ENEE 459M: Topics in Computer Engineering – Machine Learning and Data Mining

Spring 2010
Midterm III

1. **[20pt]** Consider the following instances of a dataset specified by two numeric attributes: (1,4), (2,5), (3,4), (5,1), (4,0), (6,0), (7,2), (8,4), and (9,2). Apply two iterations of the k-means algorithm, starting with the initial centers (7,0) and (4,1) for k=2. Explain the method used in each step of the algorithm, and show your work in details. Describe, at the end of each iteration, the clusters obtained (center and points in the cluster). Does the algorithm converge at the end of the second iteration – justify your answer.

2. **[20pt]** Assume that the alligator length follows a Gaussian distribution with mean 12 feet and standard deviation 2 feet while the crocodile length follows a Gaussian distribution with mean 15 feet and standard deviation 2 feet. Given that the prior probability of alligators is .6 in the total population of alligators and crocodiles, what is the probability that a length X=15 feet is from an alligator? Justify your answer.

3. **[20pt]** Given the following sample of intermixed GPA scores from a sample of the students in two High Schools A and B, each score is followed by the probability that the score comes from a student in High School A:

$$X_1 = 3.4; p_1 = .6;$$
$$X_2 = 2.8; p_2 = .3;$$
$$X_3 = 3.1; p_3 = .5;$$
$$X_4 = 3.8; p_4 = .7; .$$
$$X_5 = 2.3; p_5 = .1;$$
$$X_6 = 3.6; p_6 = .6;$$
$$X_7 = 3.4; p_7 = .4$$

Assume that the GPA scores of the students in High School A and High School B each follow a Gaussian distribution. (a) Estimate the mean of each of the distributions using the above

data. Explain the method used. (b) Estimate the number of scores that come from High School A and justify your answer.

4. **[20pt]** Given following set of training data with two numeric attributes A and B and a nominal class attribute Class:

A	B	Class
9	6	No
2	5	Yes
3	4	No
3	3	No
3	2	Yes
6	3	Yes
6	2	No
7	2	Yes
8	1	No

(a) Use instance based learning to derive the class of the instance A =5 and B=3. Show how you obtained your answer.
(b) Suppose we want to use the 3-nearest neighbor strategy. What is the class of the same instance A=5 and B=3? Again justify your answer.
(c) Suppose we want to use the same strategy as in (b) except with a weight equal to the inverse of the **distance squared**. What is the class of the same instance? Justify your answer.

5. **[20pt]** Consider again the training data used in Problem 4 and suppose we want to build a decision tree to learn the corresponding Class value. Determine the best split value of the numeric attribute A. Explain the method used in determining the split value.