

Assignment 4

CIS 453/553 Data Mining, Winter 2008

due 11:59 pm, Friday Feb 22nd

1. As we discussed in the class (slide 17 of the lecture notes of week 6), a majority function (outputs 1 only if more than half of its n inputs are 1) can be implemented by a perceptron network. Please give a workable set of values of W_j and W_0 , if $a_0 = -1$. However, a decision tree needs $O(2^n)$ nodes to represent this function. Why?

2. Given a training table with T tuples and n attributes, show that the worst-case complexity of growing a decision tree is $n \times T \times \log(T)$. (Hint: first show that the maxim depth of the tree is $\log(T)$.)

3. A bank database has five attributes for customers.

credit-ranking	age	gender	year-income	count
Excellent	< 30	Male	60k - 100k	16
Excellent	< 30	Female	> 100k	4
Excellent	> 60	Male	> 100k	15
Excellent	> 60	Female	60k - 100k	3
Good	30 - 60	Male	< 60k	16
Good	30 - 60	Female	60k - 100k	2
Good	< 30	Male	> 100k	14
Good	< 30	Female	60k - 100k	2
Fair	> 60	Male	> 100k	18
Fair	30 - 60	Female	60k - 100k	20
Fair	< 30	Male	60k - 100k	2
Fair	< 30	Female	< 60k	2

Let *credit-ranking* be the class label.

(a) How would you modify the ID3 algorithm to take into consideration the *count* of each tuple?

(b) Build a decision tree based on your algorithm.

(c) Given a customer information: age is "< 30," gender is "Male," and year-income is "> 100k", what would a naive Bayesian classification of the *credit-ranking* for the customer be?

(d) If we want to use the given data to train the a neural network, such as Figure 6.18 (Figure 7.11 in the first edition), what modification we need to do for the data tuples, considering there are only three input nodes and the neural network normally accepts numerical values?

4. State why Bagging and Boosting can improve the accuracy of the classification.

5. Based on your understanding, please compare SVM with decision tree learning about their advantages and disadvantages.

