Data Science Syllabus

Spring 2019 Cohort

Summary (about 1 month each)

1. Core Tools and techniques
2. Data Science Fundamentals and Modeling
3. Neural Networks, Computer Vision, and NLP
4. Workshops and special topics
5. Projects at companies

**Written and approved by ITC Data Science Tech Lead and Fellows Program Director

1. Core Tools and Techniques

The gist: core skills that are fundamental in most DS workflows.

Syllabus

Tools and Frameworks

This section of the course includes all the core components that are essential to practicing Data Science in industry, from **good coding practices** and some essential data **visualization** techniques, to working in **version control** in a **Linux** environment and more. Students will "get their hands dirty" with Python packages like **pandas** and acquire additional practice with tools like **numpy**.

Spark, Docker, AWS

In the **Spark** class, students learn about one of the most up-and-coming technologies in Data Science today. Spark is a valuable tool that among other things is used to handle data in a tabular format (like Pandas) at any scale and with great speed. As part of this course students will practice how to interact easily with **HDFS and S3** and be exposed to different ways of dealing with large datasets.

In the **docker** and **virtual environments** modules, students learn how to create reproducible development environments via virtual environments, and how to create and edit docker environments for an even more robust managing of their packages and other software.

Python General, Model Serving, and Databases

In this set of classes, students will start using **PyCharm** and its debugging techniques, as well as Object Oriented Programming (**OOP**) and Classes, which are key for creating data scientists that can also contribute to production code.

Emphasis will also be placed on how to put models in production, **"serving" the models**, and engineering aspects around it.

Students will learn about SQL, how Embedded Databases operate and how to work with them in Python.

Optimization

Often the backbone of many Machine Learning algorithms is an optimization technique. In these set of classes, students will learn about some of the most

commonly used optimization algorithms, such as **Linear Programming** and **Gradient Descent**, and do exercises on the Knapsack problem and Sudoku.

In addition, students will learn about the optimization techniques that provide the entire machinery behind Neural Networks and Deep Learning, such as **Stochastic Gradient Descent**, Nesterov, Adam, and more.

Data Visualization

Presenting important results and convincing the right audience with visualizations is a key skill required from data scientists. In this module, students will learn about **Matplotlib** and **Seaborn** and their respective advantages. They will also have additional sessions on **Bokeh** and creating Dashboards. This will be critical for the personal and group projects throughout the training, as well as in future employment.

Statistics

Students will learn the core components of **Statistics and Linear Algebra** that are pertinent to understand the “behind the scenes” of many models that will be taught later on in the course. This includes basic Bayesian techniques, MAP, MLE, as well as Linear Regression and how to create Statistical Tests.

2. Data Science Fundamentals and Modeling

The gist: In this section, students will learn the most important core techniques in Machine Learning and Data Science. Most of those techniques and algorithms do not involve Neural Networks but are often simpler and better choices than NNs for many problems commonly found in the industry.

Syllabus

Data Modeling

In the Data Modelling module, some of the most important concepts in Data Science and Machine Learning will be introduced: under/overfitting, risk and loss functions, bias-variance trade-off and model selection, regularization, cross-validation, different metrics, and more. In this module, algorithms like **Naive Bayes, Logistic Regression, Decision Trees, and SVMs** will also be covered in **scikit-learn** - how to think intuitively about each one and how to regularize and tune them.

Ensemble Methods

Students will learn about ensemble methods such as **Random Forests, XGBoost, LightGBM**, and get hands-on practice tuning and working with these models.

Model Evaluation, Metrics, and Model Interpretability

Measure one's model's performance and intuition is critical in understanding progress and how useful a model will be in production. For a couple of classes, students will learn about the relevant **metrics** used in Data Science and experience selecting the good metrics for the tasks at hand. In addition, students will learn about **interpretability** of their models and of their model's predictions.

Unsupervised/Linear Algebra Based Methods

Students will learn about **clustering** techniques and how to use them to derive insights from data and simplify analyses. Many matrix methods (such **SVD, PCA, ICA**, etc) and graph/network methods will also be covered.

These are all incredibly powerful tools when understood well and will allow data scientists to have a deep understanding of the data, process data prior to feeding it another algorithm or straight-out solve the problem at hand.

Timeseries, HMMs, and Anomaly Detection

In this section, students will learn the basics of **Hidden Markov Models** and some of its core algorithms, such as Viterbi and Baum Welch as well as some of its applications, and how to apply it to Geo Snapping. In addition, students will learn how to model certain **time-series** events, such as seasonal and cyclical events, and learn about auto-correlation. Finally, students will learn and practice **Anomaly Detection** via statistical techniques and supervised learning.

3. Deep Learning, Computer Vision, and NLP

The gist: the nuts and bolts of Neural Networks, Computer Vision, and Natural Language Processing

Neural Networks: Unit for Both Tracks

We start with the fundamentals of Deep Learning and Neural Networks, which include the core components that are relevant for the most common Deep Learning applications (such as in computer vision, natural language processing, time series, and more), and serve as the foundation for both tracks and future learning.

Some of the topics covered here are **backpropagation** and the optimization methods for Neural Networks, the differences between **loss** functions and metrics. In addition, students will practice the main **types of layers** used in Neural Networks, as well as **regularization** techniques, such as Dropout, L1/L2 penalties and max-norms, and more.

Students will learn **CNNs using TensorFlow and Keras** and modern architectures. This unit covers the basics of **convolutions**, convolutional layers, and how, while sharing some intuition, they can be used to learn image features and classify images. This will first be taught in TensorFlow, and then in Keras, where it will be faster. This will serve as the foundational cornerstone for the state-of-the-art Deep Learning vision algorithms that students will be exposed to in later classes and specific workshops in the CV track.

Students will learn about neural sequence models, including recurrent neural networks (**RNNs**), Long Short-Term Memory Networks (**LSTMs**), and Gated Recurrent Networks (**GRUs**). They will experience how sequence models can be used to learn embeddings of time series data namely by encoding the sequences.

Finally, students will learn about **TFLite, Tensorboard**, and other relevant tools that make training and improving models better.

CV Syllabus

Computer Vision Track: Students will learn the classic CV methods all to way to Deep Learning applications specifically focused on vision tasks. This brings students close to the state-of-the-art in these fields, and enable them to understand the main paradigms behind the famous recent papers while also understanding how to apply them in practice.

CV and Image Processing

In the **Computer Vision** Classics module, students will be exposed to the **classic Computer Vision techniques** (i.e. not DL based), such as feature extraction, convolutions, and pyramids and practice with **openCV**. These are still used in many applications, are useful in their own right, and also crucial for properly appreciating the following section on Deep Learning for Computer Vision.

Image Segmentation and Object Detection

In the Image Segmentation course, students learn the essential algorithms for classification using sliding windows, localization, segmentation and the relevant metrics for the task (intersection of the union). Students will also practice with and learn about **R-CNN** and its improvements Fast- and Faster-R-CNN, as well as **Yolo** and an overview of state-of-the-art algorithms for similar tasks. Most of these algorithms were developed in the past few years and improving upon them is a vibrant area of research.

Super Resolution

Students will learn about different algorithms for super-resolution, such as **Interpolating** (the naive approach), **Example-Based Super Resolution, Super-resolution from a single image,** and learning a Deep Convolutional Neural Networks for Image Super-Resolution.

NLP Syllabus

Natural Language Processing Track: In the **Natural Language Processing track,** students will start by learning how to build a classifier with one-hot vectorization and linear models, moving towards word **embeddings**, and language models and into layers relevant for NLP, such as **LSTM** and GRU, and how to use **RNNs** for entity extraction. Students will practice **Spacy**, **NLTK**, **Keras**, and **TensorFlow**, and through exercises and other workshops get experience working with a variety of settings for which NLP is useful.

NLP Code

Students will learn about the basic NLP models, such as **n-gram** models, “bag of words”, and word embeddings, and how to apply other ML models to classify text into different categories.

Embeddings

Embeddings have emerged as a powerful tool to analyze language and other high-dimensional discrete data. Students will learn about **word2vec** and how to use it. Exponential family embeddings use exponential families and generalized linear models to extend the main ideas of word2vec to applications beyond the text. In this part of the lecture, students will learn about exponential family distributions and generalized linear models, as well as probabilistic modeling and maximum likelihood estimation.

Entity Recognition, Tokenization, Tagging

Students will learn about Named **Entity recognition, tokenization, POS tagging,** and how to use the **Spacy** and **NTLK** packages for relevant functionality. Students will also learn how to use word representations and NER with RNNs/LSTMs.

Unsupervised Neural Machine Translation

Students will learn about unsupervised cross-lingual embeddings mappings (via a bilingual dictionary), all the way from the basic motivation to Embedding normalization, unsupervised initialization, and robust self-learning. Students will also learn about **Sequence 2 Sequence Learning** and the attention mechanism, **Learning Semantics** (Back Translation), **Learning Structure** ("Denoising" Auto-encoders), and **Adversarial Learning**

4. Workshops and Special Topics

The gist: workshops with companies are meant to expose the students to additional techniques and get a better sense of real-world challenges. They are a chance to learn more about the Data Science industry in general and about different companies in particular.

In the workshops, companies present an important problem that they face, as well as some techniques that they use to solve the problem. Then the companies provide a couple of hands-on exercises where the students will work with real data, apply the techniques that they have learned and get to try their own ideas. The workshops range between half a day to 2 days at campus.

Examples of past workshops:

- Recommender Systems: Outbrain, Taboola
- Autonomous Vehicles: Mobileye
- Cyber: Deep Instinct, Check Point
- Feature selection and engineering: SparkBeyond
- Computer Vision: DataDeg, iCarbonX
- Information Retrieval: Yahoo, Sefaria
- Anomaly Detection: Anodot
- GPUs and DL Solutions: NVIDIA
- Medical Imaging: Zebra medical
- Fin-tech: Istra
- Pre-processing and data modeling: Intel
- Text Classification and Topic Modelling: Chorus

5. Projects hosted by companies

The 5-weeks projects hosted by the companies usually focus on a proof of concept for an idea that the company wanted to check, using data sets the company has or might want to create. They serve as the biggest “hands-on” experience in the course and are an opportunity for the students to demonstrate what they learnt in the course so far and gain additional knowledge and skills. Examples for projects from previous cohorts include the following videos - Aidoc, BMW and can also be found in the ITC project mini-site.