Data Mining: Homework #1

Due Date: Sep. 22, 2010

Shumin Guo

September 27, 2010

Introduction - briefly answer the following questions

1. What is Data Mining? What is the relation between data mining and KDD?

   ANSWER: Generally, Data Mining is the process of analysing data from different perspectives and summarizing it into useful information. And technically, data mining is the process of finding correlations and patterns among dozens of fields in large relational database. Still another definitions is : a non-trivial process of identifying valid, novel, useful and ultimately understandable patterns in data. KDD (Knowledge Discovery from Database/Datasets) is a more precise definition of data mining.

2. Describe any three challenges of data mining.

   ANSWER:

   - Increasing data dimensionality and data size.
   - Various data forms.
   - New data types.

3. What is the need for data mining?

   ANSWER:

   - More and more data are generated.
   - Although a lot of data are generated, there is still a gap between stored data and knowledge, the transition won't occur automatically.
   - Due to the volume and a lot of other challenges, manual data processing is almost impossible.
   - The development of computer science and engineering generated high demands on the knowledge within the stored data.
   - Fields such as finance, business etc. have high demands on seeking knowledge from the massive data.

Data Pre-processing

1. In real-world data, tuples (instances) with missing values for some attributes are a common occurrence. Briey describe various methods for handling this problem.

   ANSWER:

   - Leave as is and treat missing value for each attribute A as a new value of A.
   - Ignore/Remove the instances with missing value.

- Manual fix - assign a value with implicit meaning.
- Statistical methods to convert missing values - majority, most likely, mean nearest neighbor.

2. Suppose the data for analysis includes the attribute Age. The age values for the data tuples (instances) are (in increasing order): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Use binning (by bin means) to smooth the above data, using 5 bins and equi-density. (If the number of data tuples is not a multiple of 5, evenly spread the extra tuples in the last few bins.) Illustrate your steps, and comment on the effect of this technique for the given data.

ANSWER: Equi-density binning expects to distribute data samples into bins evenly, which means the number of data samples should be the same for each bin, but for this example, we have 27 sample data to be distributed into 5 bins. So, we will need to assign 6 data samples for two bins and the other bins all have 5 samples.
As the data samples have been sorted increasingly, there is no need to sort it. We get the following bin distribution table.

ID#	Samples	Bin
1	13, 15, 16, 16, 19	$(-\infty, 19)$
2	20, 20, 21, 22, 22	$[19, 23)$
3	25, 25, 25, 25, 30	$[23, 31)$
4	33, 33, 35, 35, 35, 35	$[31, 35)$
5	36, 40, 45, 46, 52, 70	$[35, +\infty)$

Table 1: Data Distribution For Age Data with 5 Bins and Equi-Density Method.

COMMENTS: As in this example, age data is integer value, so I simply picked the bin boundary to be the round down integer value of the mean of the two sample values. E.g., for bin #1 and #2 the boundary should be $\frac{19+20}{2} = 19.5$, I have rounded it to 19.

Data binning is used to smooth data, and seperate data into different classes. So, the granularity of the data is increased, this can reduce or smooth the noise contained within the data, but the panelty is that it can also potentially affect the data that is of interest.

3. Using the data for Age in Question 2, answer the following:

- Use min-max normalization to transform the value 35 for age into the range [0.0; 1.0].

  ANSWER:
  For min-max normalization, we have formula:
  $v' = \frac{(v - minA) \times (newMaxA - newMinA)}{(maxA - minA)} + newMinA$
  So, for the above sample data, we have :
  $v' = \frac{(35-13) \times (1.0-0.0)}{70-13} + 0.0 \approx 0.386$
- Use z-score normalization to transform the value 35 for age. The standard deviation of the ages is 12.94.

  ANSWER:
  For zero normalization, we have formula:
  $v' = \frac{v - meanA}{stdevA}$, and the calculated mean value is $meanA \approx 29.96$, so applying this formula, we have:
  $v' = \frac{35-29.96}{12.94} \approx 0.39$.
- Use normalization by decimal scaling to transform the value 35 for age.

ANSWER:
For decimal scaling, we have formula:
$v' = \frac{v}{10^k}$, of which k is the smallest number so that $v' \in [-1, 1]$. Applying this formula to this question, we can get:
$k = 2$, and so $v' = \frac{35}{10^2} = 0.35$.

- Comment on which method you would prefer to use for the given data, giving reasons as to why.

ANSWER: Data normalization is one of the most important steps in data preprocessing, especially when dealing with parameters of different units and scales. Therefore, all parameters should have the same scale for a fair comparison between them. Actually, all the methods we used above have drawbacks, while scaling the normal data into smaller intervals, those outliers will also be affected and scaled into a scaler that is not even distinguishable to the worst case.
Also, different normalization methods are suitable for different data distribution properties. So, we need to justify before-hand what kind of distribution our sample data conforms to, and then choose the best normalization method for this kind of distribution.
Intuitively, for this example the age data will conform to Gaussian distribution, which makes zero-score normalization the most suitable method.

4. We have the following (attribute-value,class) pairs [(0,P), (4,P), (12,P), (14,N), (16,N), (16,N), (18,P), (24,N), (28,N)]. Consider two possible splits $v_1 = 13$ and $v_2 = 15$ for discretizing the above data attribute. Use entropy-based binning by binarization to find the information gain for each split and decide which split is better. [In real life, we need to consider all potential splits.]

ANSWER:
For split $v_1 = 13$:
v_1 devides the whole data set into two sets:
S_1 when $value \leq v_1$ and S_2 when $value > v_1$, and specifically, we have:
$S_1 = [(0, P), (4, P), (12, P)]$ and $S_2 = [(14, N), (16, N), (16, N), (18, P), (24, N), (28, N)]$
And the information for this split is:
$IS(S_1, S_2) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$
$= \frac{3}{9} \times 0 + \frac{6}{9} \times (-\frac{1}{6} log_2 \frac{1}{6} - \frac{5}{6} log_2 \frac{5}{6})$
$= 0 + \frac{2}{3} \times (0.4309 + 0.2191)$
$= 0.43$
The information gain of the split is:
Gain(v,S) = Entropy(S) - $IS(S_1, S_2)$
$= -\frac{4}{9} log_2 \frac{4}{9} - \frac{5}{9} log_2 \frac{5}{9} - 0.43$
$= 0.52 + 0.47 - 0.43$
$= 0.56$

For split $v_1 = 15$:
v_2 devides the whole data set into two sets:
S_1 when $value \leq v_2$ and S_2 when $value > v_2$, and specifically, we have:
$S_1 = [(0, P), (4, P), (12, P), (14, N)]$ and $S_2 = [(16, N), (16, N), (18, P), (24, N), (28, N)]$
And the information for this split is:
$IS(S_1, S_2) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$
$= \frac{4}{9} \times (-\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4}) + \frac{5}{9} \times (-\frac{1}{5} log_2 \frac{1}{5} - \frac{4}{5} log_2 \frac{4}{5})$
$= \frac{4}{9} \times (0.3113 + 0.5) + \frac{5}{9} \times (0.4644 + 0.2575)$
$= 0.3606 + 0.401$
$= 0.7616$

The information gain of the split is:
Gain(v,S) = Entropy(S) - $IS(S_1, S_2)$
$= -\frac{4}{9} log_2 \frac{4}{9} - \frac{5}{9} log_2 \frac{5}{9} - 0.7616$
$= 0.52 + 0.47 - 0.7616$
$= 0.2284$

The split v_1 has higher information gain, so it is a better choice.

5. Consider the following data in the attribute-instance format. We need to perform feature selection using Sequential Forward Selection (SFS) with consistency (cRate) as the subset evaluation metric. What will be the selected feature subset at the end of the second iteration?

F1	F2	F3	C
1	1	1	1
1	1	1	1
1	0	1	1
0	1	0	0
1	0	1	0
0	0	1	1
0	0	0	1
1	1	0	0

Table 2: Data Table for SFS

ANSWER: According to the definition of SFS, we need to sequentially iterate all the features of this data set, in the first iteration we choose feature F1, and the second iteration we will choose F2 as the feature. We can get the following consistency result:

F1	C	Consist	F2	C	Consist	F3	C	Consist
1	1	✓	1	1	✓	1	1	✓
1	1	✓	1	1	✓	1	1	✓
1	1	✓	0	1	✓	1	1	✓
0	0	✗	1	0	✗	0	0	✓
1	0	✗	0	0	✗	1	0	✗
0	1	✓	0	1	✓	1	1	✓
0	1	✓	0	1	✓	0	1	✗
1	0	✗	1	0	✗	0	0	✓

Table 3: Consistency table for iterator one

So, the cRate after first iteration is $cRate_{F1} = \frac{5}{8}$ $cRate_{F2} = \frac{5}{8}$ and $cRate_{F3} = \frac{3}{4}$. So in this iteration we will choose F3 as the starting point for next iteration.

F1	F3	C	Consist	F2	F3	C	Consist
1	1	1	✓	1	1	1	✓
1	1	1	✓	1	1	1	✓
1	1	1	✓	0	1	1	✓
0	0	0	✗	1	0	0	✓
1	1	0	✗	0	1	0	✗
0	1	1	✓	0	1	1	✓
0	0	1	✓	0	0	1	✓
1	0	0	✓	1	0	0	✓

Table 4: Consistency table of feature two

So, the cRate after the second iteration is $cRate_{F1,F3} = \frac{3}{4}$ $cRate_{F2,F3} = \frac{7}{8}$.

$\because$ $\text{cRate}_{F1,F3} < \text{cRate}_{F2,F3}$

$\therefore$ the selected feature after the second iteration will be $\{F2, F3\}$.