Summer 2017 MOOC: Data Science and Predictive Analytics

Enrollment is now open for this Summer MOOC, which starts July 01, 2017. Enrollment is limited, interested trainees should review the course syllabi, prerequisites and coverage ↓, and register.

1. Statistical Software – Pros/Cons Comparison
Getting started
Install Basic Shell-based R
GUI based R Invocation (RStudio)
RStudio GUI Layout
Help
Simple Long-to-Wide Data format translation
Data generation
I/O
Slicing and extracting data
Variable conversion
Variable information
Data selection and manipulation
Math Functions
Matrix Operations
Advanced Data Processing
Strings
Plotting
QQ Normal Probability Plots
Low-level plotting commands
Graphics parameters
Optimization and model fitting
Statistics
Distributions
Programming
Data Simulation Primer

2. Managing data with R
Saving and Loading R Data Structures
Importing and Saving Data from CSV Files
Exploring the Structure of Data
Exploring Numeric Variables
Measuring the Central Tendency - mean and median
Measuring Spread - quartiles and the five-number summary
Visualizing Numeric Variables - Boxplots
Histograms
Understanding Numeric Data – Uniform and Normal distributions
Measuring Spread - variance and standard deviation
Exploring Categorical Variables
Measuring the Central Tendency - the mode
Exploring Relationships between Variables
Missing Data
Parsing webpages and visualizing tabular HTML data
Cohort-Rebalancing (for Imbalanced Groups)

3. Data Visualization
Classification of visualization methods
Composition
Histograms and density plots
Pie Chart
Heat map
Comparison
Paired Scatter Plots
Barplots
Trees and Graphs
Correlation Plots
Relationships
Line plots using ggplot
Density Plots
Distributions
2D Kernel Density and 3D Surface Plots
Jitter plot
Appendix
Hands-on Activity (Health Behavior Risks)

4. Linear Algebra & Matrix Computing
Building Matrices
Create matrices
Adding columns and rows
Matrix subscripts
Matrix Operations
Addition
Subtraction
Multiplication
Elementwise multiplication
Matrix multiplication
Division
Transpose
Inverse
Matrix Notation
Matrix Algebra Notation
Solving Systems of Equations
The identity matrix
Vectors, Matrices, and Scalars
Sample Statistics
Mean
Variance
Applications of Matrix Algebra:
Linear modeling
Finding function extrema (min/max) using calculus
Least Square Estimation
The R lm Function
Eigenvalues and Eigenvectors
Other important functions
Linear regression
Sample covariance matrix

5. Dimensionality Reduction
Principal Component Analysis (PCA)
Independent Component Analysis (ICA)
Factor Analysis (FA)
Singular Value Decomposition (SVD)

6. Lazy Learning – Classification Using Nearest Neighbors
Understanding classification using nearest neighbors
The kNN algorithm
Calculating distance
Choosing an appropriate k
Preparing data for use with kNN
Why is the kNN algorithm lazy?
Predictive Diagnostics

7. Probabilistic Learning – Classification Using Naive Bayes
The Naive Bayes Algorithm
Assumptions
Bayes Formula
The Laplace Estimator
Case Study: Head and Neck Cancer Medication

8. Divide and Conquer – Classification Using Decision Trees
Understanding decision trees
Divide and conquer
The C5.0 decision tree algorithm
Working with Decision Trees
Choosing the best split
Pruning the decision tree
Boosting the accuracy of decision trees
Making some mistakes more costly than others
Understanding classification rules
Separate and conquer
The One Rule algorithm
The RIPPER algorithm
Rules from decision trees

…

13. Evaluating Model Performance
Measuring performance for classification
Working with classification prediction data
Evaluation: Confusion matrices
Other performance measures
Visualizing performance tradeoffs
Estimating future performance (internal statistical validation)
The holdout method

14. Improving Model Performance
Using caret for automated parameter tuning
Creating a simple tuned model
Customizing the tuning process
Improving model performance with meta-learning
Understanding ensembles
Bagging
Boosting
Random forests
Training random forests
Evaluating random forest performance

15. Data Formats and Optimization of Computation
Working with specialized data and databases
Querying data in SQL databases
Web-page Data Scraping
Downloading the complete text of web pages
Parsing JSON from web APIs
Reading and writing Microsoft Excel spreadsheets using XLSX
Generalizing tabular data structures with dplyr
Optimization and improving the computational performance
Parallel computing
GPU computing
Visualizing network data

16. Variable/Feature Selection
Variable selection methods
Case Study - ALS
Evaluating model performance

17. Regularized Linear Modeling and Knockoff Filtering
Regularized Linear Modeling
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO) Regression
Linear Regression
Assessing Prediction Accuracy
Estimating Prediction Error
Improving Prediction Accuracy
General Regularization Framework
Example: Neuroimaging-genetics study of Parkinson's Disease Dataset
n-Fold Cross Validation
Knock-off Filtering: Simulated Example
PD Neuroimaging-genetics Case-Study
Visualization

18. Big Longitudinal Data Analysis
Time series analysis
Identifying the Diff, AR and MA parameters
Structural Equation Modeling (SEM)
Case study - Parkinson's Disease (PD)
Linear Mixed model
GLMM and GEE Longitudinal data analysis

19. Text Mining & NLP
Term Frequency (TF), Inverse Document Frequency (IDF)
Document Term Matrix (DTM)
Case-Study: Job ranking

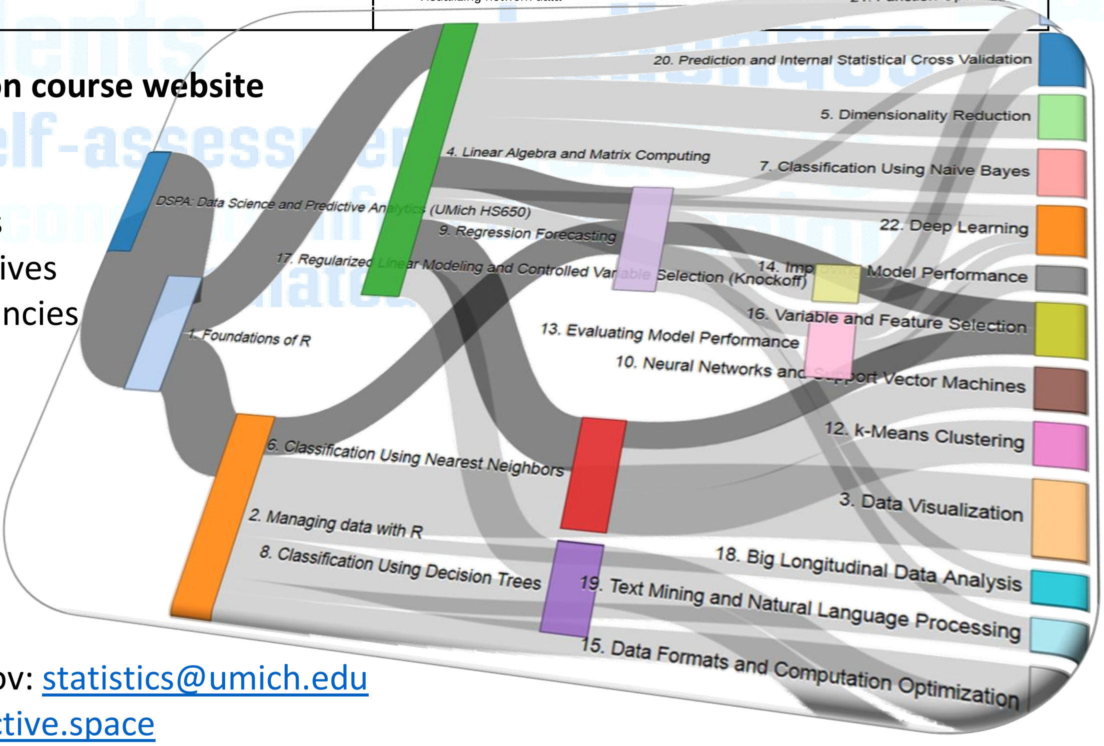
20. Prediction and Internal Statistical Cross Validation
Forecasting types and assessment approaches
Overfitting
Internal Statistical Cross-validation is an iterative process
Example (Linear Regression)
Case-Studies
Summary of CS output
Alternative predictor functions
Prediction Models
R Debugging

21. Function Optimization
Linear and Quadratic Programming
Manual vs. Automated Lagrange Multiplier Optimization
Data Denoising

22. Deep Learning
Perceptrons
Simple Neural Net: XOR and NAND
Schizophrenia Neuroimaging Study
Spirals 2D Data

Additional information on course website

- ○ Prerequisites
- ○ Enrolment logistics
- ○ Coverage & Objectives
- ○ Outcome Competencies
- ○ Certification
- ○ CMS/Canvas links
- ○ Class-notes
- ○ Video Lectures
- ○ Assignments
- ○ R code
- ○ Calendar
- ○ Instructor: Dr. Dinov: statistics@umich.edu
- ○ http://DSPA.predictive.space


21. Function Optimiza
20. Prediction and Internal Statistical Cross Validation
5. Dimensionality Reduction
4. Linear Algebra and Matrix Computing
7. Classification Using Naive Bayes
DSPA: Data Science and Predictive Analytics (UMich HS650)
9. Regression Forecasting
22. Deep Learning
17. Regularized Linear Modeling and Controlled Variable Selection (Knockoff)
14. Improving Model Performance
16. Variable and Feature Selection
1. Foundations of R
13. Evaluating Model Performance
10. Neural Networks and Support Vector Machines
12. k-Means Clustering
6. Classification Using Nearest Neighbors
3. Data Visualization
2. Managing data with R
8. Classification Using Decision Trees
18. Big Longitudinal Data Analysis
19. Text Mining and Natural Language Processing
15. Data Formats and Computation Optimization