Course Outline

Course Title : Data Mining

Class : MPhil (Information Technology) Spring 2017

Course Instructor: Muhammad Bilal Shaikh
Email : mbilal.shaikh@usindh.edu.pk

Institute of Information Communication Technology
University of Sindh,Jamshoro

Text Book : *Data Mining : Practical Machine Learning Tools and Techniques 2nd Edition*
by *Ian Witten & Eibe Frank*
University of Waikato, New Zealand

Class Timings : 1400 to 1700 Every Thursday
Venue : Multimedia Lab

Course Modules

M1: Introduction: Machine Learning and Data Mining

I. Data Flood
II. Data Mining Application Examples
III. Data Mining and Knowledge Discovery
IV. Data Mining Tasks

Study: Course Notes,
Introduction to KDD (AI Mag 1996) (KDnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf)

M2: Machine Learning and Classification

I. Machine Learning and Classification
II. Examples
III. Learning as Search
IV. Bias
V. Weka

Study: W&E, Chapter 1.

M3. Input: Concepts, instances, attributes

I. What is a concept?
II. What is an example?
III. What is an attribute?
IV. Preparing the data

Study: W&E, Chapter 2.

M4. Output: Knowledge Representation

I. Decision tables
II. Decision trees
III. Decision rules
IV. Rules involving relations
V. Instance-based representation

Study: W&E, Chapter 3.

M5. Classification - Basic methods

I. OneR
II. NaiveBayes

Study: W&E, Chapter 4

M6: Classification: Decision Trees

I. Top-Down Decision Trees
II. Choosing the Splitting Attribute
III. Information Gain and Gain ratio

Study: W&E, Chapter 4

M7: Classification: C4.5

I. Handling Numeric Attributes
   A. Finding Best Split
II. Dealing with Missing Values
III. Pruning
   A. Pre-pruning, Post-Pruning, Estimating Error Rates
IV. From Trees to Rules

Study: W&E, Chapter 5

M8: Classification: CART

I. CART Overview and Gymtutor Tutorial Example
II. Splitting Criteria
III. Handling Missing Values
IV. Pruning
   A. Finding Optimal Tree

Study: CART Tutorial, CART Manual, www.salford-systems.com

M9: Classification: more methods

I. Rules
II. Regression
III. Instance-based (Nearest neighbor)

Study: W&E, Chapter 4

M10: Evaluation and Credibility

I. Introduction
II. Classification with Train, Test, and Validation sets
   A. Handling Unbalanced Data; Parameter Tuning
   B. *Predicting Performance
III. Evaluation on "small data": Cross-validation
   A. *Bootstrap
IV. Comparing Data Mining Schemes
   A. *Choosing a Loss Function

Study: W&E, Chapter 5.

M11: Evaluation - Lift and Costs

I. Lift and Gains charts
II. *ROC
III. Cost-sensitive learning
IV. Evaluating numeric predictions
V. MDL principle and Occam's razor

Study: W&E, Chapter 5.

M12: Data Preparation for Knowledge Discovery

I. Data understanding

II. Data cleaning
III. Date transformation
IV. Discretization
V. False "predictors" (information leakers)
VI. Feature reduction, leaker detection
VII. Randomization
VIII. Learning with unbalanced data

Study: Course notes

M13: Clustering

I. Introduction
II. K-means
III. Hierarchical

Study: W&E, Course notes

M14: Associations

I. Transactions
II. Frequent itemsets
III. Association rules
IV. Applications

Study: Course notes

M15: Visualization

I. Graphical excellence and lie factor
II. Representing data in 1,2, and 3-D
III. Representing data in 4+ dimensions
    A. Parallel coordinates
    B. Scatterplots
    C. Stick figures
    D. ...

Study: Course notes

M16: Summarization and Deviation Detection

I. Summarization
II. KEFIR: Key Findings Reporter
III. WSARE: What is Strange About Recent Events

Study: KEFIR book chapter and demo,
Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks, by Weng-Keen Wong et al (about WSARE system).

M17: Applications: Targeted Marketing and Customer Modeling

I. Direct Marketing Review
II. Evaluation: Lift, Gains

III. KDD Cup 1997
IV. Lift and Benefit estimation
V. KDD Cup 1998

Study: KDD Cup 1997 report, KDD Cup 1998 report,
G. Piatetsky-Shapiro, B. Masand, Estimating Campaign Benefits and Modeling Lift, Proc. KDD-99, ACM.

M18: Applications: Genomic Microarray Data Analysis

Study: SIGKDD Explorations Special Issue on Microarray Data Mining,

I. Capturing Best Practice for Microarray Gene Expression Data Analysis, G. Piatetsky-Shapiro, T. Khabaza, S. Ramaswamy, in Proceedings of KDD-2003.

M19: Data Mining and Society; Future Directions

I. Data Mining and Society: Ethics, Privacy, and Security issues
II. Future Directions for Data Mining
III. web mining, text mining, multi-media data
IV. Course Summary

Study: Knowledge Discovery in Databases vs. Personal Privacy Symposium,
editor Gregory Piatetsky-Shapiro, IEEE Expert, April 1995.
Bayardo & Srikant, Technological Solutions for Protecting Privacy, IEEE Computer, Sep 2003.