CSE P546 Data Mining Homework 3

Due Date: May 30. We would prefer that you turn in a hard copy of your solutions at the start of class. However, you can also email them to bhushan@cs. Your file can be in any of pdf, Word or plaintext formats.

1. (20 points) Suppose the instance space is the square $0 < X < 8, 0 < Y < 8$, and you have a training set composed of positive examples at (2, 2), (2, 4), (2, 6), (6, 2), (6, 4), and (6, 6), and negative examples at (4, 2), (4, 4) and (4, 6).

   (a) (6 points) Draw the Voronoi diagram of this training set, using Euclidean distance.

   (b) (7 points) Suppose you measure the error rate of the 3-nearest neighbor algorithm on this training set using leave-one-out cross-validation (i.e., you take out each example in turn, and predict its class using the other examples). What would the measured error rate be? If some example has more than 3 examples at the same nearest distance from it such that different choices of the 3 nearest neighbors give different predictions then count this example as an error. You thus compute the worst-case error rate.

   (c) (7 points) Suppose you apply the 3-nearest-neighbor algorithm with backward elimination. Which features would be eliminated (X, Y, both, or neither)? Why?

2. (15 points) Consider learning a perceptron on nine-dimensional linearly separable data. How many training examples do you need to guarantee with 99% confidence that the learned perceptron has true error of at most 10%?

3. (20 points – 8 for your answer and 12 for the justification) Mitchell 7.5 (a). You also need to justify your answer. You don't have to give a formal proof but you must present the key ideas from which the reader can construct a formal proof if he wants.

4. (15 points – 6 for your answer and 9 for the justification) A decision stump is a decision tree with only one internal node. Which algorithm has higher bias: a decision stump or a perceptron? And which one has higher variance? Justify your answer. (Assume that all attributes are either Boolean or numeric.)

5. (20 points – 10 for each part) Han & Kamber 5.5 (this would be 6.4 in the 1st Edition).

6. (10 points) Given all the rule learning schemes we have previously seen in class, why do you think there is a need for association rule mining at all?