KASIREDDY NARAYANREDDY COLLEGE OF ENGINEERING AND RESEARCH

ABDULLAPUR (V), ABDULLAPURMET (M), R.R DIST-501505
(Approved By AICTE,New Delhi & Affliated to JNTUH,HYDERABAD)

DATA MINING LAB(CS703PC)

List of Sample Problems:

Task 1: Credit Risk Assessment

Description:

The business of banks is making loans. Assessing the credit worthiness of an applicant is of crucial importance. You have to develop a system to help a loan officer decide whether the credit of a customer is good, or bad. A bank's business rules regarding loans must consider two opposing factors. On the one hand, a bank wants to make as many loans as possible.
Interest on these loans is the banks profit source. On the other hand, a bank cannot afford to make too many bad loans. Too many bad loans could lead to the collapse of the bank.
The bank's loan policy must involve a compromise: not too strict, and not too lenient.
To do the assignment, you first and foremost need some knowledge about the world of credit.
You can acquire such knowledge in a number of ways.

1. **Knowledge Engineering. Find a loan officer who is willing to talk. Interview her and try to represent her knowledge in the form of production rules.**

2. **Books. Find some training manuals for loan officers or perhaps a suitable textbook on finance. Translate this knowledge from text form to production rule form.**

3. **Common sense. Imagine yourself as a loan officer and make up reasonable rules which can be used to judge the credit worthiness of a loan applicant.**

4. **Case histories. Find records of actual cases where competent loan officers correctly judged when, and when not to, approve a loan application.**

The German Credit Data:

Actual historical credit data is not always easy to come by because of confidentiality rules. Here is one such dataset, consisting of 1000 actual cases collected in Germany.
Credit dataset (original) Excel spreadsheet version of the German credit data.
In spite of the fact that the data is German, you should probably make use of it for this assignment. (Unless you really can consult a real loan officer!)

A few notes on the German dataset

1. **DM stands for Deutsche Mark, the unit of currency, worth about 90 cents Canadian (but looks and acts like a quarter).**

2. **owns_telephone. German phone rates are much higher than in Canada so fewer people own telephones.**

3. **foreign_worker. There are millions of these in Germany (many from Turkey). It is very hard to get German citizenship if you were not born of German parents.**

4. **There are 20 attributes used in judging a loan applicant. The goal is to classify the applicant into one of two categories, good or bad.**

Subtasks: (Turn in your answers to the following tasks)

1. **List all the categorical (or nominal) attributes and the real-valued attributes seperately. (5 marks)**
2. **What attributes do you think might be crucial in making the credit assessment? Come up with some simple rules in plain English using your selected attributes. (5 marks)**
3. **One type of model that you can create is a Decision Tree - train a Decision Tree using the complete dataset as the training data. Report the model obtained after training. (10 marks)**
4. **Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly? (This is also called testing on the training set) Why do you think you cannot get 100 % training accuracy? (10 marks)**
5. **Is testing on the training set as you did above a good idea? Why or Why not ? (10 marks)**
6. **One approach for solving the problem encountered in the previous question is using cross-validation? Describe what is cross-validation briefly. Train a Decision Tree again using cross-validation and report your results. Does your accuracy increase/decrease? Why? (10 marks)**
7. **Check to see if the data shows a bias against "foreign workers" (attribute 20), or "personal-status" (attribute 9). One way to do this (perhaps rather simple minded) is to remove these attributes from the dataset and see if the decision tree created in those cases is significantly different from the full dataset case which you have already done. To remove an attribute, you can use the preprocess tab in Weka's GUI Explorer. Did removing these attributes have any significant effect? Discuss. (10 marks)**
8. **Another question might be, do you really need to input so many attributes to get good results? Maybe only a few would do. For example, you could try just having attributes 2, 3, 5, 7, 10, 17 (and 21, the class attribute (naturally)). Try out some combinations. (You had removed two attributes in problem 7. Remember to reload the arff data file to get all the attributes initially before you start selecting the ones you want.) (10 marks)**

9. **Sometimes, the cost of rejecting an applicant who actually has a good credit (case 1) might be higher than accepting an applicant who has bad credit (case 2). Instead of counting the misclassifcations equally in both cases, give a higher cost to the first case (say cost 5) and lower cost to the second case. You can do this by using a cost matrix in Weka. Train your Decision Tree again and report the Decision Tree and crossvalidation results. Are they significantly different from results obtained in problem 6 (using equal cost)? (10 marks)**
10. **Do you think it is a good idea to prefer simple decision trees instead of having long complex decision trees? How does the complexity of a Decision Tree relate to the bias of the model? (10 marks)**
11. **You can make your Decision Trees simpler by pruning the nodes. One approach is to use Reduced Error Pruning - Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross-validation (you can do this in Weka) and report the Decision Tree you obtain? Also, report your accuracy using the pruned model. Does your accuracy increase? (10 marks)**
12. **(Extra Credit): How can you convert a Decision Trees into "if-then-else rules". Make up your own small Decision Tree consisting of 2-3 levels and convert it into a set of rules. There also exist different classifiers that output the model in the form of rules - one such classifier in Weka is rules. PART, train this model and report the set of rules obtained. Sometimes just one attribute can be good enough in making the decision, yes, just one ! Can you predict what attribute that might be in this dataset ? OneR classifier uses a single attribute to make decisions (it chooses the attribute based on minimum error). Report the rule obtained by training a one R classifier. Rank the performance of j48, PART and oneR. (10 marks)**

Mentor lecture on Decision Trees

Andrew Moore's Data Mining Tutorials (See tutorials on Decision Trees and Cross Validation)

Decision Trees (Source: Tan, MSU)

Tom Mitchell's book slides (See slides on Concept Learning and Decision Trees) Weka resources:

- **Introduction to Weka (html version) (download ppt version)**
- **Download Weka**
- **Weka Tutorial**
- **ARFF format**
- **Using Weka from command line**

Task 2: Hospital Management System

Data Warehouse consists Dimension Table and Fact Table.

REMEMBER The following

Dimension

The dimension object (Dimension):

_ Name

_ Attributes (Levels) , with one primary key

_ Hierarchies

One time dimension is must.

About Levels and Hierarchies

Dimension objects (dimension) consist of a set of levels and a set of hierarchies defined over those levels. The levels represent levels of aggregation. Hierarchies describe parent-child relationships among a set of levels.

R16 B.TECH CSE.

For example, a typical calendar dimension could contain five levels. Two hierarchies can be defined on these levels:

H1: YearL > QuarterL > MonthL > WeekL > DayL

H2: YearL > WeekL > DayL

The hierarchies are described from parent to child, so that Year is the parent of Quarter, Quarter the parent of Month, and so forth.

About Unique Key Constraints

When you create a definition for a hierarchy, Warehouse Builder creates an identifier key for each level of the hierarchy and a unique key constraint on the lowest level (Base Level) Design a Hospital Management system data warehouse (TARGET) consists of Dimensions Patient, Medicine, Supplier, Time. Where measures are 'NO UNITS', UNIT PRICE. Assume the Relational database (SOURCE) table schemas as follows

TIME (day, month, year),

PATIENT (patient_name, Age, Address, etc.,)

MEDICINE (Medicine_Brand_name, Drug_name, Supplier, no_units, Uinit_Price, etc.,) SUPPLIER :(Supplier_name, Medicine_Brand_name, Address, etc.,)

If each Dimension has 6 levels, decide the levels and hierarchies, Assume the level names suitably.

Design the Hospital Management system data warehouse using all schemas. Give the example 4-D cube with assumption names.

KASIREDDY NARAYANREDDY COLLEGE OF ENGINEERING AND RESEARCH

ABDULLAPUR (V), ABDULLAPURMET (M), R.R DIST-501505
(Approved By AICTE,New Delhi & Affliated to JNTUH,HYDERABAD)

INTERNET OF THINGS LAB (CS755PC)

List of Experiments:

1 Start Raspberry Pi and try various Linix commands in command terminal window: ls, cd, touch, mv, rm, man, mkdir, rmdir, tar, gzip, cat, more, less, ps, sudo, cron, chown, chgrp, ping etc.

2. Run some python programs on Pi like:

   Read your name and print Hello message with name

   Read two numbers and print their sum, difference, product and division.
   Word and character count of a given string

   Area of a given shape (rectangle, triangle and circle) reading shape and appropriate values from standard input

   Print a name 'n' times, where name and n are read from standard input, using for and while loops.

   Handle Divided by Zero Exception.

   Print current time for 10 times with an interval of 10 seconds.
   Read a file line by line and print the word count of each line.

3. Light an LED through Python program

4. Get input from two switches and switch on corresponding LEDs

5. Flash an LED at a given on time and off time cycle, where the two times are taken from a file.

6. Flash an LED based on cron output (acts as an alarm)

7. **Switch on a relay at a given time using cron, where the relay's contact terminals are connected to a load.**

8. **Get the status of a bulb at a remote place (on the LAN) through web.**
   The student should have hands on experience in using various sensors like temperature, humidity, smoke, light, etc. and should be able to use control web camera, network, and relays connected to the Pi.