CS548 - Knowledge Discovery and Data Mining. Spring 2015
Quiz 4 Solutions and Grading Rubric. February 24, 2015

By Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

Instructions

- **Justify** your answers. Concise and decisive answers are better than long, vague answers.
- Ask in case of doubt.
- Maximum time allowed: 20 minutes

Constructing Regression and Model Trees

The following is a general algorithm to construct *decision trees*. Show how to modify this algorithm to construct regression and model trees. Below each instruction two lines are included, one for Regression Trees (RT) and one for Model Trees (MT). You need to state how that instruction needs to be modified for the construction of either a Regression or a Model Tree. If no modification is needed, just write "*same as for decision trees*". If the modification for RT and for MT are the same, say so and write the modification only once. An example is given below.

Note: Solutions below are written in red font. We use the following convention when the answer for RT and for MT is the same:

RT: } Answer is the same for RT and for MT
MT: }

Decision Tree Construction Algorithm (S, T)

Input: S is a set of data instances with attributes $A_1,\ldots,A_m$ each of which is either numeric or nominal, and T is the nominal target.

RT: *Answer*: S is a set of data instances with attributes $A_1,\ldots,A_m$ each of which is numeric, and T is the *numeric* target attribute

MT: (Note here that the input dataset doesn't contain any nominal attributes – they have been converted to numeric beforehand.)

Output: A tree that predicts T from attributes $A_1,\ldots,A_m$

If the number of data instances in S is ≤ a given threshold OR the entropy(S,T) ≤ another given threshold

RT: [15 points] *Answer*: the number of data instances in S is ≤ a given threshold OR the standard deviation of the target attribute

MT: values of the instances in S is ≤ another given threshold

then create a leaf node with prediction value equal to the majority class of the data instances in S

RT: [15 points] create a leaf node with prediction value equal to the average of the target values of the data instances in S

MT: [15 points] create a leaf node with prediction equal to the formula obtained from applying linear regression to the data instances in S

else create an internal node. The test condition of this node is selected as follows:

1. List all the candidate test conditions, which consist of each of the nominal attributes, and each of the numeric attributes together with each of its possible split points.

2. RT: [5 points] *Answer:* List all the candidate test conditions, which consist of each of the numeric attributes together

   MT: with each of its possible split points. (Note: there are no nominal attributes as they have been converted to numeric beforehand.)

3. Calculate the entropy of each of these candidate test conditions.

   RT: [15 points] *Answer:* Calculate the Standard Deviation Reduction (SDR) of each of these candidate test conditions.
   MT:

4. Select the test condition with the lowest entropy.

5. RT: [10 points] *Answer:* Select the test condition with the highest SDR.
   MT:

6. Let's call this test condition A_i (nominal case) or $<A_i, sp>$ if A_i is numeric and sp is its selected split point.
   a. If A_i is nominal, create a child for each possible value of the A_i
      RT: [5 points] *Answer:* This case does not apply because there are no nominal attributes in the input dataset.
      MT:
   b. If Ai is numeric, create two children: one for $A_i \leq sp$ and another one for $A_i > sp$
      RT: [5 points] *Answer: same as for decision trees.*
      MT:
   c. Distribute the data instances in S in each of this node's children according to their A_i values
      RT: [5 points] *Answer: same as for decision trees.*
      MT:
   d. Repeat this process recursively for each child node. That is,

      for each child node **do**

         call Tree Construction Algorithm (set of data instances in child node,T)

         RT: [10 points] *Answer: same as for decision trees.* (Here of course the recursive call is done to this
         MT: modified function (in red) that constructs regression and model trees.)