Data Science Study Plan Version 2020

by June Li (LLD 01/20/2020)

Resources

DataCamp: www.datacamp.com

Foundations of Machine Learning by David Rosenberg: https://bloomberg.github.io/foml/#home

Coursea: www.coursera.org

Kaggle: www.kaggle.com

Udemy: www.udemy.com

Udacity: www.udacity.com

Scikit-learn: http://scikit-learn.org/stable/

Github: https://github.com

MIT Open Courses**:** https://ocw.mit.edu/courses/find-by-topic/#cat=mathematics

Website that have lots of great reading material: https://towardsdatascience.com
https://www.analyticsvidhya.com

Book: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition
https://www.amazon.com/gp/product/0387848576/ref=as_li_tl?ie=UTF8&camp=1789&creative=9325&creativeASIN=0387848576&linkCode=as2&tag=1point3acres-20&linkId=89120e90f087ad3021f401699fe775b8

Book: *Python Machine Learning*
https://www.amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow/dp/1787125939/ref=sr_1_1_sspa?ie=UTF8&qid=1539281720&sr=8-1-spons&keywords=python+machine+learning&psc=1

Book: *Deep Reinforcement Learning Hands-On*
https://www.amazon.com/gp/product/1788834240/ref=oh_aui_detailpage_o00_s00?ie=UTF8&psc=1

Book: *Practical Statistics for Data Scientists*
https://www.amazon.com/Practical-Statistics-Data-Scientists-Essential/dp/1491952962/ref=sr_1_1?ie=UTF8&qid=1539465170&sr=8-1&keywords=practical+statistics+for+data+scientists

Book: *Hands-On Machine Learning with Scikit-Learn & TensorFlow*
https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1491962291/ref=sr_1_3?ie=UTF8&qid=1541702946&sr=8-3&keywords=hands-on+machine+learning+with+scikit-learn+and+tensorflow

Book: Python Cookbook

https://www.amazon.com/Python-Cookbook-Recipes-Mastering-ebook-dp-B00DQV4GGY/dp/B00DQV4GGY/ref=mt_kindle?_encoding=UTF8&me=&qid=1548793920

Book: Cracking the Coding Interview: 189 Programming Questions and Solutions (for coding test)
https://www.amazon.com/gp/product/0984782850/ref=as_li_tl?ie=UTF8&camp=1789&creative=9325&creativeASIN=0984782850&linkCode=as2&tag=1point3acres-20&linkId=c6c3ed028a2060f7ab01a7d233dfa411

Github

https://github.com/miaojunlee

TensorFlow、Pytorch 和 Keras 的样例资源
https://zhuanlan.zhihu.com/p/51866340?utm_source=wechat_session&utm_medium=social&utm_oi=671946248773439488&from=singlemessage&isappinstalled=0

Great Resource with categories labeled: https://paperswithcode.com/area/nlp

https://aws.amazon.com/training/learning-paths/machine-learning/data-scientist/

Kaggle Cases (tabular data problems)

Corporación Favorita Grocery Sales Forecasting: https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data

Elo Merchant Category Recommendation https://www.kaggle.com/c/elo-merchant-category-recommendation

DonorsChoose.org Application Screening: https://www.kaggle.com/c/donorschoose-application-screening

TalkingData AdTracking Fraud Detection Challenge: https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/overview

Credit Card Fraud Detection: https://www.kaggle.com/mlg-ulb/creditcardfraud

Skewed value regression: https://www.kaggle.com/c/allstate-claims-severity

Kaggle Machine Learning Courses: https://www.kaggle.com/kashnitsky/mlcourse

IEEE-Fraud Case: https://www.kaggle.com/c/ieee-fraud-detection/discussion/111257

Feature Engineering: https://www.kaggle.com/learn/feature-engineering

Catboost vs lightgbm vs xgboost https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db

Part I Statistical & Mathematical Methods

Please utilize the MIT open course website to review the following topics either at undergraduate level or graduate level:

- Probability: Probability basics (axioms of probability, conditional probability, random variables, expectation, independence, etc.), multivariate distributions, introduction to concentration bounds, laws of large numbers, central limit theorem.
- Statistics: Maximum a posteriori and maximum likelihood estimation, minimum mean-squared error estimation, confidence intervals.
- Linear algebra: Vector spaces, linear transformations, singular value decomposition, eigendecomposition, principal component analysis, least squares, regression.
- Optimization: Matrix calculus, gradient descent, coordinate descent, introduction to convex optimization.
- Basic Statistics @ coursera
- An Intuitive Introduction to Probability @coursera
- Read *Practical Statistics for Data Scientists*
- Bootstrapping vs bagging: https://www.mikulskibartosz.name/bootstrapping-vs-bagging/

Part II Machine Learning Theory & Application

1. Python Basics
    a. Install Python 3.6 from anaconda
    b. Intro to Python for Data Science @ DataCamp
    c. Intermediate Python for Data Science @ DataCamp (matplotlib, dictionaries, loop)
    d. Pandas Foundations @ DataCamp (Read Data, Exploratory Data Analysis)
    e. Manipulate Data Frame with pandas (Rearrange Data and Transform Data) @DataCamp
    f. Complete Python Bootcamp: Go from zero to hero in Python 3 @udemy

2. Data Preprocessing
    a. Merging Data Frame with Pandas @ DataCamp
    b. Book Chapter 4: Building Good Training Sets: Data Preprocessing
    c. Preprocessing for Machine Learning in Python @DataCamp
    d. Python Data Science Toolbox (Part 1): Write your own functions and Lambda function
    e. After finishing the foundation courses, take Analyzing Police Activity with pandas @DataCamp. You will gain more practice cleaning messy data, creating visualizations, combining and reshaping datasets, and manipulating time series data.
    f. Outlier detection with Tukey method : https://www.kaggle.com/yassineghouzam/titanic-top-4-with-ensemble-modeling

3. Data Exploratory Analysis & Visualization
    a. Statistical Thinking in Python Part 1 @DataCamp
    b. Statistical Thinking in Python Part 2 @DataCamp
    c. Case Study in Statistical Thinking @DataCamp
    d. Statistical Simulation in Python @DataCamp
    e. Introduction to Data Visualization with Python @DataCamp
    f. Data Visualization: https://www.kaggle.com/learn/data-visualization-from-non-coder-to-coder
    g. Data Visualization: https://www.kaggle.com/learn/data-visualization

4. Dimension Reduction Techniques
    a. https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/
    b. PCA vs Factor Analysis:
    https://www.theanalysisfactor.com/what-is-a-latent-variable/
    https://www.theanalysisfactor.com/the-fundamental-difference-between-principal-component-analysis-and-factor-analysis/
    https://www.researchgate.net/post/Factor_analysis_Vs_PCA

5. Machine Learning Theory
    a. 3. Introduction to Stochastic Learning Theory @ bloomberg
    b. 4. Stochastic Gradient Descent @bloomerg
    c. Reading: A Kaggle Master Explains Gradient Descent:
    http://blog.kaggle.com/2017/01/23/a-kaggle-master-explains-gradient-boosting/
    d. 5. Excess Risk Decomposition

6. Logistic Regression & Advanced Regression Techniques
    a. Foundations of Predictive Analytics in Python (Part 1) @DataCamp
    b. Case Study: Churn Prediction @bloomberg
    c. Courseara: Practical Predictive Analytics: Models and Methods (Intuition for Regularization & Intuition for LASSO and Ridge Regression);
    d. https://stats.stackexchange.com/questions/866/when-should-i-use-lasso-vs-ridge
    e. Model Evaluation Metrics: https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/
    f. 6 L1 AND L2 Regularization @bloomberg
    g. 7 Lasso, Ridge & Elastic Net @bloomberg
    h. 8 Loss Function for Regression and Classification @bloomberg
    i. Linear Classifiers in Python @DataCamp

7. Support Vector Machine
    a. What's SVM https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
    b. Building SVM from scratch in Python https://github.com/adityajn105/SVM-From-Scratch

8. K-Nearest Neighbor
    a. Quick overview of KNN: https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7
    b. Building KNN in Python and R https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/#writing-our-own-knn-from-scratch
    c. KNN vs K-means Clustering http://abhijitannaldas.com/ml/kmeans-vs-knn-in-machine-learning.html

9. Tree-Based Models
    a. Decision Tree & Random Forest: Machine Learning with Tree-Based Models in Python @DataCamp
    b. Book Chapter 3: A Tour of Machine Learning Classifiers Using scikit-learn
    c. Supervised Learning with scikit-learn @DataCamp
    d. Bagging vs Boosting: https://becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e

10. Advanced Boosting Techniques
    a. Catboost
        - Catboost Python Package: https://tech.yandex.com/catboost/doc/dg/concepts/python-quickstart-docpage/
        - Catboost tutorial: https://github.com/catboost/catboost/tree/master/catboost/tutorials
    b. Xgboost
        - Extreme Gradient Boosting with XGBoost @ DataCamp

- Complete Guide to Parameter Tuning in XGBoost (with codes in Python): https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
- Hyperparameter Grid Search with XGBoost: https://www.kaggle.com/tilii7/hyperparameter-grid-search-with-xgboost

c. LightGBM
- LightGBM Python Package: https://lightgbm.readthedocs.io/en/latest/Python-Intro.html

d. Catboost vs Xgboost vs LightGBM https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db

e. (optional) Cost Sensitive Learning on imbalanced data https://github.com/nnikolaou/Cost-sensitive-Boosting-Tutorial

11. Hyper Parameter Tuning
a. Grid Search & Random Search http://scikit-learn.org/stable/modules/grid_search.html
b. Chapter 6: Learning Best Practices for Model Evaluation and Hyperparameter Tuning
c. Bayesian Optimization
- An Introductory Example of Bayesian Optimization in Python with Hyperopt: https://towardsdatascience.com/an-introductory-example-of-bayesian-optimization-in-python-with-hyperopt-aae40fff4ff0

d. Plotting Learning Curve http://scikit-learn.org/stable/auto_examples/model_selection/plot_multi_metric_evaluation.html#sphx-glr-auto-examples-model-selection-plot-multi-metric-evaluation-py

12. Model Evaluation: https://scikit-learn.org/stable/modules/model_evaluation.html

13. Neural Network & Deep Learning
a. Deep Learning Specialization @coursera https://www.coursera.org/specializations/deep-learning (Dr.Andrew Ng explains gradient descent and how forward and backward propagation very well. Highly recommend)
- ☐ Neural Networks and Deep Learning
- ☐ Improving Neural Networks: Hyperparameter tuning, Regularization and Optimization
- ☐ Structuring machine learning projects
- ☐ Convolutional Neural Networks
- ☐ Sequence Models

b. TensorFlow in Practice Specialization: https://www.coursera.org/specializations/tensorflow-in-practice
- ☐ Introduction to Tensorflow
- ☐ Convolutional Neural Networks in Tensorflow
- ☐ Natural Language Processing in TensorFlow
- ☐ Sequences, Time Series and Prediction

c. TensorFlow: Data and Deployment Specialization @coursera

d. Deep Learning & NN in pytorch for beginners @udemy
e. 李沐《动手学深度学习》课程视频汇总:https://www.jiqizhixin.com/articles/02111

Reading

f. Activation Functions: Neural Networks https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6
g. Epoch vs Batch Size vs Iterations: https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9
h. Optimization Algorithms: https://towardsdatascience.com/types-of-optimization-algorithms-used-in-neural-networks-and-ways-to-optimize-gradient-95ae5d39529f
i. Wide & Deep NN: https://ai.googleblog.com/2016/06/wide-deep-learning-better-together-with.html

14. Model Ensemble & Stacking
    a. Reading: Ensemble Learning to Improve Machine Learning Results https://blog.statsbot.co/ensemble-learning-d1dcd548e936
    b. https://scikit-learn.org/stable/auto_examples/ensemble/plot_voting_probas.html
    c. Reading: A Kagglers Guide to Model Stacking in Practice http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/

15. Unsupervised Learning
    a. Cluster Analysis & Dimension Reduction: Unsupervised Learning in Python @DataCamp
    b. Reading: https://towardsdatascience.com/an-introduction-to-clustering-algorithms-in-python-123438574097
    c. Reading: https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
    d. Reading: https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/ https://discuss.analyticsvidhya.com/t/dimensionality-reduction-is-good-or-bad/2444/4

16. Natural Language Processing
    a. Natural Language Processing Fundementals in Python @DataCamp
    b. Machine Learning with The Experts: School Budge @DataCamp
    c. The Fall of RNN & CNN https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0
    d. Memory, Attention, Sequence https://towardsdatascience.com/memory-attention-sequences-37456d271992

17. Hadoop, Spark & Scala
    a. Hadoop Platform & Application Framework: https://www.coursera.org/lecture/hadoop/introduction-to-apache-spark-9cq0R
    b. Big Data Analysis with Scala & Spark: https://www.coursera.org/learn/scala-spark-big-data

c. Python code @github for the two courses above: https://github.com/dangkhoadl/BigData-DistributedSystems-Courses/tree/master/Coursera_BigData_for_Data_Engineers
d. Big Data Essentials: HDFS, MapReduce & Spark https://www.coursera.org/learn/big-data-essentials (if you have time, consider taking all courses under this topic https://www.coursera.org/specializations/big-data-engineering)

18. Reinforcement Learning
    a. Book Reading: Deep Reinforcement Learning Hands-On
    b. Advanced AI: Deep Reinforcement Learning @udemy

19. Association Rule
    a. http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/
    b. https://towardsdatascience.com/association-rules-2-aa9a77241654
    c. https://www.kaggle.com/datatheque/association-rules-mining-market-basket-analysis

20. Programming: Python Beyond the Basics: Object Oriented Programming @udemy

21. Kubenetes & Kubeflow
    a. https://www.youtube.com/watch?v=vDSmAaRB07M&t=39s
    b. https://codelabs.developers.google.com/codelabs/cloud-kubeflow-pipelines-gis/index.html?index=../..index#0

Part III Data Management, Version Control & Misc

Data scientists often work with data from large scaled relational database; therefore, you should take the data management course provided by One Career instructor Y. Wu.

1. The Complete SQL Bootcamp @udemy
2. Git Complete @udemy