Intro to Data Science:

K Means

Ed Podojil Data Engineer/Scientist, Animoto

Warm Up:

Agenda

I. Homework Review/Discussion

II. Supervised Learning: What do we know?

III. Cluster Analysis

IV. K Means Clustering

V. Interpreting Results

VI. K Means Clustering in R

I. Homework Review/Discussion

Small Group Goals

- Refer to each article in small groups
- Summarize the main points, methodologies, and results
- Be prepared for a full class discussion on each!

II. Supervised Learning: What do we know?

Goals

- Review Supervised learning techniques
- Best practices: what do we use when?
- Where do we go from here?

III. Cluster Analysis

Goals

- What is a cluster?
- What is the purpose for cluster analysis?
- how do you solve a clustering problem?

	continuous	categorical
Supervised		
Unsupervised		

	continuous	categorical
Supervised	regression	classification
Unsupervised	dimension reduction	clustering

What is **a cluster**?

Definition: A group of similar data points.

The concept of similarity is central to the definition of a cluster, and therefore to cluster analysis.

Greater similarity between points = better clustering.

What is the purpose of cluster analysis?

Enhance our understanding of data by dividing the data into groups.

Clustering provides abstraction from individual data points.

Goal: extract/enhance the natural structure of the data (not to impose arbitrary structure!)

How do you solve a clustering problem?

Think of a cluster as a "potential class," then the solution to a clustering problem is to programatically determine these classes.

The real purpose of clustering is data exploration, so a solution is anything that contributes to your understanding.

Clustering - Review

A cluster is a group of data points with great similarity

We use cluster analysis to understand the natural segmentation of a dataset

Clustering problems can be solved programatically to find order in the data!

Class break

IV. K Means Clustering

Goals:

- What's K means clustering?
- What are partitions, and what do you use them for?
- How do we determine the center of a cluster?

What is K Means Clustering?

A method of cluster analysis that partitions *n* observations in a data set to *k* clusters.

Observations: rows in a data set

Partition: Set all values to have only one cluster

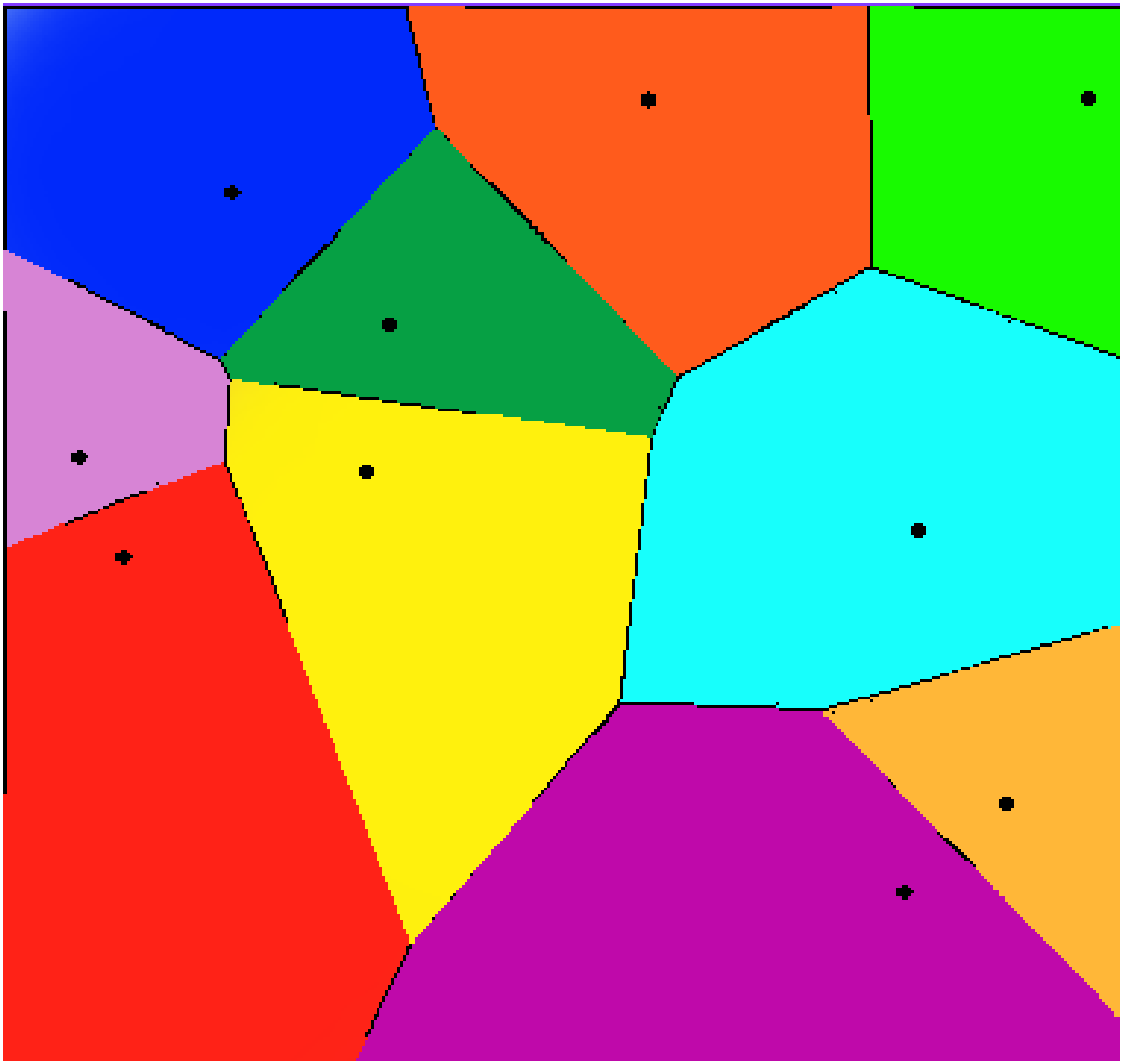
How are partitions determined?

Each point is assigned to the cluster with the nearest **centroid**

centroid: the mean/middle of the data points in a cluster

→ requires continuous features (integers, for example)

Partion Example: Voronoi Diagram



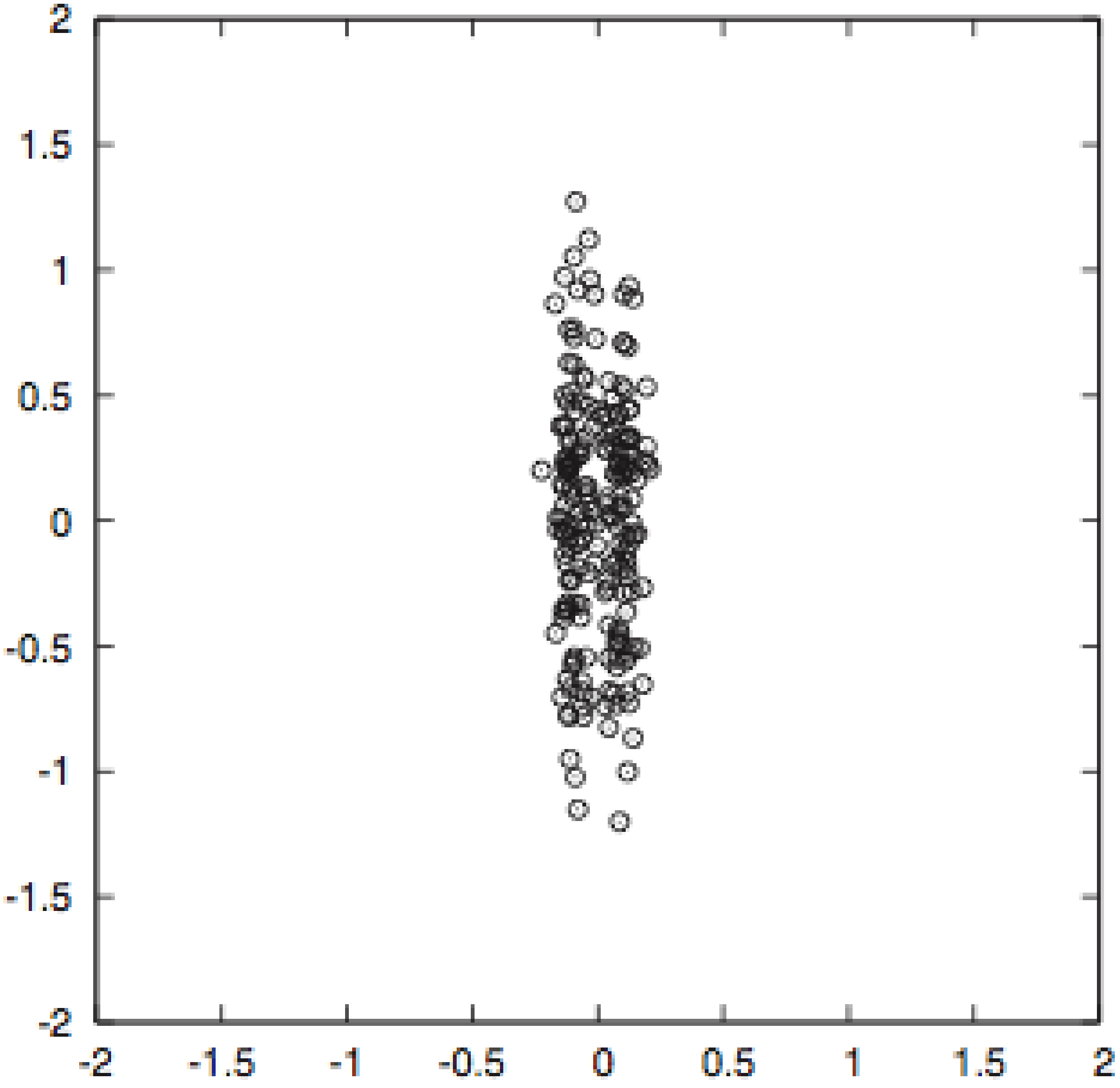
Scale Dependence

Keep in mind that partitions are variable dependent!

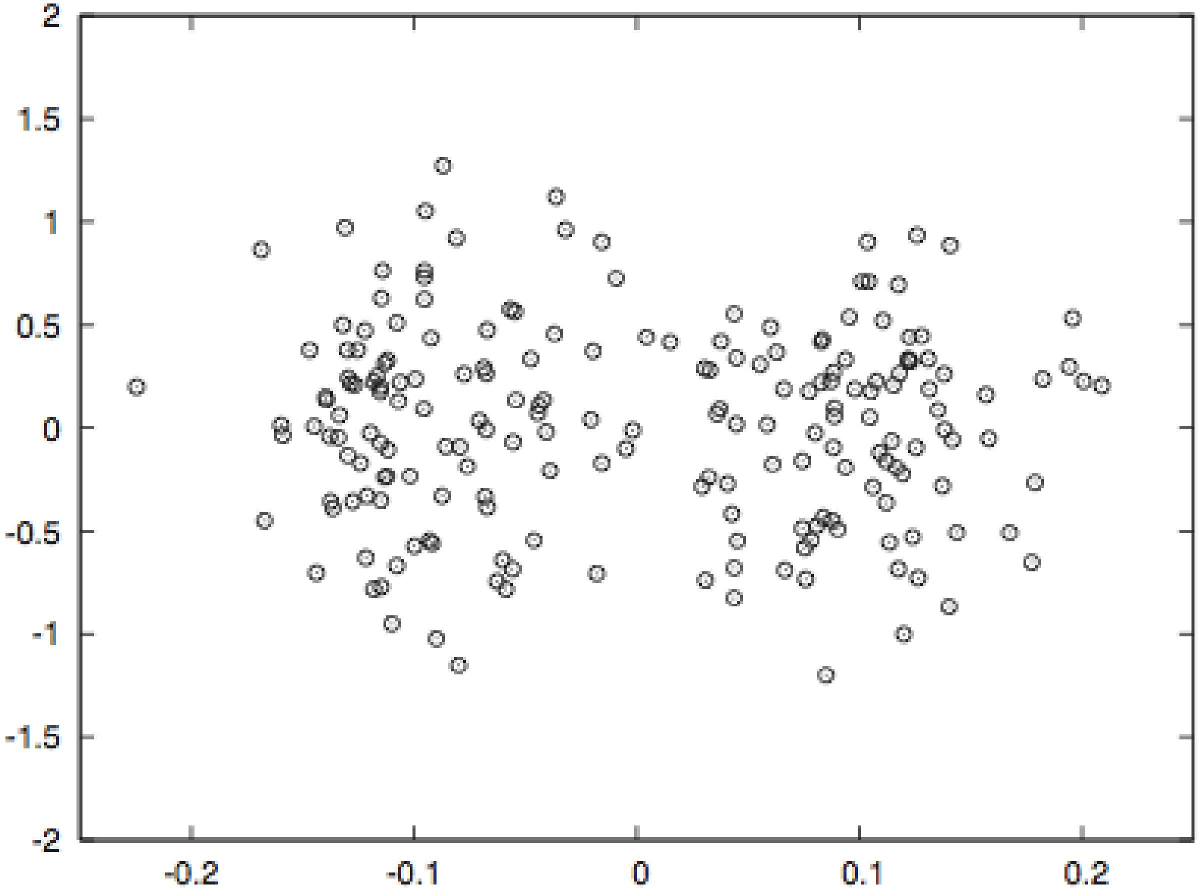
This means your data and paritions can have different results depending on your scale and variables

Therefore it’s important to think about your data representation before applying a clustering algorithm.

Consider this representation of two variables of data

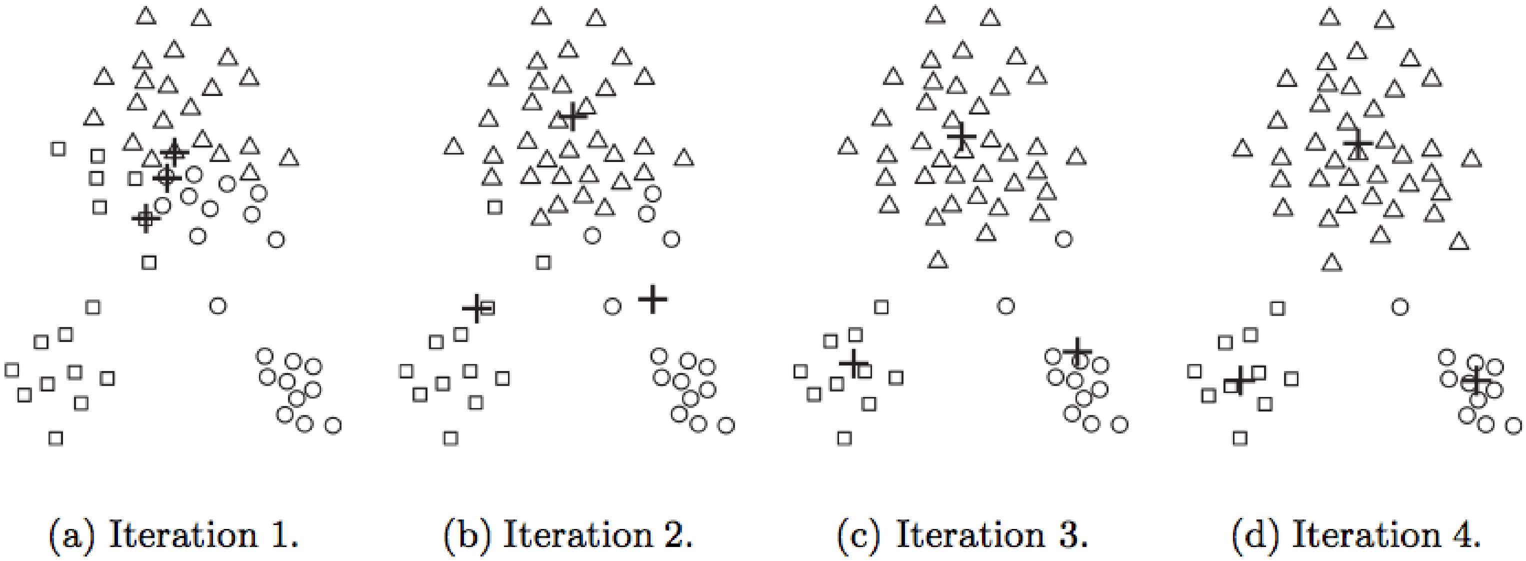


How does this new representation change our interpretation?



K Means Basics

Choose K initial centroids
For each point, find distance to nearest centroid and assign closest centroid
Recalculate centroid positions with new means/middles
Repeat until "stopping" criteria is met


(a) Iteration 1.

(b) Iteration 2.

(c) Iteration 3.

(d) Iteration 4.

Figure 8.3. Using the K-means algorithm to find three clusters in sample data.

Strengths and Weaknesses - K Means

+ Algorithmically efficient (compute time is fast!)

- Hard to work with non-convex clusters, or data with widely varying shapes and densities.

(May be overcome by increasing k and creating subclusters)

Determining Similarity

Usually we determine similarity using the *Euclidiean distance*

You might remember hearing something similar before in your math classes: *Pythagorean Distance*

For the most part, we can think of this as the distance between points using a ruler

Stopping

Like regressions, we will use the sum of squared errors to determine clustering performance

Stopping in K Means occurs when centroids change by no more than some distance or change by less than some %

Remember that different runs will converge (or stop) differently!

K Means - Review

K Means is a method of clustering for K clusters using partitions

Partitions are determined using means and Euclidean Distance

Partitions are finalized when we reach a minimum in performance over iterations

V. Cluster Validation

Goals

- What do we use to validate clusters?
- What's the silhouette coefficient?
- How do we determine k?

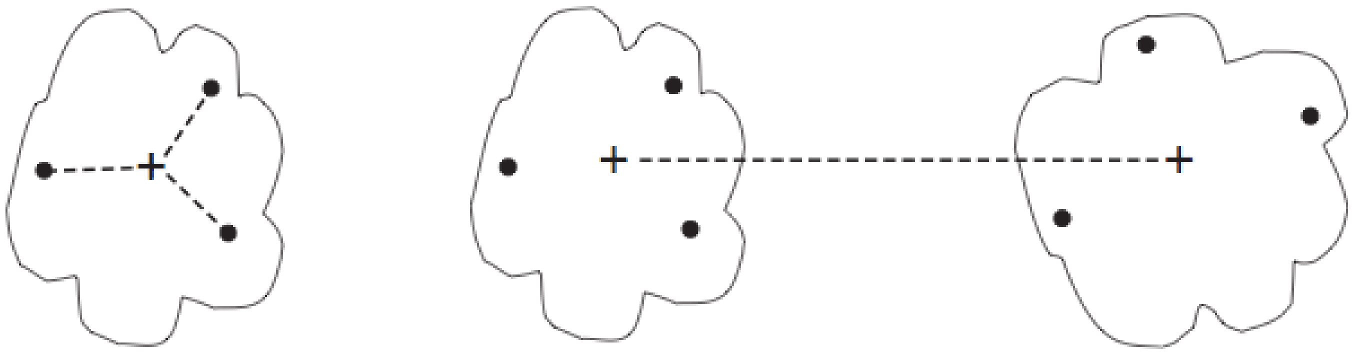
How do we validate clusters?

With K Means, we will use two validation metrics:

Cohesian: Clustering effectiveness within a cluster

Seperation: Clustering effectiveness between clusters

Cluster Validation



(a) Cohesion.

(b) Separation.

Figure 8.28. Prototype-based view of cluster cohesion and separation.

Cluster validation measures can be used to identify clusters that should be split or merged, or to identify individual points with disproportionate effect on the overall clustering.

Silhouette Coefficient

The **silhouette coefficient** combines both cohesion and seperation

It can take values between -1 and 1

We want seperation to be high (clusters further apart) and cohesian to be low (clusters close to their nodes)

When this is opposite, our clusters tend to overlap (not always bad, but usually bad)

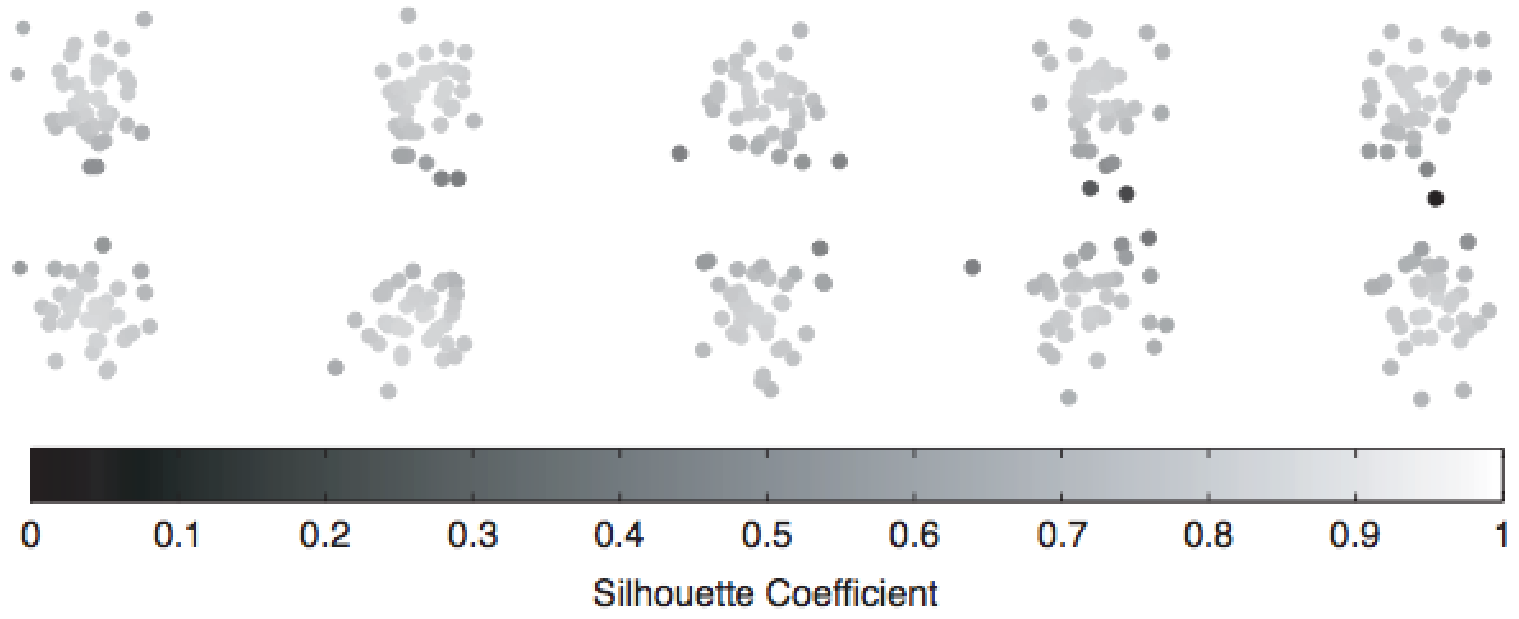


Figure 8.29. Silhouette coefficients for points in ten clusters.

What do we use this for?

The **silhouette coefficient** is very useful determine best value for k!

We can compute the best values of silhouette coefficient (or sum of squared errors) for each value of k

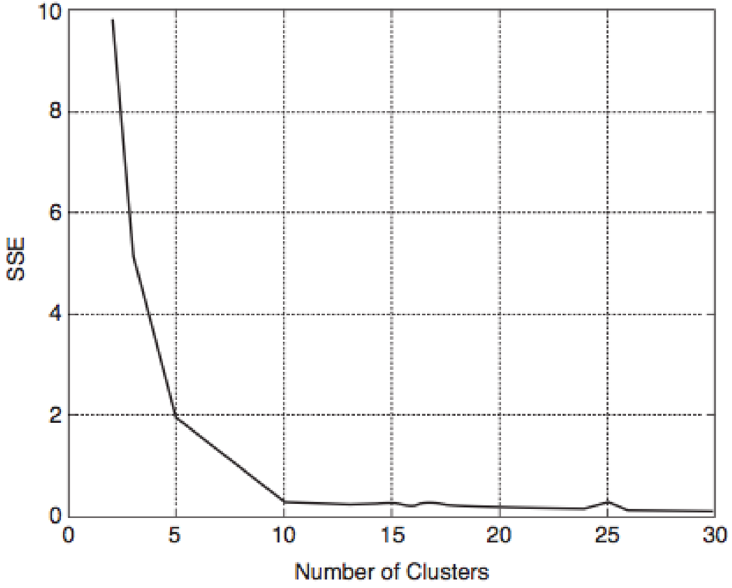


Figure 8.32. SSE versus number of clusters for the data of Figure 8.29.

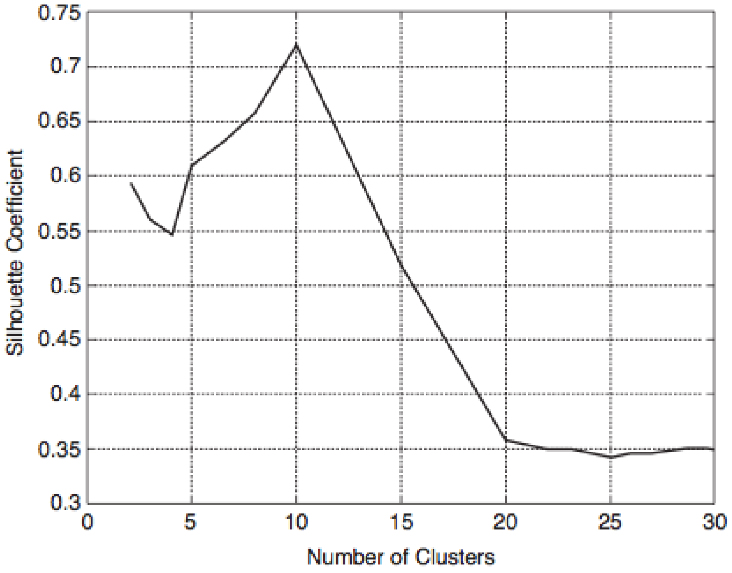


Figure 8.33. Average silhouette coefficient versus number of clusters for the data of Figure 8.29.

We can determine confidence statistically, but for now it's more important to get have a person's perspective for our clusters to be meaningful or useful to the data

Cluster Validation – Review

Clusters can be validated using seperation and cohesian

Generally we want our clusters further apart from each other, but edges closer to the centroids

We can use SSE and silhouette coefficient to determine best number of clusters

VI. K Means in R

Group Discussion

Questions to ponder about:

What are the similarities between KNN and K Means? (beyond "they both start with k"!)

Where do you think k means could be useful? Where do you think it would not be useful?

What kind of issues do you imagine running into when using K means?