CO3093 Big Data and Predictive Analytics

Credits: *20* **Convenor:** *Dr. E. Tadjouddine* **Semester:** *2nd*

Prerequisites:	*Essential: CO1008* *Desirable: CO2015, CO3091*				
Lectures:	*20*	*hours*			
Surgeries:	*10*	*hours*	**Independent Study:**	*100*	*hours*
Laboratories:	*20*	*hours*			
Assessment:	*Coursework: 60% + Two hours exam in May/June: 40%*				

Learning Outcomes **Students should be able to:**

- analyse possibly large amount of data;
- develop and back-test a predictive model;
- compare and contrast different types of predictive models;
- evaluate a predictive model;
- use a Map-Reduce approach in processing data.
- write a report on the data analysis carried out.

Explanation of Prerequisites

In addition to the pre-recquisite modules above, the basics of calculus will be helpful. The ability to program in Python is desirable but this will be part of the Lab sessions.

Module Description

As we increasingly rely upon the online environment for our daily routines, we leave behind a vast amount of information about us. Commercial and public organisations can use this information to predict our behavior. This module aims to study methods and tools enabling us to identify variables of interest and their relationships from an existing data set in order to develop a statistical model that can predict values of variables of interest. This kind of analysis should give us an insight into individual preferences, and most importantly, what someone is likely to do in a given scenario. Some of the applications include credit bank approval, marketing, stock price predictions, demand forecasting or polical campaining.

In this course, we will also study the importance of good quality data and will rely upon open libraries such as scikit-learn (http://scikit-learn.org/stable/) to implement basic models with much less programming effort. We will also learn how to compare and contrast different models for the same data and objective. As a predictive analysis does not necessarily demand a huge amount of data, we will also discuss the utility or misfortune of the so-called big data and how to process such a large amount of data efficiently by using a distributed approach as in the Apache Spark.

Syllabus

Python: the basics and relevant libraries such as numpy, scipy, and pandas and its visualisation tools will be reviewed typically during the first three Lab sessions.

Big data: its philosophy (seek, store, analyse, and act); structured and unstructured data; data sources, volume, value, and timeliness; and data cleaning and manipulation.

Basic probability and statistics: random variables, mean, variance, standard deviation; relationship between variables (correlation and regression); probability distributions including Poisson and Gaussian; sampling; hypothesis testing, confidence intervals and significance tests.

Using predictive models: setting up clear objectives, identifying predictor data and outcome data, and the decision making process. We will use the example of a credit scoring model.

Types of predictive models: regression and classification models including linear models, logistic regression, random forests, clustering, and finding similar items. These algorithms will be presented in the form of black-box and be used from within a software package, e.g., http://scikit-learn.org/stable/

Building up a predictive model: data collection, sampling, and the iterative process (understanding the data, modeling, and performance evaluation). We will run through an example e.g., stock price movements or marketing.

Processing big data: The Apache Spark tool. We will introduce the basics of how to efficiently process large data sets.

Reading List

[B] David M. Levine and David F. Stephan, *Even You Can Learn Statistics and Analytics: An Easy to Understand Guide to Statistics and Analytics (Third Edition);*
ISBN: 978-0133382662, Pearson FT Press.

[B] Steven Finlay, *Predictive Analytics, Data Mining and Big Data (Business in the Digital Economy);*
ISBN: 9781137379276, Palgrave Macmillan.

[B] Philipp K. Janert, *Data Analysis With Open Source Tools;*
ISBN: 9787564126742, O'Reilly Media.

[B] Ashish Kumar, *Learning Predictive Analytics with Python;*
ISBN 978-1783983261, PACKT Publishing.

[B] P.N. Tan and M. Steinbach and V. Kumar, *Introduction to Data Mining;*
ISBN: 9787111316701, Pearson.

[B] Anand Rajaraman and Jeffrey D. Ullman, *Mining of Massive Datasets;*
ISBN: 9781207015357, Cambridge University Press.

[B] Viktor Mayer-Schonberger and Kenneth Cukier, *Big data: A revolution that will transform how we live, work and think;*
ISBN: 9781848547926, John Murray.

[B] Frank J. Ohlhorst, *Turning Big Data into Big Money;*
ISBN: 9781118147597, Wiley.

[B] John M. Chambers, *Software for Data Analysis: Programming with R (Statistics and Computing);*
ISBN: 9781441926128, Springer.

[B] Kevin Sheppard, *Introduction to Python for Econometrics, Statistics and Data Analysis (free eBook, August 2014);*
ISBN: N/A, KevinSheppard.com.

[B] Charles Severance, *Python for Informatics: Exploring Information (free eBook, May 2014);*
ISBN: N/A, pythonlearn.com.

Convenor's Notes

There are two assignements and one final exam.

CW1 Assignment, 30% module mark

CW2 Assignment, 30% module mark