DAT8 Course on Data Warehousing and Machine Learning

Michael O. Akinde & Uffe B. Kjærulff

26 January 2002

Abstract

This document describes the contents of the DAT8 course "Data Warehousing and Machine Learning (DWML)". Together, these two disciplines comprise what is known as *data mining* or *knowledge discovery in databases (KDD)*.

1 What is Data Mining?

The term *data mining* is actually a misnomer, as it refers to the activity of extracting knowledge/patterns/hypotheses[1] from data. A more appropriate term for data mining would be *knowledge mining from data* (Han & Kamber 2000).

Data mining is an interdisciplinary field, including disciplines from database systems, statistics, machine learning, visualization, and information science. Machine learning, in turn, includes subdisciplines like decision trees, neural networks, Bayesian networks, Instance-Based Learning, genetic algorithms, learning sets of rules, etc.

The popular term *knowledge discovery in databases (KDD)* is an often used synonym for data mining. Actually, the term "data mining" is more correctly viewed as a single step in the knowledge discovery process which consists of the following steps (Han & Kamber 2000):

1. Data cleaning (removing noise and inconsistent data)
2. Data integration (combining multiple data sources)
3. Data selection (retrieving the data relevant for mining)
4. Data transformation (preparing data for mining)
5. Data mining (extraction of data patterns)
6. Pattern evaluation (identification of interesting patterns)
7. Knowledge presentation (presentation of mined knowledge)

[1] We use the terms knowledge, patterns, and hypotheses interchangeably.

To avoid confusion, however, we shall use the term "data mining" in its broader sense (i.e., comprising the entire knowledge discovery process), which is consistent with the colloquial way the term is being used in the media.

The above seven steps represent a process where the user is left mostly passive. Another step might be added, representing the situation where the user interacts with mined knowledge, possibly providing additional (maybe hypothetical) information, to perform system-assisted reasoning:

8. Knowledge manipulation (reasoning based on patterns and user input)

For example, this step represents an important activity when the mined knowledge is represented in the form of a Bayesian network (see Section 3.1).

Steps 1–4 are performed using data warehousing techniques, whereas Steps 5–8 are performed using machine learning techniques.

2 Data Warehousing

What is data warehousing, and why do we need to know about it?

2.1 "Data in Jail" - the Data Access Crisis

If there is a single key to survival for any business in the 2000s and beyond, it is being able to analyze, plan and react to changing business conditions in a rapid fashion. To do this, top managers, analysts and "knowledge workers" in an enterprise need more and better information.

Information technology itself has made possible revolutions in the way that organizations today operate throughout the world. But the sad truth is that in many organizations despite the availability of more and more powerful computers on everyone's desks and communication networks that span the globe, large numbers of executives and decision makers can't get their hands on critical information that already exists in the organization.

Every day organizations large and small create billions of bytes of data about all aspects of their business, millions of individual facts about their customers, products, operations and people. But for the most part, this data is locked up in a myriad of computer systems and is exceedingly difficult to get at. This phenomenon has been described as "data in jail".

Only a very small fraction of the data that is captured, processed and stored in the enterprise is actually available to executives and decision makers. Essentially, by far the majority of enterprises today are "information rich", but "data poor".

2.2 Data Warehousing - Providing Data Access to the Enterprise

This led, in the 1990s to the development of significant new concepts and technology. The term Data Warehouse was coined by Bill Imnon in 1990, which he defined in the

following way: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" (Imnon 1995). Essentially, data warehousing has grown out of the repeated attempts on the part of various researchers and organizations to provide their organizations flexible, effective and efficient means of getting at the sets of data that have come to represent one of the organization's most critical and valuable assets.

The data warehouse industry is one of the fastest growing industries in software engineering. As the new technology gained wider use, data analysts and researchers soon realized that data warehousing can be used in a wide variety of data- and knowledge-intensive tasks. In 1996, Ralph Kimball provided a more general definition of a data warehouse in (Kimball 1996) as "a copy of transaction data specifically structured for query and analysis". Today, data warehousing is used in everything from simple customer profiling and other business analysis to weather forecasts, fraud detection, network diagnosis, etc. It is no coincidence that data warehousing and business intelligence is one of the fastest growing business sectors in the IT world, with projected growth estimates of becoming a 113 billion dollar industry in 2002.

2.3 The course

During the course, we will attempt to look at data warehousing from an industrial and technical, rather than a deeply theoretical perspective (though naturally, some theory will be involved).

2.3.1 Multi-dimensional Modelling

The multi-dimensional data model is one of the core foundations of data warehousing. We will look at the various parts of the multi-dimensional data model as proposed by Kimball: Facts, measures, star-schemas and snowflaking. We will also consider more advanced aspects of these models, such as changes in the dimensional model as well as hierarchies.

2.3.2 Building the Data Warehouse

We will consider the practical aspects of building a data warehouse. Data cleaning and integration, and optimization and tuning of a data warehouse (pre-aggregation, indexing and partitioning).

2.3.3 OLAP and Data Mining

We will look here at complex query processing techniques (i.e., various forms of data-cubes, pivoting, MF-cubes) for advanced OLAP, and how we can satisfy the special requirements of on-line analytical mining (OLAM), and the various aspects of data mining (associations, classification, prediction, clusterings and evolution analysis).

3 Machine Learning

Mitchell (1997) defines machine learning as "the study of computer algorithms that improve automatically through experience". That is, when provided with a set of cases[2] an algorithm identifies patterns in the data, and the more data (i.e., the larger the number of cases) the higher the confidence in the patterns identified.

When the machine-learning algorithm has completed its pattern extraction process, the patterns can be subjected to various kinds of evaluations, identifying the more interesting or significant patterns and their correlations.

3.1 Machine-Learning Algorithms

There is a large range of machine-learning algorithms of which only some of the most important shall be studied in this course. These are described briefly below. The descriptions are based on similar accounts by Mitchell (1997).

Decision Trees Decision tree learning is one of the most widely used and practical methods for inductive inference. It is a method for approximating discrete-valued functions that is robust to noisy data and capable of learning disjunctive expressions.

Neural Networks Artificial neural networks (ANNs) provide a general, practical method for learning real-valued, discrete-valued, and vector-valued functions. Algorithms such as BACKPROPAGATION use gradient descent to tune network parameters to best fit a training set of input-output pairs. ANN learning is robust to errors in the training data and has been successfully applied to problems such as interpreting visual scenes, speech recognition, and learning robot control strategies.

Bayesian Learning Bayesian reasoning provides a quantitative approach to weighing the evidence supporting alternative hypotheses, where the weights are probabilities. A statistical analysis of the data reveils a conditional independence structure among the variables described in the data as well as the strengths of the dependences between the variables. Both the (in)dependence structure and the strengths (expressed as conditional probability distributions) can be represented in a Bayesian network. Thus, a Bayesian network can be viewed as a statistical model of the data.

Instance-Based Learning In contrast to learning methods that construct a general, explicit description or model of the data, instance-based methods simply store the training examples (or cases). Generalizing beyond these examples is postponed until a new case must be classified. Instance-based learning includes nearest neighbor and locally weighted regression methods that assume instances can be represented in Euclidean space. It also includes case-based reasoning methods that use more

[2] A case is a vector of values of variables, and all cases in a set of cases given as input to a machine-learning algorithm range over the same set of variables.

complex, symbolic representations. A key advantage of this kind of lazy learning is that these methods can make local estimations, only taking some of the instance space into account.

Genetic Algorithms Genetic algorithms learn hypotheses through search based on simulated evolution. The search for an appropriate hypothesis begins with a population of initial hypotheses. Members of the current population give rise to the next generation through crossover and random mutation, which are patterned after processes in biological evolution. The hypotheses in the current population are evaluated relative to a measure of fitness, with the most fit hypotheses selected as seeds for the next generation.

Learning Sets of Rules One of the most expressive and human readable representations for learned hypotheses is sets of if-then rules. Such rules can be learned from a set of cases. One such important algorithm learns rules containing variables, called first-order Horn clauses. Because sets of first-order Horn clauses can be interpreted as programs in the logic programming language Prolog, learning them is often called inductive logic programming.

3.2 Evaluating Hypotheses

Being able to empirically evaluate the accuracy of hypotheses is fundamental to machine learning. Statistical methods are available for estimating hypothesis accuracy. Three questions are in focus:

Evaluate performance of hypotheses Given the observed accuracy of a hypothesis over a limited sample of data, how well does this estimate its accuracy over additional samples?

Compare accuracy of two hypotheses Given that one hypothesis outperforms another over some sample of data, how probable is it that this hypothesis is more accurate in general? Is an observed difference statistically significant?

Compare two learning algorithms When data is limited what is the best way to use this data to both learn a hypothesis and estimate its accuracy?

4 Lecture Plan

Below is a preliminary list of headings for the individual lectures of the course. The list is subject to change.

1. Introduction to Data Warehousing
   What is a data warehouse? The multi-dimensional data model.

2. Introduction to Machine Learning (Mitchell 1997, Ch. 1–2)

3. Advanced Multi-dimensional Modelling
   More about the multi-dimensional data model.

4. Decision Tree Learning (Mitchell 1997, Ch. 3)

5. Building the Data Warehouse I
   Data cleaning and integration. Physical Design of the Data Warehouse. Optimizing the data warehouse I: Practical pre-aggregation.

6. Artificial Neural Networks (Mitchell 1997, Ch. 4)

7. Building the Data Warehouse II
   Optimizing the data warehouse II: Indexing and Partitioning, Technical Design. MOLAP vs. ROLAP.

8. Bayesian Learning (Mitchell 1997, Ch. 6)

9. Complex OLAP
   Extending the power of SQL. Query Optimization.

10. Bayesian Networks and EM Algorithm (Mitchell 1997, Ch. 6)

11. Data Mining I
    Why Data Mining? Requirements of Data Mining

12. Instance-Based Learning (Mitchell 1997, Ch. 8)

13. Data Mining II
    Introduction to Data Mining

14. Learning Sets of Rules (Mitchell 1997, Ch. 10)

15. Evaluating Hypotheses (Mitchell 1997, Ch. 5)

References

Han, J. & Kamber, M. (2000). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

Imnon, W. H. (1995). What is a data warehouse, *Prism* **1**(1).

Kimball, R. (1996). *The Data Warehouse Toolkit*, Wiley.

Mitchell, T. M. (1997). *Machine Learning*, McGraw-Hill.