

620-472 Data Mining
















Data Mining refers to the management and analysis of large data sets. As it has matured it has developed a more statistical flavour, but Data Mining still owes much of its character to disciplines such as machine learning, pattern recognition, database design and high performance computing.

Data Mining became possible with the advent of large-scale data collection and the computing power necessary to process it. Data Mining involves all of the following steps

1. Data Warehousing
2. Data Cleaning
3. Data Description and Visualisation
4. Data Analysis and Interpretation

When working with very large data sets, all four steps pose serious problems. It is not uncommon to deal with databases with 10^6 records, each with anywhere from 10 to 10^4 variables. The storage required for such data is measured in giga-bytes and tera-bytes. However, this course deals only with step 4 of the Data Mining process: data analysis and interpretation.

Data Mining has been successfully applied in many commercial situations. In large competitive markets even small improvements in customer recruitment and retention or profit margins can result in big profits. However many Data Mining techniques have now become main-stream, and are regularly used in engineering, science, medicine, government and the social sciences, as well as in commerce. In some cases the techniques are useful for analysing small data sets as much as large ones. A survey on data mining applications conducted by KDnuggets in 2002 produced the following results in answer to the question: “Where do you plan to use data mining in 2002? (choose several)?” [198 votes, 433 choices]

Banking (56)	 13%
Biology/Genetics (36)	 8%
Direct Marketing/Fundraising (47)	 11%
eCommerce/Web (43)	 10%
Entertainment (3)	 1%
Fraud Detection (46)	 11%
Insurance (27)	 6%
Investment/Stocks (16)	 4%
Manufacturing (18)	 4%
Pharmaceuticals (24)	 6%
Retail (27)	 6%
Science (25)	 6%
Security (8)	 2%
Telecommunication (34)	 8%
Other (23)	 5%

In practice Data Mining comes down to a number of specific problem types and the techniques used to solve them. The following is one possible taxonomy, which is by no means exhaustive. The ones we are going to look at in more detail are emboldened.

- Association Rules
 - **Market basket analysis**
- Classification
 - **Trees**
 - **Logistic regression**
 - **Feed-forward networks**
 - Support-vector machines
 - Bayesian classifiers
- Regression
 - Trees
 - **Splines**
 - Additive models
 - Radial basis functions
- Clustering
 - **Hierarchical**
 - **k-means**
 - Self Organising Maps
 - Density based

The themes that run through the course are

1. Model fitting and selection and how to avoid overfitting
2. Actionability and interpretability of models
3. Scalable algorithms that can be used with very large data sets
4. How to accommodate high-dimensional data

Themes 1 and 2 are classical statistical considerations. Themes 3 and 4 are particular to Data Mining. If one does have very large amounts of data, then it can be more important to find something quickly, than to extract all the information present. It also makes it possible to use some of the data for validation, without compromising the model fitting.

Prerequisites

None required, however students would benefit from having completed an introductory probability or statistics unit.

Lecturer

Dr Owen Jones, room 221 Richard Berry building.

Assessment

20% coursework (weekly assignments) 80% exam

References

- T. Hastie, K. Tibshirani and J. Friedman. The Elements of Statistical Learning, Data Mining, Inference and Prediction. Springer, 2001.
- P.N. Tan, M. Steinback and V. Kumar. Introduction to Data Mining. Addison Wesley, 2005.
- P. Giudici. Applied Data Mining: Statistical Methods for Business and Industry. Wiley, 2003.
- M. Hegland. Data Mining Techniques. Acta Numerica, pp 313-355, 2001.
- B.D. Ripley. Pattern Recognition and Neural Networks. CUP, 1996.
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufman, 2001.
- I.H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd Edition. Morgan Kaufman, 2005.
- U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy Eds. Advances in Knowledge Discovery and Data Mining. MIT Press, 1996.
- G. Piatetsky-Shapiro and W. Frawley. Knowledge Discovery in Databases. AIII Press, 1991.

Web Resources

- UCI KDD Archive [<http://kdd.ics.uci.edu/>]. Dept. of Information and Computer Science, University of California, Irvine CA.
- KDnuggets [<http://www.kdnuggets.com/>]. G. Piatetsky-Shapiro.