Business Analytics

Session 5a. Clustering

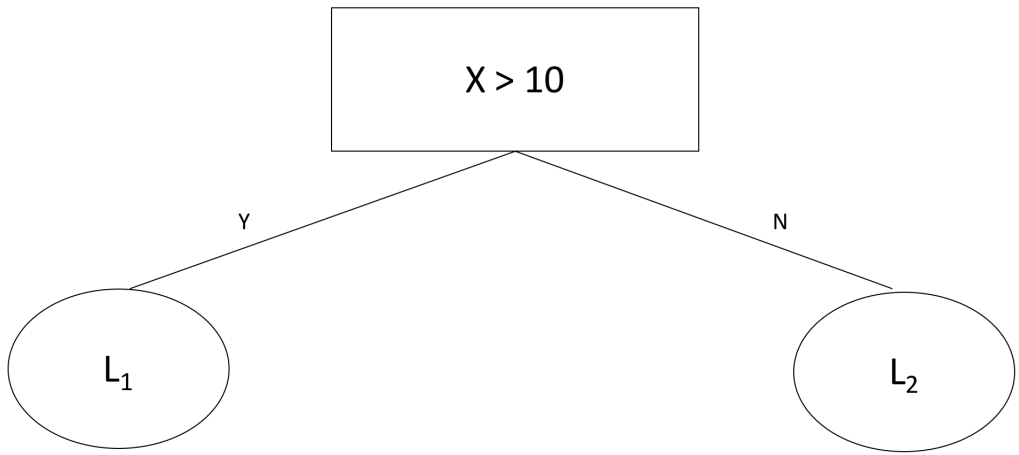
Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

Exercises

- (T/F) By constructing a tree, we can increase the Gini index of a data set.
- What is the Gini index of the following tree:



	$y = 0$	$y = 1$
L_1	10	30
L_2	30	10

How about the Gini index of the set?

Recommender System of Netflix

Clustering as Unsupervised Learning

- Unsupervised learning finds patterns/relationships/structures in data, without being optimized to solve a particular predictive task (i.e., there's no outcome).
  - Examples: Image segmentation, customer segmentation, portfolio selection, industry analysis, and many more.

- Clustering: Segment the data into a set of homogeneous clusters of observations to generate insights.
  - Observations in the *same* group should be *similar* to each other.
  - Observations in *different* groups should be as *dissimilar* to other groups as possible.

Netflix

- Online DVD rental and streaming video services.
- More than 40 million subscribers and $3.6 billion in revenue.
- Key to Netflix: Offer customers accurate movie recommendations based on a customer's own preferences and viewing history.
  - From 2006-2009, Netflix ran a contest to predict user ratings for movies.
  - A grand prize of $1 million to who could beat their algorithm, Cinematch, by 10%, measured in RMSE.

Predicting the User Ratings

- Predicting the rating is crucial to their business.

- What data can be used in the prediction?
  - Every movie has the rating from all users who rated the movie.
  - Facts about the movie itself: Actors, director, genre classification, year released, etc.

Collaborative Filtering: Using Other Users' Ratings

	Men in Black	Apollo 13	Top Gun	Terminator
Amy	5	4	5	4
Bob	3		2	5
Carl		5	4	4
Dan	4	2		

Collaborative Filtering: Using Other Users' Ratings

	Men in Black	Apollo 13	Top Gun	Terminator
Amy	5	4	5	4
Bob	3		2	5
Carl		5	4	4
Dan	4	2		

- Question: Will Carl like the movie "Men in Black"?
  - Likely to happen.
  - Amy and Carl have similar preferences over other movies, and Amy rated "Men in Black" high.

- This technique is called Collaborative Filtering.

Content Filtering: Using Movie Information

- Amy liked "Men In Black"
  - Directed by Sonnenfeld
  - Classified in the genres of action, adventure, sci-fi and comedy
  - It stars Will Smith

Content Filtering: Using Movie Information

- Amy liked "Men In Black"
  - Directed by Sonnenfeld
  - Classified in the genres of action, adventure, sci-fi and comedy
  - It stars Will Smith
- Consider recommending to Amy:
  - Barry Sonnenfeld's movie "Get Shorty"
  - "Jurassic Park", also in the genres of action, adventure, and sci-fi
  - Will Smith's movie "Hitch"
- This is called Content Filtering.
- Netflix uses both collaborative filtering and content filtering.

Data

- We use the data from MovieLens to do content filtering using the clustering technique.
- Movies are categorized into different genres.
  - Action, adventure, animation, etc.
  - Each movie belongs to many genres.
- Goal: Systematically find groups of movies with similar sets of genres.
- Methods we will cover
  - Hierarchical
  - $k-$means

Distance between Clusters

- Single linkage: Distance between points that are closest.
  $\min ||\vec{X}_i - \vec{X}_j||$ where $i \in C_1$ and $j \in C_2$
- Complete linkage: Distance between points that are farthest.
  $\max ||\vec{X}_i - \vec{X}_j||$ where $i \in C_1$ and $j \in C_2$
- Average linkage: Average distance computed over all pairs of observations between the two clusters.
  Average of $||\vec{X}_i - \vec{X}_j||$ where $i \in C_1$ and $j \in C_2$
- Average group linkage: The distance between the centers of each cluster.
  $||\vec{\mu}_1 - \vec{\mu}_2||$ where $\vec{\mu}_i$ is the center of C_i $(i = 1, 2)$
- Ward's method: How much the sum of squares will increase if we merge them.
  $\sqrt{\frac{n_1 n_2}{n_1 + n_2}} ||\vec{\mu}_1 - \vec{\mu}_2||$ where n_i is the number of data points in C_i

Hierarchical Clustering

Hierarchical Clustering Algorithm

1. Start with each data point in a cluster of its own.
2. Until there is only one cluster, do the following:
   (i) Find the closest pair of clusters
   (ii) Merge them
3. Output the tree of cluster mergers.

Illustration of the Hierarchical Method

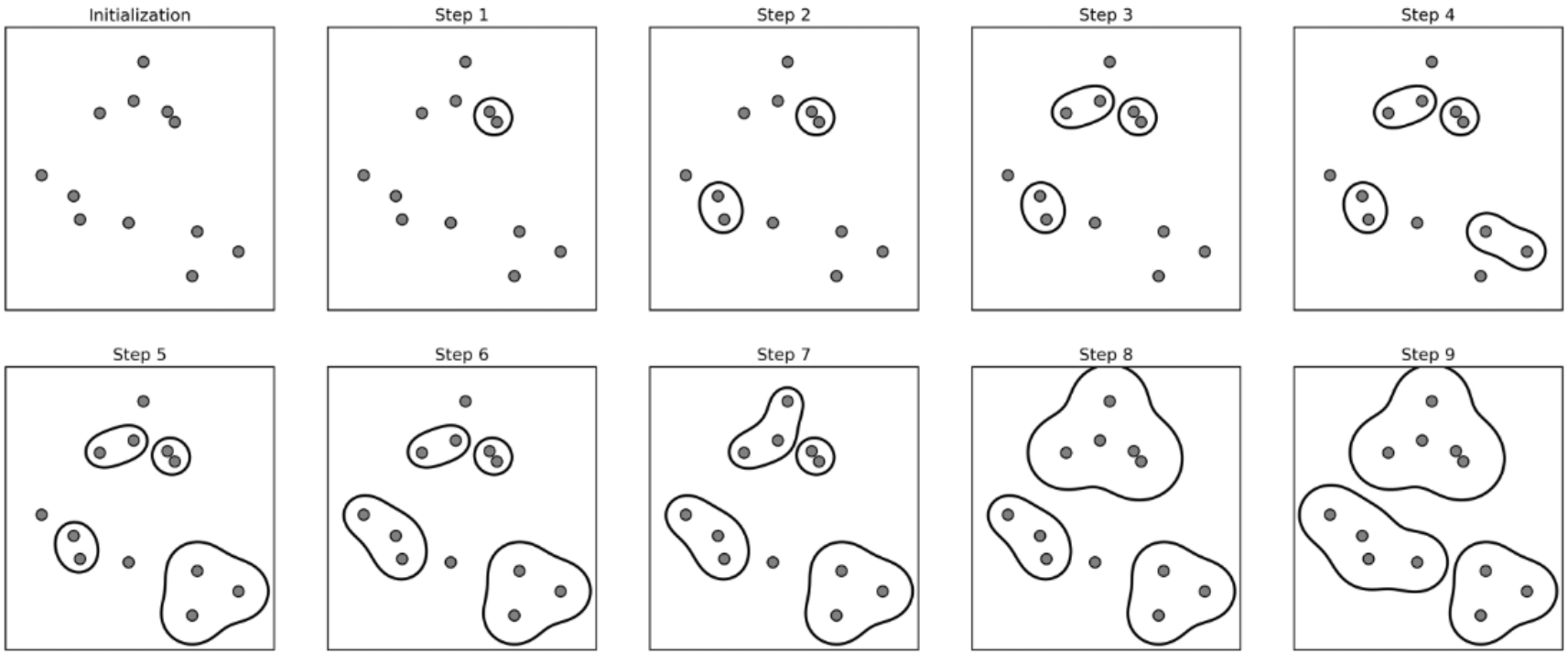
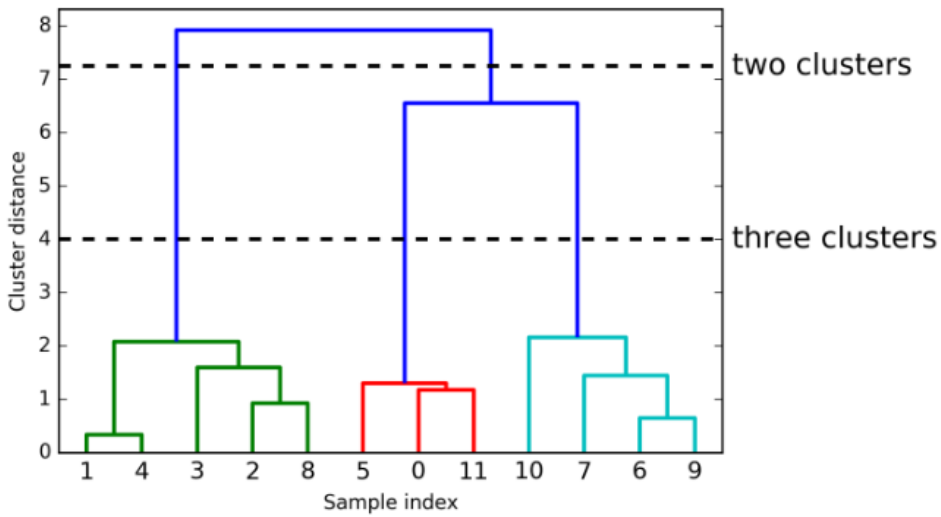
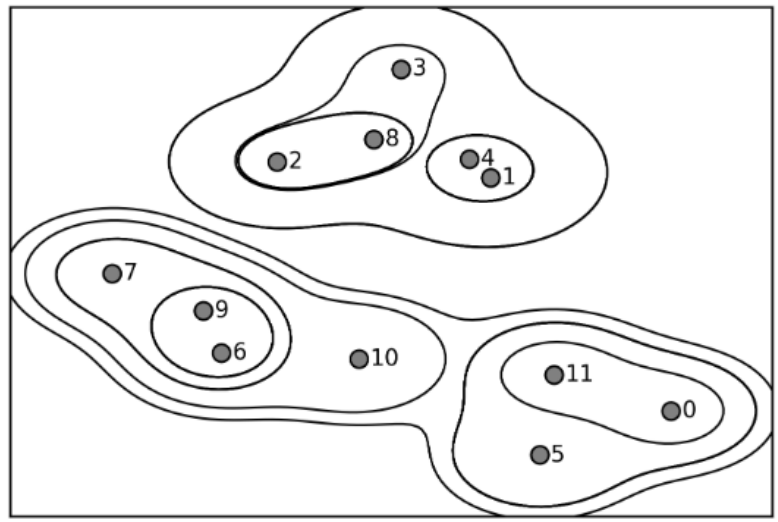


Illustration of the Hierarchical Method





Pros and Cons of Hierarchical Method

- Hight of vertical lines represents distance between points or clusters.
- Data points listed along bottom.
- Easily select how many clusters.
- May label different clusters as outcomes.

$k-$Means Clustering

$k-$Means Algorithm

1. Specify the desired number of clusters k.
2. Randomly assign each data point to a cluster.
3. Compute cluster centers.
4. Re-assign each point to the closest cluster.
5. Re-compute cluster centers.
6. Repeat steps 4 and 5 until no improvement is made.

$k-$Means Clustering

$k-$Means Algorithm

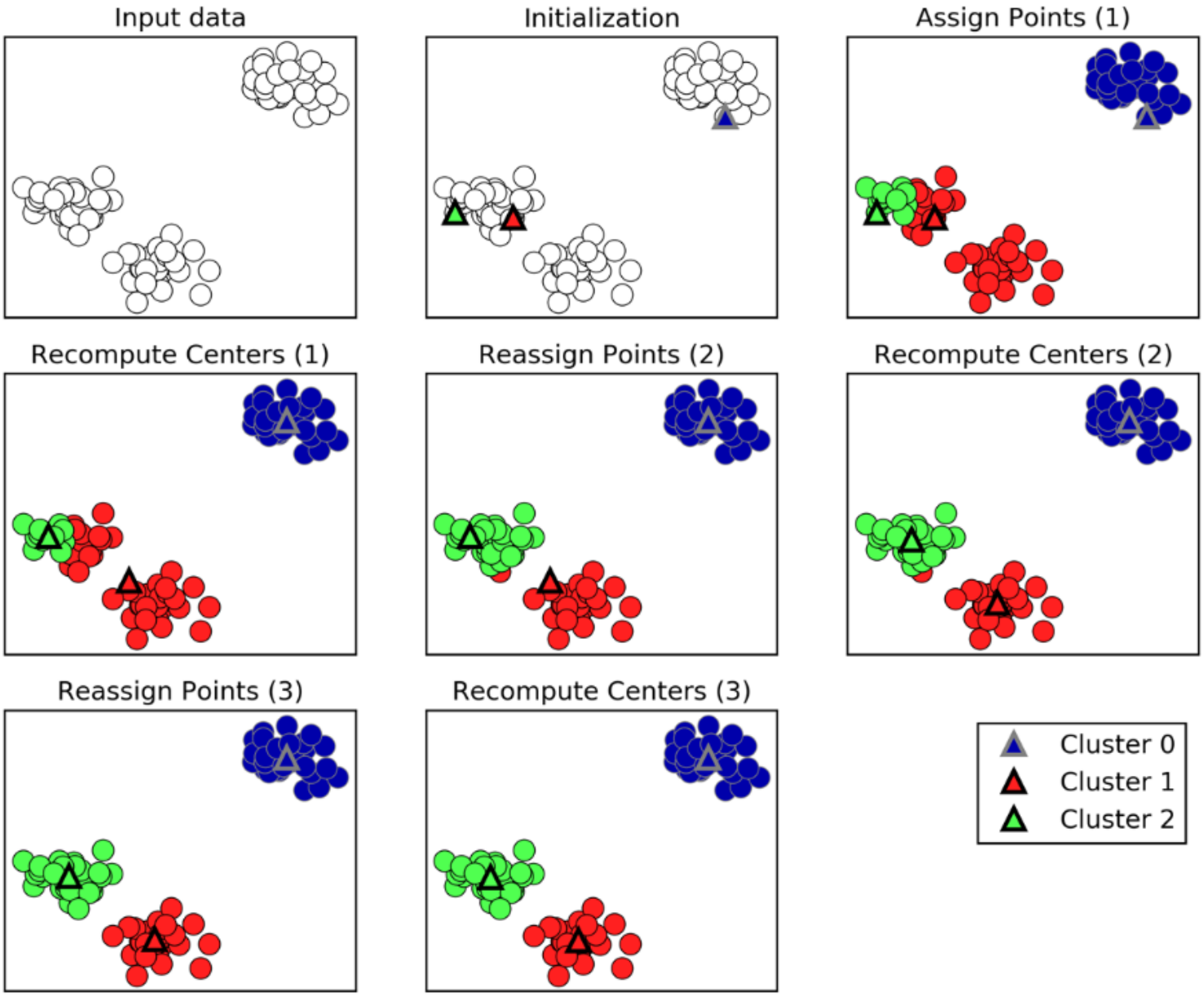
1. Specify the desired number of clusters k.
2. Randomly assign each data point to a cluster.
3. Compute cluster centers.
4. Re-assign each point to the closest cluster.
5. Re-compute cluster centers.
6. Repeat steps 4 and 5 until no improvement is made.

The objective of $k-$means is to minimize:

$$\sum_{j=1}^{k} \sum_{i \in C_j} (\vec{X}_i - \vec{\mu}_j)^2,$$

where C_j is cluster j and $\vec{\mu}_j$ is the center of C_j.

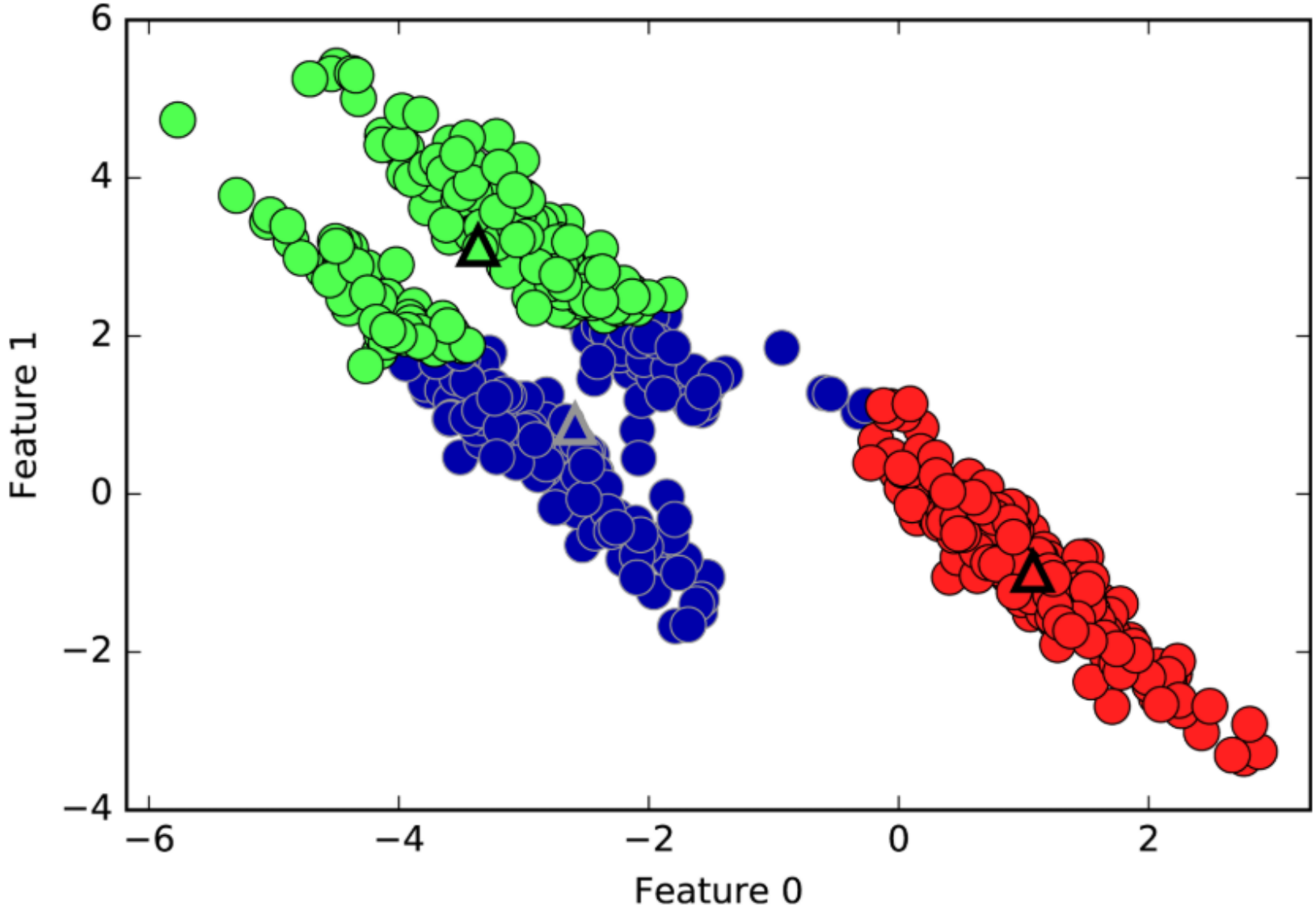
Illustration of the $k-$Means Method



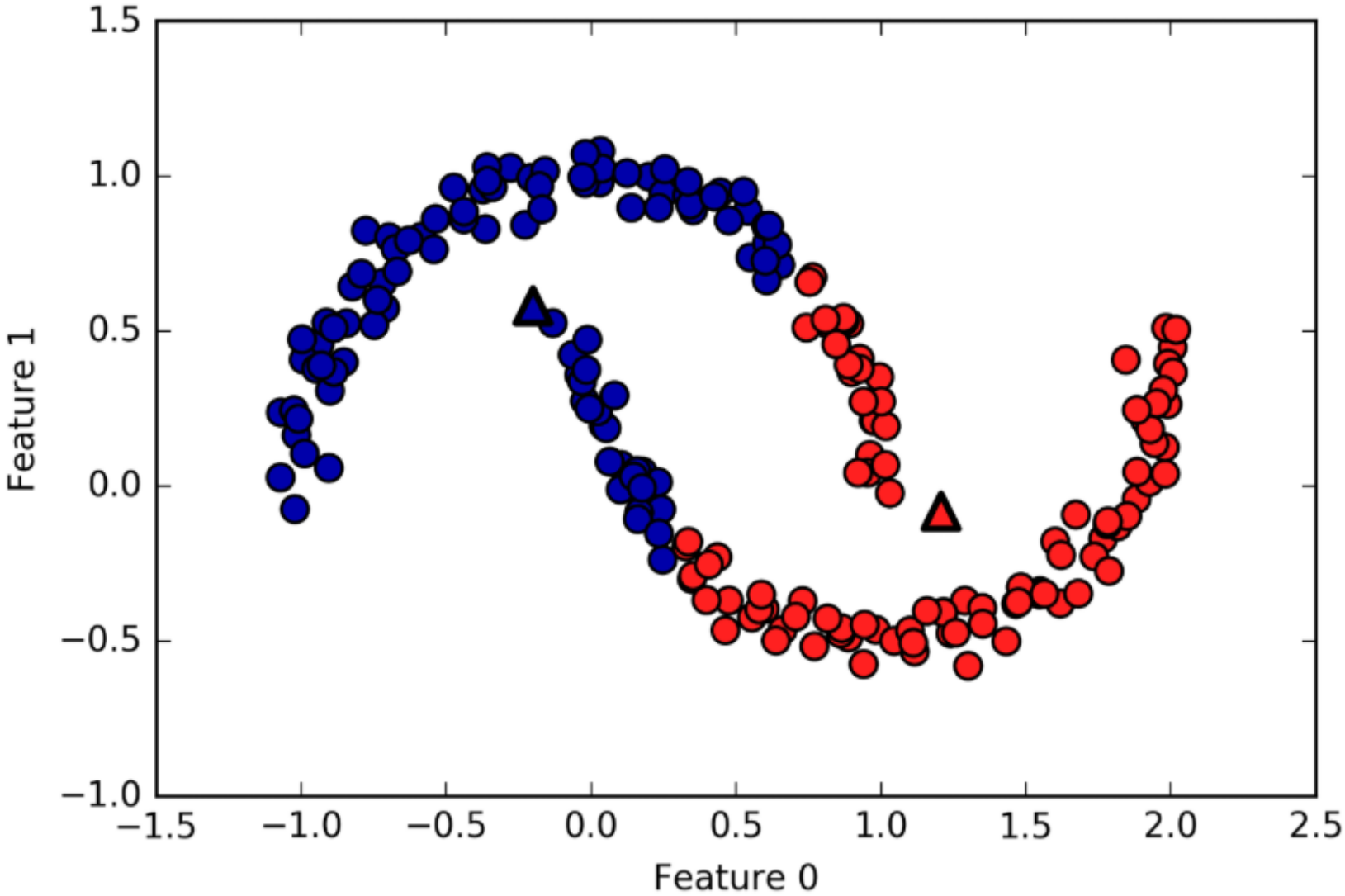
Comparisons of Two Methods

- Hierarchical clustering
  - Works well for small data sets
  - Convenient method if not sure about the number of clusters
  - Observes how clusters are nested
- $k-$means clustering
  - Suitable for a large data set if we know the number of clusters k.
  - Minimizes the squared error.
  - Cluster boundaries equidistant to centers: Cannot handle covariances well; Only simple cluster shapes.
- How do we choose k?
  - In general, the dissimilarity W will decrease as k increases.
  - Pick up k^* where there is a sharp decrease in successive differences of W.

Limitations of k–means



Limitations of k–means



Beyond Movies: Mass Personalization

- "If I have 3 million customers on the web, I should have 3 million stores."
  - Jeff Bezos, CEO of Amazon.com

- Recommender systems build models about users' preferences to personalize the user experience.
  - A new favorite band
  - An old friend who uses the same social media network
  - A book or song they are likely to enjoy

- Clustering algorithms, tailored to find similar customers and/or similar items, form the backbone of many recommender systems.