### K-Means Overview

Ahmed Alawami

Colorado State University

November 19, 2015

イロト イヨト イヨト イヨト

#### Outline

Clustering: The Problem K-Means Algorithm Illustrated Some Issues R The End

### 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.

イロト イヨト イヨト イヨト

Some Clustering Methods Approach Examples

### 1. K-Means

2. Hierarchical Clustering

◆□ ▶ ◆□ ▶ ◆ □ ▶ ◆ □ ▶ →

Ð,

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)  $% \left( {\left( {{{\rm{c}}_{\rm{c}}} \right)_{\rm{c}}} \right)_{\rm{c}} \right)$ 

- Cancer data
- Market Segmentation

ヘロト 人間 とくほど 人間とう

æ

## The Goal

Properties Minimizing Criterion Hard Problem Algorithm

The goal is to partition the data into K cluster.

Disadvantage: we need to specify the number of clusters (contrast to hierarchical clustering)

イロト イヨト イヨト イヨト

Properties Minimizing Criterion Hard Problem Algorithm

•  $C_1 \cup C_2 \cup \cdots \cup C_K = 1, \cdots, n$ 

$$C_i \cap C_j = \phi ; i \neq j$$

(日)

Properties Minimizing Criterion Hard Problem Algorithm

We want to minimize within cluster variation:

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

For Euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2$$

◆□ ▶ ◆□ ▶ ◆ □ ▶ ◆ □ ▶ ●

æ

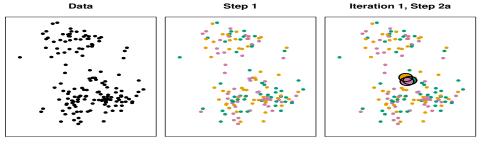
- ▶ There are *K<sup>n</sup>* ways to partition the data into *K* clusters.
- NP-Hard problem.

イロン イヨン イヨン

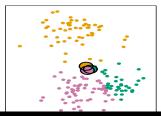
- 1. Randomly assign each observation to one of the K cluster.
- 2. Iterate until cluster assignment stop changing:
  - a. For each of the K clusters, compute the cluster's centroid.
  - b. Assign each obseration to the cluster with the closest centroid.

イロト イポト イヨト イヨト



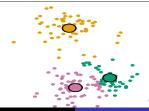


Iteration 1, Step 2b



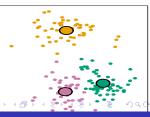
Iteration 2, Step 2a

**Final Results** 



Ahmed Alawami





Local Optimum Illustration Number of Clusters? Other Issues

## 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.

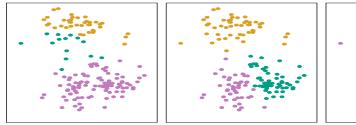
イロト イヨト イヨト イヨト





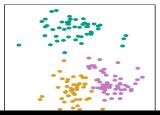
235.8





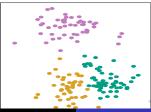






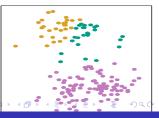
235.8

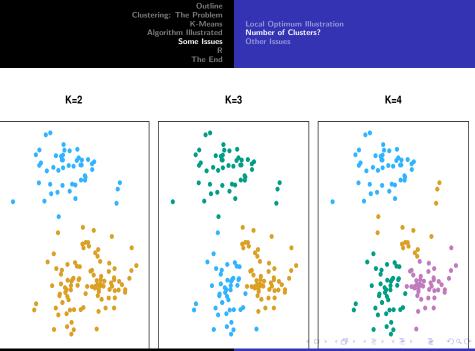
310.9





K-Means Overview





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

Outline	
Clustering: The Problem	
- K-Means	
Algorithm Illustrated	
Some Issues	
R	
The End	

# km.out = kmeans(data,number.of.clusters,num.start) plot (data,col=(km.out\$cluster+1))

References

Thank You For Listening

Questions and/or Comments

イロト イヨト イヨト イヨト

Outline Clustering: The Problem K-Means Algorithm Illustrated Some Issues R The End	References
---	------------

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

周 と く ヨ と く ヨ と …