Level 0

Date & Time: TBA (One-day seminar)

Instructor: TBA

Course Format: Lectures and discussion sessions

Course Description

This course is an overview of data science. The objective of this course is to introduce the four key areas of data science: machine learning/artificial intelligence and data mining, data engineering, visual analytics and human-computer interaction, and business analytics. After this seminar, participants are expected to have a good understanding of what data science is as well as what it can do to help in business operations.

Tentative Topics

- What is data science? What do data scientists do?
- Engineering for Data Scalability
- Effective Communications with Visual Analytics
- Big Data Analytics Success Stories

Level 1.1

Date & Time: TBA (5 classes/days delivered in 3-4 sessions, each session is 1-2 days)

Instructor: TBA

Course Format: Lectures, practical sessions, and a workshop

Course Description

This course establishes a foundation in mathematics, statistics, programming, and database required for data science. We will cover basic mathematics such as vector, matrix, and functions, as well as basic statistics. We will also get you started with basic programming in Python and SQL. At the end of this course, you should be able to perform basic data analysis such as querying data, and generating plots and summary statistics in Python.

Course Details (Tentative)

Class 1

Vector & Matrix

- What is a vector?
- Basic vector operations: addition, scalar multiplication, dot product, cross product
- What is a matrix?
- Vector & Matrix in Python

Basic Python

- Basic syntax e.g. print(), creating variables
- Basic packages, importing, calling functions from imported packages

Class 2

Statistics I

- Histogram and empirical distribution
- Measures of central tendency
- Measures of variance

Basic Visualization: Matplotlib

- Matplotlib architecture
- Basic plotting with matplotlib
- Scatter plots, line plots, and bar charts
- Histograms and box plots
- Heatmaps

Class 3

Toolbox Tutorials

- Command line
- Git

Equation, Functions, and Graphs I

- What is a function?
- Domain, range

Fundamentals of Programming I

- Variable and types
- Functions

Class 4
Equation, Functions, and Graphs II
- How to graph a linear function (intercept and slope)
- Exponentials, radicals, and logarithms
- Polynomials
- Factorization

Fundamentals of Programming II
- Control Flows

Class 5
SQL
- What are relational database and SQL?
- Basic commands: SELECT

Workshop: Mini Data Workflow

Level 1.2

Date & Time: TBA (3-4 sessions, each is 1-2 days)

Instructor: TBA

Course Format: Lectures, practical sessions, and a workshop

Course Description

This course continues to build a foundation in mathematics, statistics, programming, and database. We will cover relevant statistics, machine learning basics, and how to work with data in Python. You will also learn intermediate SQL commands in this level. At the end of this course, you should be able to query and clean data, as well as build a simple statistical/machine learning model in Python.

Course Details (Tentative)

Class 1

Python Refresher

Python: Data Structure, Numpy, Pandas

- Reading/writing data
- Lists
- Arrays in Numpy
- Data frame in Pandas

Statistics II

- Probability basics
- Random variables

Class 2

Statistics III

- Sample and sampling distribution
- Statistical inference: confidence interval and hypothesis testing
- Linear regression
- Statistical inference and linear regression in Python

Class 3

Python: Data Wrangling and Data Summary

- Missing values
- Outliers
- Merging
- Group by
- Scales/regularization
- Strings and dates

Class 4

Introduction to ML

- Supervised vs unsupervised learning
- Regression vs classification
- Training, validation, parameter tuning, feature engineering
- Overfitting, bias and variance tradeoff

Class 5
SQL

- Joins
- Aggregate

Workshop: Build your first data model.

Level 1.3 (ML/AI and Data Mining)

Date & Time: TBA (3-4 sessions, each is 1-2 days)

Instructor: TBA

Course Format: Lectures, practical sessions, and a workshop

Course Description

This course is a broad overview of machine learning algorithms. We will cover both supervised and unsupervised learning techniques. By the end of this course, students should be able to build both regression and classification models using commonly used tools such as tree-based methods and neural networks.

Course Details (Tentative)
Class 1
Supervised Learning I–Regression

- Simple/multiple linear regression
- Generalized additive models
- Ridge regression and LASSO
- Tree-based methods

Class 2
Supervised Learning II—Classification

- Introduction to classification, e.g. how to evaluate classification models
- KNN
- Support vector machine
- Logistic regression

Class 3
Unsupervised Learning

- Dimensionality reduction
- K means
- Clustering: hierarchical

Class 4
Neural Networks and Deep Learning

- Logistic regression and multilayer perceptron
- Loss function, cost function, gradient, and back-prop
- Neural networks
- Reinforcement learning
- Convolutional Neural Networks and Recurrent Neural Networks

Class 5
Workshop: Creating an (almost) intelligent agent

Level 1.3 (Data Engineering)

Date & Time: TBA (3-4 sessions, each is 1-2 days)

Instructor: TBA

Course Format: Lectures and practical sessions

Course Description

This course is a broad overview of data engineering and database management. We will cover fundamental tools in data engineering and database management such as data warehouse, ETL, Hadoop, and Spark.

Course Details (Tentative)

Class 1

- Query Processing
- Transaction Concept
- Transaction States
- Concurrent Executions
- Concurrency Control

Class 2

- Basic Motivations
- Introduction to Fragmentation
- Introduction to Replication
- Distributed Query Processing
- Introduction to Parallel Database Management System (DBMS)
- Taxonomy of Data Parallelism

Class 3

- Introduction to Hadoop
- Hadoop File System
- Data Lake
- MapReduce Programming

Class 4

- Data Pipeline Design
- What is ETL (Extract, Transform, Load)?
- ETL on Hadoop
- NoSQL Data Stores
- Data Warehousing

Class 5

- Introduction to Spark
- Writing Spark ETL Processes
- Using Spark with NoSQL Data Stores