Overview of this Short Course on Statistical Computing

November 14, 2019

Xiaoming Huo, Georgia Tech

Contents

1. Introduction; Reproducible Research: GitHub, Jupyter Notebooks
2. Classification
3. Clustering
4. Decision Trees
5. Principal Component Analysis

Contents

Platforms

- Reproducible Research: GitHub,
- Jupyter Notebooks

Supervised Learning

1. Linear Regression
2. Logistic Regression
3. Support Vector Machines
4. Nearest Neighbors
5. Decision Trees
6. Nerual Networks
7. Implementing Deep Learning Models with Pytorch
8. Naive Bayes

Unsupervised Learning

- Clustering
- Factor Analysis
- Principal Component Analysis

Applications

- Financial Asset Returns
- Capital Asset Pricing Model (CAPM) - linear regression, single index model

Within the Supervised Learning Methods

- Regression
  1. Linear Regression
  2. Nerual Networks (can be both)

- Both
  1. Decision Trees
  2. Implementing Deep Learning Models with Pytorch

- Classification
  1. Logistic Regression
  2. Nearest Neighbors
  3. Support Vector Machines
  4. Naive Bayes

What could be covered?

- Exploratory data analysis
- More statistics (seeing next few slides)
- Time series (seeing on the next few slides)

6783 What could be covered - 2

- Exploratory data analysis: Histograms and Kernel Density Estimation, Order Statistics, the Sample CDF, and Sample Quantiles, The Central Limit Theorem for Sample Quantiles, Normal Probability Plots, Half-Normal Plots, Quantile-Quantile Plots, Tests of Normality, Boxplots, Data Transformation, and Transformation Kernel Density Estimation
- Univariate distributions: Parametric Models and Parsimony, Skewness, Kurtosis, and Moments, Heavy-Tailed Distributions, Exponential and Polynomial Tails, t-Distributions, Mixture Models, Generalized Error Distributions, Likelihood Ratio Tests, AIC and BIC, Validation Data and Cross-Validation, Fitting Distributions by Maximum Likelihood, Profile Likelihood, and Robust Estimation

6783 What could be covered - 3

- **Multivariate statistical distributions:** Covariance and Correlation Matrices, Linear Functions of Random Variables, Two or More Linear Combinations of Random Variables, Independence and Variances of Sums, Scatterplot Matrices, The Multivariate Normal Distribution, The Multivariate t-Distribution, Elliptically Contoured Densities, The Multivariate Skewed t-Distributions, and The Fisher Information Matrix
- **Copulas**: definitions, special copulas, Gaussian and t-copulas, Archimedian copulas, Rank correlation

6783 What could be covered - 4

- **Time series**: Time Series Data, Stationarity, White Noise, Estimation, AR(1) Processes, AR(1) Model Revisit with R, AR(p) Process (p > 1), MA(q) Process, ARMA(p, q), ARIMA(p, d, q), Forecasting, and Nonstationary Time Series
- **GARCH Models and Time-Varying Volatility**: Some issues in Time Series, Volatility, Modeling the Volatility, Integrated GARCH Model, Exponential GARCH Model, Forecasting Future Volatilities in EGARCH(1, 1), Likelihood Inference and Implementation, and ARMA-GARCH and ARMA-EGARCH models

Main messages

- Statistical computing can be **reproducible**
- **New** paradigm in data science technologies

Thank you!

➢ Email: huo@gatech.edu