International School of Engineering

upon recommendation of the faculty and
by authority of the Board of Directors,
hereby confers upon

Anirima Ghosh

THE POSTGRADUATE CERTIFICATE IN BIG DATA ANALYTICS AND OPTIMIZATION

on successful completion of all the requirements of the 352-hour program conducted between March 25 and September 10, 2017

This program is certified for quality of content, assessment and pedagogy by the Language Technologies Institute (LTI) of Carnegie Mellon University (CMU). The program curriculum has been developed in collaboration with LTI.

INSOFE

Carnegie Mellon University
Language Technologies Institute

Dated this eighteenth day of January, two thousand and eighteen.

Dr. Dakshinamurthy V Kolluru
President

Dr. Sridhar Pappu
Executive VP-Academics

INTERNATIONAL SCHOOL OF ENGINEERING
INSOFE

Mode: Classroom Teaching
07CSE03/201709/999

Certificate Type

Certificate of Participation [✓ Assessment-based training program] Professional certification

Topics Covered

Foundations of Probability and Statistics for Data Science

Probabilistic analysis of data and models; Computing the properties of an attribute: Central tendencies (Mean, Median, Mode, Range, Variance, Standard Deviation); Expectations of a Variable; Describing an attribute; Probability distributions (Discrete and Continuous) - Bernoulli, Geometric, Binomial, Poisson and Exponential distributions; Special emphasis on Normal distribution; Central Limit Theorem; t-distribution, Chi-Square. Inferential statistics: Hypothesis Testing; Confidence Intervals, ANOVA; Describing the relationship between attributes: Covariance; Correlation.

Essential Engineering Skills in Big Data Analytics

Introduction to R – data import, querying, data subset, data manipulation, control structures, regular expressions. Data pre-processing of structured data: handling missing values, Binning, Standardization, Outliers/Noise, PCA, Type conversion. Introduction to Python - data import, querying, data subset, data manipulation, control structures, regular expressions. Data pre-processing of structured data: handling missing values, Binning, Standardization and Type conversion

Statistics and Probability in Decision Modeling

Regression (Linear, Multivariate Regression). Analyzing and interpreting regression output; Logistic Regression for classification. Time Series Analysis; Components of time series data: Trend, seasonal and random fluctuations. Smoothing techniques; Box-Jenkins method; Holt-Winters method and ARIMA

Methods and Algorithms in Machine Learning

Rule based knowledge: Logic of rules, evaluating rules, rule induction and Association rules. Construction of Decision Trees, Root node identification based on entropy, information gain; Dealing with overfitting; Pruning a Decision Tree; Unwrapping trees as rules. K-Nearest Neighbour method; Wilson editing and triangulations; K-nearest neighbours in collaborative filtering, digit recognition. Motivation for Neural Networks and its applications; Perceptron and Single Layer Neural Network, and hand calculations; Learning in a Neural Net: Back propagation and conjugant gradient techniques; Application of Neural Net in Face and Digit Recognition. Kernel methods in learning. VC (Vapnik-Chervonenkis) dimension; Shattering power of models. Algorithm of Support Vector Machines (SVM). Ensemble and Hybrid models, AdaBoost, Random Forests. Connectivity models (hierarchical clustering); Centroid models (K-Means algorithm).

AI and Decision Sciences

Concept of regularization, Lasso, Ridge and Elastic Nets, GLM
Deep learning method, auto-encoders and feature extraction methods
Convolution Neural Nets and Recurrent Neural Nets with applications in text analysis and image data. Time series analysis using LSTMs.
Planning, Thinking and architecting data science solutions
Linear Programming: Graphical analysis; Sensitivity and Duality analyses
Integer and Binary programming: Applications; Problem formulation;
Goal programming; Quadratic programming with stock portfolio allocation model application. Genetic algorithms: The algorithm and its components; fitness function, probability definitions for mutation & cross-over along with termination criteria. Monte Carlo Simulations.

Text Mining, Social Network Analysis and Natural Language Processing

Introduction to text mining and text pre-processing. Data Cleansing of unstructured data: Handling numbers, Punctuations, Stop-words, Incorrect spellings, Stemming, Lemmatization and TxD computation; Write a web crawler to collect data. Fundamentals of information retrieval; Properties of words; Vector space models; Creating Term-Document (TxD) matrices; Similarity measures; Text classification and feature selection: Matrix factorization methods. Singular Value Decomposition. .. Extend the concept of Bayes theorem in classification problems. Concept of Naive Bayes classifier and its applications. Application of Naive Bayes algorithm. Evaluation systems on the accuracy of text mining; Sentiment Analysis, Natural Language Analysis; Social network analysis; structure of Web graph; Small world model; Influence propagation in Social networks; Application of social influence analysis.

Engineering Big Data Applications with Hadoop and Spark Ecosystem

Big Data, Hadoop applications; Parallel and Distributed computing;
Introduction to algorithms; Concurrent algorithms; Linux refresher; NoSQL; GFS; HDFS; CDH4-HDFS ; Map Reduce: YARN Map Reduce Applications: Text Mining, Page Rank, Graph processing Hadoop ecosystem components: Pig, Hive, HBase, Sqoop, Mahout, H2O, Hama, Flume, Chukwa, Avro, Whirr, Hue, Oozie, Zookeeper, Kafka. Hadoop Streaming with R and Python. Spark-SQL, Streaming and ML

Building End-to-End Data Science Applications

Hands-on implementation of various state-of-the-art tools from the Hadoop ecosystem. Review the business case, data, architect the solution approach, identify the right set of tools from the stack, set –up and practice on Hadoop cluster; batch or real-time processing of data; application of machine learning methods. Deal with structured and unstructured data. solve the complete case and present.

The Art and Science of Storytelling with Data Visualizations

Why and Where of Visualizations. Communicating with data: Issues and guiding principles; Primary ingredients of data visualization; How to pick visual encodings such as colour, shape, size, etc.; Selection of chart types; How to accommodate more than 2 dimensions; Case highlighting the transition from a simple chart to a powerful visualization and importantly presentation and storytelling. Use R-ggplots and Tableau/Qliksense for visualizations

Communication, Ethical and IP Challenges for Analytics Professionals (Video)

Why is Communication important? How to communicate effectively?
Communications issues from daily life using 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.