Research School of Computer Science, Australian National University

COMP3420 Advanced Databases and Data Mining

# Tutorial 1

1. Before we go into data mining, let us review some concepts in database.

   (a) In relational database, what are base relations? What are derived relations?

   (b) What kinds of joins are commonly used in SQL?

   (c) Which one(s) of those joins is (are) considered to be relatively more efficient?

   **Answer:** Base relations are those that store data, and implemented as tables, whereas derived relations are those that do not store data, but are computed by relational operations[1].

   Commonly used joins are inner join (natural join, cross join), outer join (left outer join, right outer join, full outer join), and self join, etc..[2]

   The most naive way to implement inner join is to compute the Cartesian product of two tables and select the rows that satisfy the join predicate. Thus the computation space is large. Whereas outer joins do not involve computing Cartesian products, but just reserve one (or both) table(s) and extend the other (or both) table(s) with corresponding data, fill NULL if there is no match. This requires less computation time and space. Self joins can be any form of join (inner, outer...) that apply a table to itself. There are more intelligent ways to implement those joins, and the methods vary in different database systems.

2. What is the difference and similarity between data warehouse and database?

   **Answer:** Data warehouse is a repository of information collected from multiple sources, stored under a unified schema. The aim of data warehouse is to support decision making. The data in a data warehouse are organized around major subjects, stored to provide information from a historical perspective, and are typically summarized.

   Databases mainly deal with current data. The tasks of databases include indexing and hashing using primary keys, searching for particular records. Transactional databases support concurrent processing of multiple transactions and provide recovery mechanisms.

[1] see `http://en.wikipedia.org/wiki/Relational_database`

[2] see `http://en.wikipedia.org/wiki/Join_(SQL)`

Both data warehouse and database are used for data management, they both perform data analysis and query processing. Databases provide sources of data for data warehouses, and the latter carry out high level complicated analysis and summarization on huge volume of data.

3. In many data mining techniques, outliers are considered as noise or exceptions. But in some applications, unusual records may provide useful information.

   (a) What is outliers analysis?

   (b) What are common methods for measuring outliers? When should they be used?

   (c) Briefly give a scenario where outlier analysis is useful.

   **Answer:** (a) Outlier analysis is the analysis and detection of data objects that do not comply with the general behavior or model of the data.
   (b) There are mainly three methods for measuring outliers, namely, statistical methods, distance based methods and density based methods. Statistical methods are used when the data is assumed to be generated from a statical model, *e.g.*, assuming the data are generated from a Gaussian distribution, then statistical test will be used to test if the data come from the assumed model. When distance is used as the similarity measure, distance based methods are used, to find data points that are far from the centroid of the data. Finally, density based methods are used when considering the data coming from different regions, each regions have different densities. These are local methods, and can avoid the problem of affecting by far away data (consider different regions separately).
   (c) One example is in the bank transaction data analysis, where an outliers indicates an anomalous transaction (possibly caused by a hacker). We need to be able to detect this in order to avoid money loss.

4. Bonferroni's principle has it that (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap. A few years a go the US government publicized a project TIA (Total Information Awareness), [3] which would be achieved by creating huge databases to gather and store perseonal information of everyone in the US without any requirement of a search warrant. It is used for identifying suspicious activities, connections between people, and so on ... We hypothesize that this is a living example of

[3] see `http://en.wikipedia.org/wiki/Information_Awareness_Office`

the Bonferroni's principle, that it looks for too many vague connections that are sure to be bogus and thus violate innocent's privacy.

Consider the situation where we want to find (unrelated) two persons who stay at the same hotel on the same day for at least twice. We track $10^9$ people in 1000 days, and find that *on average*, each person stays at a hotel for 1% of the time (10 days out of 1000), and each hotel holds 100 people in a day. Suppose that everyone behaves randomly, will the data mining detect anything suspicious? [4]

(a) Can you estimate how many hotels are there in this situation?
**Answer:**

$$number\ of\ hotels = \frac{10^9 \times 10}{100 \times 1000} = 10^5$$

(b) What is the probability that two given persons $p_1$ and $p_2$ stay at the same hotel on a given day $d$?
**Answer:**

$$P_1 = \frac{1}{100} \times \frac{1}{100} \times \frac{1}{10^5} = 10^{-9}$$

(c) What is the probability that $p_1$ and $p_2$ stay at the same hotel on given days $d_1$ and $d_2$?
**Answer:**

$$P_2 = 10^{-9} \times 10^{-9} = 10^{-18}$$

(d) What is the number of pairs of days?
**Answer:**

$$N_1 = \binom{1000}{2} = \frac{1000!}{(1000-2)! \times 2!} = 500 \times 999 \approx 5 \times 10^5$$

(e) What is the The number of pairs of persons is?
**Answer:**

$$N_2 = \binom{10^9}{2} = \frac{10^9!}{(10^9-2)! \times 2!} \approx 5 \times 10^{17}$$

[4] This question originates from Section 1.2.3 of the book Mining of Massive Dataasets by Anand Rajaraman and Jeffrey Ullman. `http://infolab.stanford.edu/~ullman/mmds.html`

(f) What is the expected number of ”suspicious” pairs of persons?
**Answer:**

$$P_2 \times N_1 \times N_2 = 10^{-18} \times 5 \times 10^5 \times 5 \times 10^{17} = 250000$$

Although the probability is really low, we still end up looking for too many suspicious people!

5. What is the difference between discrimination and classification? characterization and clustering? classification and prediction?

   **Answer:** Data discrimination is a comparison of the the general features of a target class data object with the general features of objects from one or a set of contrasting classes. Classification is the process of finding a set of models (functions) that describe and distinguish data classes or concepts. The difference between these two is that discrimination compares the general features of the target class data to that of contrasting classes, whereas in classification the goal is to build models that describe and distinguish data classes from each other.

   Data characterization is a summarization of the general characteristics or features of a target class of data. While in clustering the objects are to be grouped together based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. So the difference is that in characterization the output is a set of general features, whereas in clustering it is a set of object classes.

   Classification is the process of finding a set of models that describe and distinguish data classes or concepts, while in prediction we need to analysis the data set and produce a function that can take a set of attributes as the input and predict the missing or unknown data value. In a word, classification is to predict class labels, while prediction is to predict missing data values.

6. **Getting to Know your Data**: We have the Adult dataset[5] from UCI Machine Learning repository, each data[6] has 14 attributes collected from one person, *e.g.*, age, workclass, fnlwgt, education, education-num, marital-status, $\cdots$. One question of interest is to identify which attribute(s) is indicative of whether a given person makes over 50K a year.

   (a) Have a look at the data, among these 14 attributes, which attributes is/are *nominal attribute(s)*? which is/are *binary attribute(s)*? *Ordinal attribute(s)*? *numeric attribute(s)*?

[5] http://archive.ics.uci.edu/ml/datasets/Adult

[6] http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

(b) Among these four kinds of attributes, which one/ones has/have orders? which one/ones does/do not?

(c) Think of a way to calculate the similarity between ordinal attributes.

(d) Calculate the means and variances of the *age* attribute of the first five people (by manual calculation).

(e) Calculate the means and variances of the *age* attribute of these two kinds of people (those with income over and below 50K, respectively). You can use any programming language of your choice, or load data into a spreadsheet such as the Google spreadsheet[7]. Do you think there is any relationship between the age and the income of a person?

**Answer:**

(a) **Nominal:** workclass, marital-status, occupation, relationship, race, native-country
**Binary:** sex
**Ordinal:** education
**numeric:** age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week

(b) *Ordinal* and *numeric attributes* have orders.

(c) One way is to discretized the *ordinal* attributes, mapping each value to an integer starting from 0. The lowest value in the *ordinal* attributes is equal to 0, the second lowest is equal to 1, and so on. Then the similarity between any two values of the *ordinal* attributes is defined to be the distance between the corresponding integers.

(d) mean = 41.6; variance = 101.3.

(e)
- **Less than 50K**: mean = 36.78, variance = 196.55.
- **Larger than 50K**: mean = 44.25, variance = 110.64.

It indicates elder people usually get higher pays.

7. Suppose a group of 12 *employee age* records has been sorted as follows.

$$22, 23, 25, 26, 29, 33, 34, 36, 39, 42, 51, 53$$

Partition them into three bins by each of the following methods.

(a) equal-frequency (equi-depth) partitioning

(b) equal-width partitioning.

[7] http://www.docs.google.com/

Can you use the histogram method to allocate the data in the set to 3 equal intervals?

**Answer:** We allocate the data into 3 bins with $\frac{12}{3} = 4$ members as follows.

(a) Equal-frequency partitioning

- Bin 1: 22, 23, 25, 26
- Bin 2: 29, 33, 34, 36
- Bin 3: 39, 42, 51, 53

(b) Equal-width partitioning: $min = 22$, $max = 53$, $N = 3$, and $W = \frac{max-min}{N} = 10$. Therefore, the three intervals are [22, 32], [33, 43], [44, 54].

- Bin 1: 22, 23, 25, 26, 29
- Bin 2: 33, 34, 36, 39, 42
- Bin 3: 51, 53

8. **Rattle data exploration**

This practical part consists of a self-guided step-by-step component with detailed instructions provided on the COMP3420 Wattle site at:

`https://wattlecourses.anu.edu.au/mod/page/view.php?id=833747`

You will need to use a version of the **adult** data set which has a header line (with attribute names) added. This file is available on the COMP3420 Wattle site `http://wattlecourses.anu.edu.au/mod/resource/view.php?id=556696`

The aim of this practical part is to become familiar with the **Rattle** data mining tool, and explore the **adult** data set with a variety of techniques.

We expect you to work on this part of the tutorial for around 40 to 50 minutes. If you have any questions make sure you ask one of the tutors.

9. **Further Data cleaning**: Look at a subset of the Adult dataset with dirty attribute values `https://docs.google.com/spreadsheet/ccc?key=0Asd1NNaMfJX4dDJEWGJNV0ZFWWNF0GV0Sk9wc3VBZWc#gid=0`.

(a) What are the common methods for dealing with missing values in data cleaning?

(b) Find out three dirty attributes in this example. What are they?

(c) Use the rule *Use the attribute mean for all samples belonging to the same class as the given tuple* to fill in the dirty/missing value in *age* attribute.

(d) Use the rule *Fill in the missing value manually* to fill in the dirty/missing value in *sex* attribute.

(e) Use the rule *Ignore the tuple* to deal with the dirty value in *native-country* attribute.

**Answer:**

(a) The methods include:

- Ignore the tuple.
- Fill in the missing value manually.
- Use a global constant to fill in the missing value.
- Use a measure of central tendency for the attribute (*e.g.*, the mean or median) to fill in the missing value.
- Use the attribute mean or median for all samples belonging to the same class as the given tuple.
- use the most probable value to fill in the missing value.

(b) There are three, namely, the 3rd row of attribute *age*; the 9th row of attribute *sex*; the 10th row of attribute *native-country.*

(c) 39.

(d) Male/Female?

(e) Delete.