IS 733 Homework 3, Due 3/30/2021 (midnight)

For this assignment, consider the *labor negotiations* classification dataset in Table 1, overleaf. The dataset contains the outcomes of labor contract negotiations in the business and personal services sector (teachers, nurses, university staff, police, etc) in Canada during 1987 and 1988. This is a subset of the instances and attributes. The task is to predict whether the collective agreements were accepted by both labor and management ("*good*"), or not ("*bad*"). Entries with a "?" are missing. Points

1. Suppose that we are training a decision tree model on our labor negotiations dataset. To keep things simple, for the purposes of this question we will treat missing ("?") as an extra attribute value for each attribute. For example, the *bereavement assistance* attribute now has values $\{yes, no, ?\}$.

   (a) Select the attribute to split on at the root of the tree, using the information gain criterion. Show your working. Measure the information gain in bits, i.e. use base 2 logarithms. Recall that the relevant formulae are: (20)

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

   (b) Select the attribute to split on at the root of the tree, using the gain ratio criterion. Does this select a different attribute than information gain does in this case? Show your working. Recall that the gain ratio is calculated as: (8)

$$\text{GainRatio}(A) = \text{Gain}(A)/SplitInfo(A) \ , \ SplitInfo(A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} \ .$$

2. Suppose that we want to use the PRISM algorithm to construct a set of classification rules that cover the instances in this dataset. Find the first test (i.e. *attribute* = *value*) for the first rule when covering the class *bad*. As in the previous question, treat missing ("?") as an extra value for each attribute. Show your working. (20)

3. (a) Consider a new instance, with attribute values $[generous, no, none, no]$. Using the naive Bayes classifier, calculate the predicted probability that this instance is labeled *good*. Show your working, including any intermediate probabilities that you needed to calculate. Use Laplace smoothing (i.e. add one to all of the counts when calculating probabilities, to ensure that none of the probabilities are zero). Ignore missing values when calculating probabilities (note that this is different to the previous questions, where missing values were treated as an extra value). Only calculate the probabilities you actually need to compute the class probabilities for this particular instance. (20)

   (b) Explain the "naive" assumption that the naive Bayes method uses. Is this assumption realistic for this dataset? (8)

Table 1: Labor Negotiations Data

vacation $\{below\_average, average, generous\}$	longterm disability assistance $\{yes, no\}$	contribution to dental plan $\{none, half, full\}$	bereavement assistance $\{yes, no\}$	class $\{bad, good\}$
below_average	yes	full	yes	good
average	yes	full	?	good
?	yes	half	?	good
generous	?	half	?	good
below_average	no	none	?	bad
below_average	?	full	?	good
below_average	yes	half	yes	bad
below_average	?	half	yes	bad
generous	yes	none	yes	bad
below_average	yes	half	?	bad
generous	?	?	yes	good
average	?	full	yes	good
average	?	?	yes	bad
below_average	no	half	?	bad
average	?	?	yes	good
generous	yes	none	yes	good
average	?	?	yes	bad
below_average	no	half	?	bad
average	?	?	yes	bad
below_average	yes	?	?	good
average	no	none	no	bad
?	yes	half	yes	good
below_average	yes	half	?	bad

4. (a) Having applied the ID3 decision tree induction algorithm and the rule learning algorithm PRISM to this dataset, we would like to evaluate the classification accuracy of each algorithm. Why is it not a good idea to simply calculate their classification accuracy on the training set? Outline the steps of the cross-validation method for evaluating classifiers, and explain the benefits of this procedure. (12)

   (b) Suppose we would like to solve a classification problem where all of the attributes are numeric, unlike in the task we have been considering (where the attributes are all nominal). Name two classification algorithms that are applicable in this setting. Briefly explain the representations that these methods use, the objective functions that they aim to optimize, and the search algorithms that they use to accomplish this. What are the pros and cons of the two methods that you have described, relative to each other? (12)

Total: 100