Data Scientist Nanodegree Syllabus

Before You Start

Prerequisites: The Data Scientist Nanodegree program is an advanced program designed to prepare you for data scientist jobs. As such, you should have a high comfort level with a variety of topics before starting the program. In order to successfully complete this program, we strongly recommend that the following prerequisites are fulfilled. If you do not have the necessary prerequisites, Udacity has courses and programs that prepare you for this Nanodegree program.

- Programming:
    - Python Programming: Writing functions, logic, control flow, and building basic applications, as well as common data analysis libraries like NumPy and pandas
    - SQL programming: Querying databases using joins, aggregations, and subqueries
    - Comfortable with using the Terminal, version control in Git, and using GitHub
- Probability and Statistics
    - Descriptive Statistics: Calculating measures of center and spread, estimation distributions
    - Inferential Statistics: Sampling distributions, hypothesis testing
    - Probability: Probability theory, conditional probability
- Mathematics
    - Calculus: Maximizing and minimizing algebraic equations
    - Linear Algebra: Matrix manipulation and multiplication
- Data wrangling
    - Accessing database, CSV, and JSON data
    - Data cleaning and transformations using pandas and Sklearn
- Data visualization with matplotlib
    - Exploratory data analysis and visualization
    - Explanatory data visualizations and dashboards
- Machine Learning
    - Feature Engineering
    - Supervised Learning: Regression, classification, decision trees, random forest
    - Unsupervised Learning: PCA, Clustering

The following programs can prepare you to take this nanodegree program. There are also several free courses that you can use to prepare.

- Programming for Data Science with Python
- Data Analyst Nanodegree Program
- Intro to Machine Learning Nanodegree Program

Educational Objectives: The ultimate goal of the Data Scientist Nanodegree program is for you to learn the skills you need to perform well as a data scientist. As a graduate of this program, you will be able to:

- Use Python and SQL to access and analyze data from several different data sources.

- Use principles of statistics and probability to design and execute A/B tests and recommendation engines to assist businesses in making data-automated decisions.
- Deploy a data science solution to a basic flask app.
- Manipulate and analyze distributed datasets using Apache Spark.
- Communicate results effectively to stakeholders.

Estimated Length of Program: 4 months
Program Structure: Self-paced
Textbooks required: None
Textbooks optional: Elements of Statistical Learning, Machine Learning: A Probabilistic Perspective, Python Machine Learning
Instructional Tools Available: Video lectures, mentor-led student community, forums, project reviews

Syllabus

Project 1: Write a Data Science Blog Post

In this project, you will choose a dataset, identify three questions, and analyze the data to find answers to these questions. You will create a GitHub repository with your project, and write a blog post to communicate your findings to the appropriate audience. This project will help you reinforce and extend your knowledge of machine learning, data visualization, and communication.

Supporting Lessons: Solving Problems with Data Science

Supporting Lessons	Learning Outcomes
THE DATA SCIENCE PROCESS	➔ Apply the CRISP-DM process to business applications ➔ Wrangle, explore, and analyze a dataset ➔ Apply machine learning for prediction ➔ Apply statistics for descriptive and inferential understanding ➔ Draw conclusions that motivate others to act on your results
COMMUNICATING WITH STAKEHOLDERS	➔ Implement best practices in sharing your code and written summaries ➔ Learn what makes a great data science blog ➔ Learn how to create your ideas with the data science community

Project 2: Build Pipelines to Classify Messages with Figure Eight

Figure Eight (formerly Crowdflower) crowdsourced the tagging and translation of messages to apply artificial intelligence to disaster response relief. In this project, you'll build a data pipeline to prepare the message data from major natural disasters around the world. You'll build a machine learning pipeline to categorize emergency text messages based on the need communicated by the sender.

Supporting Lessons: Software Engineering for Data Scientists

Supporting Lessons	Learning Outcomes
SOFTWARE ENGINEERING PRACTICES	➔ Write clean, modular, and well-documented code ➔ Refactor code for efficiency ➔ Create unit tests to test programs ➔ Write useful programs in multiple scripts ➔ Track actions and results of processes with logging ➔ Conduct and receive code reviews
OBJECT ORIENTED PROGRAMMING	➔ Understand when to use object oriented programming ➔ Build and use classes ➔ Understand magic methods ➔ Write programs that include multiple classes, and follow good code structure ➔ Learn how large, modular Python packages, such as pandas and scikit-learn, use object oriented programming ➔ *Portfolio Exercise*: Build your own Python package
WEB DEVELOPMENT	➔ Learn about the components of a web app ➔ Build a web application that uses Flask, Plotly, and the Bootstrap framework ➔ *Portfolio Exercise*: Build a data dashboard using a dataset of your choice and deploy it to a web application

Supporting Lessons: Data Engineering for Data Scientists

Supporting Lessons	Learning Outcomes
ETL PIPELINES	➔ Understand what ETL pipelines are ➔ Access and combine data from CSV, JSON, logs, APIs, and databases ➔ Standardize encodings and columns ➔ Normalize data and create dummy variables ➔ Handle outliers, missing values, and duplicated data ➔ Engineer new features by running calculations ➔ Build a SQLite database to store cleaned data
NATURAL LANGUAGE PROCESSING	➔ Prepare text data for analysis with tokenization, lemmatization, and removing stop words ➔ Use scikit-learn to transform and vectorize text data ➔ Build features with bag of words and tf-idf ➔ Extract features with tools such as named entity recognition and part of speech tagging ➔ Build an NLP model to perform sentiment analysis
MACHINE LEARNING PIPELINES	➔ Understand the advantages of using machine learning pipelines to streamline the data preparation and modeling process ➔ Chain data transformations and an estimator with scikit-learn's Pipeline ➔ Use feature unions to perform steps in parallel and create more complex workflows

	➔ Grid search over pipeline to optimize parameters for entire workflow ➔ Complete a case study to build a full machine learning pipeline that prepares data and creates a model for a dataset

Project 3: Design a Recommendation Engine with IBM

IBM has an online data science community where members can post tutorials, notebooks, articles, and datasets. In this project, you will build a recommendation engine, based on user behavior and social network in IBM Watson Studio's data platform, to surface content most likely to be relevant to a user.

Supporting Lessons: Experiment Design

Supporting Lessons	Learning Outcomes
EXPERIMENT DESIGN	➔ Understand how to set up an experiment, and the ideas associated with experiments vs. observational studies ➔ Defining control and test conditions ➔ Choosing control and testing groups
STATISTICAL CONCERNS OF EXPERIMENTATION	➔ Applications of statistics in the real world ➔ Establishing key metrics ➔ SMART experiments: Specific, Measurable, Actionable, Realistic, Timely
A/B TESTING	➔ How it works and its limitations ➔ Sources of Bias: Novelty and Recency Effects ➔ Multiple Comparison Techniques (FDR, Bonferroni, Tukey) ➔ *Portfolio Exercise*: Using a technical screener from *Starbucks* to analyze the results of an experiment and write up your findings

Supporting Lessons: Recommendations

Supporting Lessons	Learning Outcomes
INTRODUCTION TO RECOMMENDATION ENGINES	➔ Distinguish between common techniques for creating recommendation engines including knowledge based, content based, and collaborative filtering based methods. ➔ Implement each of these techniques in python. ➔ List business goals associated with recommendation engines, and be able to recognize which of these goals are most easily met with existing recommendation techniques.
MATRIX FACTORIZATION FOR RECOMMENDATIONS	➔ Understand the pitfalls of traditional methods and pitfalls of measuring the influence of recommendation engines under traditional regression and classification techniques. ➔ Create recommendation engines using matrix factorization and FunkSVD ➔ Interpret the results of matrix factorization to better understand latent features of customer data ➔ Determine common pitfalls of recommendation engines like the cold start problem and difficulties associated with usual tactics for assessing the

	effectiveness of recommendation engines using usual techniques, and potential solutions.

Project 4: Data Science Capstone Project

In this capstone project, you will leverage what you've learned throughout the program to build a data science project of your choosing. You will define the problem you want to solve, identify and explore the data, then perform your analyses and develop a set of conclusions. You will present the analysis and your conclusions in a blog post and GitHub repository. This project will serve as a demonstration of your ability as a data scientist, and will be an important component of your job-ready portfolio.

Supporting Lessons: Data Science Projects

Supporting Lessons	Learning Outcomes
ELECTIVE 1: DOG BREED CLASSIFICATION	➔ Use convolutional neural networks to classify different dogs according to their breeds ➔ Deploy your model to allow others to upload images of their dogs and send them back the corresponding breeds ➔ Complete one of the most popular projects in Udacity history, and show the world how you can use your deep learning skills to entertain an audience!
ELECTIVE 2: STARBUCKS	➔ Use purchasing habits to arrive at discount measures to obtain and retain customers. ➔ Identify groups of individuals that are most likely to be responsive to rebates.
ELECTIVE 3: ARVATO FINANCIAL SERVICES	➔ Work through a real-world dataset and challenge provided by Arvato Financial Services, a Bertelsmann company ➔ Top performers have a chance at an interview with Arvato or another Bertelsmann company!
ELECTIVE 4: SPARK FOR BIG DATA	➔ Take a course on Apache Spark and complete a project using a massive, distributed dataset to predict customer churn ➔ Learn to deploy your Spark cluster on either AWS or IBM Cloud
ELECTIVE 5: YOUR CHOICE	➔ Use your skills to tackle any other project of your choice