Chapter 1

Introduction

Many large collections of structured information are stored in a relational format. Most often they are stored in a relational database, but other relational representations include object-oriented databases and semi-structured representations such as XML. These large collections of structured data are common: they occur in diverse areas including retail sales, telecommunications, insurance, medicine and scientific domains.

Recently there has been a growing interest in "mining" this data. In fact, this is the impetus for the newly emerging discipline of data mining. The goal in data mining is to discover unsuspected relationships and to summarize data in ways that are both understandable and useful to the owner of the data [Hand et al., 2001]. One way in which this goal can be accomplished is by building a statistical model describing the data.

The construction of statistical models is a task that has been well studied in both statistics and machine learning communities. There is a huge body of work on building both predictive models (models that are optimized for accurate prediction of a subset of the features) and descriptive models (models that describe the data or describe how the data is generated). Our focus here will be almost entirely on descriptive modeling, in particular density estimation. Many algorithms for density estimations exist. They vary in the types of models learned. Some algorithms construct parametric models, in that they assume a particular form for the distribution, and the task is parameter estimation. Examples include estimating the parameters of Gaussian densities, Poisson distributions or multinomial distributions. Other algorithms, for example kernel estimators, use non-parametric models and no assumptions are made about the functional form. A third category of models are mixtures of parametric models; most commonly these are mixtures of Gaussians but any parametric model may be chosen.

However, few of these density estimation algorithms are capable of handling data in its relational form. The input to the algorithm is assumed to be in the form of a collection of instances all of which have the same set of attributes. Depending on the problem, this input may be considered a random sample of the population, or it may be the entire population. However, in both cases, the assumption is made that the structure of the items in the input is identical. If the input is coming from a relational database, then we see that the the algorithms essentially only work on relational databases with a single table. However, the majority of relational databases have more than one table!

This thesis describes our approach to learning statistical models from relational data. Our goal is to build statistical models that are able to capture the important inter-table correlations in the data and exploit the information available in relational structure of the data. Our hope is that these statistical models will more accurately capture the dependencies and correlations in the domain than previous approaches and prove useful for both data exploration and summarization in relational domains. Our contributions include a collection of statistical models applicable in relational settings and automated induction algorithms for the models. In addition to the theoretical descriptions of the models and learning algorithms, we demonstrate their application on real-world databases and provide some evidence that our hope has been met.

1.1 Motivation

Consider a simple relational database that describes customer transactions. The database may include a number of tables. Suppose we have a customer table containing customer information including demographic data such as income or education, an item table describing the items that may be purchased, and a table describing purchases, linking customers to the items that they have purchased. From this data, we may be interested in figuring out which items customers tend to buy together (basket analysis or affinity analysis), or the categories of customers that make certain types of purchases (customer profiling). The data is clearly relational: there is a many-many relationship between customers and the items they purchase.

Unfortunately, few inductive learning algorithms are capable of handling data in its relational form. Most are restricted to dealing with a flat set of instances, with a homogeneous set of attributes. To use these methods, one typically "flattens" the relational data, removing its richer structure, treating it as a collection of fixed-length vectors of attribute-value pairs stored in a single table. In our case, the data may be flattened into a table that records for each purchase, the characteristics of the customer and the characteristics of the item purchased. Alternatively, we may have a single row for each customer with an attribute for each potential purchase. A third alternative is to have a row for each item, and an attribute for each potential consumer. These last two approaches have the disadvantage that we must fix either the number of customers or the number of items in advance.

This flattening has several important adverse implications. From a database perspective, we will either need to materialize the flattened view, storing it as a single table, or perform the required join operations each time we query the database for statistics. Each of these has disadvantages. Converting the data to a single table is undesirable because it means we cannot mine our database directly. It introduces redundancy and potentially high consistency-maintenance overhead. Databases are typically organized to avoid redundancy; in order to save space and reduce maintenance overhead, tables are normalized to remove repeated information. Converting a set of tables into a single table can introduce duplication and we lose our compact representation. Alternatively, executing the join each time we need to gather statistics from the database can be expensive and time consuming. In addition, we may need to build auxiliary indexes in order to perform the required joins repeatedly.

More importantly, however, this flattening process loses information which might be crucial in understanding the data. From a statistical perspective, storing the data in a single table can corrupt the integrity of our results. First there is the danger of introducing statistical skew when we flatten the data. In our transaction data, if we flatten the data into a vector of people and item attributes, then people who make a lot of purchases will be over-represented in this data. If we try to infer characteristics of people by computing statistics from this flattened data, we will get incorrect results.

Another form of information loss occurs if we lose the links between related objects; we may be able to infer important properties of an object based on the objects to which it is linked. For example, if two people live in the same household, they may make similar purchases. Obviously we cannot make these inferences if the links are not even modeled. A third form of information loss occurs when we repeatedly copy related objects without maintaining their shared identity. Suppose two purchases are made by the same person. If we simply make two copies of the person, we are making a false independence assumption. If we make inferences about unobserved attributes of the person, we have lost the requirement that these inferred attributes must be equal for each copy of the person. All of these drawbacks severely limit the ability of current statistical methods to mine relational databases.

1.2 Our Approach

In this thesis we provide the tools for constructing statistical models from relational data. Our goal is to learn structured probabilistic models, that represent statistical correlations both between the properties of an entity and between the properties of related entities. These statistical models can then be used for a variety of tasks including knowledge discovery, exploratory data analysis, data summarization and anomaly detection.

The starting point for our work is the structured representation of probabilistic