Overview and Generic Advice

STOR 565

Andrew Nobel

April, 2021

Paradigm Shift

Traditional Scientific Method: Hypothesis Driven

- Formulate a hypothesis
- Collect data to confirm/refute hypothesis

Modern Scientific Method: Data Driven

- Acquire data from high-throughput measurement technologies
- Mine the data for possible hypotheses
- Use the data again to test selected hypotheses

Free Parameters

Fact: Most ML methods involve one or more free parameters

- PCA: number of components
- hierarchical clustering: selecting subtree of dendrogram
- k-means clustering: number of clusters k
- k-nearest neighbors: distance measure, number of neighbors
- naive Bayes: distribution families for individual components
- ridge and LASSO regression: penalty parameter λ
- SVM: choice of slack penalty, kernel
- decision trees: impurity, size of initial tree T_0, penalty for pruning
- bagging: number of bootstrap samples
- boosting: number of components in model

Less Visible Choices

- data preprocessing: imputing missing values, normalization
- filtering of features
- transformation of features
- definition of new features, e.g., as functions of existing features

Methods Overview

Unsupervised: Finding structure in data

- Principal component analysis
- Clustering: hierarchical and k-means

Supervised: Building predictive models

- Classification: k-NN, LDA, LogReg, N-Bayes, SVM, histograms, trees
- Regression: OLS, ridge, LASSO, trees
- Aggregation: bagging and boosting

Theory Overview

Background

- matrix and linear algebra
- calculus: partial derivatives, gradients, Hessians, Taylor expansions
- probability: mean and variance, conditional expectations, covariance
- statistics: distributions, maximum likelihood, CDF method

Covered

- order, minima, and maxima
- convex sets, convex and concave functions, basic properties
- random vectors, multivariate normal
- classification: Bayes risk and Bayes rule
- probability inequalities: Markov, Chebyshev, Hoeffding

Generic Advice on Consulting or Collaboration

1. Familiarize yourself with the subject area
2. Get your client/collaborator to state the problem in their own terms
3. Familiarize yourself with the data
4. Do exploratory analyses before undertaking supervised analyses
5. Try simple methods before complicated ones
6. Try more than one method