1st Semester

24PC1001 DATA SCIENCE (3-0-0)

Objectives: The objective of this course to comprehend the idea of Linear Methods for Regression and Classification, Model Assessment and Selection, Additive Models, Support Vector Machines(SVM).

Outcomes: Upon successful completion of this course students will able to

1. Learn the basic idea of regression models and least squares, Multiple regression, Logistic regression.
2. Know Model Assessment and Selection and Apply Boot strap methods, Bayesian approach.
3. Learn Boosting methods-exponential loss, Numerical Optimization via gradient boosting, Cluster analysis.
4. Apply Support Vector Machines (SVM), SVM for regression.

MODULE – I

Linear Methods for Regression and Classification: Overview of supervised learning, Linear regression models and least squares, Multiple regression, Subset selection, Ridge regression, least angle regression and Lasso, Linear Discriminant Analysis, Logistic regression .

MODULE – II

Model Assessment and Selection**:** Bias, Variance and model complexity, Bias-variance trade off, Optimism of the training error rate, Estimate of In-sample prediction error, Effective number of parameters, Bayesian approach and BIC, Cross-validation, Boot strap methods, Confusion matrix and ROC curves.
Dimensionality reduction (PCA, Kernel PCAfeature Selection, Non-negative matrix factorization and collaborative filtering).

MODULE – III

Additive Models, Trees and Boosting: Generalized additive models**,** Regression and classification trees, Boosting methods-exponential loss and AdaBoost, Numerical Optimization via gradient boosting, Examples (Spam data, California housing , New Zealand fish, Demographic data), random forests.
Unsupervised Learning, Cluster analysis (k-means, Hierarchical clustering, spectral clustering).

MODULE – IV

Support Vector Machines(SVM) and K-nearest Neighbor: SVM for classification, Reproducing Kernels, SVM for regression.
K-nearest –Neighbour classifiers(Image Scene Classification), Gaussian mixtures and EM algorithm.

Books Recommended:

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning-Data Mining, Inference and Prediction, 2nd Edition, Springer Verlag, 2009.
2. G. James, D. Witten, T. Hastie, R. Tibshirani: An Introduction to Statistical Learning with Applications in R, Springer, 2013.

Book for References:

1. C. M. Bishop: Pattern Recognition and Machine Learning, Springer,2006
2. L. Wasserman: All of Statistics.
3. T.M. Mitchell, Machine Learning, Mc. Graw Hill, 1997.