

CSE5230/DMS/2004/2

Data Mining - CSE5230

Data Mining and Statistics Machine Learning

CSE5230 - Data Mining, 2004 Lecture 2.1

Lecture Outline

- ◆ **Data Mining and Statistics**
 - ❖ A taxonomy of Data Mining Approaches
 - » Verification-driven techniques
 - » Discovery-driven techniques
 - Predictive
 - Informative
 - ❖ Regression
 - ❖ Exploratory Data Analysis
- ◆ **Machine Learning**
 - ❖ Concept Learning
 - ❖ Hypothesis Characteristics
 - ❖ Complexity of Search Space
 - ❖ Learning as Compression
 - ❖ Minimum Message Length Principle
 - ❖ Noise and Redundancy

CSE5230 - Data Mining, 2004 Lecture 2.2

Lecture Objectives

- ◆ **By the end of this lecture, you should:**
 - ❖ understand the link between the type of pattern being sought and the DM approach chosen
 - ❖ be able to give examples of verification and discovery-driven DM techniques, and explain the difference between them
 - ❖ be able to explain the difference between supervised and unsupervised DM techniques
 - ❖ give an example of the use of regression
 - ❖ describe the empirical cycle
 - ❖ explain the terms “complete” and “consistent” with respect to concept learning
 - ❖ describe the characteristics of a useful hypothesis
 - ❖ use the “kangaroo in the mist” metaphor to describe search in machine learning
 - ❖ explain the Minimum Message Length principle

The Link between Pattern and Approach - 1

- ◆ **Data mining aims to reveal knowledge about the data under consideration**
- ◆ **This knowledge takes the form of *patterns* within the data which embody our understanding of the data**
 - ❖ Patterns are also referred to as structures, models and relationships
- ◆ **The approach chosen is inherently linked to the pattern revealed**
 - ❖ e.g. you won't find a decision tree if you use clustering algorithm

The Link between Pattern and Approach - 2

- ◆ It is not expected that all the approaches will work equally well with all data sets
- ◆ *Visualization* of data sets can be combined with, or used prior to, modeling and assists in selecting an approach and indicating what patterns might be present

A Taxonomy of Approaches to Data Mining - 1

Verification-driven	Discovery-driven	
Query and reporting Statistical analysis	Predictive (Supervised)	Informative (Unsupervised)
	Regression Classification	Clustering Association Deviation detection (outliers)

Verification-driven Data Mining Techniques - 2

- ◆ **Verification data mining techniques require the user to postulate some hypothesis**
 - ❖ Simple query and reporting, or statistical analysis techniques then confirm this hypothesis
- ◆ **Statistics has been neglected to a degree in data mining in comparison to less traditional techniques such as**
 - ❖ neural networks, genetic algorithms and rule-based approaches to classification
- ◆ **Many of these “less traditional” techniques also have a statistical interpretation**

Verification-driven Data Mining Techniques - 3

- ◆ **The reasons for this are various:**
 - ❖ Statistical techniques are most useful for well-structured problems
 - ❖ Many data mining problems are not well-structured:
 - » the statistical techniques breakdown or require large amounts of time and effort to be effective

Problems with Statistical Approaches - 1

- ◆ **Traditional** statistical models often highlight linear relationships but not complex non-linear relationships (e.g. correlation)
- ◆ Exploring all possible higher dimensional relationships, often (usually) takes an unacceptably long time
 - ❖ the non-linear statistical methods require knowledge about
 - » the type of non-linearity
 - » the ways in which the variables interact
 - ❖ This knowledge is often not available in complex multi-dimensional data mining problems

Problems with Statistical Approaches - 2

- ◆ Statisticians have traditionally focused on *model estimation*, rather than *model selection*
- ◆ For these reasons less traditional, more exploratory, techniques are often chosen for modern data mining
- ◆ The current high level of interest in data mining centres on many of the newer techniques, which may be termed *discovery-driven*
- ◆ Lessons from statistics should not be forgotten. Estimation of uncertainty and checking of assumptions is as important as ever!

Discovery-driven Data Mining Techniques - 1

- ◆ **Discovery-driven data mining techniques can also be broken down into two broad areas:**
 - ❖ those techniques which are considered predictive, sometimes termed supervised techniques
 - ❖ those techniques which are termed informative, sometimes termed unsupervised techniques
- ◆ **Predictive techniques build patterns by making a prediction of some unknown attribute given the values of other known attributes**

Discovery-driven Data Mining Techniques - 2

- ◆ **Informative techniques do not present a solution to a known problem**
 - ❖ they present interesting patterns for consideration by some expert in the domain
 - ❖ the patterns may be termed “informative patterns”
 - ❖ often the goal is to approximate the probability distribution of the data set
- ◆ **Some of the main predictive and informative techniques are:**
 - ❖ Regression
 - ❖ Classification
 - ❖ Clustering
 - ❖ Association

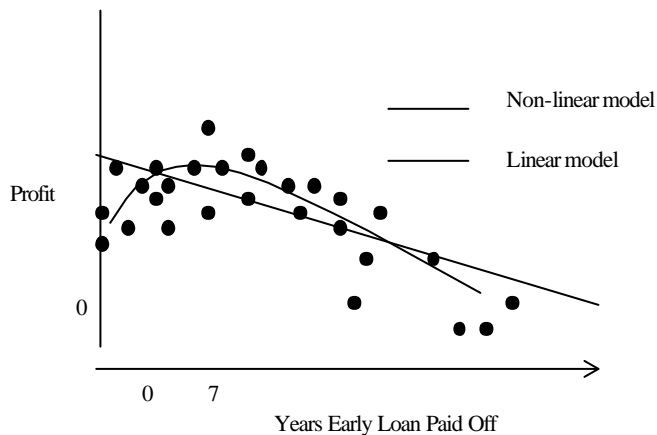
Regression

- ◆ Regression is a predictive technique which discovers relationships between input and output patterns, where the values are continuous or real valued
- ◆ Many traditional statistical regression models are linear
- ◆ Neural networks, though biologically inspired, are in fact non-linear regression models
- ◆ Non-linear relationships occur in many multi-dimensional data mining applications

An Example of a Regression Model - 1

- ◆ Consider a mortgage provider that is concerned with retaining mortgages once taken out
- ◆ They may also be interested in how profit on individual loans is related to customers paying off their loans at an accelerated rate
 - ❖ For example, a customer may pay an additional amount each month and thus pay off their loan in 15 years instead of 25 years
- ◆ A graph of the relationship between profit and the elapsed time between when a loan is actually paid off and when it was originally contracted to be paid off appears on the next slide

An Example of a Regression Model - 2



CSE5230 - Data Mining, 2004

Lecture 2.15

An Example of a Regression Model - 3

- ◆ The linear regression model (linear in the variables) does not match the real pattern of the data
- ◆ The curved line represents what might be produced by a non-linear model (perhaps a neural network, or linear regression on a known non-linear function which is linear in the variables) (see example on whiteboard)
- ◆ This curved line fits the data much better. It could be used as the basis on which to predict profitability
 - ❖ Decisions on exit fees and penalties for certain behaviors may be based on this kind of analysis

CSE5230 - Data Mining, 2004

Lecture 2.16

Exploratory Data Analysis (EDA)

- ◆ **Classical statistics has a dogma that the data may not be viewed prior to modeling**
 - ❖ aim is to avoid choosing biased hypotheses
- ◆ **During the 1970s the term Exploratory Data Analysis (EDA) was used to express the notion that both the choice of model and hints as to appropriate approaches could be data-driven**
- ◆ **Elder and Pregibon (1996) describes the dichotomy thus:**

“On the one side the argument was that hypotheses and the like must not be biased by choosing them on the basis of what the data seemed to be indicating. On the other side was the belief that pictures and numerical summaries of data are necessary in order to understand how rich a model the data can support.”

EDA and the Domain Expert - 1

- ◆ **It is a very hard problem to include “common sense” based on some knowledge of the domain in automated modeling systems**
 - ❖ chance discoveries occur when exploring data that may not have occurred otherwise
 - ❖ these can also change the approach to the subsequent modeling

EDA and the Domain Expert - 2

- ◆ **The obstacles to entirely automating the process are:**
 - ❖ It is hard to quantify a procedure to capture “the unexpected” in plots
 - ❖ Even if this could be accomplished, one would need to describe how this maps into the next analysis step in the automated procedure
- ◆ **What is needed is a way to represent meta-knowledge about the problem at hand and the procedures commonly used**

An Interactive Approach to DM

- ◆ **A domain expert is someone who has meta-knowledge about the problem**
- ◆ **An interactive exploration and a querying and/or visualization system guided by a domain expert goes beyond current statistical methods**
- ◆ **Current thinking on statistical theory recognizes such an approach as being potentially able to provide a more effective way of discovering knowledge about a data set**

Machine Learning

- ◆ “A general law can never be verified by a finite number of observations. It can, however, be falsified by only one observation.”

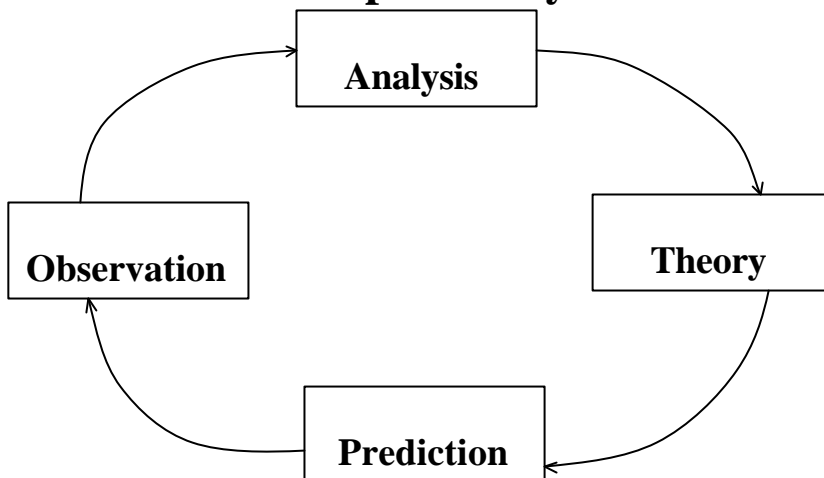
Karl Popper

- ◆ Many algorithms now used in data mining were developed by researchers working in machine learning and artificial intelligence, e.g.:
 - ❖ neural networks, self-organizing maps, decision trees, many clustering algorithms, Bayesian classifiers, Bayesian networks
- ◆ The patterns that machine learning algorithms find can never be definitive theories
 - ❖ They might be supported by the data sets used in training and testing, but does not mean that they are “true”
 - » Popper said that this was true of all scientific theories
 - not every one agrees with Popper
- ◆ Any results discovered *must* to be tested for statistical relevance

CSE5230 - Data Mining, 2004

Lecture 2.21

The Empirical Cycle



CSE5230 - Data Mining, 2004

Lecture 2.22

Concept Learning (1)

- ◆ **Example: the concept of a wombat**
 - ❖ a learning algorithm could consider the characteristics (features) of many animals and be advised in each case whether it is a wombat or not. From this a definition would be deduced.
- ◆ **The definition is**
 - ❖ *complete* if it recognizes all instances of a concept (in this case a wombat).
 - ❖ *consistent* if it does not classify any negative examples as falling under the concept.

Concept Learning (2)

- ◆ **An incomplete definition is too narrow and would not recognize some wombats.**
- ◆ **An inconsistent definition is too broad and would classify some non-wombats as wombats.**
- ◆ **A bad definition could be both inconsistent and incomplete.**

Hypothesis Characteristics

- ◆ **Classification Accuracy**
 - ❖ 1 in a million wrong is better than 1 in 10 wrong.
- ◆ **Transparency**
 - ❖ A person is able understand the hypothesis generated. It is then much easier to take action
- ◆ **Statistical Significance**
 - ❖ The hypothesis must perform better than the naïve prediction. Imagine a situation where 80% of all animals considered are wombats. A theory that all animals are wombats would be is right 80% of the time! But nothing would have been learnt about classifying animals on the basis of their characteristics.
- ◆ **Information Content**
 - ❖ We look for a rich hypothesis. The more information contained (while still being transparent) the more understanding is gained and the easier it is to formulate an action plan.

Complexity of Search Space

- ◆ **Machine learning can be considered as a *search problem*. We wish to find the correct hypothesis from among many.**
 - ❖ If there are only a few hypotheses we could try them all but if there are an infinite number we need a better strategy.
 - ❖ If we have a measure of the quality of the hypothesis we can use that measure to select potential good hypotheses and based on the selection try to improve the theories (hill-climbing search)
- ◆ **Consider the metaphor of the kangaroo in the mist** (see example on whiteboard).
 - ❖ This demonstrates that it is important to know the complexity of the search space. Also that some pattern recognition problems are almost impossible to solve.

Learning as a Compression

- ◆ We have learnt something if we have an algorithm that creates a description of the data that is shorter than the original data set
- ◆ A knowledge representation is required that is incrementally compressible and an algorithm that can achieve that incremental compression



- ◆ The file-in could be a relation table and the file-out a prediction or a suggested clustering

Types of Input Message (File-in)

- ◆ Unstructured or random messages
- ◆ Highly structured messages with patterns that are easy to find
- ◆ Highly structured messages that are difficult to decipher
- ◆ Partly structured messages
 - ❖ Most data sets considered by data mining are in this class. There are patterns to be found but the data sets are not highly regular

Minimum Message Length Principle

- ◆ MML says that the best theory to explain data set is the one that minimizes the sum of the length, in bits, of the description of the theory, plus the length of the data when encoded using the theory.

01100011001001101100011010101111100100110

00110011000011

110001100110000111

- ◆ i.e., if regularity is found in a data set and the description of this regularity together with the description of the exceptions is still shorter than the original data set, then we have found something of value.

Noise and Redundancy

- ◆ The distortion or mutation of a message is the number of bits that are corrupted
- ◆ making the message longer by including redundant information can ensure that a message is received correctly even in the presence of noise
- ◆ Some pattern recognition algorithms cope well with the presence of noise, others do not
- ◆ We could consider a database which lacks integrity to contain a large amount of noise
- ◆ patterns may exist for a small percentage of the data due solely to noise

References

- ◆ [BeL1997] Berry, J.A. & Linoff, G. **Data Mining Techniques: For Marketing, Sales, and Customer Support**, John Wiley & Sons, Inc., 1997
- ◆ [EIP1996] Elder, John F. and Pregibon, Daryl, A **Statistical Perspective on KDD**, In **Advances in Knowledge Discovery and Data Mining**, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Cambridge, Mass., 1996.
- ◆ [Han1999] Hand, D.J., **Statistics and Data Mining: Intersecting Disciplines**, SIGKDD Explorations, 1(1), pp. 16-19, 1999.
- ◆ [HMS2001] Hand, D., and Mannila, H., and Smyth, P. **Principles of Data Mining**, The MIT Press, Ch. 1, 2001.
- ◆ [GMP1997] Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. **Statistical Themes and Lessons for Data Mining**, *Data Mining and Knowledge Discovery*, 1(1), pp. 11-28, 1997.
- ◆ CSE5230 web site links page