INTERNATIONAL SCHOOL OF ENGINEERING

INSOFE

PGP in Big Data Analytics and Optimization

Course Contents

CSE 7315c: Foundations of Probability and Statistics for Data Science

This five-day module is aimed at preparing you for the very essential skill of "thinking like a statistician. Financial times identified statistical thinking as one of the top 10 skills every educated person should have, and hence you'll be learning a very important and essential skill. This course thoroughly trains you on:

- Probability theory and related algorithms
- Simulations
- Descriptive statistical methods
- Inferential statistical methods

From a tools perspective, you will gain confidence with tools like R and Excel.

Day 1

On this day, there will be an introduction to random variables, probability theory, conditional probability, and to a most powerful algorithm in probability theory - Bayes Theorem.

Day 2

On this day, the focus is to show you what you should do when you are given any data set – large or small. You will learn to look at one variable closely and gain intuition. The topics will cover:

- Understanding the properties of an attribute: Central tendencies (Mean, Median, Mode); Measures of spread (Range, Variance, Standard Deviation); Basics of Probability Distributions; Expectation and Variance of a variable

Day 3

You will get into the deeper aspects of various distributions. You will understand the parameters that define the probability distributions and differences between discrete and continuous distributions.

- Discrete probability distributions: Bernoulli, Binomial, Geometric, Poisson and properties of each.
- Continuous probability distributions: Exponential; Special emphasis on Normal distribution; t-distribution

Day 4

You will also start inferring about populations from samples.

- Inferential statistics: How to learn about the population from a sample and vice versa; Central Limit Theorem; Sampling distributions; Confidence Intervals, Hypothesis Testing
- Analyzing multivariate normal distributions; an example on anomaly detection.

Day 5

Till this point you will have received the complete picture how to understand the data, attributes, distributions, sample versus population, and procedure for statistical testing, etc. While you continue the analysis of a variable, you will extend that understanding to analyze the relationship between variables.

On this day, you will learn how to conduct a statistical hypothesis testing and will be introduced to various methods such as chi-square test, t-test, z-test, F-test and ANOVA methods in detail.

You will also learn to describe the relationship between attributes using Covariance and Correlation as a precursor to Regression basics.

Hands-on implementation of each of these methods will be conducted in R.

CSE 7212c: Essential Engineering Skills in Big Data Analytics Using R and Python

This is a three-day module where we introduce data reading, pre-processing and then move to designing, evaluating and implementing predictive models using the most widely used tools R and Python. This module will help you become hands-on in identifying and applying the right set of techniques, analyze and present the thought process.

In the first two sessions of this module, candidates will be given an exposure the following techniques:

- R and Python basics, understanding data structures, functions, control structures, data manipulations, date and string manipulations, etc.
- Pre-processing Techniques: Binning, Filling missing values, Standardization & Normalization, type conversions, train-test data split, ROCR1
- Hands-on implementation of all the pre-processing techniques in R and Python.

The third session of this module will be driven completely by a business case analysis

- The objective of this session is to give the breadth and depth of solving a Data Science problem and defend your analysis.

- We provide a business case in advance in which you will be required to apply all the data pre-processing steps and prepare the input for ML algorithms learnt thus far.
- The lab is designed such that everyone participates in a discussion, design the solution approach for the given business case and defend the analysis approach.

CSE 7302c: Statistics and Probability in Decision Modeling

This five-day module is aimed at teaching the most widely used statistical techniques in Data Science.

You will learn very powerful supervised learning methods, viz., Linear Regression, Naïve Bayes classifier and Logistic Regression, which are used to solve problems in Prediction and Classification.

- Linear Regression - probabilistic interpretation
  - Relationship between multiple variables: Regression (Linear, Multivariate Linear Regression) in prediction.
  - Understanding the summary output of Linear Regression from Excel
  - Residual Analysis
  - Identifying significant features, feature reduction using AIC, multi-collinearity check, observing influential points, etc.
  - Non-normality and Heteroscedasticity
  - Hypothesis testing of Regression Model
  - Confidence intervals of Slope
  - R-square and goodness of fit
  - Influential Observations - Leverage
  - Multiple Linear Regression
  - Polynomial Regression
  - Categorical Variables in Regression

Classification

- Naïve Bayes classifier
  - Revisit probability distributions, Joint and conditional probabilities
  - Model Assumptions, Probability estimation
  - Required data processing
  - M-estimates, Feature selection: Mutual information
  - Classifier
- Feature Reduction/Dimensionality reduction
  - Background: Eigen values, Eigen vectors, Orthogonality
  - Principal components analysis
- Regularization methods
  - Lasso, Ridge and Elasticnets
- Logit function and interpretation

- Hands-on R session on Logistic Regression using a business case. Types of error measures to look for. ROCR.
- Logistic Regression in classification; output interpretations

A powerful approach to analyze financial data and other forms of data based on their time dependent past values is known as the Time series analysis. On Day 5, the focus is on analyzing and understanding Time Series with financial markets as the case study.

- Trend analysis
- Cyclical and Seasonal analysis
- Smoothing; Moving averages; Auto-correlation; ARIMA; ARIMAX
- Applications of Time Series in financial markets

CSE 7305c: Methods and Algorithms in Machine Learning

This module discusses the principles and ideas underlying the current practice of data mining and introduces to a powerful set of useful data analytics tools. For each of the techniques, both the traditional approach and the Big Data approach are taught.

At the end of the course, the student will be able to answer questions like “Which machine learning technique is likely to work under which situation?”, “How to handle fraud detection?” and “How to build a powerful recommendation engine?”

From techniques perspective, the student learns:

- Rule based approach, distance based approach, mathematical modeling, etc.

Day 1

You will begin to learn some foundations on the rule pattern, construction of rule based classifier from data, turning trees into rules, rule growing strategy, rule evaluation and stopping criteria, several business metrics such as action ability, explicability, etc. and later turn towards association rules and cover them in detail.

Classification rules

- Indirect: from decision trees
- Direct: sequential covering

Perhaps the most common applications of data mining for most users are Associations, Market Basket, Recommendation Engines, etc. On this day, you will learn a powerful technique for accomplishing all these tasks – Apriori.

Association rules

- How to combine clustering and classification; How to measure the quality of clustering; Outlier analysis
- A mathematical model for association analysis; Large itemsets; Association Rules
- Apriori: Constructs large itemsets with mini sup by iterations
- Interestingness of discovered association rules; Application examples; Association analysis vs. classification
- Fp-trees

Day 2

You will learn one of the top 10 machine algorithms. With the knowledge of direct and indirect rules, you will further learn the details of the rule based classifiers.

- Discuss a use case and manually derive the rules.
- Top down induction of decision trees (TDIDT)
- Attribute selection based on information theory approach.
- Recursive partitioning (binary search)
- ID3, C4.5, C5.0 for pattern recognition problems, avoid overfitting, converting trees to rules.

Day 3

You will be introduced to instance based classifier namely k-Nearest Neighbors which is simple yet one of the powerful supervised learning techniques. Another popular technique called collaborative filtering will be taught. The details of each are as follows:

- Computational geometry; Voronoi Diagrams; Delaunay Triangulations
- K-Nearest Neighbor algorithm; Wilson editing and triangulations
- Aspects to consider while designing K-Nearest Neighbor
- Hands-on example of K-Nearest Neighbor using R
- Collaborative filtering and its application areas

Day 4

According to many researchers, Support Vector Machines (SVM) is the most elegant technique developed in the last two decades. You will learn about this extremely powerful, cutting-edge technique on this day.

- Linear learning machines and Kernel space, Making Kernels and working in feature space
- Demonstrate the working of SVM classification and regression problems using a business case in R.

Day 5

In machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any one single algorithm. The basics of ensembles are bagging & boosting that will be covered in detail and later progress with machine learning methods that use either or both approaches to build ensemble models.

- Bagging & boosting and its impact on bias and variance
- C5.0 boosting
- Random forest
- Gradient Boosting Machines and XGBoost which are the very popular winning recipe of data science competitions.

Day 6

This module aims at teaching the principles and ideas underlying the current practice of data mining and introduces to a powerful set of unsupervised data analytics tools such as Clustering to find hidden structures within the unlabeled data. Before we delve into the details we conduct a math refresher on topics such as linear algebra, matrices and distance metrics.
You will learn the most commonly used unsupervised learning algorithm – Clustering.

- Different clustering methods; review of several distance measures
- Iterative distance-based clustering;
- Dealing with continuous, categorical values in K-Means
- Constructing a hierarchical cluster, K-Medoids, k-Mode and density based clustering to handle different data types in practice.
- Test for stability check of clusters.
- Hands-on implementation of each of these methods will be conducted in R.

Day 7

Business case analysis

- The objective of this session is to provide an application and end-to-end view of solving a Data Science problem and defend your analysis.
- We provide a business case in advance in which you will be required to apply all the data pre-processing steps and prepare the input for one or more ML algorithms learnt thus far.
- The lab is designed such that everyone participates in the discussion, design the solution approach for the given business case and defend the analysis approach.

CSE 7322c: Engineering Big Data with Hadoop and Spark Ecosystem

Companies collect and store large amounts of data during daily transactions. This data is both structured and unstructured. The volume of the data being collected has grown from

MB to TB in the past few years and is continuing to grow at an exponential pace. The very large size, lack of structure and the pace at which it is growing characterize the “Big Data”.

To analyze long-term trends and patterns in the data and provide actionable intelligence to managers, this data needs to be consolidated and processed in specialized processes; those techniques form the core of the module.

This is a five-day module which provides a detailed understanding of several important aspects in Big data.

This course thoroughly trains candidates on the following techniques:

- Sql querying (with a focus on statistical analysis)
- Hadoop and Map Reduce methods of programming
- Designing columnar databases
- Spark and H2O

From a tools perspective, this course introduces you to Hadoop. You will learn one of the most powerful combinations of Big Data, viz., Hadoop – R & Spark.

The course gives an exciting motivation for learning Big Data. A host of problems where traditional techniques fail will be introduced. Overview of this course, Intro to Hadoop, NoSQL, In Memory Computing, etc.

- World uses more Big Apps than you realize – A taxonomy and demonstrations of apps
- Discuss projects, watch more demos, solve assignments to understand high volume, high velocity and high variety data & solution stacks available for them
- An architectural overview of the business problem (e.g., customer analysis) where scale becomes important and why traditional approaches fail
- How Hadoop is designed to solve big data problems; Its components and the ecosystem

Data center as a computer: From Cells and Grids to Master-Slave Clouds - Evolution of clusters

- Design Considerations: Cost, failure
- What's so special about Hadoop?
- Sequential and Concurrent algorithms design, metrics

You will learn to setup your own Hadoop VM, Start accessing and using INSOFE cluster, practice problems on task dependency graphs and parallel implementations

You will get to know what's so special about Hadoop.

- GFS, HDFS, Next Generation HDFS

You will work with HDFS and learn to build capacity planning calculators

- Rapidly ingesting & organizing unstructured data
  - Chukwa, Flume, Avro
  - NoSQL: Big Table, HBase, Document stores, Graph stores, Key-Value stores

There will be a demo of NoSQL databases, practice assignments in Flume and Avro. Pick one ecosystem component to master

- Processing frameworks on clusters
  - Map reduce
  - Yarn, MR2, R-Hadoop, Hadoop Streaming with Python
  - BSP, Spark

Learn for many simple problems, how to come up with parallel algorithms. Write & execute small R-Hadoop / Hadoop Streaming programs. Write and execute basic scripts and build machine learning models on Spark processing framework, observe differences in programming paradigms of MR and BSP.

SQL on Hadoop

- Sqoop, Hive and variants

You will learn to use Sqoop, Hive to ingest and query non-trivial relational data sets. Use Hadoop-as-a-service platforms like Qbole and Xplenty

- PIG and other hadoop family
  - PIG programming, Oozie, Zookeeper and Mahout

Write & execute small PIG programs. Execute Oozie workflows. Build recommender engine, classification and clustering models in Mahout.

CSE 7323c: Building End-to-End Data Science Applications

This is a six-day complete hands-on module designed to leverage the theory covered in the entire course. The goal is to solve some real-world problems using all the concepts learnt thus far. Particularly, architecting the solution approach, choosing the right tool set, situational based analysis, relevant programming techniques and skills will be emphasized.

Example:

The typical use cases for the program are “analyzing a customer in near real-time”, “recommendations in near real-time”, etc., as applied in Retail, Banking, Airlines, Telecom or Gaming industries.

At the end of the program, the participants will have solved a couple of case studies completely and more importantly will be able to articulate the solution approach and solve a data science problem and how to utilize the tool-set.

CSE 7124c: Foundations of Text Mining and Search

This module aims at teaching the principles and ideas underlying text mining and social network analytics. **Text mining**: Unstructured data comprises more than 80% of the stored business information (primarily as text). This helped text mining emerge as a leading-edge technology.

This course aims to provide a conceptual and practical understanding of various aspects of text processing. In this module, several applications of text classification tasks. The course gives an exposure to the basics of search engines like indexing, crawling, and ranking.

The lab activities will be done in Python

Day 1

- Introduction to the Fundamentals of information retrieval;
  - TF and IDF
  - Thinking about the math behind text; Properties of words; Vector Space Model
- Matrix factorization: SVD
- Text Indexing
  - Inverted Indexes
  - Boolean query processing
  - Handling phrase queries, proximity queries
  - LSA

Day 2

On day 2, you will learn Relevance ranking and Link analysis algorithms.

- Relevance Ranking
  - Need for Relevance Ranking
  - Evaluation Metrics for Ranking
- Link Analysis Algorithms
  - PageRank

PREFERRED TEXT BOOKS:

Manning, C., Raghavan, P., and Schutze, H. (2007). An Introduction to Information Retrieval. Cambridge University Press.

Chakrabarti, S. (2002). Mining the Web: Discovering knowledge from hypertext data. Morgan-Kaufman.

Introduction to Information Retrieval, by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

REFERENCE BOOKS:

- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Pearson Education.
- Witten, I. H., Moffat, A., and Bell, T. C. (1999). Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan-Kaufman.
- Grossman, D. A. and Frieder, O. (1998). Information Retrieval: Algorithms and Heuristics. Kluwer.
- Croft, B., Metzler, D., and Strohman, T. (2009). Search Engines: Information Retrieval in Practice. Pearson Education.
- Search Engines: Information Retrieval in Practice, by Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.
- Also, we will refer many papers from recent WWW, WSDM, SIGIR, CIKM and ICWSM proceedings.

CSE 7321c: AI and Decision Sciences

This is six-day module where you will learn advance topics that give you exposure to keep you up-to-date with the current research trends and latest happenings in Data Science. Additionally, this module is designed to enhance your decision capabilities when confronted with strategic choices.

Deep Learning is the fastest growing field in Machine Learning, an approach to AI that has been revolutionizing several industries and playing a major role in changing the way we live. The areas of application of Deep Learning are wide spread due to its intelligent systems and ability to learn. These systems learn and classify several components in a similar fashion to that of a human brain, but with much higher precision and speed. It uses Deep Neural Networks to learn to extract & epitomize data that comes in several forms such as images, text and sound. Rather than using modality-specific traditional Machine Learning techniques, today's advanced deep neural networks can learn at high speed owing to GPU's computational power and the big data algorithms.

Take-way from this module will be advanced machine learning models which can process both structured and unstructured data (images, text, etc.) for performing tasks such as classification, regression and more. The models in many contexts can also overcome the limitations of other linear and shallow models.

This module will be organized in a workshop mode. A combination of theory and hands-on will be practiced for entire day.

Another fascinating machine learning technique is Neural Networks. There are ardent followers and equally passionate haters of this technique! You will learn one of the most commonly used and important types of Neural Networks.

Day 1: Artificial Neural Networks

- Perceptron model and its limitations
- Multi-layer perceptron and non-linear data
- Learning using Back-propagation
- ANNs for classification and regression of structured data

Day 2: Deep learning

- Auto-encoders and unsupervised learning
- Stacked auto-encoders and semi-supervised learning
- Regularization - Dropout and Batch normalization
- Image classification and hyper-parameter tuning

Day 3: Convolutional neural networks

- Large scale image classification
- Text classification
- Very deep architectures

Day 4: Recurrent neural networks

- Long short-term-memory cells
- Text classification
- Sentiment analysis
- Time-series

Day 5:

Additionally, students will have increased ability to turn real-world problems into mathematical and spreadsheet models. This module also teaches three classes of models: Optimization, Simulation and Statistical. The application areas originate from problems in finance, marketing and operations. The course also teaches techniques of quick experimentation and scientific guessing. The course will work on both analytics-based and intuitive abilities to help engineers take sound strategic decisions.

At the end of the program, you will be able to answer questions like “Should I outsource a service or do it in-house?”, “How to optimize a supply chain?”, “How to price a product when faced with demand uncertainty?" and "How to price a derivative product using a Monte Carlo simulation?"

This course thoroughly trains candidates on the following techniques:

- Linear Programming and Sensitivity Analysis
- Recognize the optimization problem
- Setting up systematically
- Recognizing that linear optimization is relevant across several domains
- Process of optimizing
  - Identify the decision variables
  - State the objective function as a linear combination of the decision variables
  - Hidden constraints
  - Corner points
  - Unbounded and infeasible solutions
- Integer and binary programming using Case studies such as Capital budget allocations
- Assignment problem and Transportation Problem and it's applications on Product mix problems and Production scheduling
- Solving and analyzing using R

Day 6:

On this day, you will learn how to find the best solution given that most of the ML methods give you a suboptimal solution.

- Guided Random Searches
- Summary of Biology
- Understanding Gene, Chromosomes Population and Reproduction and mutation
- GA at a snapshot
- Hand working few examples
- Operational Aspects
  - How do we decide mating pool?
  - How do we determine reproduction?
- Mutation scheme
- Convergence of a GA
- Hands-on implementation using several business cases.

Additionally, methods such as Simulated Annealing and Monte Carlo Simulations will be taught.

Hands on implementation in R

Planning & Thinking Skills for Architecting Data Science Solutions:

The emphasis is not on engineering but on consulting and architecting solutions. We teach frameworks and expose the students to a variety of problems. Practitioners who do not need hands-on experience in analytics but need to pre-sell or architect will also benefit from the module. This module also helps bringing all the concepts that will be taught in other modules into perspective, helping students provide end-to-end solutions to business problems.

At the end of the program, the students will be able to answer important questions like “What is the big data analytics ecosystem?”, “What are the different forms in which the data is available?”, “What are the standard techniques to solve several classes of problems?”, “How to analyze the problem through a systematic framework?”, “What are the suitable error measures?” and “What are the most efficient ways to present analytics results?”.

CSE 7120c: The Art and Science of Storytelling with Data Visualizations

- Why and Where of Visualizations
- Storytelling- A great art and science (examples)
- Communicating with data: Issues and guiding principles; Primary ingredients of data visualization; How to pick visual encodings such as colour, shape, size, etc.; Which chart to use when; How to accommodate more than 2 dimensions
- Case highlighting the transition from a simple chart to a powerful visualization, complete with storytelling
- Using R-ggplots/Tableau/Qliksense for visualizations

CSV 1103: Communication, Ethical and IP Challenges for Analytics Professionals (Video access)

This module emphasizes the importance of communication for Analytics professionals, especially since they are expected to deal with technical and non-technical users more closely than in any other discipline.

Students also learn to appreciate the importance of ethical, legal and IP issues given that regulations are still sketchy in this field where adoption is increasing at rapid rates. Students learn to appreciate how to avoid ethical and legal pitfalls and what issues to be aware of when dealing with data.

- Why is Communication important?
- How to communicate effectively: Telling stories
- Communications issues from daily life with examples using audio, video, blogs, charts, email, etc.
- Seeing the big picture; Paying attention to details; Seeing things from multiple perspectives
- Challenges: Mix of stakeholders, Explicability of results, Visualization
- Guiding Principles: Clarity, Transparency, Integrity, Humility
- Framework for Effective Presentations; Examples of bad and good presentations
- Writing effective technical reports
- Difference between Legal and Ethical issues
- Challenges in current laws, regulations and fair information practices: Data protection, Intellectual
- property rights, Confidentiality, Contractual liability, Competition law, Licensing of Open Source
- software and Open Data
- How to handle legal, ethical and IP issues at an organization and an individual level
- The "Ethics Check" questions

Disclaimer: *The document gives day-wise topic details. While the overall structure will remain as discussed, the faculty reserves the right to make minor modifications (additions/deletions of topics, or increasing/reducing the scope of individual topics) as they deem fit to make the course as relevant as possible.*