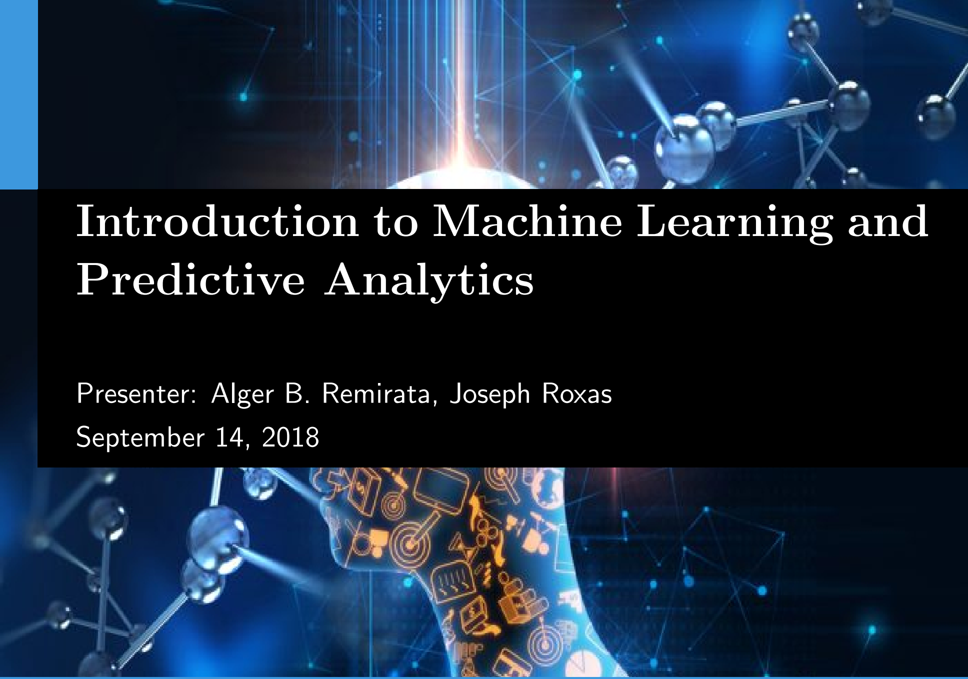

Introduction to Machine Learning and Predictive Analytics

Presenter: Alger B. Remirata, Joseph Roxas

September 14, 2018

Predictive Modeling

Predictive modeling is the process of developing a mathematical tool or model that generates an accurate prediction.

There are a number of common reasons why predictive models fail, e.g,

- inadequante pre-processing of the data
- inadequate model validation
- unjustified extrapolation
- over-fitting the model to the existing data
- explore relatively few models when searching for relationships

Some Key Concepts of Predictive Modeling

Prediction versus Interpretation

The trade-off between prediction and interpretation depends on the primary goal of the task. The unfortunate reality is that as we push towards higher accuracy, models become more complex and their interpretability becomes more difficult.

Key Ingredients of Predictive Models

The foundation of an effective predictive model is laid with *intuition* and *deep knowledge of the problem context*, which are entirely vital for driving decisions about model development. The process begins with *relevant* data.

Terminology

- The *sample*, *data point*, *observation*, or *instance* refer to a single independent unit of data
- The *training* set consists of the data used to develop models while the *test* or *validation* set is used solely for evaluating the performance of a final set of candidate models.
  NOTE: usually people refer to the *validation* set for evaluating candidates and divide *training* set using cross-validation into several *sub-training* and *test* sets to tune parameters in model development.
- The *predictors, independent variables, attributes, or descriptors* are the data used as input for the prediction equation.
- The *outcome, dependent variable, target, class, or response* refer to the outcome event or quantity that is being predicted.

Overview

Part 1. General Strategies

- A short tour of the predictive modeling process
- Data pre-processing
- Over-fitting and model tuning

Part 2. Regression Models

- Measuring performance in regression models
- Linear regression and its cousins
- Nonlinear regression models
- Regression trees and rule-based models
- A summary of solubility models
- Case study: compressive strength of concrete

Overview

Part 3. Classification Models

- Measuring performance in classification models
- Data pre-processing
- Over-fitting and model tuning
- Classification trees and rule-based models
- A summary of grant application models
- Remedies for severe class imbalance
- Case study: job scheduling

Part 4. Other Consideration

- Measuring predictor importance
- An introduction to feature selection
- Factors that can affect model performance