En route to Data Science

Vitali Avagyan

Data Scientist
TurinTech
London, UK

updated on August 15, 2022

Essentials

- Communication skills
  - simplifying
  - explaining
  - demonstrating
  - visualising
- Networking
  - Big Data London
  - LinkedIn
  - reading blogs (*e.g.* MIT Technology Review) and writing blogs (*e.g.* Medium)
- reading conference papers and journals
  - NeurIPS, ICMLC, *etc.*
  - IEEE Transactions on Pattern Analysis and Machine Intelligence, Elsevier Pattern Recognition, IEEE Transactions on Neural Networks and Learning Systems, Journal of Machine Learning Research, *etc.*
  - google scholar, google-scholar alerts
  - Connected Papers

Python

- Python Fundamentals incl. Object-Oriented Programming (OOP) in Python
    - Modules, packages and paths
        - pip, Path (pathlib library), os, typing
        - pickle and json
    - Class vs Object; Encapsulation, Inheritance (+ Abstraction); Polymorphism
        - built-ins
        - DRY principle
        - f-string
        - properties, decorators
    - Hashable vs non-hashable data structures
        - list, dict, comprehensions, tuple, set
    - Iterators, generators and recursion
    - Exception handling and testing
        - assert; try ... except ...finally
        - mypy, pytest, unittest

Python for Data Science

- Python for Data Science
  - conda (anaconda), venv (python) -> environment management
  - Jupyter notebook (AWS SageMaker is a plus)
  - Pandas, Numpy, SciKit-Learn, Scipy
  - Tensorflow/Keras and Pytorch
  - Plotly, Bokeh, Matplotlib, Seaborn -> visualisation
  - Flask, Streamlit, Dash -> web applications

R

- Functional programming in R
    - Tools/Packages you should know on top of Base R
        - dplyr, data.table (dtplyr)
        - lubridate and hms
        - purrr, stringi/stringr, zoo
        - keras and reticulate (to work with Python)
        - ggplot2
        - DBI, RODBC and dbplyr (for database manipulations)
        - plumber and shiny (web applications)

SQL, IDEs, Linear Algebra

- Other languages: SQL (MySQL or Microsoft SQL Server, PostgreSQL is a plus)
- IDE/Interpreter
  - Visual Code (+ extensions), Posit (+ Posit Server for cloud), PyCharm
  - Databases; JSON, YML (YAML), flat files
- Linear Algebra
  - Vectors, Matrices, and Tensors, eigenvectors/eigenvalues, singular value decomposition

Statistics and Machine Learning

- Normal, Poisson, and Exponential distributions, Mean, SD, Percentiles; Gaussian and Markov-Chain processes
- Linear & Logistic Regressions; T-test/ANOVA (+ their non-parametric equivalents), error metrics (MSE, RMSE etc.); Monte-Carlo simulation; Sampling with (and without) replacement; Copulas
- Maximum Likelihood Estimator, Bayes rule, Prior/Posterior distributions, naive Bayes, K-means, KNN, Markov Chain, HMM, Decision Trees (+ Random Forest, XGBoost, LightGBM and CatBoost); Gaussian mixture models
- Neural Networks (feed-forward, CNNs, RNN variants (e.g. LSTM and GRU))
- Clustering (k-means); dimensionality reduction (PCA, UMAP AutoEncoders and t-SNE)

Times-series and Forecasting

- Weak and Strong Stationarity
  - Dickey-Fuller Test (Augmented Dickey Fuller Test)
  - Unit Roots, White Noise
  - Box-Jenkings and Yeo-Johnson transformations
- Autocorrelation and Partial Autocorrelation Function, multicolinearity
- Traditional models: ARIMA (SARIMA, SARIMAX), ETS, Prophet, Naive
- Deep Learning models: RNN variants (e.g. LSTM and GRU)), NeuralProphet
- Granger Causality, moderators, mediators, colliders and confounders
- Vector Auto Regression (VAR)
- ARCH and GARCH processes
- Anomaly Detection

Math, Linux and Containers

- Differential and Integral Calculus
  - Derivatives, Partial Derivatives
  - Interpolation, Taylor expansion
  - Optimisation (especially gradient descent-based algorithms); Lagrange multipliers; Constrained and Unconstrained Programming
- Linux/GNU and command-line proficiency
  - Understanding of namespaces
  - Use of SSH, use of Bash (or any other shell), and RegEx (it will save you tons of time)
    - re (Python)
    - awk and sed (shell)
  - Vi/Vim
  - PowerShell (preferable over CMD if on Windows)
  - git (the industry standard)
- Containerisation and Virtualisation
  - docker (understanding of microservices architecture and container orchestration (Kubernetes) is a plus)
  - Virtual Machines, Oracle VirtualBox, WSL (if on Windows 10)

Competitive Edge

- Cloud
  - Hands-on either with AWS, Azure, or GCP
  - cloud instances (ec2 etc.) and cloud storage systems (s3 etc.)
  - command-line interfaces (aws cli etc.), use of SDKs (paws, Boto3, etc.)
- Natural Language Processing
  - Bag-of-words
  - Transformers
  - Big pre-trained models (Word2Vec, GloVe, BERT, Albert) & Transfer Learning
- Reinforcement Learning
- DevOps and Infrastructure as a Code
  - Code review/debugging/testing
  - Ansible and Terraform
  - Configuration management (Chef or Puppet)
  - Automation (Jenkins)
- Computer Vision (CNN's and Transformers)

Big Data; Distributed Computing and AutoML

- Big Data
  - Distributed/parallel computing
  - map-reduce
  - apply family (base R) and map family from purrr library R
  - programming in Scala, Apache Spark (understanding of Hadoop helps), Apache Parquet file format
  - foreach, doParallel, doSnow (R)
  - joblib (Python); Ray; Dask (Python); PySpark
  - Understanding of RAM and computation in memory
- AutoML/Pipelines
  - EvoML (TurinTech)
  - DataRobot
  - Databricks
  - PyCaret
  - pickle, mlflow

Recommended Books - Machine Learning

- Machine Learning: An Applied Mathematics Introduction by Paul Wilmott (**foundational**)
- Machine Learning Engineering by Andriy Burkov (**foundational**)
- Deep Learning with Python, 2nd Edition by Francois Chollet (**foundational and practical**))
- Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow by Aurelien Geron (**practical**)

Recommended Books - Statistics

- Statistical Rethinking: A Bayesian Course with Examples in R and Stan, *Second Edition* by Richard McElreath (**foundational and practical**))
- Engineering Statistics by Douglas C Montgomery, George C Runger and Norma F Hubele (**practical**)
- Pattern Recognition and Machine Learning by Christopher Bishop (**theoretical**)
- The Elements of Statistical Learning by Hastie et al. (**theoretical**)

Recommended Books - Econometrics/Time-Series

- Basic Econometrics by Damodar Gujarati (**foundational**)
- Econometrics by Example, 2nd Edition by Damodar Gujarati (**practical**)
- Deep Learning for Time Series Forecasting by Jason Brownlee (**practical**)

Recommended Books - Software Engineering

- Expert Python Programming: Master Python by Learning the Best Coding Practices and Advanced Programming Concepts, *4th Edition* by Michał Jaworski, Tarek Ziadé
- Python Object-Oriented Programming, *Fourth Edition* by Steven F. Lott & Dusty Phillips
- The Pragmatic Programmer, *20th Anniversary Edition* by Andy Hunt and Dave Thomas

Other Resources

- Machine Learning Mastery (python)
- Statistical Rethinking lectures by Richard McElreath (statistics)
- ritvikmath (statistics)
- SoloLearn (programming)
- 3blue1brown (math)
- Kaggle (competitions)
- Abishek Thakur (practical ML)
- Super Data Science Podcast with Jon Krohn (podcast)
- Two Minute Papers (research)
- NeEDS - Network of European Data Scientists (research)
- Jay Alammar (NLP)
- KodeKloud (DevOps)

Thank you

Connect with me in LinkedIn