BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

INSTRUCTION DIVISION

SECOND SEMESTER 2020-2021

Course Handout Part II

Date: 16/01/2021

In addition to part-I (General Handout for all courses appended to the timetable) this portion gives further specific details regarding the course.

Course No. **: CS F320**
Course Title **: Foundations of Data Science**
Instructor-in-charge **: NAVNEET GOYAL (goel@)**
Course TA **: Ayushi Gaur (p20190023@)**

Catalog Description

Data Science is the study of the generalizable extraction of knowledge from data. Unprecedented advances in digital technology during the second half of the 20th century and the data explosion that ensued in the 21st century is transforming the way we do science, social science, and engineering. Application of data science cut across all verticals. A data scientist requires an integrated skill set spanning mathematics, probability and statistics, optimization, and branches of computer science like databases, machine learning etc. The course aims at providing the mathematical and computer science foundations required for data science.

Text & Reference Books:

T1. Foundations of Data Science - Avrim Blum, John Hopcroft, Ravi Kannan, January, 2018.
R1. Machine Learning: An Algorithmic Perspective – Stephen Marsland, CRC Press, 2e, 2015.
R2. Pattern Recognition & Machine Learning – Christopher M Bishop, Springer, 2006.
R3. An Introduction to Data Science – Jeffrey Saltz and Jeffrey Stanton, Sage Publications, September 2017.

LECTURE PLAN (42)

Topic	Topic Details	No. of Lectures	Chapter Reference
Course Overview & **Introduction to Data Science**	1. Motivation/course objectives 2. Some motivating applications 3. Data Dimensionality 4. Types of Data	2	T1 – Ch. 1 Class Notes + web resources
High-dimensional data & Curse of Dimensionality	1. Characteristics of High-dimensional data a. Law of large numbers b. Geometry of HD c. Vectors in HD d. Gaussians in HD 2. Curse of Dimensionality (CoD) a. Sparsity b. Nearest Neighbor search c. Concentration of Lp-norm 3. Dimensionality Reduction Techniques – PCA & SVD	4	T1 – Chs. 2, 3

Frequentist vs. Bayesian approach to Probability	1. Frequentist Approach 2. Bayesian Approach 3. Prior to Posterior – Bayes' Theorem 4. MLE vs. MAP	2	Class Notes + https://sites.google.com/site/bayestutorial/
Probability Distributions and Mixture Models	1. Exponential family of distributions (Bernoulli, Beta, Binomial, Dirichlet, Gamma, & Gaussian) 2. Mixture Models – Mixture of Gaussians	2	R2 – Ch.2, Appendix B
Optimization Techniques	1. Unconstrained/Constrained optimization 2. Convex Optimization & Lagrange Multipliers 3. Quadratic Programming 4. Primal/dual 5. Kernels 6. Gradient Descent & its variants	3	Class Notes
Function Approximation Techniques	1. Basis Functions 2. Splines 3. Mixture Models	2	R1: Chs. 5,7
Tensors	1. Introduction to Tensors 2. Tensor Algebra 3. Tensor Calculus 4. Modeling multidimensional data using Tensors	3	http://web.iitd.ac.in/~pmvs/courses/mcl702/tensors.pdf
Machine Learning Basics	1. Supervised Learning a. Regression (polynomial, linear basis function models) b. Classification (Naive Bayes, Decision Tree, SVM, NN) 2. Unsupervised Learning a. K-means Clustering b. Expectation Maximization Clustering c. Self-Organizing Maps (SOMs) 3. Anomaly Detection 4. Machine Learning, Function Estimation, & Optimization 5. Model Underfitting & Overfitting 6. Model Selection & Complexity a. Occam's Razor b. VC dimension c. Structural Risk Minimization d. Bias Variance Decomposition	8	T1 – Chs. 5,7 R1 – Chs. 2-4,8,12,14 R2 – Chs. 1,3,9 R3 – Ch. 18
Markov Chain Monte Carlo (MCMC) Methods	1. Gibbs Sampling 2. Metropolis Hastings Algorithm	2	T1: Ch. 4 R1: Ch. 15 R2: Ch. 11
Probabilistic Graphical Models	1. Markov Models 2. Hidden Markov Models (HMM) 3. Bayesian Belief Networks (BBN) 4. Markov Random Field (MRF)	3	T1: Ch. 9 R1: Ch. 16 R2: Ch. 8

Time-series Data & Analytics	1. Importance & Characteristics of time series data 2. Sources of time series data 3. Similarity Metrics & Dynamic Time Warping 4. Multi-variate Time Series Data – IoT Data 5. Time Series analytics: a. Regression b. Classification c. Clustering d. Anomaly Detection	5	Class Notes + web resources
Big Data & Big Data Analytics	1. Introduction to Big Data Analytics 2. Big Data - sources & applications 3. Social Media Data	1	T2 – Ch. 20
Distributed Computing Frameworks	1. MapReduce and its variants 2. Spark	2	Class Notes + web resources
Data Visualization	1. Visualization Foundations 2. Visualization Pipeline 3. Scalar, Vector, & Tensor Visualization 4. Visualization Techniques for Spatial, Geospatial, & Time-series Data 5. Role of SOMs in Data Visualization	3	T2 – Chs. 12,13

Evaluation Scheme:

Component	Duration	Weightage	Date (Time)
Midsem Test (Closed Book)	90 Mins.	30%	TBA
Assignment(s)/Lab. Test/Quiz	TBA	30%	TBA
Comprehensive Exam (partly open)	120 Mins.	40%	13/05 (FN)

Labs. on R: No structured lab. sessions, but students will be provided with Lab. sheets on important topics.
Notices: All notices will be uploaded on NALANDA only.
Chamber Consultation Session: Online session once a week (M-10). Interested student(s) need to inform apriori if a session is required.
Makeup Policy: To be granted only in case of serious illness or emergency.
Email Policy: Communication through email is highly discouraged. If you want to discuss anything, attend the chamber consultation session. Academic queries/doubts can be posted on NALANDA (a discussion forum will be created)
Plagiarism Policy: If any student is found involved in any kind of plagiarism in any of the evaluation components, the matter will be directly reported to the Examination Committee.
NC Policy: Students securing 10% or less marks will get an NC grade. Students in the [10-15%] bracket are also likely to get NC.

Instructor-in-charge
CS F320