Statistical Methods in Data Mining

Exercises

Professor Dr. Gholamreza Nakhaeizadeh

Exercises part 1

1. Discuss the relation between Data Mining and

- **Statistics,**
- **Machine Learning**
- **Database Technology**

2- How can you convince the head of a pharmaceutical company about the importance of Data Mining for his company ? Describe your argumentation.

3- Give an example for the case that a discovered pattern is informative although it is not new and is already known.

4- Besides WEKA, there are other Data Mining Tools that are available on Internet to every one for free. Conduct a search in Internet, find two of them and provide a short tool description for each of them.

5- Besides the three large databases mentioned in the course find two others and describe them in summary.

6- Elaborate the difference between implicit and explicit patterns.

7- Besides Statistics, AI, Database technology, visualization and privacy mention two other disciplines (or concepts) that can be used in Data Mining

8- Elaborate three data mining applications in which privacy issues are important

9- Describe the importance of pattern understandability by an example.

10- Elaborate three examples of the application of data mining in finance

Exercises part 2

1- Which phases of the DM-Process are in your opinion more important and why ?

2- Go to: *the UC Irvine Machine Learning Repository* (http://archive.ics.uci.edu/ml/), then go to View ALL Data Sets . Select three datasets. Describe for each of them the attribute type

3- Describe the attribute type in the following datasets:

ID	Income in three years ago	Education	Age	Income
1A	24552	High School	32	27026
2A	88282	BSc	52	93725
3B	82902	PhD	41	82356
4A	39838	High School	56	36828
5C	53542	PhD	32	62542
6M	63826	MS	28	64882
7D	82783	MA	43	89025
8A	72886	High School	33	74925
9Q	21383	BA	37	62572
1R	63552	BA	41	66427
1T	62522	High School	25	63552
1E	65254	PhD	56	67252

Customer ID	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

Exercises part 2 (continues)

4 - Give examples for : Time series data, Spatial data, Spatial-tempo data

5 - Find another example for an interval-scaled attributes besides temperature in degrees

6 - Select 10 different words as attribute from the following documents and construct the representation (document) data matrix for each of them. Source of documents is Reuter

1. SUMITOMO BANK AIMS AT QUICK RECOVERY FROM MERGER
Sumitomo Bank Ltd <SUMI.T> is certain to lose its status as Japan's most profitable
bank as a result of its merger with the Heiwa Sogo Bank, financial analysts said
Osaka-based Sumitomo, with desposits of around 23.9 trillion yen, merged with
Heiwa Sogo, a small, struggling bank with an estimated 1.29 billion dlrs
in unrecoverable loans, in October
But despite the link-up, Sumitomo President Koh Komatsu told Reuters he is confident his bank can
quickly regain its position"We'll be back in position in first place within three years," Komatsu said in
an interview He said that while the merger will initially reduce Sumitomo's profitability and efficiency,
7it will vastly expand Sumitomo's branch network in the Tokyo metropolitan area where it has been relatively weak.
…… (REUTER)

2. BOND CORP STILL CONSIDERING ATLAS MINING BAIL-OUT
Bond Corp Holdings Ltd <BONA.S> and Atlas Consolidated Mining and Development Corp <ATLC.MN> are still
holding talks on a bail-out package for the troubled mining firm, an Atlas statement said Atlas, the Philippines' biggest
copper producer, said it had been hit by depressed world copper prices It reported a net loss of 976.38 mln pesos
in the year ending December 1986, compared with a net loss of 1.53 billion in 1985 The company said it had
been able to cut its losses because its scaled-down copper operations in the central island of Cebu
started in the second half of 1986 Atlas said negotiations were continuing on the acquisition by
Bond of the company's existing bank loans and their restructuring into a gold loan A memorandum of
understanding signed by the two sides in October last year said Bond would acquire Atlas' total loans of 275
mln dlrs, to be repaid by the mining company in gold….. (REUTER)

3. CRA SOLD FORREST GOLD FOR 76 MLN DLRS - WHIM CREEK<
Whim Creek Consolidated NL> said the consortium it is leading will pay 76.55 mln dlrs for the acquisition
of CRA Ltd's <CRAA.S> <Forrest Gold Pty Ltd> unit, reported yesterday CRA and Whim Creek
did not disclose the price yesterday Whim Creek will hold 44 pct of the consortium, while
<Austwhim Resources NL> will hold 27 pct and <Croesus Mining NL> 29 pct, it said in a statement
As reported, Forrest Gold owns two mines in Western Australia producing a combined 37,000 ounces of gold a year
It also owns an undeveloped gold project
REUTER

Exercises part 2 (continues)

7- Consider the following data describing the weights of 10 individual in KG:

76 65 52 89 63 75 90 295 58 49

Why is for this dataset the mean no appropriate measure of location ? Which measure would be suitable ?

Compute the both measures.

8- Compute the mode for the attributes “Education” and “car” in the table of the exercise 3.

9- Is the "range" an appropriate measure of spread for the data containing outliers ? Give an example.

10- Describe two situations where stratifies sampling would be more adequate as simple random sampling

Exercises part 3

1. How the duplicate attributes can be identified ?

2. Suppose that you want to reduce the number of the attributes in the regression

   $$Y = a_0 + a_1 X1 + a_2 X_2 + a_3 X_3 + a_4 X_4$$

   How can you do it by using the wrapper approach ?

3. Describe two examples (algorithms) for attribute reduction based on embedded approach

4. Describe two criteria for attribute ranking in the case of continuous-valued and nominal attributes

5. Outline a situation in which identification numbers would be useful for prediction

Exercises part 4

1. **In which situation cross-validation is useful?**

2. **Describe three Data Mining applications that can be handled by supervised learning and three applications that can be handled by unsupervised learning.**

3. **What is the “default” accuracy rate in a classification task ? Give two examples.**

4. **Mention two criteria for evaluation of classification models.**

5. **Describe the difference between the test dataset and the validation dataset by an example.**

6. **What is the meaning of “Overfitting” ?**

7. **How can be determined that the model prediction is better than the prediction achieved just by the mean of the target variable ?**

Exercises part 5

1. From the dataset: below compute
   a- the Information gain for the attributes age, income, Student and Credit Rating
   b- the Gain Ratio for the attribute income

Nr.	Age	Income	Student?	Credit Rating	Buys Computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

- Source: Jiawei Han, et al. 2006

2. Elaborate the strengths and weakness of Decision Trees

Exercises part 5 (continues)

3. **From the dataset below compute the Gini-Index for attributes income, car, gender and customer ID.**

Customer ID	Income >2000	Car	Gender	Credit Rating
Customer 1	no	yes	F	bad
Customer 2	no statement	no	F	bad
Customer 3	no statement	yes	M	good
Customer 4	no	yes	M	bad
Customer 5	yes	yes	M	good
Customer 6	yes	yes	F	good
Customer 7	no statement	yes	F	good
Customer 8	yes	no	F	good
Customer 9	no statement	no	M	bad
Customer 10	no	no	F	bad

4. **Why is pruning in Decision Trees important ?**
5. **The attribute A has the values:**

100 100 200 700 700 200 600 600

Elaborate the discritization of A by using the successive average method

Exercises part 6

1. From table 1:

Table 1

TID	Items
1	bread, milk
2	bread, meat, orange juice, eggs
3	milk, meat, orange juice, cola
4	bread, milk, meat, orange juice
5	bread, milk, meat, cola

a- compute the support and confidence for A-rules :
{ Meat } → { Orange juice }, {eggs} → {cola}
{meat, orange juice } → {eggs}
{ Milk } → { meat }, {bread} → {milk}
{ milk, meat } → {orange Juice }

b- From the itemset { orange juice, meat, bread } generate all possible A-rules. Identify between all generated rules those rules with sup_min = 0,60 and conf_min = 0,80.

2. Elaborate the first and second Apriori-Principal

3. Elaborate two phases of generating the AR

4. Represent table 2 in a binary form

5. What does the brute-force strategy mean

Table 2

TID	Items
1	bread, tea, cola, eggs, cheese
2	meat, orange juice, eggs
3	milk, tea, chips , cola
4	tea, milk, meat, orange juice
5	bread, milk, meat, tea
6	orange juice, cola, bread

Exercises part 7

1. Elaborate the strengths and weaknesses of ANN.
2. Describe " Stop Training" in connection with ANN.
3. Elaborate the learning process in ANN for the case of supervised learning.
4. Describe the following terms in connection with ANN: Input Function, Activation Function and Output Function.
5. Elaborate the Gradient descent learning rule.
6. Discuss the different categories of learning methods.
7. Elaborate the differences between the Regression Analysis and ANN
8. Compute the learning rule in the case of a Perceptron network
9. Elaborate the differences between a Perceptron and a BP-Network
10. How does the Backpropagation of the errors mean ?
11. Compute the learning rule by using Gradient Decent Principe in a BP-Network for the output layer.
12. In the case of coding of continuous-valued data by using transformation function suppose that

X_n^{old} : original value (n=1, 2, … , N)

X_n^{new} : new, transferred value (n=1, 2, … , N)

X_{max}^{old} : maximal original value

X_{min}^{old} : minimal original value

By using these values compute a linear transformation function

Exercises part 8

1. **Regression Analysis (RA) is based on supervised learning. Why ?**

2. **Discuss three business problems that can be solved by using RA**

3. **Suppose that in the Single-Equation Linear Regression the explanatory variable changes on one unit. Compute the change of the dependent variable. Compare the result with the result of Multivariate Regression Equation**

4. **Compute the OLS-estimators of the coefficients In a Single-Equation Linear Regression for:**

$\sum X_i = 229,9$, $\sum y_i = 1242,9$, $\sum x_i^2 = 1569,2$, $\sum X_i y_i = 7279,7$ **Number of observation = 10**

5. **Show that for the regression** $y = \beta_0 + \beta_1 x + u$ **the relations:** $E(u|x)=0$ and $V(u|x) = \sigma^2$ **lead to**

$E(y|x) = \beta_0 + \beta_1 x$ and $V(y|x) = \sigma^2$

6. **Examine the scale change on the OLS-estimators in the Single-Equation Linear Regression**

Exercises part 8 (continues)

7. Cov (x,y) is defined as :

$$\text{Cov}(x,y) = 1/n \sum_{i=1}^{n} (x_i - x)(y_i - y)$$

Show that in the Single-Equation Linear Regression $\hat{ß}_1 = \dfrac{\text{Cov}(x,y)}{\bar{X}^2 - \bar{X}^2}$

8. Show the validity of the following relations in the Single-Equation Linear Regression : $y = ß_0 + ß_1 x + u$

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} \hat{u}_i x_i = 0 \qquad \sum_{i=1}^{n} \hat{u}_i \hat{y}_i = 0 \qquad \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$$

9. Show that the estimated regression Equation Y = ß0 + ß1X

goes through the point (X, Y)

10. Show the validity of the following relation for the Single-Equation Linear Regression

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

Exercises part 8 (continues)

X	Y
10	2
9	2
15	3
21	5
18	4
18	3
22	5
23	5
20	4
25	6

11. For the figures for X and Y in the table:

- Compute the OLS- estimators, R^2, adjusted R^2

- Test the validity of the relations of the exercise 8 empirically

12. Compute the OLS-estimators for the trend equation $y = ß_0 + ß_1 t$