CSC 5800: Intelligent Systems

Homework 3

Due Date: November 1st, 2018 **Total: 100 Points**

Problem 1. Cost Matrix

(10 Points)

Consider Problem 6 in Homework 2. Which attribute would be chosen as the first splitting attribute if the following cost matrix is used? Use the Cost function value as your measure for splitting.

Cost Matrix	Attribute Value		
		T	F
ACTUAL CLASS	+	-1	100
	-	1	0

Problem 2. Evaluation Measures

(20 Points; 2.5 points each)

For the Confusion Matrix shown below, compute the following values:

Confusion Matrix	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	98	20
	-	37	143

(a) Accuracy (b) Precision (c) Recall (d) F-measure (e) Cost (f) Sensitivity
(g) Specificity (h) False Positive Rate

Problem 3. Model Evaluation

(10 Points; 4 + 6)

A database contains 100 documents out of which only 10 documents are relevant for a given query. Two search engines, A and B, report the following documents.

System A: **RRNNNRNNNR RNNRNNNNRN NNRNNNNNRN** and so on
System B: **RRRNNNNNNN NRRNNNNNNR NNNNNRNNNN** and so on

where **R** represents relevant document and **N** represents Non-relevant document. A search engine expert wants to estimate the accuracy of both these systems by computing the following information, so that, he can know how many documents to display.

(a) Using the PR-Curve info, compute the Precision at 40% Recall.
(b) If the expert decides to display only the first 15 documents for both of these search engines, calculate Precision, Recall and F-measure.

Problem 4. R Exercise

(20 Points; 10 points each)

Load the iris data into R.

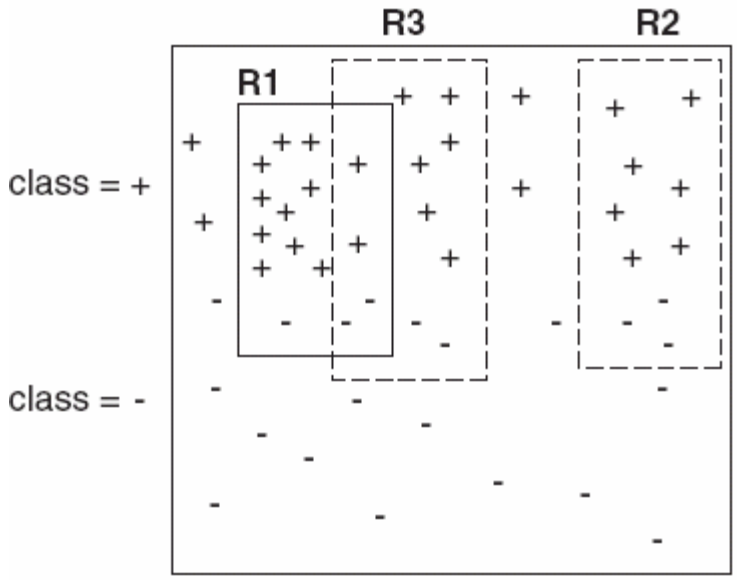
(a) Construct and display the following decision trees
   a. An unpruned decision tree
   b. A Tree with a maximum of 5 leaf nodes

(b) Split the data into training, validation and test sets. Report the error rates using 3-fold, 5-fold and 10-fold cross-validation schemes.

Problem 5. Rule-based Classification

(30 Points; 5 points each)

The following Figure illustrates the coverage of the classification rules R1, R2, and R3. Determine which is the best and the worst rule according to:

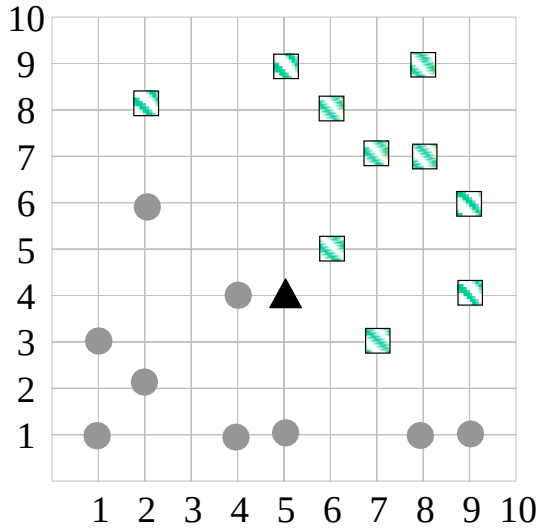


(a) FOIL's Information Gain.
(b) The Laplace measure.
(c) The m-estimate measure (with $k = 2$ and $p+ = 0.58$).
(d) The rule accuracy after $R1$ has been discovered, where none of the examples covered by $R1$ are discarded.
(e) The rule accuracy after $R1$ has been discovered, where only the positive examples covered by $R1$ are discarded.
(f) The rule accuracy after $R1$ has been discovered, where both positive and negative examples covered by $R1$ are discarded.

Problem 6. Nearest-Neighbor Classification

(10 Points; 4 + 6)

Consider the following 2-dimensional dataset. Classify the test point (triangle) using:
(Treat the squares as + and circles as –)



(a) 3- and 5- nearest neighbor
(b) Euclidean and Manhattan distance weighted 3-nearest neighbor