21-805-0302: Probability and Statistics for Data Science

Core/Elective: **Core** Semester: **3** Credits: **4**

Course Description

The aim of this course is to introduce fundamental concepts in probability and statistics from a data-science perspective. The aim is to become familiarized with probabilistic models and statistical methods that are widely used in data analysis.

Course Outcomes (CO)

After the completion of the course, the students will be able to:

CO1: Learn the core concepts of probability theory.
CO2: Understand basic principles of statistical inference in estimation and testing.
CO3: Understand the connection between statistical theory and statistical practice.
CO4: Understand the collection, analysis, interpretation, and presentation of data.
CO5: Evaluate problems on discrete and continuous probability distributions.
CO6: Explore certain statistical concepts in practical applications of data science domain.

Course Content

1. Probability theory: probability spaces, conditional probability, independence – Random variables: discrete and continuous random variables, functions of random variables, generating random variables – Multivariate random variables: joint distributions, independence, generating multivariate random variables, rejection sampling – Expectation: Mean, variance and covariance, conditional expectation.

2. Random process: definition, mean and autocovariance functions, iid sequences, Gaussian and Poisson process , random walk – Convergence of random process: types of convergence, law of large numbers, Central limit theorem, monte carlo simulation – Markov chains: recurrence, periodicity, convergence, markov-chain monte carlo- Gibbs sampling, EM algorithm, variational inference.

3. Descriptive statistics: histogram, sample mean and variance, order statistics, sample covariance, sample covariance matrix – Frequentist statistics: sampling, mean square error, consistency, confidence intervals, parametric and non-parametric model estimation.

4. Bayesian statistics: Bayesian parametric models, conjugate prior, bayesian estimators – Hypothesis testing: testing framework, parametric testing, permutation test, multiple testing –

Mixture models: Gaussian mixture models, multinomial mixture models.

5. Linear regression: linear models, least-squares estimation, interval estimation in simple linear regression, overfitting – Multiple linear regression models: Estimation of model parameters, MLE – Nonlinear regression: Non linear least squares, transformation to linear model – Generalized linear models: logistic regression models, Poisson regression.

References

1. Michael Mitzenmacher and Eli Upfal; Probability and Computing, 2e, Cambridge University Press, 2017.

2. Alan Agresti, Christine A. Franklin and Bernhard Klingenberg; Statistics: The Art and Science ofLearning from Data, 4e, Pearson, 2017.

3. Sheldon M Ross; A First Course in Probability, 10e, Pearson, 2018.

4. Robert V Hogg, Joseph W McKean and Allen T Cralg; Introduction to Mathematical Statistics, 8e, Pearson, 2018.

5. Douglas C Montgomery, Elizabeth A Peck and G Geoffrey Vining; Introduction to Linear Regression Analysis, 5e, Wiley-Blackwell, 2012.