Data Science Course

Background

The world is quickly being transformed by data-driven insights and prediction algorithms. The basic skills required for data analytics on and off the cloud, statistical modeling, and proficiency with a complex ecosystem of tools and platforms – span many fields and are not easy to obtain through conventional learning. Come and explore the basic techniques of data science, and algorithms for data mining (e.g., clustering), and basic statistical modeling (e.g., linear and non-linear regression). We beleive your analysis workflow should be interactive and achievable using free and Open Source software.

This course will enable you and/or your teams to become more productive in their analysis workflow and produce visual interactive insights.

All the tools, platforms and software that we use during this course is Open Source and will run on Windows, Linux and Mac. You will even get your own cloud server to interact with during the course.

Note

This is not a programming course but programming knowledge will surely be to your advantage. The course will focus mostly on Python and R. You will however see very little of Excel!

We are working on a R and Python spesific programming course. Please be in touch for more detail.

Duration:

5 days

Format:

Instructor lead class with 50% lecture and 50% hands-on labs. Students can bring their own laptops and require either Chrome browser 8.0 or better or Firefox 3.6 or later.

Course outline

1. Setting up the Data Scientist's Toolbox and Environment
2. Programming (Python,R)
3. Getting, Storing and Cleaning Data
4. Exploratory Data Analysis
5. Reproducible Research
6. Statistical Inference
7. Regression Models
8. Practical Machine Learning
9. Developing and Publishing Data Products

Setting up the Data Scientist's Toolbox and Environment

Upon completion you will be able to identify and classify data science problems. You will also have created your Github account, created your first repository, and pushed your first markdown file to your account.

Basic Programming

Overview of data types and objects, reading and writing data,Control structures, functions, scoping rules, dates and times,Loop functions, debugging tools, simulation, code profiling

Getting, Storing and Cleaning Data

Upon completion you will be able to obtain data from a variety of sources. You

will know the principles of tidy data and data sharing. Finally, you will understand and be able to apply the basic tools for data cleaning and manipulation.

Exploratory Data Analysis

After completing this you will be able to make visual representations of data using the plotting systems, apply basic principles of data graphics to create rich analytic graphics from different types of datasets, construct exploratory summaries of data in support of a specific question, and create visualizations of multidimensional data using exploratory multivariate statistical techniques.

Reproducible Research

In this course you will learn to write a document using R markdown or the IPython notebook, integrate live code into a literate statistical program, compile markdown documents using, and organize a data analysis so that it is reproducible and accessible to others.

Statistical Inference

In this section you will learn the fundamentals of statistical inference. Students will receive a broad overview of the goals, assumptions and modes of performing statistical inference. Students will be able to perform inferential tasks in highly targeted settings and will be able to use the skills developed as a roadmap for more complex inferential challenges.

Regression Models

You will learn how to fit regression models, how to interpret coefficients, how to investigate residuals and variability. Students will further learn special cases of regression models including use of dummy variables and multivariable adjustment. Extensions to generalized linear models, especially considering Poisson and logistic regression will be reviewed.

Practical Machine Learning

Upon completion of this module you will understand the components of a machine learning algorithm. You will also know how to apply multiple basic machine learning tools. You will also learn to apply these tools to build and evaluate predictors on real data.

Developing and Publishing Data Products

Students will learn how communicate using statistics and statistical products. Emphasis will be paid to communicating uncertainty in statistical results. Students will learn how to create simple web applications of their analysis.

Sofware

Install the free software stack from insightStack

- Anaconda Python Distribution
- Orange machine learning environmet
- Each learner will also get a personal Virtual Private Server to use during the duration of the course and will be active for at least one month after the course.

Course price:

TBA per student (ex vat). (Please request a quote for the latest price.) This includes lunch, tea and course material.

Other resources

Online courses

- Course ideas - coursera

- Udacity - Exploratory Data Analysis
- Udacity - Statistics

Interesting problems and examples

- Classification - Predicting Titanic survival
- Classification - Digits Recongniser
- Recomender - Beer recomender
- Customer Churn - Predicting Customer Churn

Books

- Comprehensive list of books - Forked from Github

Free and Open Source Tools

(adapted from this post) * Anaconda * Orange * Weka * knime