Guide to a Self-Taught DATA SCIENTIST

Shared by The Ravit Show

Skills for Data Scientist

- Mathematical and Statistical knowledge
- Knowledge of Machine Learning algorithms
- Programming Language knowledge. Ex: Python, R
- Handling large Datasets
- Domain Knowledge
- Problem-Solving ability
- Data Wrangling
- Database Management
- Data Visualization
- Cloud Computing
- Microsoft Excel
- DevOps

MATHEMATICS BASICS

- **Multivariable Calculus**
- Functions of several variables
- Derrivatives and gradients
- Step function, Sigmoid function, logit function, ReLU
- Cost Function
- Plotting of Functions
- Minimum and Maximum values of a function

- **Linear Algebra**
- Vectors Matrices
- Transpose of a Matrix
- The inverse of a Matrix
- The determinant of a Matrix
- Dotproduct
- Eigenvalues
- Eigenvectors

MATHEMATICS BASICS

- **Probability and Statistics Basics**
- Mean, Median, Mode
- Standard Deviation & Variance
- Correlation coefficient and the covariance
- MatrixProbability distributions(Binomial, Poisson, Normal)
- p-value, Baye's Theorem, Confusion Matrix, ROC Curve
- A/B Testing
- Monte Carlo Simulation

- **Optimization Methods**
- Cost function/Objective function
- Error function
- Gradient Descent and it's variants (e.g., Stochastic Gradient Descent Algorithm)

PROGRAMMING BASICS

- **Python**
- Basic Python
- OOPs Concept
- Jupyter Notebook
- Python Libraries such as
- Numpy, Pylab, Seaborn
- Matplotlib, Pandas
- Scikit-Learn
- PyTorch
- etc..

- **R**
- Basic R syntax
- Foundational R Programming Concepts such as Data Types, vectors arithmatic, indexing and Data Frames
- How to perform operations in R including sorting, data wrangling using dplyr, and data visualization with ggplot2
- R studio

MACHINE LEARNING BASICS

- **Supervised Learning**
- Basic Regression
- Multi Regression Analysis
- Regularized Regression
- Logistic Regression Classifier
- Support Vector Machine (SVM)
- K-nearest neighbour (KNN) Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Naive Bayes
- Gradient boosting
- etc

- **Unsupervised Learning**
- K Means clustering
- K-Median
- DBScan
- Hierarchical clustering
- etc..

EDA BASICS

- **Learn Data Basics**
- Learn how to manupulate data in various format, for example, CSV file, PDF file, TEXT file, etc..
- Learn how to clean data, impute data, scale data, import and export data, and scrap data from internet.
- Some packages of interest are pandas, MumPy, pdf tools, stringr, etc..
- Aditionally, R and Python contain several inbuilt datasets that can be used for practice.
- Learn data transformation and dimentionality reduction techniques such as covariance matrix plot, Principal Component Analysis (PCA) and Linear Descriminant Analysis (LDA)

- **Data Visualization Basics**
- Data Component
- Geometric Component
- Mapping Component
- Scale Component
- Labels Component
- Ethical Component

Build up your Online Presence

- Write Blogs
- Do projects and upload them on GitHub
- Public Speaking
- YouTube Tutorials
- Share your experience on Social Media
- Write Books
- Create a Course
- Podcast

NETWORKING

- Make friends
- Meet experts and talk with them
- Learn from experts
- Meet a mentor
- Make yourself visible to outside world
- It also helps you to get a good job in your dream companies.