Introduction to Data Mining
lecture notes

Lecturer: Amos Fiat

scribe: Asaf Koren

October 15 2002

Abstract

A short overview of some of the terms, problems and algorithms that fall under the definition of the evolving field of Data Mining.

1 Machine Learning

Data mining is closely related to Machine Learning. Machine learning can be divided into two categories: Supervised and Unsupervised.

1.1 Supervised Learning

The user gives the machine an example of the calculation it needs to perform. For example: giving samples of written letters to a hand writing recognition program.

1.2 Unsupervised Learning

Machine deduces general properties of data without additional external "explanations". For example: Clustering problem - a program sorts a given a set of data packets into related groups.

2 Market Basket Analysis and Association Rules

2.1 Market Basket Analysis

Market Basket Analysis considers data similar to the content of shopping baskets of a sample of shoppers coming out of a market.
An example for Basket data:

Apples	Beer	Chocolate
1	1	1
0	1	0
1	1	0
0	1	1

Each row represents one sampled basket content, "1" means item present "0" means absence from the basket. From this kind of table we may try and predict the content of other baskets. We can find the likeliness of some item to be present, or the relation of different products to one another. When analyzing the data we assume that the baskets content behaves according to a certain probability distribution and that the baskets listed are a valid sample.

2.2 Association Rules

We define:

$$\text{Support}(S) = \frac{\text{Number of lines containig } S}{\text{Number of total lines}}$$

where S is a set of possible items of the Market Basket. The support number is related to the probability of the set of items to be found in any basket. We can bound the probability of a difference larger than ϵ between the Support(S) and the Probability of S (Δ_S) using the Chernoff bound. According to this bound the probability decays exponentially with then number of sampled baskets.

We define:

$$\text{Confidence}(\alpha \to \beta) = \frac{\text{Support}(\alpha \cup \beta)}{\text{Support}(\alpha)}$$

Confidence can be interpreted as the likelihood of the items in the β set to be present in a basket in which a α set is found.

Examples:

- From the table:

$$\text{Support}(\{\text{Beer}, \text{Apples}\}) = \frac{2}{4} = 0.5$$

$$\text{Confidencae}(\text{Beer} \to (\{\text{Apples}, \text{Chocolate}\}) = \frac{1/4}{1} = 0.25$$

- The HIV positive "paradox"
  The probability of a False Positive result of a HIV test is 5% (meaning, a healthy person taking an HIV test has 5% of receiving the bad news from his doc). The statistics of HIV infected people is assumed to be 1 in 10000. If we had a perfect sample of 10000 people taking the HIV test (the people are baskets, HIV and HIV Positive are the possible items of the basket) we would get the following "surprising" information:

$$\text{Support}(\{\text{HIVPositive}\}) = 500/10000$$

$$\text{Support}(\{\text{HIV}, \text{HIVPositive}\}) \ = 1/10000$$

$$\text{Confidence}(\{\text{HIVPositive}\} \Rightarrow \{\text{HIV}\}) = 1/500$$

Meaning - if you are tested positive for HIV, the chances are $1/500$ that you are really infected.

2.3 problems and algorithms

Example of problems:

- Given a sample of baskets find all the sets of items which satisfy Support$(S) > P$.
- Find all the rules which satisfy Confidence(Rule) $> P$.

2.3.1 Naive algorithms

- Support problem - calculate the Support for each possible set, report those which exceed P. Complexity $O(k * 2^n)$, k being number of sampled baskets and n the number of type of items.
- Confidence problem - calculate the Confidence of each possible rule, report those with confidence exceeding P. The number of possible rules is 3^n (each item can be on the left or right side of the rule or not present at all). Complexity is $O(k * 3^n)$.

2.3.2 Apriori Algorithm

Agrawal and Srikant's (1994) Apriori algorithm solves these problems efficiently but with hyper-polynomial complexity.

3 Prediction problems

We are given a table of the following form:

X_1	X_2	X_3	...	X_n	Target

According to this data we should predict Target when a new set of parameter $X_1 \cdots X_n$ is given. When Target is an item out of a discrete set the problem is called Classification. When Target is Real the problem is called Regression. Examples:

- Classification problem -$X_1 \cdots X_n$ are physical parameters of a human (weight, height, age etc.) and target can be out of $\{$man, woman$\}$.
- Regression problem - Data is physical parameters, Target is number of years to live.

In both cases the problem will be to predict the target for a new set of numbers $X_1 \cdots X_n$.

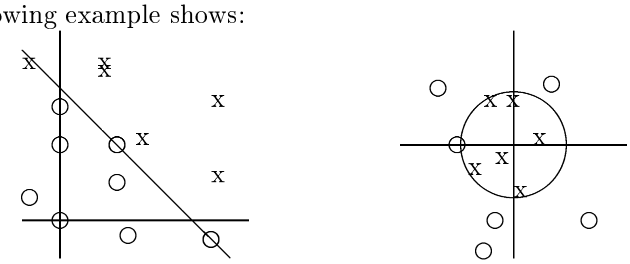
3.1 Linear Regression and relatives

One way for predicting targets is by using linear regression. Linear regression produces the parameters $(\alpha_1 \cdots \alpha_n)$ of a linear equation that predicts the target.

$$\sum_{i=1}^{n} \alpha_i X_i = \text{prediction}$$

For classification problem we can compare the prediction to a number (β) for deciding the class. These kinds of equations produce hyper-planes (line 2D, Surface 3D ..) that divide the space into two parts. The prediction for a Sample, $X_1 \cdots X_n$, is decided according to the side of the space in-which it resides.

In some cases, Hyper-planes can not serve as good predicators, as the following example shows:



For the left graph a line is a "good" predicator for the target, but in the right one a line will never suffice.

To solve the case of the right graph we transform the feature space of the data

$$(x, y) => (x, y, x^2, y^2, x^2 + y^2)$$

If we perform linear regression on the one-dimensional feature space of $(x^2 + y^2)$ we get a good predicator (the distance of the point).

This kind and much more advance transformation for using linear regression are some of the ideas behind the following approaches for prediction problems:

- Support Vector Machines (SVM)
- Kernel Techniques

The challenge is therefore finding the right transformation for the given data.

3.2 Generalization Error and Structural Minimization

Picking the right type of function for transformation can affect the error of the prediction. For example: picking a polynomial of a high order for predicting a linear phenomenon (with noise) can result in good matching for the samples but with terrible matching for new samples. The Generalization Error theorem gives a bound for the error of a prediction that depends on the family of functions we use. F_1 is a family of functions (e.g. first order polynomials)

$$F_1 = \{f_1, f_2, f_3 \cdots f_n\}$$

We define:

$$E_{\text{generalization}} = \text{The target prediction error of a new sample}$$

$$E_{\text{obj}} = \text{Error of the linear regression in respect to given data (samples)}$$

We use VC(Vapnik - Chervonenkis) dimension of F_1 which is a measurement of the computational capacity of a function family for the bound:

$$E_{\text{generalization}} \leq E_{\text{obj}} + \text{VC}(F)$$

The bound is useless when dealing with infinite dimension space since then $\text{VC}(F)$ is infinite. For this case we can use the Fat Shattering Dimension algorithm.

Structural minimization deals with the different hypothesis for F and their effect on the error bound.

4 Dimension Reduction

Reducing the complexity of prediction problems can be achieved by reducing the rank of the data matrix.
Suppose we have a reduced rank matrix. How do we measure the closeness of one matrix to its rank reduced sister matrix ?
The following are two different distance definitions we can use:
Frobenius norm:

$$|A - B|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - b_{ij})^2}$$

Operative norm:

$$|A|_1 = \max_{||x||_2=1} ||Ax||_2$$

Suppose we need to measure the distance between two matrices with different dimensions. How do we compare two matrixes with different dimension?
We can transform both to a canonic representation using SVM (Singular Value Decomposition).
We define the SVM of M as:

$$M = U \Sigma V^{\mathrm{T}}$$

where U and V are orthonormal and Σ is diagonal.
The Values of Σ are called the Singular Values of M.