.
1. K-Means
2. Hierarchical Clustering
We will partition the data into distinct groups such that the observations within each group are similar to each other, while observations in different groups are different from each other.

We can cluster on either observations or features. (e.g. genome data)
Cancer data
Market Segmentation
The goal is to partition the data into $K$ clusters.

Disadvantage: we need to specify the number of clusters (contrast to hierarchical clustering)
\[ C_1 \cup C_2 \cup \cdots \cup C_K = 1, \cdots, n \]
\[ C_i \cap C_j = \emptyset ; i \neq j \]
We want to minimize within cluster variation:

$$\min_{C_1, \ldots, C_K} \sum_{i=1}^{K} W(C_i)$$

For Euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} \|x_i - x_j\|^2$$
There are $K^n$ ways to partition the data into $K$ clusters.

NP-Hard problem.
1. Randomly assign each observation to one of the K cluster.
2. Iterate until cluster assignment stop changing:
   a. For each of the K clusters, compute the cluster’s centroid.
   b. Assign each observation to the cluster with the closest centroid.
Clustering: The Problem

K-Means Algorithm Illustrated

Some Issues

The End

Data

Step 1

Iteration 1, Step 2a

Iteration 1, Step 2b

Iteration 2, Step 2a

Final Results

Ahmed Alawami

K-Means Overview
Local Optimum

It is important to run the algorithm several times with different initial cluster assignment.

We choose the solution with the smallest value to our objective function.
Clustering: The Problem
K-Means Algorithm Illustrated
Some Issues
R
The End

Local Optimum Illustration
Number of Clusters?
Other Issues

K=2

K=3

K=4

Ahmed Alawami
K-Means Overview
Should we standardize observations (or features)?

Validation: Are we clustering noise or are these true subgroups?

K-means forces every observation into a cluster (Mixture models)

Not robust to perturbation to the data
km.out = kmeans(data, number.of.clusters, num.start)
plot(data, col=(km.out$cluster + 1))
Thank You For Listening

Questions and/or Comments
References

- Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning.
- James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning