Data Science and Analytics: Advances in data Sciences and Architecture

INFO 7390

Fall 2017

Sri Krishnamurthy, CFA, CAP

Email   analyticsneu@gmail.com

Office Hours   By appointment

Class   Saturday 9.00 –12.30 pm

Location

Course Objectives:

In this course, we will explore the key techniques used in business intelligence, analytics and visualization. We will learn a variety of methods and software for finding patterns (such as regression, neural networks, association rules, CART, text analysis, geo-spatial analysis, forecasting etc.), building models, and ultimately making decisions using data sets. We will address questions such as:

- How to detect anomalies in large datasets?
- How do credit-card companies detect fraud?
- How to forecast credit worthiness?
- What challenges does Big Data pose to companies and how to handle these challenges?

In addition to understanding the techniques using case studies and in-class exercises, we will explore visualization tools using Tableau and discuss current topics of interest in this space such as Cloud computing, Big Data etc.

This is a hand-on course. So you will have to get your hands dirty. As you start looking at data sets, you will realize a couple of things. One, data is dirty and requires cleaning. We will look into various cleansing techniques and methods used in the industry. Second, information is hidden. We will look into ways in which you can explore data and try methods to extract information from data sets. Third, there is no correct answer. The tools we will demonstrate will give you methodologies to explore data sets. You will have to decide whether the information you have gathered makes sense and whether it can be used for decision making. We will work on many examples to understand the power of these techniques and learn ways of analyzing results.

Software:

We will use R, Python for exercises and to illustrate concepts. Students can use R or Python for final projects.

Lectures:

Lectures would include discussion and illustration of methodologies. I will be posting required and optional reading a week before the class. We will also have guest lectures and in-class exercises.

Grading:

4 Case studies: 40% (Group); Midterm Project: 20% (Group); Final Project: 25% (Group); Class Presentation:10%; Class participation: 5%

Case Studies:

You will work on four case studies to demonstrate your understanding of the topics covered in class.

Midterm Project:

This will be similar to the two case studies but will be an extended case study. You will work in groups. A 10+ page report that discusses your analysis, results and findings will be the deliverable.

Final Project:

The format is similar to the midterm project but you will have the opportunity to choose the topic you want to work on. Additional information will be provided as the class progresses.

Class Presentation:

You will be presenting one topic and an associated example in class. You will have 20 minutes to present. A list of topics will be posted

Class Participation:

You are expected to attend all lectures and participate in class. If you plan to miss a class for a genuine reason, you must email the instructor of your absence. If you miss a class, to get class participation credit, you must submit a 2 page report on the class missed. If you don't submit the report, and you are absent more than 2 classes without reasonable excuse, you will automatically lose 5% of the class participation credit.

Fall 2017– Course Outline

The following is a tentative outline of this course. Depending on the pace and progress, some materials is subject to change. Reading materials are very limited in this outline. As we progress, more readings and references will be added.

Session	No	Topics		In-class presentations
Saturday Sep 9	1	Overview: An intuitive understanding to data science and machine learning	**R tutorials posted**	
Saturday Sep 16	2	Data Wrangling and Missing Data Analysis	**Python tutorials posted**	**1**
Saturday Sep 23	3	Data Exploration and Feature Engineering 1. PowerBI 2. Tableau 3. R-shiny	**Tableau & PowerBI tutorials posted** **Case study 1 given**	**2**
Saturday Sep 30	4	Predictive Analytics – Supervised Learning (Regression, Neural Networks, Decision Trees)		**3**
Saturday Oct 7	5	Predictive Analytics – Unsupervised Learning (K-means, Hierarchical Clustering etc.)	**Case study 1 due** **Case study 2 posted**	**4**
Saturday Oct 14	6	Model Selection and Evaluation		**5**
Saturday Oct 21	7	Automating Data science - Microsoft Azure Studio - Scikit Learn - Pipelining (Luigi, Airflow)	**Microsoft Azure Studio Tutorial posted** **Case study 2 due** **Mid-term case study posted**	**6**
Saturday Oct 28	8	Machine Learning APIs		**7**
Saturday Nov 4th	9	Mid-term case studies	**Mid-term reports due + 15-minute presentation per team** **Case study 3 posted**	**8**

Date		Topic		
Saturday Nov 11th		No class	**Veteran's day**	
Saturday Nov 18th	10	NLP & Text Analytics	**Case study 3 due** **Case study 4 posted**	**9**
Saturday Nov 25th		No class	**Thanksgiving**	
Saturday Dec 2nd	11	Time Series Analysis	**Case study 4 due** **Final project proposals**	**10**
Saturday Dec 9th	12	Geo-spatial analytics	**ArcGIS tutorial posted**	**11**
Saturday Dec 16th	13	Final presentations	**Final project reports due** **Final in-class presentations**	