DATA SCIENCE | CURRICULUM

Immersive program covers all the necessary tools and concepts used by data scientists in the industry, including machine learning, statistical inference, and working with data at scale. Students use SQL and NoSQL tools as they advance in the course to build richer predictive models. On graduation, they will a good grasp of contemporary, practical, and relevant tools and techniques and will have built numerous data science applications.

WEEK 1

MODULE ONE
DATA SCIENCE FOUNDATIONS, DATA WRANGLING AND EPLORATORY DATA ANALYSIS

Students will learn to setup the process of Data science through:

- Cleanup of datasets using Python language and Pandas library
- Exploratory data analysis to generate hypotheses and intuition
- Communication of results through visualization, stories, and summaries

FUNDAMENTAL CONCEPTS

- Version control - Fork repository, push & pull code
- Pair programming and Test Driven Development
- Data analysis - types of statistics and analytical methods and their relationship
- Where and how to acquire data, methods for evaluating source data, and data transformation and preparation
- Use Python's Requests package to obtain data from web pages
- Use Python's Beautiful Soup to parse the content of a web page to find useful data for subsequent analysis

EXEMPLARY TECHNIQUES

- Python, Pandas, GitHub, UNIX Bash scripts, SQL
- Optional – coverage of contemporary Web scraping and Data wrangling tools.

PROJECT 1

DIVERGENCE ACADEMY

AMAZON RECOMMENDER

In the first week, students work in small groups using Amazon Reviews dataset to apply the Exploratory Data Analysis, Data Wrangling and basic Feature Engineering concepts to answer a few sentiment analysis questions from the product review data for a product category of student's choice.

WEEK 2 MODULE TWO
STATISTICAL MODELING AND INFERENCE

Students will learn to draw conclusions based on data. Upon completion of this module, students will be able to describe:

- Approaches to performing inference, and acceptance of results
- Concepts in causal inference and motivate the need for experiments
- Statistical tools to help plan experiments: exploratory analysis, power calculations, and the use of simulation
- Statistical methods to estimate causal quantities of interest and construct appropriate confidence intervals
- Scalable methods suitable for "big data", including working with weighted data and clustered bootstrapping.

Students will also be able to:

- Design, plan, implement, and analyze online experiments using contemporary tools
- Implementation of basic "A/B tests", within-subjects designs and sophisticated experiments
- Make and interpret predictions from a Bayesian perspective.
- Understand the Explore-Exploit strategies related to Multi-armed Bandits

FUNDAMENTAL CONCEPTS

- Contexts in which inference is desirable
- Modeling for Inference vs Modeling for Prediction
- Key statistics concepts – Distributions, Sampling, Confidence Intervals, Hypothesis Testing
- Statistical model selection
- Applied Probability for Statistical Inference
- Understand the cycle: model, apply, predict, setup experiments and observe

EXEMPLARY TECHNIQUES

- Python packages - NumPy, SciPy, PyMC
- Optional – coverage of contemporary A/B Testing tools.

PROJECT 2

MULTI-ARMED BANDITS

Multi-armed bandit approach to Internet display advertising to maximize sales; or find the best treatment out of many possible treatments while minimizing losses.

WEEK 3 MODULE THREE
REGRESSION AND CLASSIFICATION

Students will learn to draw conclusions based on data. Upon completion of this module, they will be able to apply:

- Modeling Lifecycle – Specification, Fit, Accuracy, and Reliability.
- Feature Selection - finding "optimal" model parameters based on data
- Linear Regression - Bias-variance Tradeoff
- Logistic Regression including multiclass modeling (Multinomial, Bernoulli, and Gaussian).

Students will also be able to:

- Implement training and testing of datasets
- Implement K-fold and leave-one-out cross-validation approaches
- Understand variances, hetero / homoscedasticity, Multi-collinearity – two or more predictor variables

FUNDAMENTAL CONCEPTS

- Feature Engineering – Selection, Extraction, and Transformation
- Choosing the goal for data mining - Objective function and Loss function.
- Generalization - Fitting and over-fitting and Complexity control.
- Linear regression, Logistic regression, Support-vector machines, and Regularization
- Model Evaluation & Hyper-parameter tuning

EXEMPLARY TECHNIQUES

- Python Package - Scikit-learn
- Optional – coverage of contemporary Machine Learning tools.

PROJECT 3

DIVERGENCE ACADEMY

CITY BIKESHARE SYSTEM FORECAST

Kaggle in Class is a service provided by Kaggle to host competitions as part of class projects. Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated bia network of kiosk locations throughout a city. Students in the class are asked to combine historical patterns with weather data to forecast bike rental demand.

Students will be equipped to:

- Apply visualization of model performance under various kinds of uncertainty; further consideration of what is desired from data mining results using Decision Trees, Random Forests, and Ensembles.
- Implement Natural Language Processing (NLP) processes into projects and software applications.
- Programmatically extract data stored in common formats
- Audit data quality (validity, accuracy, completeness, consistency, and uniformity)
- Critically assess options for cleaning data in different contexts
- Store, retrieve, and analyze data using NoSQL databases

FUNDAMENTAL CONCEPTS

- Using trees for classifications and predictions through Bayesian Classifiers, and Classification and Regression Trees (CART)
- Growing and pruning the tree.
- Use Python's Natural Language Toolkit and TextBlob library to perform natural language analyses on text data
- Algorithms including KD-trees and locality sensitive hashing are learned.
- Understand N-Gram language models of Natural Language Processing. Other topics include Tokenization, Vectorization

EXEMPLARY TECHNIQUES

- Python Package - Scikit-learn, PyMongo, Twitter API, NLTK and TextBlob
- Optional – coverage of contemporary Graphical tools like py2neo for network analysis or Node2XL

PROJECT 4

DIVERGENCE ACADEMY

MID-TERM – HEALTHCARE ANALYTICS

Develop an application that consumes a Logistic Regression and Natural Language Processing based model to determine two classes of labeled twitter data i.e. depressed and not-depressed. Store the tweets in NoSQL database and plot data on a map.

WEEK 5

MODULE FIVE
UNSUPERVISED LEARNING: CLUSTERING AND DIMENSION REDUCTION

Students will learn to apply integrated supervised and unsupervised Methods, such as:

- Feature selection – Filtering and wrapping algorithms, and Tradeoffs – speed, relevance, and usefulness
- Unsupervised methods in predictive analytics
- Unsupervised methods used in network and text analytics
- Dimension reduction of predictor space
- Predictive models on subsets of homogeneous records
- Graphing analysis algorithms for clustering (community detection in graph networks)

FUNDAMENTAL CONCEPTS

- Cluster Analysis – basic clustering problem, k-means clustering, k-means in Euclidean space, and k-means as optimization
- Feature transformation - Principal Components Analysis, Independent Components Analysis, Cocktail Party Problem
- Dimension reduction techniques – Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF)
- Detection in graph networks – Breadth-first search (BFS), Depth-first search (DFS), A* search (based of Dijkstra)

EXEMPLARY TECHNIQUES

- Python Package - Scikit-learn
- Optional – coverage of contemporary Machine Learning tools in Clustering, Decision Trees, and Graphical visualization.

PROJECT 5

DIVERGENCE ACADEMY

FOREST COVER TYPE CLASSIFICATION

Kaggle in Class is a service provided by Kaggle to host competitions as part of class projects. Students are asked to predict forest cover type from cartographic variables. The data is in raw form (not scaled) and contains binary columns of data for qualitative independent variables such as wilderness areas and soil types.

Students will learn to use the Big data infrastructure to preprocess and consume large datasets for Machine Learning models. This will include learning to:

- Leverage Hadoop ecosystem for Pre-processing, Exploratory Data Analysis and Predictive Modeling
- Program Mappers, Reducers and jobs using Hive, SQOOP, and PIG scripting.
- Hadoop data workflows and jobs with Python
- Read and write data to HDFS
- Apply the next generation framework i.e. Spark (in-memory), for Filtering, Aggregating and Searching

FUNDAMENTAL CONCEPTS

- Use Hadoop via Python bindings to write customized map-reduce jobs from scratch and run in Hadoop cloud environment
- Understand the distributed computing environment.
- Hadoop Anatomy: HDFS, Name nodes, Job trackers, Data nodes

EXEMPLARY TECHNIQUES

- Python Packages - Scikit-learn, Pig, Hive, Sqoop, Spark - SQL, MLLib, GraphX, Clusters on Amazon Web Services (AWS) and/or Azure, Mrjob, Pydoop
- Optional – coverage of contemporary Machine Learning tools built on top of Hadoop infrastructure.

PROJECT 6

SPARK WITH CRAIGSLIST

Build a model that classifies the unstructured text data of a job title to a given job category.

WEEK 7

MODULE SEVEN
DEEP LEARNING & DATA VISUALIZATION

Students will learn to apply deep learning approaches and draw conclusions based on data. Upon completion of this module, students will be able to:

- Describe loading and saving models to plot intermediate results for supervised optimization models for Deep learning.
- Feed-forward neural net trained with backpropagation.
- Conduct unsupervised learning, applying deep belief network and restricted Boltzmann machine models.

FUNDAMENTAL CONCEPTS

- Supervised optimization for Deep learning
- Learning a classifier – Zero-One loss, Negative Log-likelihood loss, Stochastic Gradient Descent (SGD)
- Regularization – L1 and L2, Early stopping.
- Unsupervised learning – generative modeling.

EXEMPLARY TECHNIQUES

- Python packages – Python Image Library, MATPLOTLIB, seaborn, plearn and NumPy.
- Optional – coverage of contemporary ML tools that generate Deep learning based models.

PROJECT 7

DIVERGENCE ACADEMY

DEEP LEARNING AT THE GROCERY STORE

Build an application that provides information on a packaged food product based on an image taken with a smartphone. Steps include finding similar foods, extracting features of images using Deep Learning model, and querying the catalog using nearest neighbor model.

WEEK 8 MODULE EIGHT

RECOMMENDATION SYSTEMS & TIME-SERIES FORECASTING

Students will develop recommender systems to help people find products, information and even other people. Upon completion of this module, students will be able to understand the various real-world handoffs between Business Analysts, Data Scientists and Data Engineering Teams. Additionally, students will be able to:

- Combine conceptual understanding and practical implementation of recommenders
- Implement basic recommenders from scratch
- Use software libraries and tools to implement more advanced recommenders
- Develop REST API for predictive models
- Deploy models into production using various methods including Predictive Modeling Markup Language (PMML)
- Develop web applications that consume predictive models
- Understand Platform-as-a-service offerings to deploy web applications
- Review additional uses cases such as Anomaly Detection and Customer Churn

FUNDAMENTAL CONCEPTS

- Nearest distance algorithms – Manhattan, Euclidean, Minkowski, Pearson correlation coefficient, Cosine similarity, and k-nearest neighbors
- Time-series for forecasting application – trend and seasonality

EXEMPLARY TECHNIQUES

- Python packages –NumPy, SciPy, Scikit-learn, Pandas
- Optional – coverage of contemporary ML tools that serialize models, and automate deployment of models to Cloud platforms.

PROJECT 8

DIVERGENCE ACADEMY

BEER RECOMMENDER

Use the data from Beer Advocate to recommend users other varieties of beers which are graded on appearance, aroma, palate, and taste plus users “overall” grade. Use the nearest distance algorithms to model the recommender, and deploy the model to a cloud platform.

WEEK 9-12

MODULE NINE
FINAL PROJECT

Students integrate Data Science skills through an application to a project focusing on real-world open data. The course serves as the capstone of the student's 8-weeks of learning. The student works alone with support from staff to tailor the data science process steps to develop a minimum viable data product within two weeks. The students are evaluated on their problem hypothesis, statistical model, insights delivered through use of the model, flexibility of the model including bias and variance, and communication of the end-to-end approach through an oral presentation.

FUNDAMENTAL CONCEPTS

- Use the design process to isolate an appropriate problem to solve
- Evaluate the computational feasibility of the problem
- Choose data sources that can be used to address the problem
- Design and implement an appropriate computational architecture
- Design and implement an appropriate set of analysis steps
- Design and develop a data visualization to clearly convey the results of the analysis to a layperson
- Assemble final portfolio and present project at Career Day

MORE ABOUT PROJECTS

Data science projects at Divergence Academy are focused on developing and deploying predictive models in production. While the topics in the class cover statistical modeling for explanation, the intent is to have students be ready for real-world application where they are constantly making trade-off decisions. The immersive program considers the tradeoffs as dimensions of business domain, design, data, algorithms, tools, and communication. Each module covers certain content from several dimensions, which are reinforced in that module's project.

The rigor with which the program drives the topics covered in the immersive program allow us to sleep soundly at night. We are confident that our graduates haven't just learned the tools and techniques that the data scientists use but by the time they leave the classroom, our graduates are data scientists. They are ready to approach the problem space in their new careers and assemble the suite of tools and methods to answer insightful questions and communicate comprehensible results. They are competent, capable, confident, and ready to work.

CAPSTONE

PASSION PROJECT

Students are free to use anything covered in the class or learn something new to answer specific question that they want to address. The goal here is to deliver a Data Product. Every student works intensely to create something cool, interesting, useful or worthwhile.