****

**FIRST SEMESTER 2022-2023**

# Course Handout Part II

Date: 29.08.2022

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: Prof. N.L.Bhanu Murthy

### 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, Springer - 2013.

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

**Reference Books:**

R1. Tom M. Mitchell: Machine Learning, The McGraw-Hill Companies, Inc., 1997

R2. Kevin P Murphy: Machine Learning, a probabilistic perspective, MIT Press, 2012

R3. David Barber: Bayesian Reasoning and Machine Learning, CUP, 2012

**4. Course Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| Lecture No | Learning Objectives | Topics to be covered | Chapter in the Text Book |
| 1 | To introduce the course | Introduction and significance of the course for data science discipline | Class Notes |
| 2 | To introduce data science pipeline and models | Data Science pipeline, learning models | Class Notes |
| 3 – 4 | To review and learn probability theory from data science perspective | Review of Probability – Continuous and Discrete Random Variable, Probability density and mass functions, Expectation, Variance/Covariance of random variables, Gaussian distribution, | T1 – 1.2. (excluding 1.2.5 and 1.2.6) |
| 5 - 10 | To understand building regression models and probabilistic curve fitting | Introduction to Regression, Polynomial curve fitting, Gradient descent algorithms, overfitting, regularization, probabilistic perspective of Polynomial Curve Fitting | 1.1, class notes, 1.2.5 |
| 11 - 14 | To understand Maximum likelihood and Bayesian Inference of Bernoulli Distribution, Bayesian curve fitting | Beta distribution, Bernoulli distribution - Maximum likelihood estimation and Bayesian inference,  Bayesian Curve Fitting | 2.1, 1.2.6 |
| 15 – 18 | 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 |
| 19 – 20 | To understand probability bounds that are necessary for data science | Probability Bounds (Markov, Chebyshev, and Chernoff Bounds) | Class Notes |
| 21 – 22 | To understand non-parametric methods of density estimators | Nonparametric Methods - Kernel density estimators, Nearest-neighbour methods | T1 – 2.5 |
| 23 – 26 | 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 |
| 27 – 33 | 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 |
| 34 – 38 | 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 |
| 39 – 40 | To apply the Data Visualization techniques | Basic Data Visualization Techniques - Mapping Data to Graphical Elements, Histograms, Pie Charts, Box Plot  Percentile Plots and Empirical Cumulative Distribution Functions, Scatter Plots, Visualizing Spatio-temporal Data  OLAP and Multidimensional Data Analysis | T2 – Chap 3, Class Notes |
| 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%** | 05/11 1.30 - 3.00PM | Closed |
| Assignments (2-3) | - | **30%** | TBA | Open |
| Comprehensive | 3 hours | **40%** | 18/12 FN | Closed |

Note: At least 40% of the evaluation components for Mid-semester grading.

**6. CHAMBER CONSULTATION HOUR:** Tuesday 5PM – 6PM

**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**