CE / CZ 4073

Data Science for Business
Semester 2 | 2017-2018

Instructor Information

Instructor	Email	Office
Sourav SEN GUPTA	sg.sourav@ntu.edu.sg	N4-02b-66 | 6790 4587

General Information

Description

“Data is the new Oil” in the modern era of Information. It is of paramount importance in each and every sector of Business to collect, store, maintain, visualize and analyze data, in every form and shape, to garner crucial information regarding its scope, relevance, performance and decisions.

This course will introduce the students to the basic tools and techniques of data manipulation, data visualization, statistical modelling, and inference, to enable them to make data-driven decisions in various Business scenarios.

Outcome

By the end of the semester, the students should be able to formulate data-oriented problems in various Businesses, handle sufficient data relevant to the problems, visualize the problems in terms of data, perform exploratory statistical analysis on the data, build machine learning models for prediction, classification, clustering, forecasting from the available data, and finally, build an engaging “data-story” to communicate the original Business problem, its technical formulation, and the data-driven solution.

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

If you get excited by nerdy data cartoons, please do join the class! ;-)

If you want to “try out” the course, come to the first lecture on 19 January, Friday, 8:30 AM @ LT11 (NS2-04-15)

Structure

20 Regular Lectures (one hour each)	8:30 AM to 10:30 AM on Fridays	Week 01 to 13 (minus holidays)
+9 Regular Tutorials (one hour each)	2:30 PM to 3:30 PM on Fridays	Week 02 to 13 (minus holidays)
+2 Lectures (one hour) for Revision	8:30 AM to 10:30 AM on Fridays	Week 14 (only on 20th of April)
+2 Tutorials (one hour) for extra help	2:30 PM to 3:30 PM on Fridays	Week 01 and 14

Evaluation

Written Examination (40%) at the end of the Semester + Three Assignments (20% x 3 = 60%) during the Semester

The written examination will be conducted similar to other SCSE end-semester written examinations. The assignments will be data science related computing assignments in R. In total, four assignments will be posted during the semester, two before the recess week and two after, and the best three (out of the four) will be counted towards the final grade.

Course Material

Required Material

There is no single textbook for the course. The students are expected to be mature enough to follow the lectures and refer to multiple resources (mostly online), as and when required. The only mandatory component of the course is to learn R as the computing framework. It may be expected that the students will install R and R-Studio on their computers.

References and Resources

An Introduction to Statistical Learning (http://www-bcf.usc.edu/~gareth/ISL/) : James, Witten, Hastie, Tibshirani

Data Science for Business (http://data-science-for-biz.com/DSB/Home.html) : Provost and Fawcett

R Package (https://www.r-project.org/) and **RStudio** (https://www.rstudio.com/) : Download and Install, if possible

Additional resources, if required, will be shared with the students from time and again, in the Lectures and/or Tutorials. Almost all of these resources will either be online (free and open source) books or online (freely available) lecture videos.

Course Schedule

Week	Lec / Tut	Topic	Remarks
01 (19/1)	Lecture 01	Motivation and Introduction - What is Data Science?	
	Lecture 02	Basic concepts of Statistics and Data Handling in R	
	Tutorial 00	*Installation of R and R-Studio*	*Optional session*
02 (26/1)	Lecture 03	Prediction in Business - Introduction to Linear Models	
	Lecture 04	Linear Regression - Training, Estimation and Inference	
	Tutorial 01	*Introduction to R for Statistics and Data Visualization*	*Hands-on Demonstration*
03 (02/2)	Lecture 05	Classification in Business - Motivation for Linear Models	
	Lecture 06	Logistic Regression - Classification using Linear Models	
	Tutorial 02	*Linear Regression - Review of Concepts, Applications*	*Hands-on session in R*
04 (09/2)	Lecture 07	Classification using Naïve Bayes and Support Vectors	
	Lecture 08	Performance of a Classification Model - Accuracy, ROC	
	Tutorial 03	*Logistic Regression - Review of Concepts, Applications*	*Hands-on session in R*
05 (16/2)	None	None	Chinese New Year (holiday)
06 (23/2)	Lecture 09	Decisions in Business - Tree Models, Classification Rules	
	Lecture 10	Aggregating Models - Regression and Classification Forests	
	Tutorial 04	*Performance of Classification Models - Linear, Bayes, SV*	*Hands-on session in R*
07 (02/3)	Lecture 11	Choosing a Model - How to avoid Overfitting?	e-Learning (recorded lecture)
	Lecture 12	Bias-Variance Trade-off and Cross-Validation	e-Learning (recorded lecture)
	Tutorial 05	*Trees and Forests for Regression and Classification*	*e-Learning (code samples in R)*
08 (09/3)	None	None	Recess Week

Week	Lec / Tut	Topic	Remarks
09 (16/3)	Lecture 13	Notion of Distance, Nearest Neighbors and Prediction	
	Lecture 14	Neighborhoods to Clusters – k-Means and Dendograms	
	Tutorial 06	*Choosing a Model - Training, (Cross-)Validation and Test*	*Hands-on session in R*
10 (23/3)	Lecture 15	Visualizing Multivariate Data and Dimensionality Reduction	
	Lecture 16	Summary of Supervised and Unsupervised Models in Business	
	Tutorial 07	*Nearest Neighbors and Clustering (k-Means, Dendograms)*	*Hands-on session in R*
11 (30/3)	None	None	Good Friday (holiday)
12 (06/4)	Lecture 17	Forecasting in Business – Fundamentals of Time Series	
	Lecture 18	Time Series – Trends, Seasonality and Cycles in Data	
	Tutorial 08	*Introduction to R for Time Series handling and visualization*	*Hands-on session in R*
13 (13/4)	Lecture 19	Data Analytic Thinking – Practical examples from Business	
	Lecture 20	Data Analytic Thinking – Practical examples from Business	
	Tutorial 09	*Time Series Analysis for Business Forecasting Problems*	*Hands-on session in R*
14 (20/4)	Lecture 21	Revision of all concepts discussed during the course	
	Lecture 22	Revision of all concepts discussed during the course	
	Tutorial 10	*Revision of all concepts discussed during the course*	*Optional session*

There is no Laboratory assigned for the course, as yet. It is being arranged, and the students will be notified in the class. Laboratory sessions will run as ‘free-access laboratory’, without supervision, and the students may use it to work with R. In case the students want to install R and R-Studio on their own laptops (highly recommended), they will get assistance.

Extra Office Hours will be regularly arranged by Sourav Sen Gupta, as required, in consultation with the students.