Bootcamp training in programming, statistics and data science
June 2021

1 Introduction

This is a training designed for Accointing Guatemala that covers topics in computer programming, math and statistics, data science and machine learning. It includes the explanation of the basic theory necessary to implement simple machine learning models, as well as instruction in how to handle Python programming in most common environments, that covers various popular modules and libraries in data science.

Course:	Data Science Bootcamp
Main topics:	Mathematics for data science Descriptive and inference statistics Python programming Basic machine learning models
Time:	Four weeks, 5 days per week, 2 hour each session for a total of 40 hours of session work.
Instructor:	Alan Reyes-Figueroa, Ph.D. Professor of Mathematics and data science at Universidad del Valle, Data scientist at Bam.

2 Skills to Develop

- Learn what data science is, the various activities of a data scientist's job, and methodology to work as a data scientist.
- Develop hands-on skills using the theory and tools, languages, and Python libraries used by professional data scientists.
- Import and clean data sets, analyze and visualise data, and build and evaluate machine learning models and pipelines using Python.
- Apply various data science skills, techniques, and tools to complete a project and publish a report.

3 Syllabus

1. Installation and set up of Python working environments for data science: Anaconda Python, Jupyter notebooks and Jupyter-lab, VSCode and Jupyter built in.
2. Libraries and Python modules for data science: Numpy, Pandas, Matplotlib, Seaborn, Plotly, Statsmodels, Scikit-learn. SQL basic queries and modules for working with SQL in Python.
3. Math and statistical concepts: Vectors and matrix calculus and linear algebra, plotting functions, basic probability, distributions, descriptive statistics, inference statistics, hypothesis testing.
4. Python programming: Variables and functions, conditionals, cycles *for*, *while* and exceptions. Basic structures: lists, tuples, dictionaries, strings. Vectors and matrices.
5. Data science: Read and modify datasets with Pandas. Deal with missing information. Data exploration: histograms, covariance analysis, visualisation and plots. Principal components analysis.

6. Machine learning models: Scikit-learn, K-means and other clustering algorithms, K-nn nearest neighbours, logistic regression, Support vector machines. Evaluation metrics and cross-validation. Linear regression with Statsmodels. Other models such as trees and random forests.

4 References

- P. Bruce, H. Bruce (2020). *Practical Statistics for Data Scientists*. O'Reilly.
- A. Martelli, A. Ravenscroft, D. Ascher (2005). *Python CookBook*. O'Reilly.
- M. Harrison, T. Petrou (2020). *Pandas CookBook*. Packt.
- C. Bishop (2000). *Pattern Recognition and Machine Learning*. Springer
- T. Hastie, R. Tibshirani, J. Friedman (2013). *The Elements of Statistical Learning*. Springer.

5 Proposed Calendar

	Content		Content		Content		Content		Content
1	Installation and Setup	2	Python crash course I	3	Python crash course II	4	Numpy	5	Visualisation
	Anaconda Jupyter Manage environments VSCode + Jupyter		Variables, data types Conditionals, for, range While, break, exception Functions		List, comprehension Dictionaries, tuples Strings Files I/O		Vectors and matrices Tensors Operations, reshape Basic plots		Matplotlib Seaborn Pandas built-in Plotly
6	Pandas I	7	Pandas II	8	SQL	9	SQL in Python	10	Probability
	Dataframes Access cells Select subdataframes Combine and merge		Filter commands Handle missing data Dummy variables Handle time and dates		Basic queries Aggregate functions Group by, order by Join and merge tables		SQLite MySQL		Probability basics Bayes' law Discrete distributions Continuous distrib.
11	Descriptive statistics	12	Inference statistics	13	Hypothesis testing	14	Data exploration	15	Principal components
	Histograms Mean, median, modes Variance Covariance, correlation		Confidence intervals Bootstrap, sampling Visualisation tools QQ-plots, densities		Hypothesis tests: to test normality compare distributions to compare samples		Distributions Covariance analysis Stats and summaries Visualisation		PCA Dimension reduction Data exploration Biplots
16	Clustering	17	Regression I	18	Regression II	19	Classification I	20	Classification II
	K-means t-SNE Other methods		Linear regression Residual analysis Categorical variables Predictions		Hypothesis testing What if assumptions do not hold? Ridge and LASSO		K-nn Naive Bayes Logistic regression Train and test		SVM Decision trees Random forests Cross-validation