CISC-483-683
Homework 8: Model Trees and Classification Rules

Due: Wednesday, November 6, 2019

60 points

NOT Accepted After The End of the Grace Period

1. (5 pts.) Bob and Coleen are each entering a Netflix competition to develop an outstanding movie recommendation system. Bob argues that his model is better than Coleen's and Coleen argues that her model is better than Bob's. Let us refer to Bob's model as M$_{bob}$ and Coleen's model as M$_{coleen}$. You agree to run both models on 15 different sets of test data, and you get the following accuracy rates for the 15 test sets:

Test-set	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M$_{bob}$	84	73	66	82	66	73	74	77	69	80	66	68	66	78	83
M$_{Coleen}$	86	72	67	75	62	74	66	70	70	75	63	65	67	70	76

Determine whether one of the models is really better than the other model. Let your desired confidence level in saying that they are different be 90% Please show the details of your work and give your conclusion about whether the difference between the two models is statistically significant.

2. (5 points) Consider the following dataset, where the class or response variable is mpg (miles per gallon):

Instance	**Cylinders**	**HP**	**Pounds**	**ACC**	**Year**	**mpg**
1.	8	130	3504	12	2005	18
2.	4	90	2950	17.3	2000	27
3.	4	88	2395	18	2003	34
4.	4	94	2372	15	1990	24
5.	4	75	2155	16.4	1996	28
6.	4	96	2665	13.9	2004	32
7.	4	46	1835	20.5	1995	26
8.	4	70	1937	14.2	1997	29
9.	4	52	2130	24.6	2008	44
10.	4	85	4615	14	2000	10
11.	4	79	2625	18.6	2004	28
12.	4	76	2400	17.5	2006	25
13.	8	149	4335	14.5	1985	16
14.	8	125	4000	15.0	2012	15

Consider the attribute HP (horsepower). Which of the following is a better split point for this attribute: 77.5 or 95? Show your work in making the decision.

3. (10 points) For this problem, you will be using the dataset called **California-housing-19.csv** found on the class web site; this dataset represents aggregated data collected from neighborhoods in the 1990 census. The leftmost attribute **Value** is the target or class attribute and represents the value of a house in units of $100,000.

Use Weka to build a linear regression model (use LinearRegression in Weka), a Model Tree, and a Regression Tree (the method for creating a Model Tree or a Regression Tree is called **M5P** in Weka) from the training data, with the evaluation done on the training data. (Be sure to set the parameter *Attribute Selection Method* to *No Attribute Selection* in LinearRegression so that Weka does not discard any of the attributes.) For the Regression Tree and the Model Tree, use smoothing and pruning and set 4 as the minimum number of instances in a leaf node.

(a) What is the model produced by linear regression?

(b) Draw the top three levels of the Regression and Model Trees (they are the same).

(c) Examine the leaf nodes of the regression tree and the leaf nodes of the model tree. What do you see as the difference between the leaf nodes of the regression tree and the leaf nodes of the model tree?

(d) Compare the three models on the basis of Root Mean Squared Error. Which model is best?

4. (15 points) Suppose that you have the following dataset

Instance	STABILITY(ST)	DEVIATION(D)	WIND(W)	VISIBILITY(V)	CLASS
1.	yes	.5	tail	poor	land
2.	yes	1.0	head	good	delay
3.	no	.9	head	poor	land
4.	yes	0	head	poor	land
5.	no	.2	head	good	land
6.	yes	.1	tail	good	dlay
7.	no	.4	head	good	delay
8.	yes	.6	head	poor	land
9.	yes	.1	head	poor	land
10.	yes	.5	head	good	delay
11.	no	.6	tail	good	land
12.	no	.4	tail	good	delay
13.	no	.6	tail	good	delay
14.	no	.8	head	good	delay
15.	no	.2	head	good	land
16.	no	.4	head	good	delay
17.	yes	.6	tail	poor	land
18.	no	.5	tail	good	land
19.	yes	.1	tail	poor	land
20.	yes	.2	head	good	delay
21.	yes	.8	tail	good	delay
22.	yes	.9	tail	good	delay

You are going to use the Prism method (that is, accuracy as the rule quality measure) to construct a set of perfect classification rules for `CLASS=delay`. (If there is a tie, use coverage to break the tie.) At each step, please list the instances that are being considered (you can just list the instance numbers from the table), the conjuncts that are considered as the new addition to the rule and the aqccuracy of the rule if that conjunct is added to the rule, the conjunct that was selected, and the reason that one was selected. We will use local discretization of the numeric attribute **DEVIATION**, but to avoid a lot of computation, we will only consider .3 and .7 as possible split points. Thus the possible conjuncts for the DEVIATION attribute are DEVIATION $\leq .3$, DEVIATION $> .3$, DEVIATION $\leq .7$, and DEVIATION $> .7$.

(a) Develop the first rule for the class **delay**.

(b) Develop the second rule for the class **delay**.

(c) Develop the third rule for the class **delay**.

5. (15 points) Suppose that you have developed the following rule:

```
IF      VISIBILITY = good
        and DEVIATION ≤ .7
        and DEVIATION > .3
        and STABILITY = no
        and WIND = tail
THEN Class = delay
```

and you have the following pruning set:

Instance	STABILITY(ST)	DEVIATION(D)	WIND(W)	VISIBILITY(V)	CLASS
1.	no	.1	head	poor	land
2.	yes	.6	tail	good	delay
3.	no	.1	head	poor	land
4.	no	.6	tail	good	delay
5.	no	.5	tail	good	land
6.	yes	.6	tail	good	delay
7.	no	.4	tail	good	land
8.	yes	.6	tail	good	land
9.	no	.6	tail	good	delay
10.	no	.1	head	poor	land
11.	no	.5	head	good	delay
12.	no	.4	head	good	delay

Use incremental reduced error pruning and the evaluation measure used by RIPPER ($\frac{p-n}{p+n}$) to prune the above rule. Please show your work and how you made your decisions. Give the final rule that results.

6. (10 points) Weka: Use the College-HW-19.csv dataset that can be found on the class web site. This dataset includes a variety of information on colleges in the United States. First discretize numeric attributes using Discretize (supervised) with the default settings — recall that you must select *Filtered Classifier* under *meta.* and then click on *FilteredClassifier* next to the *Choose* box in order to select the discretization scheme and to select the classifier to be used.

   Use PRISM, Ripper (JRIP) and decision trees (J48) and 10-fold cross validation to develop and evaluate models for classifying instances as PUBLIC or PRIVATE institutions. Discuss the differences in the models in terms of accuracy, F-measure, the number of rules or number of leaf nodes of the tree (if there are a very large number of rules, you need not count them but just say *large number of rules*). Do the models appear to be similar or quite different?