General Assembly
Course Curriculum

DATA SCIENCE

GA.CO/DS

Data Science
Table of Contents

OVERVIEW

THE FRAMEWORK

Ever wonder how the NetFlix recommendation engine works or how Amazon.com determines what items "you may also like?" All of these things are driven by training a computer how to learn using the large amounts of data that exist in these systems.

The 11-week data science course is a practical introduction to the interdisciplinary field of data science and machine learning which is at the intersection of computer science, statistics, and business. You will learn to use the programming languages, tools, and technologies, such as Python to help you acquire, clean, parse, and filter your data. A significant portion of the course will be a hands-on approach to the fundamental modeling techniques and machine learning algorithms that enable you to build robust predictive models of real-world data and test their validity. You will also gain practice communicating your results and insights about how to build systems that are more intelligent and take advantage of the data that you have (think recommendations systems or targeted ads). By the end of the course, students will be able to:

- Acquire, clean, and parse large sets of data using Python
- Choose the appropriate modeling technique to apply to your data
- Programmatically create predictive data models using machine learning techniques
- Apply probability and statistics concepts to create and validate predictions about your data
- Communicate your results to an appropriate audience

STUDENTS

SOFTWARE DEVELOPERS OR THOSE WITH SOME PROGRAMMING EXPERIENCE

This course provides individuals with programming experience the skills required to solve problems using computation that involve large data sets such as predicting user behavior on their website, making decisions, or the best way to classify content. Individuals learn how to apply the code necessary to be able to make predictions and create models.

PROJECTS

FINAL PROJECT

For the Data Science final project, you will work individually and asked to address a data-related problem in your professional field or in a field you interested in. Picking a subject you are passionate about is important as it will make it more fun and help you produce a better project! You will be asked to acquire a real-world data set, form a hypothesis about it, clean, parse, and apply modeling techniques and data analysis principles to ultimately create a predictive model. Students present their results and each write a technical paper that includes the following:

- Clearly articulated problem statement
- Summary of data acquisition, cleaning, and parsing stage
- Qualitative and quantitative market research
- Clear presentation of your predictive model and the processes you took to create it
- Presentation style appropriate to the audience

Your instructional team will help you scope out your project so that you choose something that is feasible to accomplish given the skills you acquire in the course. You will alternatively have the option of participating in a Kaggle competition, online competitions on data prediction, which will require an application of a similar skillset.

UNITS

UNIT 1: THE BASICS	Introduction to Data Exploration	Lesson 1
	Introduction to Machine Learning	Lesson 2
UNIT 2: FUNDAMENTAL MODELING TECHNIQUES	K-Nearest Neighbors Classification	Lesson 3
	Naive Bayes Classification	Lesson 4
	Regression and Regularization	Lesson 5
	Logistic Regression	Lesson 6
	K-Means Clustering I	Lesson 7
UNIT 3: FURTHERING MODELING TECHNIQUES	K-Means Clustering II	Lesson 8
	Ensemble Techniques	Lesson 9
	Decision Trees and Random Forests	Lesson 10
	Support Vector Machines	Lesson 11
	Dimensionality Reduction	Lesson 12
	Recommendation Systems	Lesson 13
UNIT 4: OTHER TOOLS	Database Technologies	Lesson 14
	Network Analysis	Lesson 15
	Map-Reduce	Lesson 16
	Final Project Working Session	Lesson 17
	Final Project Working Session	Lesson 18
	Where To Go Next	Lesson 19
	Final Project Working Session	Lesson 20
	Final Project Presentations	Lesson 21
	Final Project Presentations	Lesson 22

1 THE BASICS

1 INTRODUCTION TO DATA EXPLORATION

- Describe the data mining workflow and the key traits of a successful data scientist.
- Extract, format, and preprocess data using UNIX command-line tools
- Explore & visualize data

2 INTRODUCTION TO MACHINE LEARNING

- Explain the concepts and applications of supervised and unsupervised learning techniques
- Describe categorical and continuous feature spaces, including examples and techniques for each
- Discuss the purpose of machine learning and the interpretation of predictive modeling results

2 FUNDAMENTAL MODELING TECHNIQUES

3 K-NEAREST NEIGHBORS CLASSIFICATION

- Describe the setting and goal of a classification task
- Minimize prediction error using training and test sets, optimize predictive performance using cross-validation
- Understand the kNN classification algorithm, its intuition, and implementation
- Implement the "hello world" of machine learning (kNN classification of iris dataset)

4 NAIVE BAYES CLASSIFICATION

- Outline the basic principles of probability, including conditional probability and Bayes' theorem
- Describe inference in the Bayesian setting, including the prior and posterior distributions and the likelihood function
- Understand the naive Bayes classifier and its assumptions
- Implement a spam filter using the naive Bayes technique

5 REGRESSION AND REGULARIZATION

- Explain the concepts of regression models, including their assumptions and applications
- Discuss the motivation and use for regularization techniques and their use
- Implement a regularized fit

2 FUNDAMENTAL MODELING TECHNIQUES (CONTINUED)

6 LOGISTIC REGRESSION

- Describe the applications of logistic regression to classification problems and probability estimation
- Introduce the concepts underlying logistic regression, including its relation to other regression models.
- Predict the probability of a user action on a website using logistic regression

7 K-MEANS CLUSTERING I

- Explain the purpose of exploratory data analysis, its applications in continuous and categorical feature spaces, and the interpretation and use of clustering results
- Discuss the importance of the distance function in cluster formation, as well as the importance of scale normalization
- Implement a k-means clustering algorithm

3 MARKETING ANALYTICS SITE AND CONTENT

8 ENSEMBLE TECHNIQUES

- Describe general ensemble techniques such as bagging and boosting
- Build an enhanced classification algorithm using AdaBoost

9 DECISION TREES AND RANDOM FORESTS

- Describe the use and construction of decision trees for classification tasks
- Create a random forest model for ensemble classification

10 DIMENSIONALITY REDUCTION

- Explain the practical and conceptual difficulties in working with very high-dimensional data
- Understand the application and use of dimensionality reduction techniques
- Draw inferences from high-dimensional datasets using principal components analysis

11 RECOMMENDATION SYSTEMS

- Explain the use of recommendation systems, and discuss several familiar examples
- Understand the underlying concepts, including collaborative & content-based filtering
- Implement a recommendation system

4 OTHER TOOLS

12 DATABASE TECHNOLOGIES

- Introduce concepts and use of relational databases, alternative database technologies such as NoSQL, and popular examples of each

13 NETWORK ANALYSIS

- Describe the use of graphs and graph theory to analyze problems in network analysis
- Explore network visualization

14 MAP-REDUCE

- Describe the use of graphs and graph theory to analyze problems in network analysis.
- Introduce the map-reduce framework and popular implementations including Hadoop
- Implement and explore examples of map-reduce tasks

15/16 FINAL PROJECT WORKING SESSION

- In-class working session

17 WHERE TO GO NEXT

- Review of concepts and examples from preceding weeks
- Discussion of resources & tools for further study

18 FINAL PROJECT WORKING SESSION

- In-class working session

19/20 FINAL PROJECT PRESENTATIONS

- Final project presentation and discussion

FAQS

WHY IS THIS COURSE RELEVANT TODAY?

Given the prevalence of technologies and the amount of data available in the online world about users, products, and the content that we generate, businesses could be making so much more well-informed decisions if this vast amount of data was more deeply analyzed through the use of data science. The data science course provides the tools, methods, and practical experience to enable you to make accurate predictions about data, which ultimately leads to better decision-making in business, and the use of smarter technology (think recommendation systems or targeted ads).

WHAT PRACTICAL SKILL SETS CAN I EXPECT TO HAVE UPON COMPLETION OF THE COURSE?

This course provides you with technical skills in machine learning, algorithms, and data modeling which allow you to make accurate predictions about your data. You'll create your models using Python and gain a good grasp of the programming language. Furthermore, you will learn how to parse and clean your data which can take up to 70% of your time as a data scientist.

WHO WILL I BE SITTING NEXT TO IN THIS COURSE?

Individuals who have a strong interest in manipulating large data sets, finding patterns in data, and making predictions. Software developers who want to solve problems that involve large data sets, such as predicting user behavior on their website, making decisions, or the best way to classify content. Individuals with a good grasp of programming, a solid knowledge of statistics and probability but missing the intersection of them both. Prerequisites:

- Good grasp of college-level statistics and probability
- Ability to program in a scripting language such as Python or R

FAQS

WHAT CAN I EXPECT BY THE END OF THE COURSE?

By the end of the course, you can expect to be able to acquire, parse, clean, and apply various modeling techniques to your data to make predictions. You should also be able to communicate your findings to both a non-technical and technical audience in both written and verbal formats.

WILL THERE BE ANY PRE-WORK?

Yes. You will be required to complete approximately 8 - 10 hours of pre-work. This includes some tutorials and practice using the command line.

SHOULD I COME EQUIPPED WITH ANYTHING?

Yes. Please come prepared with a laptop (Mac OSX is preferred but not required).