First Course Handout

MTH443A: Statistical & AI Techniques in Data Mining

1. **Objectives:** The objective of the course is to study and learn important statistical and AI/ML techniques used in data mining. The course would cover fundamental concepts and techniques associated with unsupervised and supervised modes of learning. Topics and concepts to be covered in the course includes: Dimension Reduction and Visualization Techniques; Chernoff faces, principal component analysis. Measures of similarity/ dissimilarity. Cluster Analysis: hierarchical and non-hierarchical techniques. Density estimation techniques; parametric and Kernel density estimation approaches. Association rule mining; Market Basket Analysis. Classification and Discriminant Analysis Tools; classification based on Fisher's discriminant functions, Bayes classifier, TPM and ECM minimizing classification rules, logistic discrimination rules, perceptron learning and Support Vector Machines. Statistical Modelling; estimation and inferential aspects of multiple regression, Kernel regression techniques. Tree based methods; Classification and Regression Trees. Neural Networks; multi-layer perceptron, feed-forward and recurrent networks, supervised ANN model building using back-propagation algorithm, ANN model for classification. Genetic algorithms, neuro-genetic models. Self-organizing Maps. At the end of the course, it is expected that a student who has done the course would know the mathematical foundation of the concepts and would also have the confidence to handle and analyse real life data.

2. **Prerequisites:** MSO201A/MSO205A/MTH418A or equivalent MTH department course(s)

3. **Course Contents:** (approx. 40 lectures)
   Introduction and preliminary concepts- 2 lectures
   Principal component analysis – 3
   Measures of similarity/ dissimilarity -2 lectures
   Cluster Analysis – 4 lectures
   Density estimation techniques-3 lectures
   Association rule mining-4 lectures
   Discriminant Analysis & Classification - 9 lectures
   Statistical Modelling- 3 lectures
   CART- 4 lectures
   ANN- 4 lectures
   GA-2 lectures

4. **Lecture/Lab hour**:

   Lecture: M (L03) W (L03) F (L03) 10:00-11:00
   Lab: Th 17:00-18:00 (Computer Lab 204H, DJ Building)

5. **Office Hours:** For any doubts, students are free to email the instructor and schedule a meeting to clear doubts.
6. **Evaluation Components:**
   Quizzes (02): 20 marks
   Midsem: 25 marks
   Endsem: 40 marks
   Course Project (group of 4): 15 marks
   Course project presentations along with project report submission will be held during the last week of instructions. Attendance during course project presentation would be mandatory.
   Final grade will be awarded out of a total of 100 marks.

7. **Course Policies:**
   Attendance: Students are **strongly advised to attend** the lecture and lab sessions.
   Honesty Practices: No act of dishonesty during exams or plagiarism in project report will be tolerated and prompt disciplinary action (**deregistration, awarding F grade, grade penalty, reporting to SSAC**, as may be deemed appropriate) will be taken.
   Withdrawal/drop: As per DOAA Guidelines.

8. **Books & References:**

a) T. Hastie, R. Tibshirani and J, Friedman: The elements of statistical learning: Data Mining, Inference and Prediction; Springer Series in Statistics, Springer.
b) R. A. Johnson and D.W. Wichern: Applied multivariate statistical analysis, Pearson.
c) Andrew Webb: Statistical Pattern Recognition, John Wiley & Sons.
d) S. S. Haykin: Neural Networks: A comprehensive foundation; Prentice Hall.