

CS 536: Machine Learning

Homework 1

We are going to use Weka in some assignments. And assignment 1 helps to familiarize you with Weka.

1. Run decision tree classifier on Weka. After you download and install Weka on your computer, run Weka under “Explorer” mode; and do the following for each of the data sets “iris”, “contact-lenses” and “weather.nominal”:
 - (a) Load the data under “Preprocess”.
 - (b) Run the data using weka.classifiers.Id3 under “Classify”.

For each data set, give the learned tree structure, and the error rate on the training data and the error rate using 2-fold, 5-fold, and 10-fold cross-validation. Would they be the same?

Note: Although decision trees can be applied to continuous attributes, Id3 of Weka only runs on nominal attributes. When you run the above three data sets, could you see the difference? After loading data, how to check each attribute’s detailed information (numeric or nominal, any missing value, etc.)?

2.
 - (a) Generate a dataset with n binary-valued attributes for which a decision learning algorithm without pruning must generate a complete decision tree.
 - (b) Generate a dataset with n binary-valued attributes for which a decision-tree learning algorithm without pruning must generate a decision tree with only one internal node.
 - (c) For both (a) and (b), can you characterize the size of the resulting tree as a function of n? Can you characterize the size of the dataset that would create the above mentioned decision trees on n attributes?
 - (d) Let n=5, run your data through Weka, using weka.classifiers.Id3. Do you generate the desired trees? Show your resulting classifier.
3. Create a new learning algorithm, GINI, by modifying the Weka program for Id3 to select the attribute as follows, rather than the best attribute with the max Gain value.

The gini index function can be used to evaluate the goodness of all the potential split points along all the attributes. Consider a dataset S consisting of n records, each belonging to one of the c classes. The gini index for the set S is defined as:

$$\text{gini}(S) = 1 - \sum_{j=1}^c p_j^2$$

where p_j is the relative frequency of class j in S . If S is partitioned into two subsets S_1 and S_2 , the index of the partitioned data $\text{gini}^D(S, \text{sc})$ can be obtained by:

$$\text{gini}^D(S, \text{sc}) = \frac{n_1}{n} \text{gini}(S_1) + \frac{n_2}{n} \text{gini}(S_2)$$

where n_1 and n_2 are the number of examples of S_1 and S_2 , respectively, and sc is the splitting criterion. Here the attribute with the min gini index is chosen as the best attribute to split.

Run Id3 and GINI on data sets. And give scatter plots of the error rate, and tree size of Id3 versus GINI, respectively. What conclusions can you reach about the relative merits of these two algorithms?

In order to get more data sets, you can download the data sets from <ftp://ftp.cs.waikato.ac.nz/pub/ml/datasets-UCI.jar>, which contains a number of benchmark problems from the UC Irvine repository of machine learning data sets. Do NOT use the following data sets for decision trees: balance-scale, breast-w, colic.ORIG, colic, diabetes, glass, heart-statlog, ionosphere, iris, labor, letter, segment, sonar, soybean, vehicle, vote, vowel, and waveform. For the other data sets use Weka to delete the numeric attributes and all attributes with missing values: Under “Preprocess”, deselect the numeric attributes, and then click on “Apply Filters”.