Birla Institute of Technology & Science, Pilani
Hyderabad Campus

Summer Term -2022
Course Handout Part II

28.05.2022

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

Course No.: CS F320
Course Title: Foundations of Data Science
Instructor-in-Charge: Dr. Rajib Ranjan Maiti

1. Scope and Objectives

This course lays down the necessary foundations of data science for insightful and deeper understanding of courses like Machine Learning, Data Mining and Information Retrieval etc. It emphasizes probabilistic, statistical and computational foundations of data science. The curse of dimensionality and relevant dimension reduction techniques like PCA are discussed. The pre-processing techniques like data wrangling, feature extraction, feature selection, cleansing, standardization etc. are also be discussed in the course. The data visualization techniques like boxplots, scatter plots, heat maps, histograms etc. are explored in this course. This course also introduces Big Data and Analytics to students and how it is different from non-Big Data.

Having successfully completed this course, students will be able to demonstrate fundamental knowledge and understanding of

- Necessary computational, mathematical, or statistical techniques and models to build data science applications.
- Dimensionality reduction techniques and its consequences.
- Data Pre-processing techniques
- Data Visualization techniques and tools
- Big Data & Analytics

2. Pre requisites:

MATH F113 – Probability and Statistics

3. Text Books

T1: Pattern Recognition and Machine Learning – Christopher M. Bishop, Pearson, 1st Ed. - 2013.

T2: An Introduction to Data Mining – Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, Pearson, 2nd Ed. – 2005.

Reference Books:

R1. Tom M. Mitchell: Machine Learning, The McGraw-Hill Companies, Inc.
R2. Kevin P Murphy: Machine Learning, a probabilistic perspective
R3. David Barber: Bayesian Reasoning and Machine Learning

4. Course Plan

Lecture No	Learning Outcomes	Topics to be covered	Chapter in the Text Book
01	To introduce the course	Introduction and significance of the course for data science discipline	Class Notes
2 – 7	To review and learn probability theory from data science perspective	Review of Probability - Continuous and Discrete Random Variable, Probability density and mass functions, Expectation and, Guassian distribution, probabilistic perspective of Polynomial Curve Fitting	T1 – 1.2.1 to 1.2.4
8 – 12	To understand Binary and Multinomial variable and distribution, Maximum likelihood and Bayeisan Inference of Gaussian Distribution	Conditional Gaussian distributions, Marginal Gaussian distributions, Bayes' theorem for Gaussian variables, Maximum likelihood and Bayesian Inference for the Gaussian, covariance, Bayesian probabilities, Beta Distribution, Dirichlet distribution, Student's t-distribution, Bayesian Curve Fitting	T1 – 2.1, 2,2, 2.3.1 to 2.3.7, 1.2.6.
13 – 15	To understand Information Theory and Decision Theory fundamentals that are necessary for Data Science	Minimizing Misclassification rare & expected loss, The reject option, Inference and decision, Loss functions for regression, Relative Entropy and Mutual Information, Decision Tree	T1 – 1.5 and 1.6
16 – 17	To understand probability bounds that are necessary for data science	Probability Bounds (Markov, Chebyshev, and Chernoff Bounds)	Class Notes
18 – 19	To understand non-parametric methods of density estimators	Nonparametric Methods - Kernel density estimators, Nearest-neighbour methods	T1 – 2.5
20 – 24	To understand Computational foundations that are necessary for data science	Unconstrained/Constrained optimization, equality/inequality constraints, convex optimization, Lagrange multiplier, primal/dual concept, building linear regression models using kernels	Class Notes, T1 - 6.1, T1 – Appendix E
25 – 31	To understand the curse of dimensionality and relevant techniques like PCA etc.	Curse of Dimensionality, Principal Component Analysis	T1 – 1.4., 12.1, Class Notes
32 – 37	To apply Data Preprocessing techniques to build accurate prediction models	Types of Data, Data Quality, Data Pre-processing, Measures of Similarity and Dissimilarity, Data wrangling techniques	T2 – Chap. 2
38 – 40	To apply the Data Visualization techniques	Basic Data Visualization Techniques - Mapping Data to Graphical Elements, Histograms,	T2 – Chap 3, Class Notes

		Pie Charts, Box Plot Percentile Plots and Empirical Cumulative Distribution Functions, Scatter Plots, Visualizing Spatio-temporal Data OLAP and Multidimensional Data Analysis	
41 – 42	To evaluate characteristics of Big Data & Analytics and how it is different from non-Big Data	Introduction to Big Data & Analytics	Class Notes

5. Evaluation Scheme

Component	Duration	Weightage	Date&Time	Nature of Component
Mid-Semester Test	90 mins	**30%**	23/06/2022 3.30 - 5.00PM (90Mins)	Closed
Assignments (2 Nos.)	-	**15% + 15% = 30%**	TBA	Open
Comprehensive	180 mins	**40%**	20/07 AN (180 mins)	Closed

Note: 40% of the evaluation to be completed by mid-sem grading.

6. CONSULTATION HOUR: Tuesday, Thursday and Saturday (10:030 AM **– 11:00 AM**)

7. Make-up: Make-up will be granted only to genuine cases with prior permission only.

8. NOTICES: All notices will be put up in CMS and students are strongly advised to log in to CMS and look for notices quite often.

9. Academic Honesty and Integrity Policy: Academic honesty and integrity are to be maintained by all the students throughout the semester and no type of academic dishonesty is acceptable.

Instructor-in-charge
CS F320