Stat 154: Modern Statistical Prediction and Machine Learning
Course Outline, Fall 2019

(Tentative topics & dates, subject to change depending on the pace of the course)

Instructor: Gaston Sanchez (gasigiri [at] berkeley [dot] edu)
Lecture: TuTh 3:30-5:00pm in 390 Hearst Mining
Exams:

- **Exam 1**: Thursday Oct-17th, during class time
- **Exam 2**: Thursday Dec-5th, during class time

Textbook acronyms:

- **ISL**: An Introduction to Statistical Learning (by James et al, 2015)
- **ESL**: The Elements of Statistical Learning (by Hastie et al, 2009)

Principal Components Analysis (PCA)

We begin with **Principal Components Analysis**, one of the unsupervised learning topics of this course. Simply put, PCA allows us to study the systematic structure of a data set (of quantitative variables). Although PCA can be approached from multiple perspectives, we will approach it from a data visualization perspective, and a decisive geometric flavor.

Reading: ISL 10.1 and 10.2; ESL 14.5
Dates: Aug-29 / Sep-05

Preamble for PCA

- The duality of the data matrix: rows (individuals) and columns (variables)
- Common operations for individuals: the average individual, distance between individuals, multivariate dispersion, and inertia
- Common operations for variables: variables as vectors, length of a vector, vector and scalar projections, angle between vectors, variance, covariance, correlation

Fundamentals of PCA

- PCA from three perspectives: projected inertia, maximized variance, data decomposition
- PCA solution with EVD of cross-products X^TX and XX^T

Application of PCA (anatomy of PCA solution)

- How many components to retain?
- How can a component be interpreted?
- What visualizations can be obtained, and how to read them?
- Some practical considerations

Digression on Matrix Decompositions

- Matrix decompositions: Eigenvalue Decomposition (EVD)
- Singular Value Decomposition (SVD) and lower rank approximations
- Relationship between EVD and SVD

Linear Regression: Introduction

After PCA we shift gears to supervised learning methods that have to do with predicting a quantitative response. We begin with **Linear Regression models** which are the stepping stone for all supervised learning methods. We will study the general regression framework by paying attention to the algebraic and geometric aspects, while postponing the discussion of the learning elements for later (to be covered in *Concepts of Learning Theory*).

Reading: ISL chapter 3; ESL 3.1-3.3
Dates: Sep 10-12

Introduction to Linear Regression

- Motivating an intuitive feeling for regression problems
- The regression function: conditional expectation
- Classic framework of Ordinary Least Squares (OLS)

Theoretical core of OLS and Optimization

- Geometries of Least Squares: individuals, variables, and parameters perspectives
- Gradient descent algorithm
- Linear regression from a probabilistic approach: Maximum Likelihood

Issues with Least Squares

- Issues with OLS and potential solutions
- Multicollinearity issues

Concepts of Learning Theory

In this part of the course we review the notion of ***Learning*** *(from a supervised point of view)* and other related concepts: What are the conceptual pieces of a learning problem? What do we mean by learning? How do we measure the learning ability of a machine/model? We will also talk about several aspects concerning Learning theory: model performance, bias-variance tradeoff, overfitting, validation, and model selection.

Reading: ISL chapters 1, 2 and 5; ESL 7.1-7.5, 7.10-7.11
Dates: Sep 17-26

A framework for Supervised Learning

- Supervised Learning Diagram: anatomy of supervised learning problems
- The meaning of Learning, and the need for Training and Test sets
- Types of errors and their measures: In-sample and Out-of-sample errors
- Noisy targets and conditional distributions

Bias-Variance trade-off

- Derivation of the Bias-Variance decomposition formula
- Case study: data simulation to illustrate the bias-variance decomposition
- Interpretation of the bias-variance trade-off

Overfitting

- What is overfitting? Why should we care about it? What causes overfitting?
- Case study: data simulation to illustrate overfitting

Model Validation and Model Selection

- Validation: What is it? Why do we need it?
- Validation holdout framework, and repeated validation, bootstrap validation
- Three-way holdout frmework
- Cross-Validation: k-fold CV, leave-one-out CV
- Some general considerations

Linear Regression: Regularization Methods

Having introduced the basic concepts of Learning Theory, we'll expand our discussion of linear regression. The classic Least Squares solution for linear models is not always feasible or desirable. One main idea to get a better solution is by **regularizing the regression coefficients**. This can be done in a couple ways: 1) by transforming the predictors and reducing the dimensionality of the input space; or 2) by penalizing the criterion to be minimized via restricting the size of the regression coefficients.

Reading: ISL 6.2-6.5; ESL 3.4-3.6
Dates: Sep-26 / Oct-08

Regularization via Dimension Reduction methods

- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLSR)

Regularization via Penalized methods

- Ridge Regression (RR)
- Other methods: lasso, elastic net, and cousins
- Geometries of penalized parameters

Regression: Moving Beyond Linearity

Linear models can be quite useful but they have some limitations that force us to move beyond linearity. The main idea in this section is to relax the linearity assumption while still trying to obtain interpretable models. We'll learn about various approaches that allow us to augment/replace the input variables with an array of transformations such as polynomial regression, step functions, splines, local regression, RBF, etc.

Reading: ISL chapter 7; ESL 5.1-5.3, 6.1-6.2
Dates: Oct 10-15

Limitation of linear models, and notion of linearity
Some nonlinear approaches

- Polynomial regression

- Stepwise regression
- Basis functions (basis-expansion)
- Splines
- Local regression
- Radial Basis Functions

Midterm 1 (Oct-17)

Exam 1: Thursday Oct-17th, during class time

- *PCA, EVD, SVD*
- *Linear Regression*
- *Dimension Reduction methods*
- *Penalized methods*
- *Nonlinear methods*
- *Concepts of Learning Theory*

Introduction to Classification

The other major type of supervised learning problems covered in this course has to do with classification methods: predicting a qualitative response. We begin with **Logistic Regression** which provides a nice bridge between linear regression and classification ideas.

Reading: ISL chapter 4; ESL chapter 4
Dates: Oct 22-24

Introduction to Classification

- Limitations of the classic regression model
- Motivation for logistic transformation

Logistic Regression

- Model
- Error Measure
- Algorithm(s)

Discriminant Analysis

An important set of classification methods have to do with Discriminant Analysis (DA). The origins and foundations these techniques are based on Ronald Fisher's geometric approach commonly known as Canonical Discriminant Analysis. This method can be considered to be a classification with an unsupervised touch. We will also study the so-called generative classification methods: Linear DA, Quadratic DA, and Naives Bayes.

Reading: ISL chapter 4; ESL chapter 4
Dates: Oct 24-31

Discriminant Analysis

- Canonical Discriminant Analysis (CDA)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Naive Bayes

Other Methods

- K-Nearest Neighbors (kNN)
- Support Vector Machines (SVM) *time permitting*

Model Performance with Classification Methods

How do we measure the learning ability of a classification model? Similar to what we did for regression models, we will discuss how to assess model performance in a classification setting. We will also talk about concepts like: confusion matrices, sensitivity and specificity, true positives and false positives, as well as Receivier Opearting Characteristics (ROC) curves.

Reading: ISL chapter 4; ESL chapter 4
Dates: Nov 05-07

Classification Performance Measures

- Error measures
- Misclaffication error
- Sensitivity and Specifcity
- Receiver Operating Characteristic (ROC) Curve

Clustering Methods

Simply put, **Clustering** has to do with finding groups in data. This is the second unsupervised topic of the course, covering partition methods as well as hierarchical agglomerative techniques.

Reading: ISL 10.3; ESL 14.3
Dates: Nov 12-14

Preliminary Concepts

- Decomposition of total spread
- Proximity measures: distance and disimilarity measures

Partition methods

- K-means

- K-medoids

Hierarchical (agglomerative) methods

- Single linkage
- Complete linkage
- Average linkage
- Centroid linkage

Tree-based Methods

The last section of the course involves decision trees (classification and regression trees) and derived methods. Measures on entropy. Ensemble methods (aggregrating individual learners) such as Boosting and Bagging.

Reading: ISL chapter 8; ESL 7.1-7.5
Dates: Nov 19-29

Tree-based Methods

- Entropy
- Gini impurity
- Classification Trees
- Regression Trees
- Random Forest
- Ensemble Methods and Aggregation

Midterm 2 (Dec-05)

Exam 2: Thursday Dec-05th, during class time

- *Cumulative with an emphasis on the second half of the course*