CHAPTER 1: INTRODUCTION

Overview of a Data Analytics Pipeline

A typical data analytics pipeline consists of several major pillars. In the example shown in Figure 1.1, it has four pillars: sensor and devices, data preprocessing and feature engineering, feature selection and dimension reduction, modeling and data analysis. While this is not the only way to present the diverse data pipelines in real-world, they more or less resemble this arctiteture.

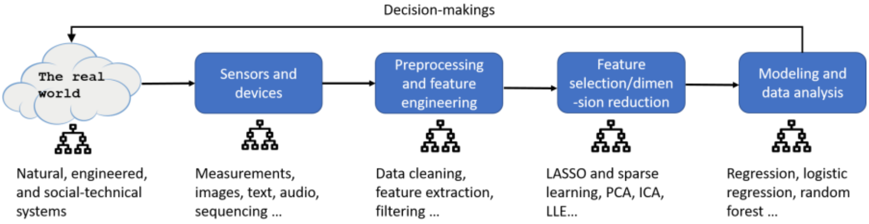
Decision-makings
The real world
Sensors and devices
Preprocessing and feature engineering
Feature selection/dimen -sion reduction
Modeling and data analysis
Natural, engineered, and social-technical systems
Measurements, images, text, audio, sequencing ...
Data cleaning, feature extraction, filtering ...
LASSO and sparse learning, PCA, ICA, LLE...
Regression, logistic regression, random forest ...

Figure 1.1: Overview of a data analytics pipeline

The pipeline starts with a real-world problem, for which we are not sure about the underlying system/mechanism, but we are able to characterize the system by defining some variables. Then, we could develop sensors and devices to acquire measurements of these variables. These measurements, we call as data, are objective evidences that we can use to explore the statistical principles or mechanistic laws regulating the system behaviors. But, before analyzing the data and building models using the data, in practice, the data preprocessing and feature engineering are important. For example, some signals acquired by sensors are not interpretable or not easily compatible with human sense, such as the signal acquired by MRI scanning machines in the Fourier space. Data preprocessing also refers to removal of outliers or imputation of missing data, detection and removal of redundant features, to name a few. After the preprocessing, we may conduct feature selection and dimension reduction to distill or condense signals in the data and reduce noise. Finally, we are ready to conduct modeling and data analysis on the prepared dataset to gain knowledge and build prediction models of the real-world system. Decision-makings such as prediction, intervention, and control policies can be derived based on the fitted models to optimize and control the real-world system.

This book focuses on the last two pillars of this pipeline, the modeling, data analysis, feature selection, and dimension reduction methods. But it is helpful to keep in mind of the big picture of a data analytics pipeline. Because in practice, what works is the whole pipeline.

Structure of the Chapters

The structures of the Chapters follow the same manner.

- Each chapter will introduce two or three techniques. In most cases, one technique is about regression model while another one is about tree model.
- For each technique, we will highlight the intuition and rationale behind it.

- Then, we articulate the intuition, use math to formulate the learning problem, and present the full version of the analytic formulation. But, it is always important to remember its intuitive underpinning.
- Then, we use R to implement the technique on both simulated and real-world dataset, present the analysis process (together with R code), show the dynamics in the analysis process, and comment on the results.
- Some remarks are also made to enhance understanding of the techniques, reveal their different natures by other perspectives, reveal their limitations, and mention existing remedies to overcome these limitations.

Topics in a Nutshell

Data models – regression based techniques:

- Chapter 2: Linear regression, least-square estimation, hypothesis testing, why normal distribution, its connection with experimental design, R-squared.
- Chapter 3: Logistic regression, generalized least square estimation, iterative reweighted least square (IRLS) algorithm, approximated hypothesis testing, Ranking as a linear regression
- Chapter 4: Bootstrap, data resampling, nonparametric hypothesis testing, nonparametric confidence interval estimation
- Chapter 5: Overfitting and underfitting, limitation of R-squared, training dataset and testing dataset, random sampling, K-fold cross validation, the confusion matrix, false positive and false negative, and Receiver Operating Characteristics (ROC) curve
- Chapter 6: Residual analysis, normal Q-Q plot, Cook's distance, leverage, multicollinearity, subset selection, heterogeneity, clustering, gaussian mixture model (GMM), and the Expectation-Maximization (EM) algorithm
- Chapter 7: Support Vector Machine (SVM), generalize data versus memorize data, maximum margin, support vectors, model complexity and regularization, primal-dual formulation, quadratic

programming, KKT condition, kernel trick, kernel machines, SVM as a neural network model

- Chapter 8: LASSO, sparse learning, L1-norm and L2-norm regularization, Ridge regression, feature selection, shooting algorithm, Principal Component Analysis (PCA), eigenvalue decomposition, scree plot
- Chapter 9: Kernel regression as generalization of linear regression model, kernel functions, local smoother regression model, k-nearest regression model, conditional variance regression model, heteroscedasticity, weighted least square estimation, model extension and stacking

Algorithmic models – tree based techniques:

- Chapter 2: Decision tree, entropy gain, node splitting, pre- and post-pruning, empirical error, generalization error, pessimistic error by binomial approximation
- Chapter 4: Random forest, Gini index, weak classifiers, probabilistic mechanism why random forest works
- Chapter 5: Out-of-bag (OOB) error in random forest
- Chapter 6: Importance score, partial dependency plot, residual analysis
- Chapter 7: Ensemble learning, Adaboost, sampling with (or without) replacement
- Chapter 8: Importance score in random forest, regularized random forests (RRF), guided regularized random forests (GRRF)
- Chapter 9: System monitoring reformulated as classification, real-time contrasts method (RTC), design of monitoring statistics, sliding window, anomaly detection, false alarm
- Chapter 10: Integration of tree models and regression models in inTrees, random forest as a rule generator, rule extraction, pruning, selection, and summarization, confidence and support of rules, variable interactions, rule-based prediction

In this book, we will use lower-case letters, e.g., x, to represent scalars, bold-face lower-case letters, e.g., $\boldsymbol{v}$, to represent vectors, and bold-face upper-case letters, e.g., $\boldsymbol{W}$, to represent matrices.