Business Data Mining

Course Objectives

To develop an understanding of the strengths and limitations of popular data mining techniques and to be able to identify promising business applications of data mining.

Syllabus

Machine Learning, Data Mining, Concepts, attributes and Output, Classification, Evaluation & Credibility and Lifts & Costs, Clustering, Association, Visualizations and Summarization and Applications.

Expected Outcomes

Upon completion of this course, the students will be able to: 1. Understand the basic theory and models used in machine learning. 2. Recognize when machine learning and data mining tools are applicable. 3. Understand and apply a wide range of clustering, estimation, prediction, and classification algorithms, including k-means clustering, and regression trees, the C4.5 algorithm, logistic Regression, k-nearest neighbor, multiple regression. 4. Understand and apply the most current data mining techniques and applications, such as text mining, mining genomics data, and other current issues. 5. Plan and execute successful machine learning and data mining projects, including selecting an adequate process for your specific task and avoiding the main machine learning pitfalls

References

1. S.K. Shinde and Uddagiri Chandrasekhar, Data Mining and Business Intelligence, DreamtechPress, 2015

2. Hand, Principles of Data Mining, PHI Learning Private Limited-New Delhi, 2004

3. Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques Paperback, Elsevier, 2010

4. Berry and Linoff. Mastering Data Mining, Wiley India Private Limited, 2008

5. Kumar, Data Mining: Principles and Techniques, Elsevier, 2012

6. Shmueli, Patel, and Bruce, Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Wiley-Blackwell, 2007.

7. Delmater and Hancock. Data Mining Explained, Digital Press, 2001

8. Introduction to KDD (AI Mag 1996) (KDnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf)

9. Weng-Keen Wong et al, Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks, https://www.aaai.org/Papers/AAAI/2002/AAAI02-034.pdf

10. G. Piatetsky-Shapiro, T. Khabaza, S. Ramaswamy, Capturing Best Practice for Microarray Gene Expression Data Analysis, in Proceedings of KDD-2003.
http://dl.acm.org/citation.cfm?id=956797 Study: Knowledge Discovery in Databases vs. Personal Privacy Symposium, editor Gregory Piatetsky- Shapiro, IEEE Expert, April 1995.
http://www.kdnuggets.com/gpspubs/ieee-expert-9504-priv.html

Units Topics

1 Machine Learning, Data Mining, Concepts, attributes and Output Data Flood; Data Mining Application Examples; Data Mining and Knowledge Discovery; Data Mining Tasks; Machine Learning and Classification, Examples; Learning as Search; Bias, Weka; Preparing the data;

Knowledge Representation - Decision tables; Decision trees; Decision rules; Rules involving relations; Instance-based representation.

2 Classification Basic Methods – OneR, NaiveBayes; Decision Trees - Top-Down Decision Trees, Choosing the Splitting Attribute, Information Gain and Gain ratio; C4.5 - Handling Numeric Attributes, Finding Best Split Dealing with Missing Values, Pruning, Pre-pruning, Post-Pruning, Estimating Error Rates, From Trees to Rules; CART - CART Overview and Gymtutor Tutorial Example, Splitting Criteria, Handling Missing Values, Pruning, Finding Optimal Tree; Other Methods – Rules, Regression, Instance-based (Nearest neighbour).

First Internal Examination

3 Evaluation & Credibility and Lifts & Costs

Definition, Classification with Train, Test, and Validation sets Handling Unbalanced Data; Parameter Tuning, Predicting Performance, Evaluation on "small data": Cross-validation

Bootstrap, Comparing Data Mining Schemes, Choosing a Loss Function;

Lifts & Costs - Lift and Gains charts, ROC, Cost-sensitive learning, Evaluating numeric predictions, MDL principle and Occam's razor;

Data Preparation for Knowledge Discovery -

Data understanding, Data cleaning, Date transformation, Discretization, False "predictors" (information leakers), Feature reduction, leaker detection, Randomization, Learning with unbalanced data.

Second Internal Examination

4 Clustering, Association, Visualizations and Summarization

Clustering – Definition, K-means, Hierarchical; Association – Transactions, Frequent itemsets, Association rules, Applications; Visualization – concept, Graphical excellence and lie factor,

Representing data in 1,2, and 3-D, Representing data in 4+ dimensions - Parallel coordinates, Scatterplots, Stick figures;

Summarization and Deviation Detection – Summarization, KEFIR: Key Findings Reporter, WSARE: What is Strange About Recent Events.

5 Applications Targeted Marketing and Customer Modelling - Direct Marketing Review, Evaluation: Lift, Gains, Lift and Benefit estimation;

Genomic Microarray Data Analysis – Definition and techniques;

Data Mining and Society; Future Directions; Data Mining and Society: Ethics, Privacy, and Security issues; Future Directions for Data Mining, web mining, text mining, multi-media data.

Final Examination