Unsupervised Learning

Josh Goldstein

July 17, 2019

Unsupervised vs Supervised Learning

Machine Learning techniques applied to data without a response/treatment (**unlabeled**).

- Clustering: Find groups in a population that share similar attributes
- Principal Components Analysis (Dimensionality Reduction)
  - Find patterns in data features
  - Visually represent high-dimensional data
  - Pre-processing step before supervised learning

No fixed analysis goal in unsupervised learning. Exploratory analysis to get new insights into the data, requires some creativity.

Cluster Analysis

Cluster Analysis

Clustering: Find groups in a population that share similar attributes

- Several approaches to defining clusters, no consensus on 'best method'; different approaches provide different insights
- **k-means** Clustering: Assumes a fixed number of clusters
- **Hierarchical** Clustering: Assumes the number of clusters is unknown

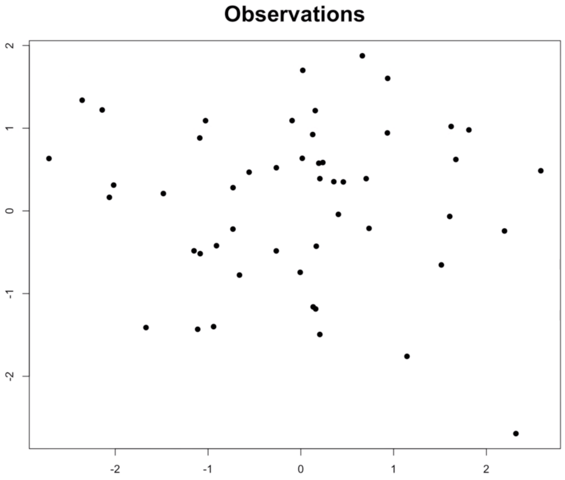
k-means Clustering

Starts by assuming a fixed number of clusters. Algorithm:

- Randomly assign each point to a cluster
- Calculate the centers of all points in each cluster
- Reassign points to new clusters based on their closest center
- Recalculate centers; iterate until no points change cluster assignment

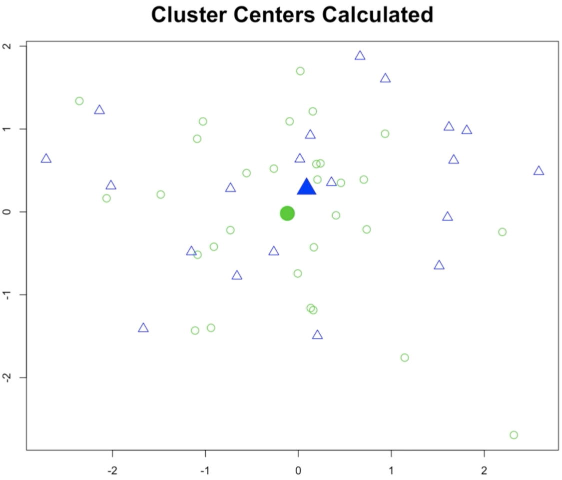
In R: **kmeans()** in the stats package.

k-means Clustering

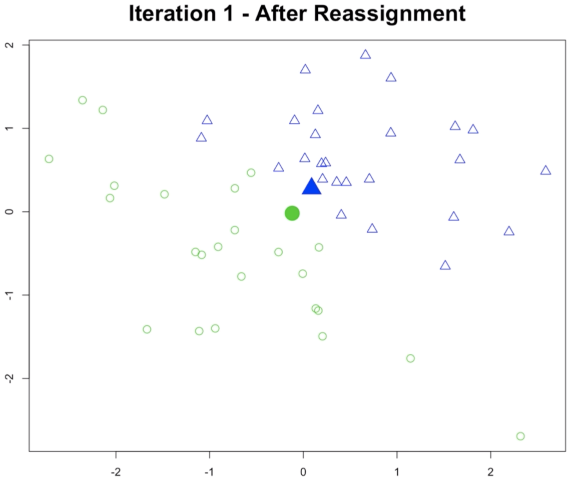


k-means Clustering

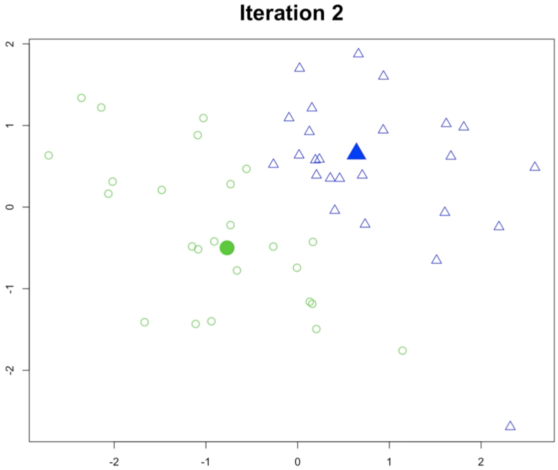
k-means Clustering



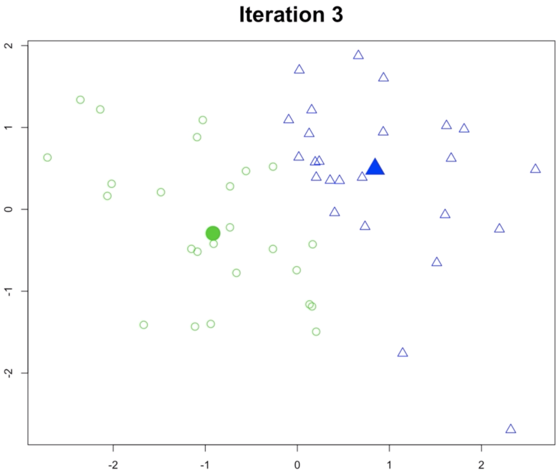
k-means Clustering



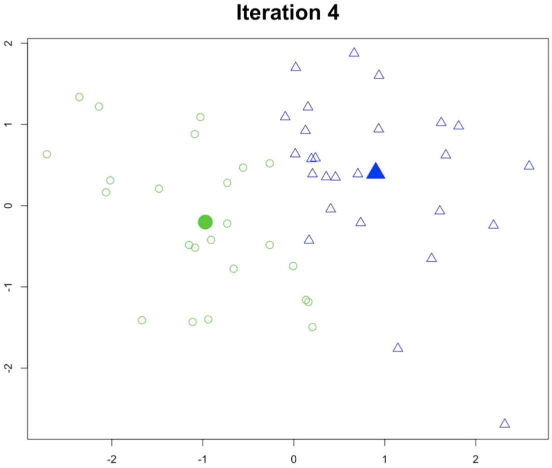
k-means Clustering



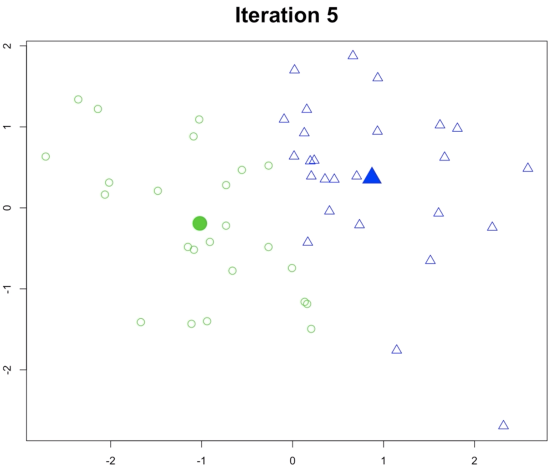
k-means Clustering



k-means Clustering



k-means Clustering



k-means Clustering issues

- Choose the number of clusters: heuristic choice based on the 'within-cluster sum of squares' (sum of squared distance from points to cluster centers).
- Stochastic method based on initial assignment of points to clusters. Run multiple times and choose the best outcome.
- Appropriate to rescale the data when variables are on different measurement scales.

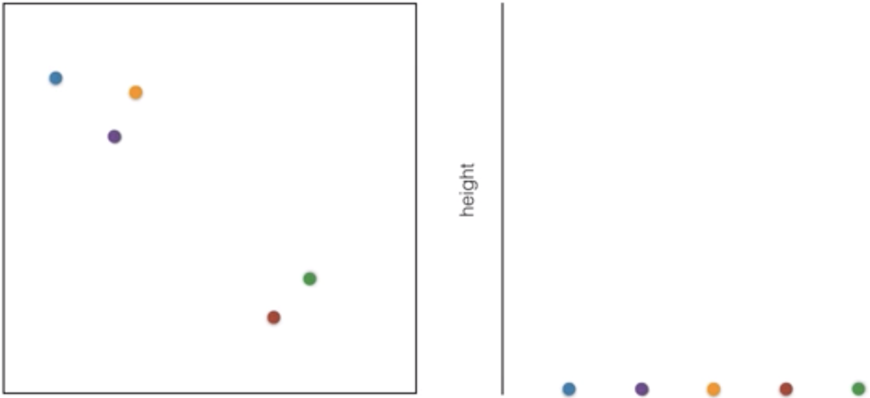
Hierarchical Clustering

Assumes the number of clusters is unknown. Hierarchical clustering can be **agglomerative** or **divisive**. Agglomerative ('bottom-up') clustering:

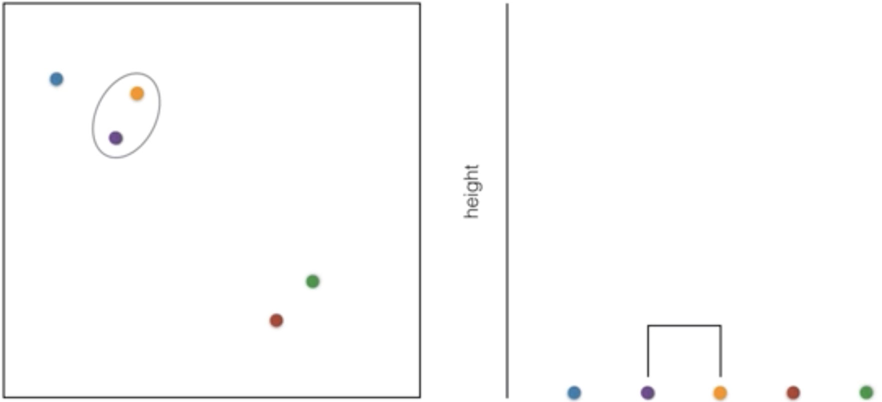
- Start by assigning each point to its own cluster
- Then merge the 'closest' two clusters using some distance metric
- Repeat until all points are in a single cluster

Divisive clustering starts with all points in a single cluster and iteratively splits them.
In R: **hclust()** in the stats package.

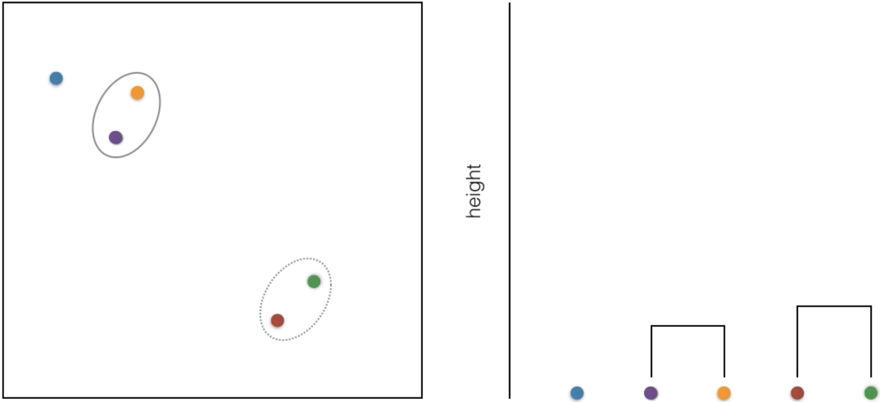
Hierarchical Clustering



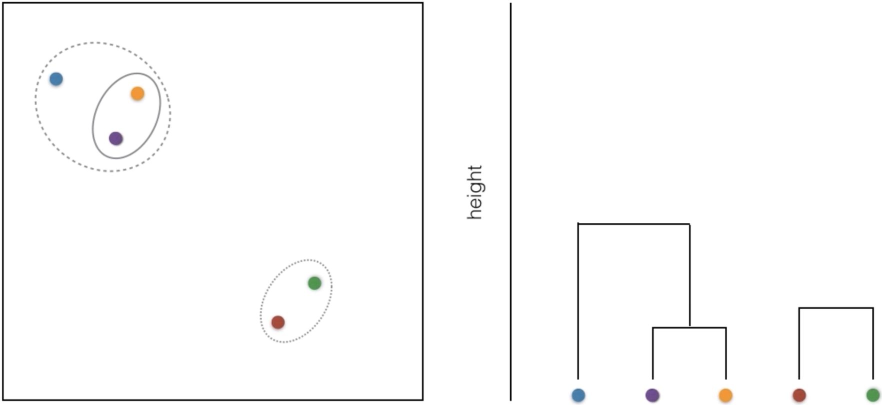
Hierarchical Clustering



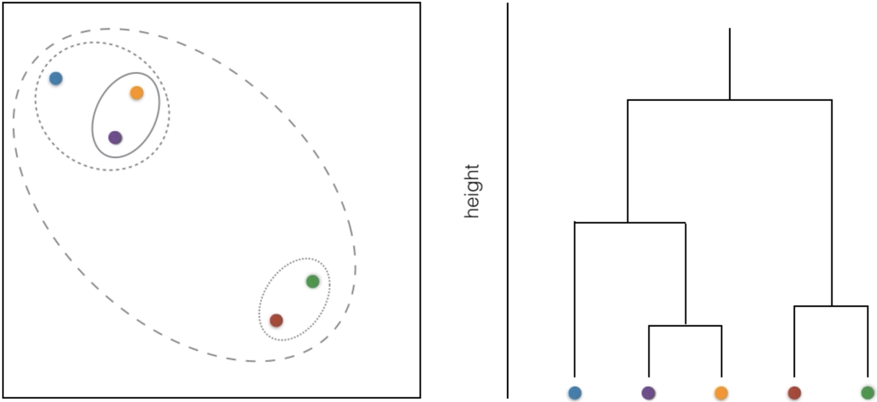
Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering

Possible distance metrics:

- complete: largest pairwise distance between all observations
- single: smallest pairwise distance
- average: average of pairwise distances
- centroid: difference between cluster centroids

Complete and average are most common. Single produces unbalanced trees where clusters are formed one observation at a time.

Bayesian Hierarchical Clustering

- **bclust** R package[1]
- Combines agglomerative clustering with variable selection, useful for high dimensional datasets
- Assumes key information on clustering may be hidden in a small subset of the variables, downweights noise variables using 'spike and slab' prior

[1] Nia, V. P., & Davison, A. C. (2012). High-Dimensional Bayesian Clustering with Variable Selection: The R Package bclus. Journal of Statistical Software, 47, 1-22.

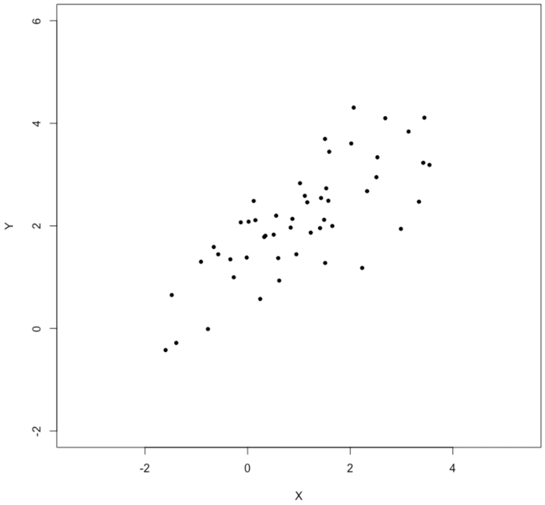
Principal Components Analysis

Principal Components Analysis

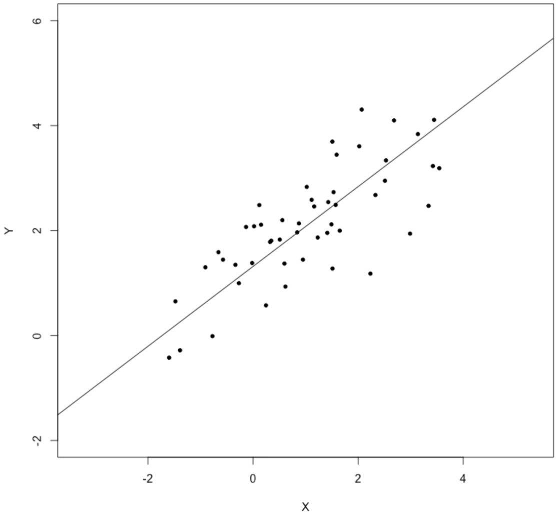
Dimensionality Reduction Technique. Goals: Find structure in features, aid in visualization. Principal components are:

- Linear combinations of variables
- Uncorrelated with one another (corthogonal)
- Constructed to maintain the most possible variance in the data

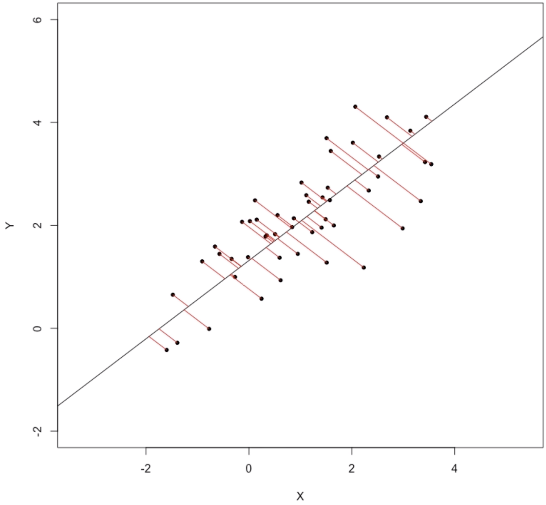
Principal Components Analysis: 2D exampleg



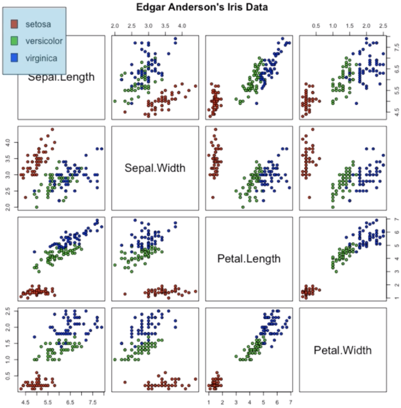
Principal Components Analysis: Regression line

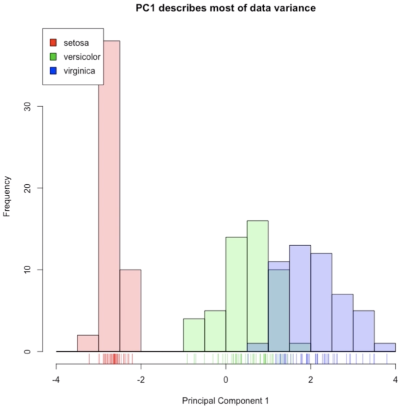


Principal Components Analysis: Projected component scores



PCA: Visualizing high-dimensional data





Principal Components Issues

- Scaling: Usually necessary. Otherwise variance of features with larger values overwhelms the rest
- Handling Missing values
  - Drop observations with any missing features (MAR assumption)
  - Impute missing values
- Handling Categorical data
  - Encode numerically
  - Other methods e.g. Multiple Factor Analysis