En route to Data Science

Vitali Avagyan

Data Scientist
TurinTech
London, UK

updated on March 31, 2022

R

For aspiring data scientists:

- Communication skills: simplifying, explaining, demonstrating, visualising etc.
- 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 and shiny
    - DBI, RODBC and dbplyr (for database manipulations)
    - plumber (web applications)

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)
  - Jupyter notebook (AWS SageMaker is a plus)
  - Pandas, Numpy, SciKit-Learn, Scipy
  - Tensorflow/Keras and Pytorch
  - Matplotlib, Seaborn and Bokeh (visualisation)
  - Flask, Streamlit (web applications)

SQL, IDEs, Linear Algebra and Stats

- Other languages: SQL (MySQL or Microsoft SQL Server, PostgreSQL is a plus)
- IDE/Interpreter
  - Visual Code (with its excellent extensions), RStudio (also RStudio Server for cloud), PyCharm
  - Databases; JSON, YML(YAML), flat files
- Linear Algebra
  - Vectors, Matrices, and Tensors
- Stats/ML
  - Normal, Poisson, and Exponential distributions, Mean, SD, Percentiles; Gaussian and Markov-Chain processes
  - Linear & Logistic Regressions; Forecasting;T-test/ANOVA, error metrics (MSE, RMSE etc.); Monte-Carlo simulation
  - Maximum Likelihood Estimator, Bayes rule, naive Bayes, K-means, KNN, Markov Chain, Decision Trees (Random Forest is a plus); Gaussian mixture models
  - Neural Networks (at least feed-forward neural networks, CNNs, RNNs (e.g. LSTM and GRU) and AutoEncoders)

Math, Linux and Containers

- Differential and Integral Calculus
  - Derivatives, Partial Derivatives
  - Interpolation, Taylor expansion
  - Optimisation (especially gradient descent-based algorithms)
- 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 (git has become the industry standard)
- Containerisation and Virtualisation
  - docker (understanding of microservices architecture and 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
  - Transfer Learning
- Reinforcement Learning
- DevOps and Infrastructure as a Code
  - Code review/debugging/testing
  - Ansible and/or Terraform
  - Configuration management (Chef or Puppet)
  - Automation (Jenkins)
- Computer Vision

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)
  - Dask in Python; Ray; PySpark
  - Understanding of RAM and computation in memory
- AutoML/Pipelines
  - EvoML (from TurinTech)
  - DataRobot
  - Databricks
  - PyCaret
  - pickle, mlflow

Recommended Books

- Machine Learning: An Applied Mathematics Introduction by Paul Wilmott; Machine Learning Engineering by Andriy Burkov; Basic Econometrics by Damodar Gujarati (foundational)
- Deep Learning with Python, 2nd Edition by Francois Chollet; Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow by Aurelien Geron (practical)
- Pattern Recognition and Machine Learning by Christopher Bishop; The Elements of Statistical Learning by Hastie et al (theoretical)
- Python Object-Oriented Programming, Fourth Edition by Steven F. Lott & Dusty Phillips; The Pragmatic Programmer, 20th Anniversary Edition by Andy Hunt and Dave Thomas (programming)

Other Resources

- Machine Learning Mastery (python)
- SoloLearn (programming)
- 3blue1brown (math)
- Kaggle (competitions)
- Abishek Thakur (kaggle)
- Super Data Science Podcast with Jon Krohn (podcast)
- Two Minute Papers (research)
- NeEDS - Network of European Data Scientists (research)
- Krish Naik
- Jay Alammar
- KodeKloud (DevOps)

Thank you

Connect with me in LinkedIn