Statistical Learning Methods for Big Data

Xu (Sunny) Wang
Ph.D. in Statistics

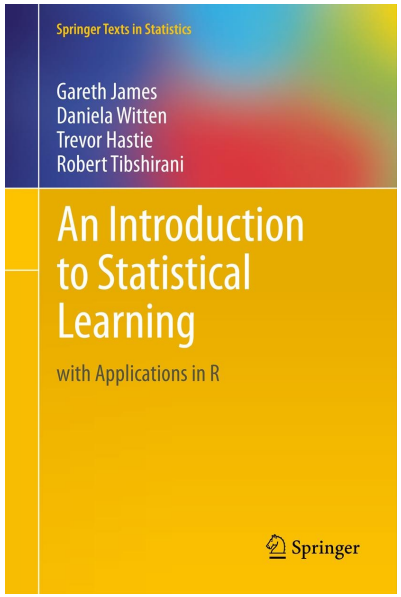
Department of Mathematics
Wilfrid Laurier University, Waterloo, ON

October 30, 2023

Banff International Research Station
for Mathematical Innovation and Discovery

Declaration

- By no means, this review is very comprehensive
- Textbook "An Introduction to Statistical Learning with Application in R by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani".



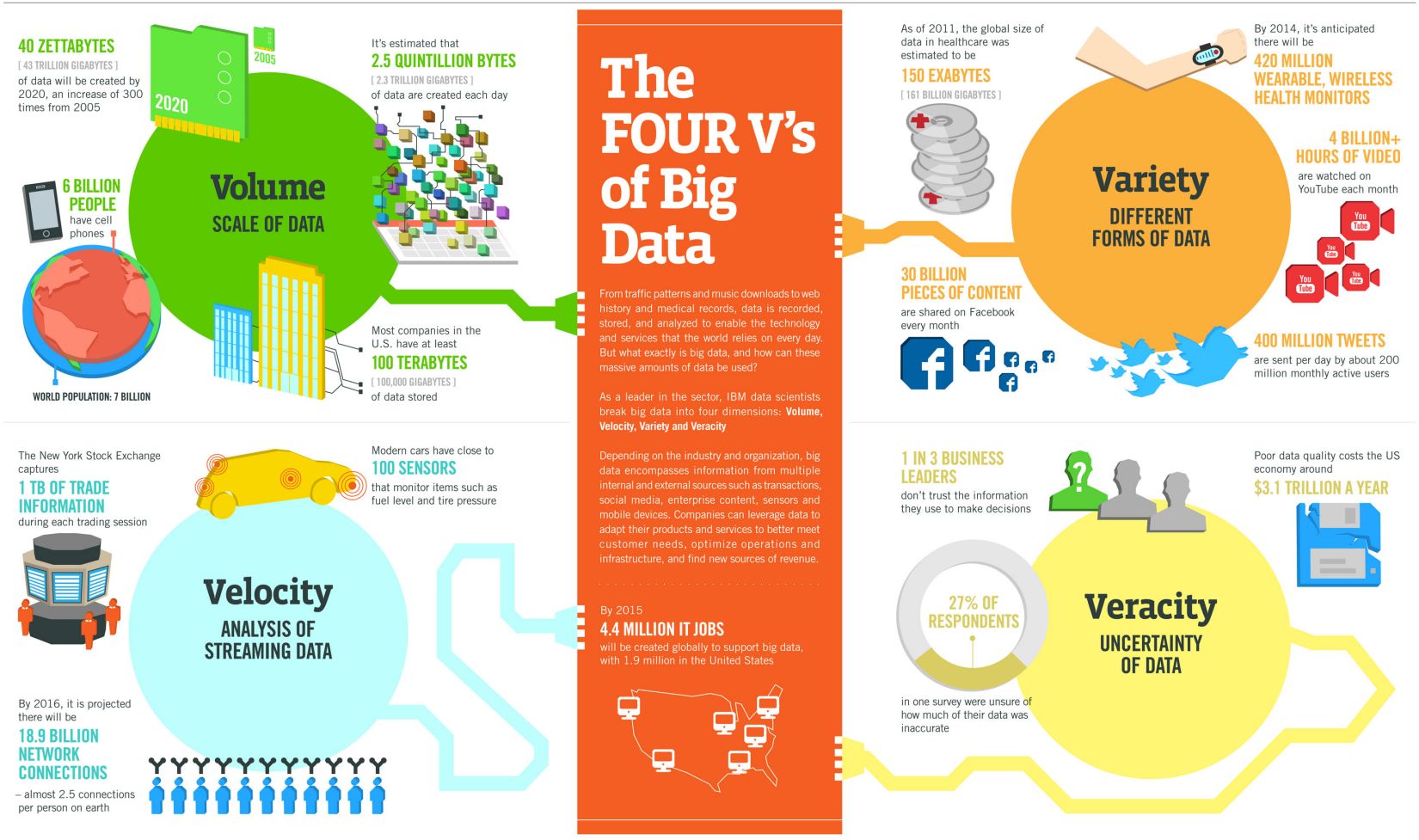
- My own Statistical Learning lecture notes
- Credit: Presentations made by Drs. Hugh Chipman, Trevor Hastie, Nancy Reid etc.

What are Big Data?

What are Big Data?

- From Wikipedia: "**Big data** is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. "

Characteristics of Big Data



Summary

Supervised Learning

- Predict a response Y using predictors $\mathbf{X} = \{X_1, X_2, \ldots, X_p\}$.
- A training sample of $(\mathbf{X}, Y)$ pairs.
- Continuous response $\Rightarrow$ "regression"
- Categorical response $\Rightarrow$ "classification"

Unsupervised Learning

- Discover structure in $\mathbf{X}$ without Y values.
- Clustering, dimensional reduction methods etc.

Two Quotes

Two quotes by famous Statisticians

"Essentially, all models are wrong, but some are useful"
George Box

"The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"
Fred Mosteller and John Tukey

Statistical Learning Methods

Supervised learning methods

- K-nearest Neighbour
- Generalized Additive Model
- Tree-based methods (recursive partitioning, Bagging, Random Forest, Boosting)
- Support Vector Machine
- Neural Networks

Unsupervised learning methods

- Principal Component Analysis
- K-means
- Hierarchical Clustering Analysis

General Framework

$$Y = f(\mathbf{X}) + \varepsilon$$

- Y = response variable
- $\mathbf{X} = \{X_1, X_2, \ldots, X_p\}$ = predictor variable(s)
- $f(\mathbf{X})$ is an unknown function
- ε is a random error

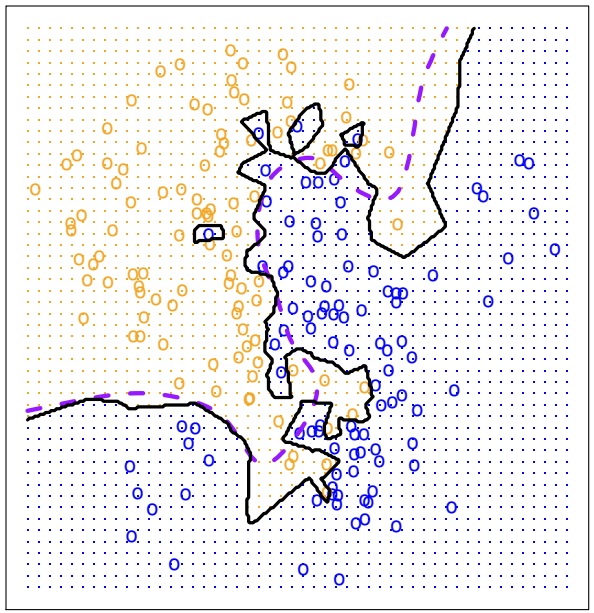
$$Y = \text{signal} + \text{noise}$$

Statistical learning typically focuses on estimation of "signal", with minimal attention given to "noise".

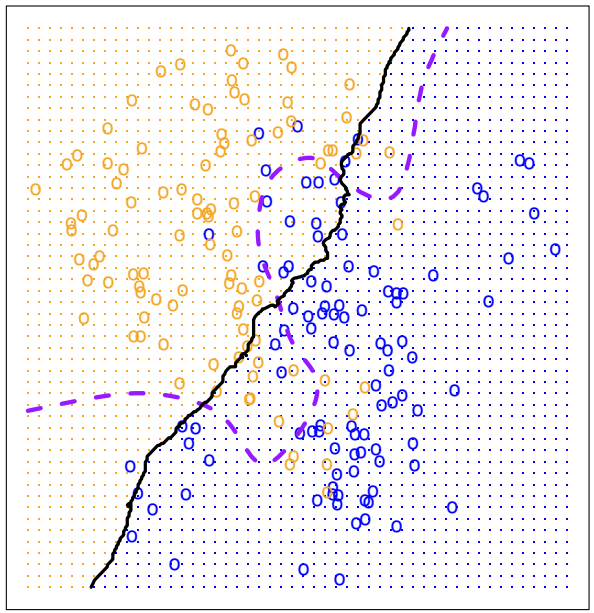
K-nearest Neighbour

$$\hat{f}(\mathbf{X}) = Ave(Y|\mathbf{X} \in N(\mathbf{X}))$$

KNN: K=1



KNN: K=100



K-nearest Neighbour

- KNN is good for small p, i.e. $p \leq 4$ and large N.
- KNN is "not good" when p is large due to the curse of dimensionality.
- Nearest neighbours tend to be far way in high dimensions.

Generalized Additive Models

Strong assumption of linear regression: Effect of varying X_1 does not depend on value of other X's.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \varepsilon$$

Generalize to have additive model with univariate functions:

$$Y = \beta_0 + g_1(X_1) + g_2(X_2) + \ldots g_p(X_p) + \varepsilon$$

- Allow for flexible nonlinearities in several variables, and retains the additive structure of linear models
- Easy interpretation.
- Estimation of p separate univariate functions much easier than estimation of a single $f(X_1, X_2, \ldots, X_p)$.
- Extension: allow some low-order interactions

Neural Networks

Nonlinear models with linear regressions at their core...

They have the functional form

$$f(X) = \Psi\left[\alpha_0 + \sum_i \alpha_i \Phi(\beta_{i0} + \sum_j \beta_{ij} X_j)\right]$$

with Ψ, Φ known, nonlinear functions.

- Estimate the coefficients (β's and α's).
- Nonlinear regression with many parameters.

A linear combination of...

A nonlinear transformation of ...

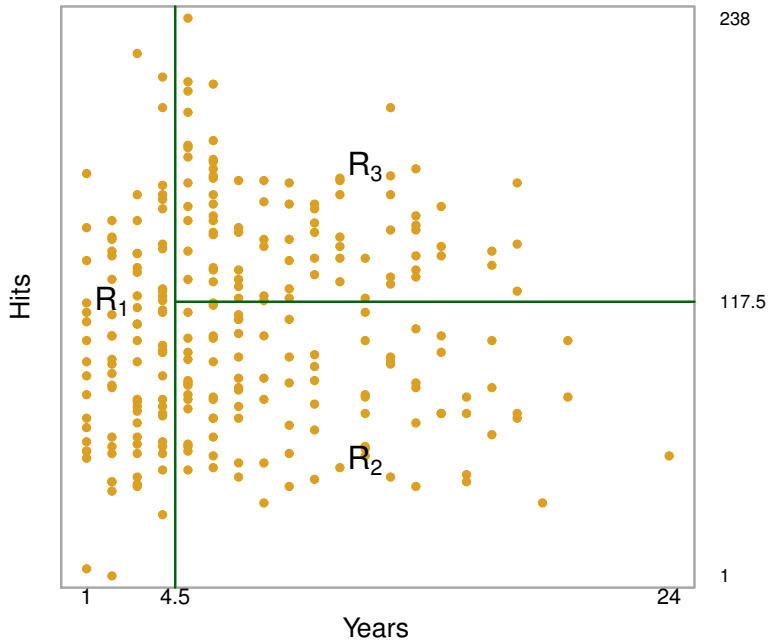
A linear combination of ...

the original variables

Decision Trees

Recursively partition the X space into rectangular regions to make them as homogeneous as possible.

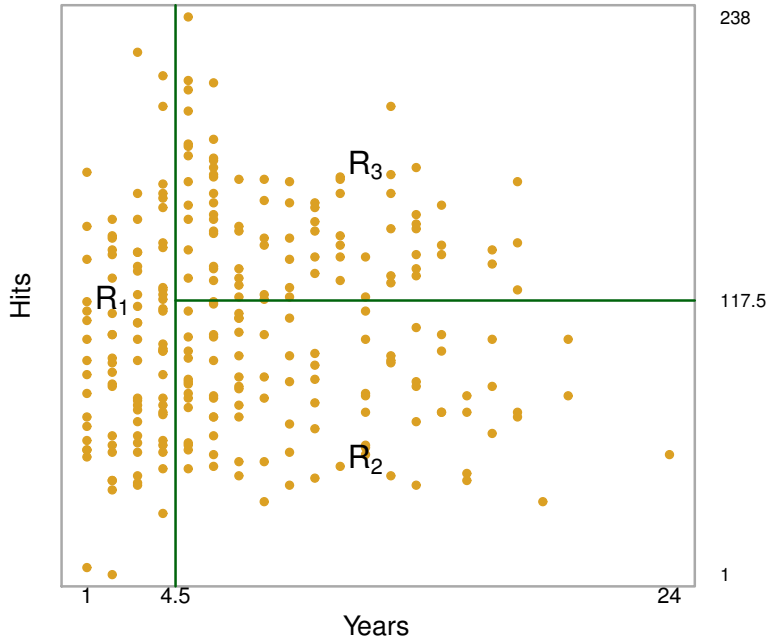
Example: Predict (log) `Salary` of baseball player, given `Years` in major leagues and `Hits` made last year.



- Learn the "local structure", similar to KNN
- Learn variables used, split values, depth.

Decision Trees

Decision trees are interpretable, flexible, detecting interactions and automatically perform variable selections



But they're sensitive to noise, "not good" at representing additive structures and allow variables dominate the tree structure

Ensemble Models

Overcome the limitations of a single tree by fitting a "sum of trees" model.

- Let $(T_1, M_1), \ldots, (T_m, M_m)$ identify a set of m trees and their terminal node μ's.

$$Y = g(x; T_1, M_1) + g(x; T_2, M_2) + \ldots + g(x; T_m, M_m) + \varepsilon$$

- For an input value x, each $g(x; T_i, M_i)$ outputs a corresponding μ
- The prediction is the sum of the μ's
- Random Forests (Breiman 2001) and Boosting (Freund & Schapire 1997) are two algorithms for building this type of models.

Ensemble Models

Breiman's **random forests** (2001) use randomized search at each split and the bootstrap samles

- Uses noise sensitivity of trees to build a stable model
- De-correlate individual trees

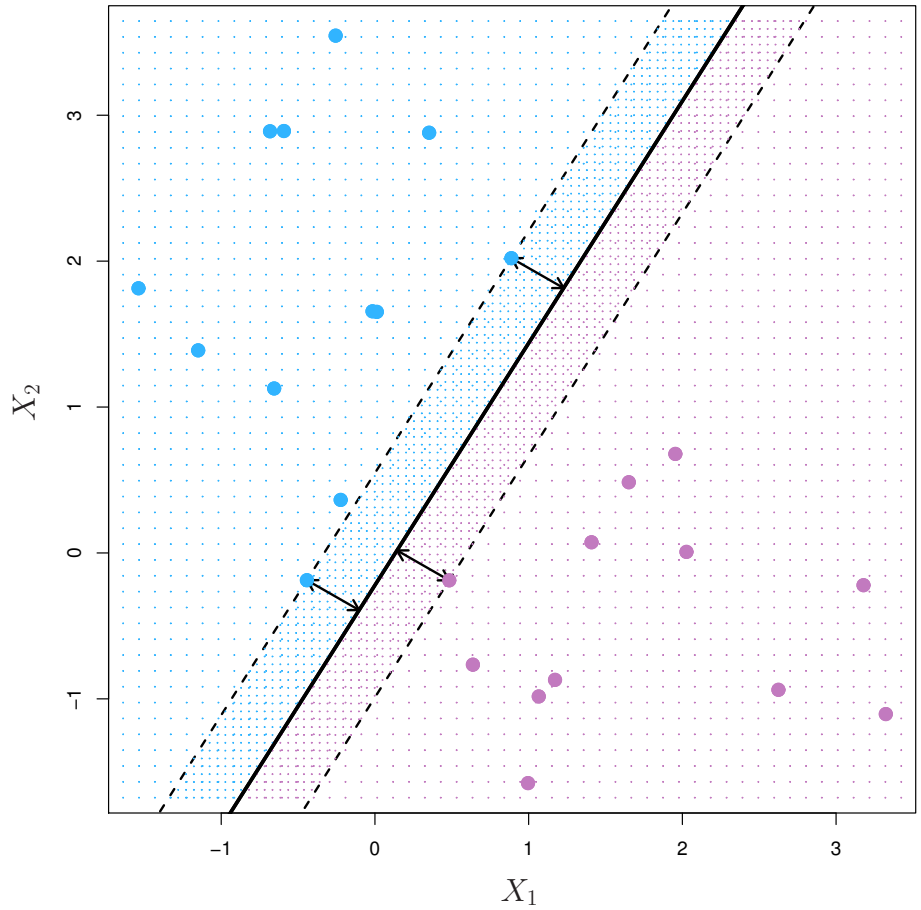
Freund and Schapire's **boosting algorithm** (1997) encourages each tree to fit structure not captured by the other trees - fitting trees sequentially.

- Enables an additive model to be fit.

Random forests and boosting are among the state-of-the-art methods for supervised learning.
However their results can be difficult to interpret.

Support Vector Machines

Originated as a 2-class classification problem (Vapnik, 1996).
Approach: find a hyperplane that separates the input space into two regions, maximally separating two classes.



Support Vector Machines

Two other key ideas:

1. Allow some misclassifications (amount is a tuning parameter).
2. Transform input vector $\mathbf{X}$ into a higher-dimensional space where a hyperplane is more likely to separate classes (often a parametrized transformation).

Comments on point 2:

- A "kernel trick" avoids the need to actually compute the high-dimensional mapping.
- Expensive algorithm - $O(n^2)$ for n observations.

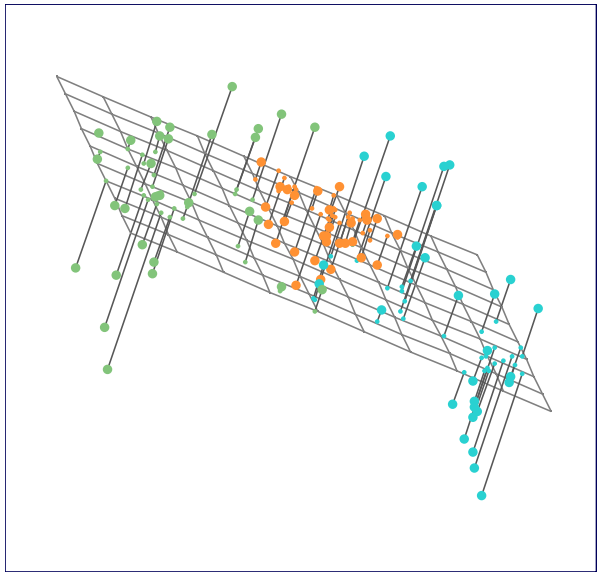
SVM is one of many **Kernel methods** for learning.

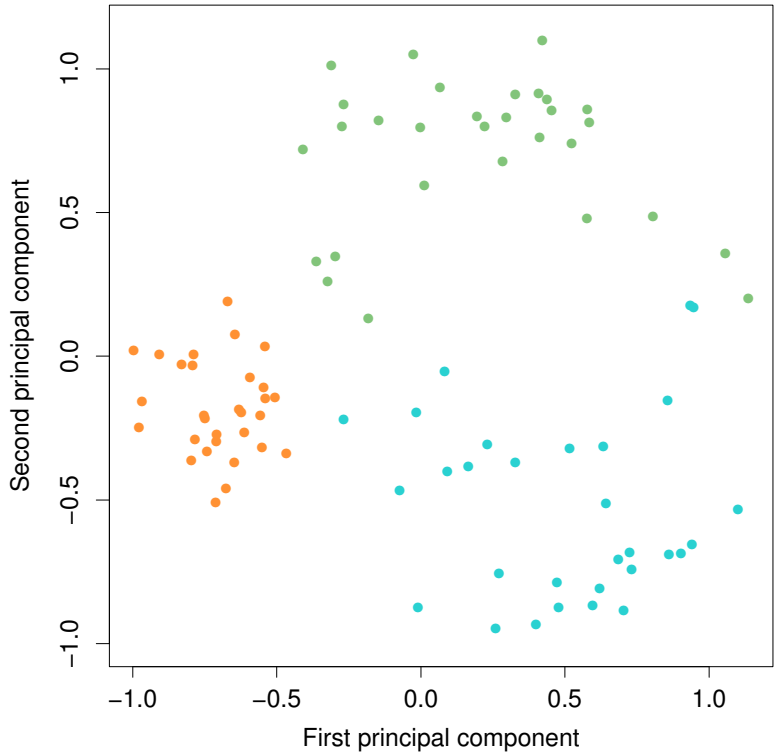
Dimension Reduction

$$y = f(g(x)) + \varepsilon$$

- The function g maps a high-dimensional input vector $\mathbf{X}$ to a lower-dimensional space.
- Principal component analysis (PCA) seeks projections $\alpha_1^T \mathbf{x}, \alpha_2^T \mathbf{x}, ...$ with maximal variance
- PCA is a data visualization or data pre-processing tool before supervised techniques are applied.
- Similar approach in "deep learning": estimate functions of inputs without using the response until the final learning step.

Principal component analysis





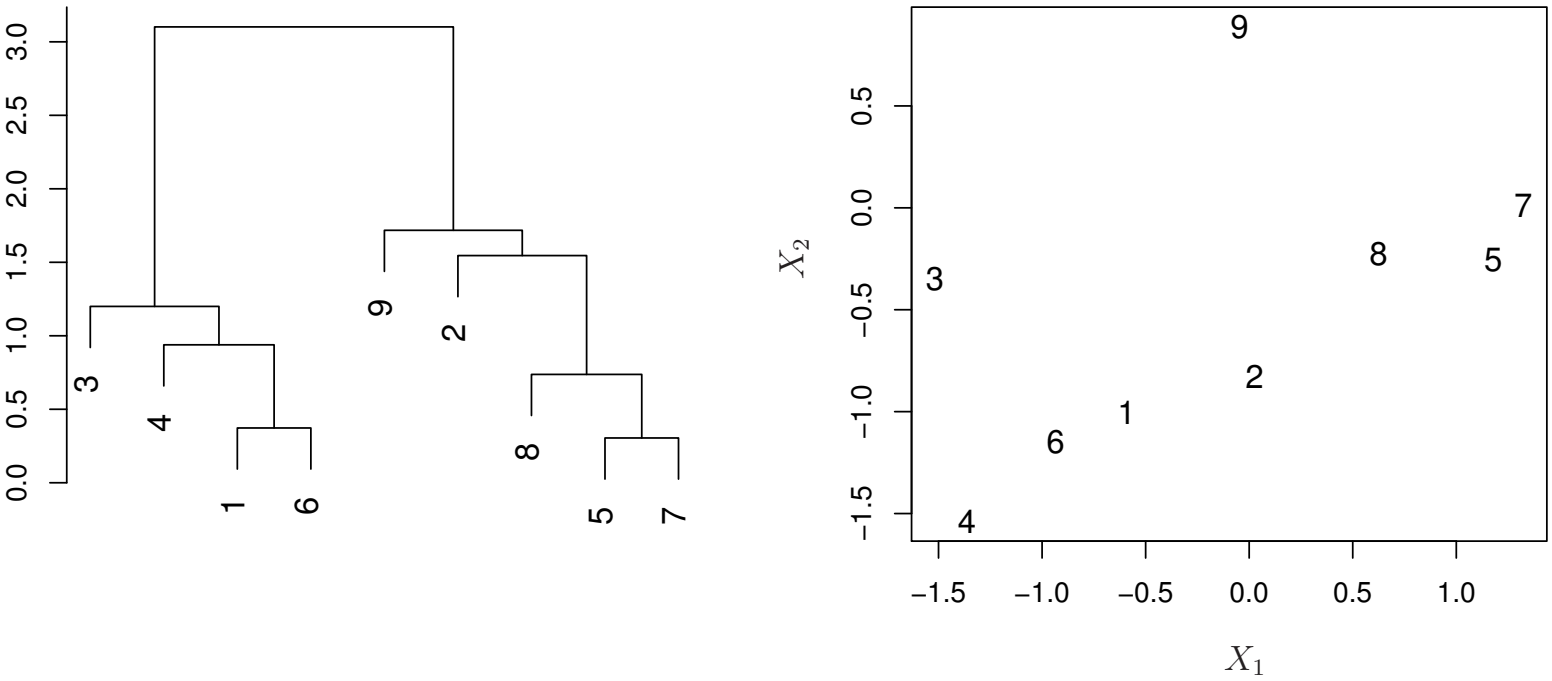
The first two principal components of a data set span the plane that is closest to the n observations, in terms of average squared Euclidean distance.

Clustering

A broad class of methods for discovering unknown subgroups in data

- Hierarchical clustering analysis: do not know in advance how many clusters; a tree-like dendrogram
- K-means clustering: seek to partition the observations into a pre-specified number of clusters.

Hierarchical clustering



Linkage (Inter-cluster dissimilarity): Complete, Single, Average, Centroid

K-means

variations: k-medoid, bisecting k-means, X-means clustering, and G-means

Other Techniques

- Variable selection or regularization (Ridge, LASSO, Elastic net)
- Internal validation: training/test split, k-fold cross validation

Some reference books:

- *An introduction to Statistical Learning with Applications in R* by James, Witten, Hastie and Tibshirani
- *Statistical Learning and Data Mining*, Hastie, Tibshirani and Friedman
- *Pattern Recognition and Machine Learning*, Bishop
- *Bayesian Methods for Nonlinear Classification and Regression*, Denison, Holmes, Mallick and Smith
- *Applied Functional Data Analysis: Methods and Case Studies*, James O Ramsay, Bernard W. Silverman.

Thank you!
Questions & Comments