ADVANCED DATA MINING APPLICATIONS - MKTG5883

Syllabus

1. GENERAL INFORMATION:

Instructor name: Email:

Credit: 4 (3 lecture, 1 lab).

Prerequisite: CS 5423

2. COURSE INFORMATION:

- Course description:

Data mining is the science of discovering structure and making predictions in data sets (typically, large ones) and involves a good deal of both applied work (programming, problem solving, data analysis) and theoretical work (learning, understanding, and evaluating methodologies).

Providing knowledge about the potential exploitation of knowledge in the application database. Learners learn the knowledge exploitation process, common problems and association rules, sequence problem, the problem of classification, clustering and applications of data mining in real life.

- Course objectives:

To provide students with an understanding of the fields of statistics and computer science, adopt a statistical perspective the majority of the course (programming, problem solving, data analysis). Student should be able to tackle new data mining problems, by: (1) selecting the appropriate methods and justifying your choices; (2) implementing these methods programmatically and evaluating the results (applications); (3) explaining the results to a researcher outside of statistics or computer science.

3. BOOK AND MATERIALS:

- Required textbook and references:

- Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", Third Edition, Morgan Kaufmann Publishers, 2012.
- David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.
- Hastie T., Tibshirani R., and Friedman J., "The Elements of Statistical Learning", 2rd Ed.

4. GRADING PROCEDURES:

Assignments: 20%

Computer-based testing: 30%

Final Examination: 50%

5. COURSE OUTLINE:

Week	Topic
1, 2	**Unsupervised problems.** - Information retrieval and PageRank. Finding documents relevant to a given query. Representation by bag-of-words, measures of similarity (or distances), searching by similarity, evaluating performances, incorporating user feedback. Applying this to the web, and exploiting link structure via the PageRank algorithm.
3, 4	- Clustering. Dissimilarity and scatter. K-means clustering, K-medoids clustering. Hierarchical clustering, interpreting clustering trees, different linkages, top-down and bottom-up. Determining the number of clusters.
5, 6	- Dimension reduction. Principle component analysis. Directions of maximal variance, or equivalently, approximating a matrix by another matrix with a given (smaller) rank. Intepretation of principal components, usages, limitations. Multidimensional scaling, isomap, local linear embedding.
7, 8	- Correlation analysis. Correclation. Canonical correclation analysis. Zero correlation versus independence. Short comings of correlation for nonlinear relationships. Rank correlation, maximal correlation, distance correlation.
9, 10	**Supervised problems :** - Linear regression. Univariate and multivariate linear regression, viewing multivariate regression from simple univariate viewpoint. The assumptions underlying linear regression, and the corresponding optimality properties (best linear unbiased estimate) and inferential properties. Weighted linear regression.
11	- Regularized regression. The bias-variance tradeoff. Outperforming linear regression: shrinkage and ridge regression. The importance of variable selection. Best subset selection, forward and backwards stepwise regression, lasso, least angle regression.
12	- Model selection and validation. Training error and optimism. The

	validation set approach. Leave-one-out cross-validation, K-fold cross-validation. The one standard error rule. The bootstrap.
13, 14	- Classification. Nearest neighbor classification. Linear regression of an indicator vector. Linear discriminant analysis, reduced rank discriminant analysis and Fisher's linear discriminant. Logistic regression, and regularized logistic regression.
15	- Trees and boosting. Classification and regression trees. Boostrap sampling and bagging. Boosting, and the connection to regularized regression.

6. ACADEMIC INTEGRITY POLICIES:

- Student may not use Vietnamese languague in class, or will be reduced 2% final marks

- Be punctual to come and leave the class.

- Maximum cancellation time per semester is 6 hours per class.

Instructor's Signature