CS2545 - Data Science for Big Data Analytics

Fall 2018 - Course Outline

Course description

Data science enables one to bring structure to large quantities of data and make analysis possible. The purpose of the course is to introduce students to the fundamentals of data science and prepare them in dealing with the challenges of Big Data analytics. It covers advanced Python programming and Python libraries for data analysis. It presents data visualization techniques and statistical methods, as well as data exploration techniques such as data cleaning and munging, manipulating data, rescaling and dimensionality reduction. It includes an introduction to machine learning with linear regression, classification and clustering and presents special data analysis topics of time-series analysis. Also, it introduces data analysis approaches with relational databases and big data frameworks such as Dask.

Topics

1. Introduction to Python

i) Basics
ii) Advanced concepts
iii) Python tools and libraries for data analysis, such as, IPython notebook, NumPy and SciPy

2. Data Wrangling and Exploration

i) Introduction to Pandas
ii) Working with data: Cleaning and Munging, Manipulating Data
iii) Transformation, Rescaling, Dimensionality Reduction

3. Data Visualization

i) Visualization fundamentals, Infographics, Interactive Visualization
ii) matplotlib, Bar Charts, Line Charts, Scatterplots

4. Statistics

i) Statistics basics: Describing a set of data, Central tendencies, Outlier
ii) Probability basics, Pmf, Cdf, Pdf, Modeling distributions, Estimation
iii) Relationships between variables, Correlation and causation
iv) Hypothesis and Inference: Statistical Hypothesis Testing, Confidence Intervals

5. Machine Learning Introduction

i) Basics of machine learning
ii) Machine learning techniques:
Linear Regression
Classification: k-Nearest Neighbors
Clustering

6. Data Engineering: Data Manipulation at Scale

i) Accessing data from relational databases
ii) Scaling data analysis with Dask

7. Special Topics in Data Analysis

i) Time-series analysis (*time permitting*)
ii) Geospatial analysis (*time permitting*)