Introduction to Data Science

Course Overview

This course is about learning from data, in order to gain useful predictions and insights. Separating signal from noise presents many computational and inferential challenges, which we approach from a perspective at the interface of computer science and statistics. Through real-world examples of wide interest, we introduce methods for five key facets of an investigation:

- data munging/scraping/sampling/cleaning in order to get an informative, manageable data set;
- data storage and management in order
- to be able to access data - especially big data - quickly and reliably during subsequent analysis;
- exploratory data analysis to generate hypotheses and intuition about the data;
- prediction based on statistical tools such as regression, classification, and clustering; and
- communication of results through visualization, stories, and interpretable summaries.

Learning Outcomes

After successful completion of this course, you will be able to:

- Use Python and other tools to scrape, clean, and process data
- Use data management techniques to store data locally and in cloud infrastructures
- Use statistical methods and visualization to quickly explore data
- Apply statistics and computational analysis to make predictions based on data
- Apply basic computer science concepts such as modularity, abstraction, and encapsulation to data analysis problems

Week - 1

1. Introduction to Data Science (What is Data Science and Data Science Process, What are Data, Data Sources, Types of Data, Data Formats, Messy Data)
2. Data Exploration (Basics of Sampling, Biases in Sampling, Measures of Centrality, Measures of Spread, Principles of Visualizations, Histograms, Scatter plots, Pie Charts, Stacked Area Graphs, Box Plots, 3D data)
3. Data Engineering (Tabular Data, Pandas and Scraping)
4. Exploratory Data Analysis and Effective Data Visualizations
5. Code Camp-1 (Web Scraping and EDA: Scraping Wikipedia Billboard for Top 100 : Homework 1) https://canvas.harvard.edu/courses/29726/assignments/167915

Week - 2

6. Introduction to Regression (Statistical modeling, predicting a variable, Regression Vs Classification, Error, Loss functions, Line of Best Fit, K nearest neighbours)
7. Linear Regression (Linear Regression, Comparing Models, Evaluating Model, Model Uncertainty, Bootstrapping for estimating sampling error, Model Fitness R^2, Training Vs Testing sets,
8. Multiple Linear Regression - 1 (Multiple Linear Regression, Evaluating Significance of Predictors, Hypothesis testing, R^2, Information Criteria, AIC/BIC)
9. Multiple Linear Regression - 2 (Comparing parametric and nonparametric models, Multiple Linear Regression with Interaction Terms, Polynomial Regression)
10. Code Camp-2 (Multiple Linear Regression: Forecasting Bike Sharing Usage: Homework 3) https://canvas.harvard.edu/courses/29726/assignments/172398

Week - 3

11. Model Selection (overfitting, model selection, variable selection, Cross Validation)
12. Regularization (Bias Vs Variance, Regularization: LASSO and Ridge,
13. PCA (High Dimensionality, Dimensionality Reduction, PCA, Using PCA for Regression)
14. Visualization for Communication (Visualization Goals, Effective Visualizations, Tools for interactive graphics, Structure of Communicative Graphics, Application to modeling)
15. Code Camp-3 (Model Selection, Regularization, PCA: Forecasting Bike Sharing Usage: Homework 4) https://canvas.harvard.edu/courses/29726/assignments/172400

Week - 4

16. Classification, Logistic Regression - 1 (Classification, Binary Response & Logistic Regression, Bayes classifier, Model Diagnostics in Logistic Regression, Multiple Logistic Regression, Classification Boundaries)
17. Classification, Logistic Regression - 2 (Multiple Logistic Regression, Classification Boundaries, ROC curve, k-NN for Classification, Choice of k, k-NN with Multiple Predictors)
18. Missing Data (Dealing with Missing Data, Naively handling missingness, Types of Missingness, Sources of Missingness, Imputation Methods, Handling missing data)
19. Decision Trees (Geometry of Data, Interpretable Models, Decision Trees, Numerical vs Categorical Attributes, Splitting Criteria, Gini Index, Information Theory, Entropy, Stopping Conditions & Pruning)
20. Code Camp-4 (Logistic Regression, ROC, Data Imputation: Automated Breast Cancer Detection:) Homework 6 https://canvas.harvard.edu/courses/29726/assignments/175288

Week - 1

Module - 1

Introduction to Data Science

Module Overview:

Learning Objectives:

Python recap

List Comprehension

Lambda Functions

Tasks :

Lab0

Module - 2

Data Exploration

Module Overview:

Learning Objectives:

Intro to Numpy, Matplotlib

Tasks :

Lab1

Module - 3

Data Engineering

Module Overview:

Learning Objectives:

Numpy continuation

Intro to Pandas

Tasks :

Lab1

Module - 4

Exploratory Data Analysis and Effective Data Visualizations

Module Overview:

Learning Objectives:

Completing of the cs109a_hw0

Tasks :

Lab2, cs109a_hw0

Module - 5

Code Camp - 1

Module Overview:

Learning Objectives:

Requests

Beautiful Soup

Tasks :

Lab2

Module - 6

Week - 1 Exam

Tasks :

cs109a_hw1

Week - 2

Module - 7

Introduction to Regression

Module Overview:

Learning Objectives:

Module - 8

Linear Regression

Module Overview:

Learning Objectives:

Module - 9

Multiple Linear Regression - 1

Module Overview:

Learning Objectives:

Module - 10

Multiple Linear Regression - 2

Module Overview:

Learning Objectives:

Module - 11

Code Camp - 2

Module Overview:

Learning Objectives:

Module - 12

Week - 2 Exam

Week - 3

Module - 13

Model Selection

Module Overview:

Learning Objectives:

Module - 14

Regularization

Module Overview:

Learning Objectives:

Module - 15

PCA

Module Overview:

Learning Objectives:

Module - 16

Visualization for Communication

Module Overview:

Learning Objectives:

Module - 17

Code Camp - 3

Module Overview:

Learning Objectives:

Module - 18

Week - 3 Exam

Week - 4

Module - 19

Classification, Logistic Regression - 1

Module Overview:

Learning Objectives:

Module - 20

Classification, Logistic Regression - 2

Module Overview:

Learning Objectives:

Module - 21

Missing Data

Module Overview:

Learning Objectives:

Module - 22

Decision Trees

Module Overview:

Learning Objectives:

Module - 23

Code Camp - 4

Module Overview:

Learning Objectives:

Module - 24

Week - 4 Exam