Data Science Essentials

Data Science Onramp Essentials aim to build and enhance your data science skills to meet the prerequisite of our data science program.

Most of the mini courses are written in text format and few in video format. We provide Teaching Assistant (TA) support and office hours. If you encounter any problems, please feel free to reach our TAs either at their office hours or schedule an appointment which fits better to your schedule. Each mini course has its own grading policy. In general, grading is based on assignments/projects, online discussions, and quizzes. If you select more than one mini course, the average of each mini course grade will be counted towards your final grade.

It includes the following mini courses: Intro to R, Basics of Python, Intro to SQL, Basics of Java, Intro to C++, Basic Probability, MongoDB, and Linear Algebra.

Introduction to R

The goal of this course is to teach you how to program in R and how to use R for effective data analysis. We will also cover in detail the plotting systems in R as well as some of the basic principles of constructing data graphics. The course consists of 3 parts with exercises, quizzes in each chapter and a final project of student’s choice.

1. First we cover basics of R programming - Data Structures supported by R, Control structures, Functions, Debugging techniques. These topics will help you get started with R and help you in writing good reusable R codes. This will also serve as basis for what is taught in the later sections.

2. The second part is the major focus of the course, exploratory data analytics and visualization. In this section you will learn how to manipulate the data. You will also learn how to use base plotting system of R, lattice and ggplot packages to plot graphs. We will not only teach you how to plot graphs in R but also provide general rules you can follow to build appropriate graphs and be able to tell the story of what is happening in the data using graphs.

3. In this section we will teach how to showcase your work. First you will learn how to use R markdown to make documents that can be made into a webpage. Then you will learn to make web apps and showcase your work using shiny package. By the end of this course you will be making your very own web app!

Course Structure:

- Getting started with R (installing R and R studio, writing some basic commands, and installing libraries in R)
- Data structures supported by R
• Operators, control structures, and functions
• Introduction of some utilities, including basic statistics in R and basic debugging
• How to tidy and manipulate the data sets in R
• How to make plots in R
• The R work showcase-R markdown and shiny package

**Basics of Python**

In this course, we are going to learn about the basics of python. In industries, most of the computer programmers use two important approaches to writing complex applications, recursive approach, and iterative approach. You will learn about these important concepts from modules as well as from programming assignments. It is good to have hands on experience of the concepts you have learned so, at the end, you will get introduced to one of the most commonly performed tasks of classification in data mining and machine learning and we will implement a recommendation system using Scikit-learn package. In this course, we ill also introduce you to other useful packages, like Matplotlib and Scikit-learn, which are being widely used in the industry.

**Course Structure:**

• What is Python and how to install it
• Basic operators in python
• Data structures and metrics in Python
• Some basic algorithm concepts including statements, functions and recursion
• Two basic python packages: Numpy and Pandas
• How to visualize in Python
• Basic machine learning concepts
• Run regression model ins Python

**Introduction to SQL**

This is a course on SQL(Structured Query Language) which covers the basic concepts of databases and SQL as a language for accessing databases. This course is intended for beginners and assumes no background whatsoever in databases/SQL. We will start the course by introducing the need to use databases, building entity-relationship models, installing MySQL and then we will proceed to programming(querying) the database using SQL. At the end of the course, you will work to create a sample database, input data and perform complex SQL operations on the database to retrieve the desired output. The assignments are included to provide a working knowledge of the concepts taught throughout the course.

**Course Structures:**

• Introduction to databases
Basics of Java

Java is one of the most fundamental programming languages used in industry. In all programming rankings, Java always ranks in the top 3, which can clearly show its importance. One advantage for Java is that Java is an object-oriented programming language, which makes it easier to integrate Java codes together. As it runs on JVM, it is so easy to set up the running environment, which means it is good for code immigration.

This is an intro-level Java course. It is intended for people without any Java programming experience. The course is an interweaving of two parts. The first part is knowledge about basic programming with Java, and the second part is to use the knowledge to work on a project. The two parts will not be strictly separated. Once we are equipped with enough knowledge to take a small step, we make some progress in our project. In the process, we may encounter some confusion and wonder what could be done next. Then we learn something new that helps us resolve the problem, and make further progress. This cycle repeats.

Course Structure:

- Install Java and run Eclipse
- Basic concepts such as variable and command in Java
- How to write a method in Java
- Basic data structures in Java
- Practice your knowledge with the first project
- Practice your knowledge with the second project
- Latent semantic analysis concepts & code in Java
Introduction to C++

R, Python, Java, and MATLAB are the primary languages used by data scientists, but sometimes they are just not fast enough! One of the strategies to optimize them is to rewrite the most time critical code in a language closer to the hardware, including C/C++. Among these languages, Cython and Rcpp are two main methods to speed up Python and R by C/C++. While Cython is mainly a dialect of Python, Rcpp is to inline C/C++ program into a R program, so Rcpp users really need to write a fraction of programs in C/C++.

This course requires no prior knowledge of any programming languages, and we will not cover programming languages other than C/C++. However, we will go through the C++ appendix of Eddelbuettel's "Seamless R and C++ integration with Rcpp," which provides the necessary knowledge of C/C++ for reading how to use Rcpp in the main text. After learning a solid knowledge of standard C/C++, you will have a basic idea how to write a C/C++ program. Whenever you need to optimize their R and Python, you will have enough knowledge to start learning Rcpp and Cython by yourselves.

Course Structure:

- Why C++ is important
- How to compile and debug C++ program
- Control structures in C++
- Basic C++ functions
- Pointers and memory management in C++
- What is object oriented and how to use it in C++
- Generic programming and the STL

MongoDB

This course is for those students who are not familiar with DataBase concepts and want to learn NoSQL databases like MongoDB. Students who know MongoDB and wants to work on a project using python and MongoDB to develop their cross-technology skills will also be benefited by taking this course. In this course, we will work on yelp data set which contains more than 1.5 million records and hence is ideal to learn MongoDB MapReduce concept.

Course Structure:

Part A: Basics of MongoDB and Json

Module 1

- Introduction to MongoDB : SQL vs NoSQL, Advantage and disadvantage
Module 2

- Playing around with Database: DB creation, collection creation
- Data Type supported by MongoDB
- Insert operation
- Accessing MongoDB using pyMongo

Module 3

- What is Python and IDE?
- Installation of Python and pyCharm IDE
- Yelp data Insertion in the MongoDB using

Module 4

- More complex find query: And, Or operations
- Working with list
- Projection, Limiting, Sorting

Module 5

- Count, Indexing of records
- Regular expression in MongoDB
- Text search, Aggregation and pipeline

Module 6

- Map-Reduce concept in theory: Map function and Reduce function
- Concurrency in map reduce.

Part B: Python Refresher 2 days

Module 7

- Introduction To Python Data Structure: Lists, Tuples, Dictionary
- Operation on Data Structures: Sorting, Searching etc.
- String Matching and Manipulation: string Concatenation, Comparison, Slicing, and Splitting.
- Date Time manipulation: Concept of package import.

Module 8
- Conditional Statements: If, If ... else, If ... elif ... else
- Looping construct: for, while
- Functions: Simple functions, Parameter passing, default parameters, return values.
- Exception handling

**Part C: Data Analysis of Yelp data using pyMongo and MongoDB**

**Module 9**

- Accessing MongoDB via Python: pyMongo
- Retrieving yelp data from MongoDB via python: Find, Aggregate etc.
- Updating Records using pyMongo
- MapReduce using pyMongo

**Part D: Data Visualization**

**Module 10**

- Introduction to matplotlib and basemap
- Pi chart, Histogram, Scatter plots, Bar Graph and Box Plots
- Generation of Word Cloud: Using wordcloud package

**Linear Algebra**

In this course, we will cover basic Linear Algebra and Calculus used in Machine Learning with Python. Machine Learning with Python course breaks the topics of machine learning into different Pages. Each page uses different levels of Linear Algebra and Calculus from easy to hard. Thus, we organize our course in different parts. By understanding each part (with their previous parts), we understand the Linear Algebra and Calculus used in a particular page in Machine Learning with Python course. In particular, Part A and B cover one variable differentiable Calculus used in Parameter ESTIMATION. Part C covers basic matrix operations. In particular, we will explain the details when a matrix is invertible. Part D explains multi-variable differentiable Calculus which is used in Curve Fitting and Error Function.

**Course Structure:**

**Part A: Linear Algebra used in Parameter ESTIMATION**

**Unit 1: Vectors**

- Mathematical Definitions of Vectors
- Represents and Plot Vectors in Python
Unit 2: Sets
- Mathematical Definitions of Sets
- Represents and Plot Sets in Python

Part B: Calculus used in Parameter ESTIMATION

Unit 3: Functions
- Mathematical Definitions, Python Code, and Graph of Functions
- Polynomial Function and Solve a Polynomial Equation
- Exponential and Monotonic Function
- Logarithmic and Inverse Function

Unit 4: Derivatives
- Slope
- Tangent Line
- Derivative

Unit 5: Computing Derivatives
- Computing Derivatives

Unit 6: Optimization Problems
- Maxima and Minima
- Optimization Problems
- Second Derivative Test

Project 1: Use what we learned to explain what and why we did in 7.6 Parameter ESTIMATION: Map in Machine Learning with Python Course.

Part C: Linear Algebra used in Linear Regression

Unit 7: Matrices
- Mathematical Definitions of Matrices
- Transpose and Common Matrices
- Represent Matrices in Python

Unit 8: Matrices and Vectors Operations
- The Idea of Matrices Multiplication
- The Idea of Rank of Matrices
- Transpose
- Compute Matrix Operations in Python

Unit 9: Determinant, Rank and Inverse (I)
• The Idea of Determinant of Matrices
• The Idea of Inverse of Matrices

Unit 10: Determinant, Rank and Inverse (II)

• Existence of the inverse matrices and Cramer’s Rule
• How to get an inverse Matrices
• From Normal Equation to Estimates

Project 2: Use what we learned to explain what and why we did in 8.2 Linear Regression in Machine Learning with Python Course.