SE488
Individual Assignment #1

Description:
The goal of this assignment is to practice the supervised learning algorithm to create a decision tree and to use Weka tool to create automatically decision trees.

Question 1

Training	fever	vomiting	diarrhea	shivering	**Classification**
d_1	no	no	no	no	healthy (H)
d_2	average	no	no	no	influenza (I)
d_3	high	no	no	yes	influenza (I)
d_4	high	yes	yes	no	salmonella poisoning (S)
d_5	average	no	yes	no	salmonella poisoning (S)
d_6	no	yes	yes	no	bowel inflammation (B)
d_7	average	yes	yes	no	bowel inflammation (B)

1.1 Calculate the entropy of the dataset
1.2 Calculate the information gain (IG) for each of the attribute
1.3 Which attribute should be considered as the first node (root)
1.4 Create the whole Decision Tree based on the dataset

Question 2

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Consider the training examples shown in for a binary classification problem.

2.1 What is the entropy of this collection of training examples
2.2 What are the information gains of a1 and a2 relative to these training examples?
2.3 What is the best split (among a1 and a2) ?

Question 3

The goal of this exercise is to predict the **risk** of a car driver.

The dataset contains the following information :

Attribute	Description	Values
time	time since obtaining a drivers license in years	{1-2, 2-7, >7}
gender	gender	{male, female}
area	residential area	{urban, rural}
risk	the risk class	{low, high}

Below is the data :

ID	time	gender	area	risk
1	1-2	m	urban	low
2	2-7	m	rural	high
3	>7	f	rural	low
4	1-2	f	rural	high
5	>7	m	rural	high
6	1-2	m	rural	high
7	2-7	f	urban	low
8	2-7	m	urban	low

3.1 Create a decision tree based on the training data. The stopping criteria of the decision tree is when all instances in the branch have the same class.
3.2 Predict the result of the following driver's information:

ID	time	gender	area
A	1-2	f	rural
B	2-7	m	urban
C	1-2	f	urban

Question 4 (using WEKA tool)

Download the attached bankinfo.csv file
Use Weka tool to transform it into a .arff file
Save the file and name it bank.arff .
Load the data of the bank.arff into WEKA.
Use J48 classifier in WEKA with the default parameters and using the 10-fold cross-validation as an evaluation approach.
Q4.1 Take a screenshot of the obtained WEKA results
Q4.2 What are the obtained Accuracy, Precision and Recall results ?
Q4.3 Explain from the confusion Matrix how the Precision, Recall and Accuracy was obtained?
Q4.4 Visualize the tree, take a screenshot and attach it to your answersheet.
Download the attached "bank-data test.arff" and use it as a "supplied test set" in WEKA
To test your model and to know how your model managed to classify the test instances.

Q4.5 save the results in file called: “bank predicted.arff” Interpret the predicted results compared to the actual results. What is your conclusion ?
Deliverables

Deliverables:

- Your completed **AnswerSheet**.pdf, which must exclusively be in PDF format. This file should contain your answers to all four exercises. If you prefer, you may write your answers using pen and paper, photograph them, and include these images in your PDF.

Note: For questions Q4.1 and Q4.2, please include screenshots from WEKA to support your answers. Ensure that all content is clear and easily readable.

- The ***bank.arff*** and ***bank predicted.arff*** files

Due Date : 09/29/2024