Roadmap to
DATA SCIENCE (DS)

Introduction

So now the very first question arises is, "**What is Data Science?**" Data science means different things for different people, but at its gist, **data science is using data to answer questions.** This definition is a moderately broad definition, and that's because one must say data science is a moderately broad field!

Data science is the science of analyzing raw data using statistics and machine learning techniques with the purpose of drawing conclusions about that information.

So briefly it can be said that Data Science involves:

- Statistics, computer science, mathematics
- Data cleaning and formatting
- Data visualization

The roadmap is divided into 6 weeks of content and each week progressively builds upon a new concept. The roadmap includes a lot of material and assignments, and with the right amount of enthusiasm and determination, one can ace Data Scientist role easily.

Week 1 (Reviewing Python and R)

Day 1 & 2 – Learn required programming language!

There are many programming languages out there, of which only two are suitable for Data Science, namely Python and R.

1) PYTHON –

Python provides a vast selection of libraries, namely NumPy, pandas, sklearn, TensorFlow, PyTorch, etc., which are super helpful and require little effort for data science.
Before starting, it is important to setup python in your device, using this as a reference.

Learning python is not hard. Here are a few resources which will teach you about the language swiftly:-

- Medium Blog
- YouTube Video by Free Code Camp

In case you come across a weird syntax or want to find a solution to a problem, the official documentation is the best way to resolve the issues!

2} R -

Revise R language from here.

Day 3 – Explore libraries of R

- dplyr
- ggplot2
- Tidyr
- Shiny

Day 4 – Explore Numpy,Pandas

The best way to learn about libraries is via their official documentation.
Other resources are as follows:-

- Video By Free Code Camp
- Numpy in 15 minutes

Data is what drives machine learning. Analyzing, visualizing, and cleaning information is an essential step in the process. For this purpose, Panadas comes to the rescue!
Pandas is an open-source python package built on top of Numpy and developed by Wes McKinney**.**
Like NumPy, Pandas has official documentation, which you may refer to here. Other resources are as follows:-

- Medium Blog by Paritosh Mahto
- Pandas in 15 minutes

Day 5 – Explore MatPlotLib

Matplotlib is a powerful library that provides tools (histograms, scatter plots, pie charts, and much more) to make sense of data.
The best source to refer to is the documentation in case of discrepancies.
Below are the links to some valuable resources covering the basics of Matplotlib:-

- Code With Harry
- Free Code Camp

Day 6 & 7 - Play around in Kaggle

Use this day as a practice field, to utilize all your skills you learnt. Head over to Kaggle and download any dataset you like. Apply the skills you procured and analyze trends in different data sets. Here is a brief walkthrough of the UI.
All about Kaggle

Week 2 & 3 (Basic Mathematics for ML, KNN, Linear Regression)

FOR WEEK 2 –

Focus on these topics :

1) Mathematics

Math skill is very important as they help us in understanding various machine learning algorithms that play an important role in Data Science.

- **Part 1:**
    - Linear Algebra
    - Analytic Geometry
    - Matrix
    - Vector Calculus
    - Optimization
- **Part 2:**
    - Regression
    - Dimensionality Reduction
    - Density Estimation
    - Classification

2) Probability

Probability is also significant to statistics, and it is considered a prerequisite for mastering machine learning.

- Introduction to Probability
- 1D Random Variable
- The function of One Random Variable
- Joint Probability Distribution
- Discrete Distribution
    - Binomial (Python | R)
    - Bernoulli
    - Geometric etc
- Continuous Distribution
    - Uniform
    - Exponential

    - Gamma
- Normal Distribution (Python | R)

FOR WEEK 3 –

Focus on these topics :

3) Statistics

Understanding of **Statistics** is very significant as this is a part of Data analysis.

- Introduction to Statistics
- Data Description
- Random Samples
- Sampling Distribution
- Parameter Estimation
- Hypotheses Testing (Python | R)
- ANOVA (Python | R)
- Reliability Engineering
- Stochastic Process
- Computer Simulation
- Design of Experiments
- Simple Linear Regression
- Correlation
- Multiple Regression

Week 4 (Basic ML Algorithms)

Machine Learning

ML is one of the most vital parts of data science and the hottest subject of research among researchers so each year new advancements are made in this. One at least needs to understand basic algorithms of **Supervised and Unsupervised Learning**. There are multiple libraries available in Python and R for implementing these algorithms.

- **Introduction:**
    - How Model Works
    - Basic Data Exploration
    - First ML Model
    - Model Validation
    - Underfitting & Overfitting
    - Random Forests (Python | R)
    - scikit-learn
- **Intermediate:**
    - Handling Missing Values
    - Handling Categorical Variables
    - Pipelines
    - Cross-Validation (R)
    - XGBoost (Python | R)
    - Data Leakage

Week 5 (Getting familiar with Deep Learning)

Deep Learning

Deep Learning uses TensorFlow and Keras to build and train neural networks for structured data.

- Artificial Neural Network
- Convolutional Neural Network
- Recurrent Neural Network
- TensorFlow
- Keras
- PyTorch
- A Single Neuron
- Deep Neural Network
- Stochastic Gradient Descent
- Overfitting and Underfitting
- Dropout Batch Normalization
- Binary Classification

Week 6 (Getting familiar with transforming raw data into useful features)

Feature Engineering

In Feature Engineering discover the most effective way to improve your models.

- Baseline Model
- Categorical Encodings
- Feature Generation
- Feature Selection

Natural Language Processing

In NLP distinguish yourself by learning to work with text data.

- Text Classification
- Word Vectors