ARCHIVED

Problem Set 5

Assigned: March 27
Due: April 5

Problem 1

Suppose that you are trying to carry out classification learning where M is the classification attribute and the rest are predictive attributes. You are given the following data set:

ID	A	B	D	M	Number of instances
1.	1	1	2	1	1
2.	1	1	2	2	3
3.	1	1	2	3	2
4.	1	2	1	1	5
5.	1	2	1	3	16
6.	1	2	2	1	15
7.	1	2	2	2	7
8.	1	2	2	3	3
9.	1	2	3	2	1
10.	1	2	3	3	6
11.	2	2	1	2	5
12.	2	2	1	3	1
13.	2	2	2	2	9
14.	2	2	3	1	4
15.	2	2	3	3	22

A. How does Naive Bayes classify the instance $A = 1$, $B = 1$, $D = 1$? (You do not have to use the Laplacian correction.)

B. What is the maximum accuracy attainable on this training set by any possible classifier in predicting M from A, B, D?

C. The simplest baseline classifier is the most common category; that is, the classifier always predicts the same value of M. What is the accuracy of that classifier over this dataset?

D. What is the accuracy of the following linear classifier:

if $A+B+D > 6.5$ then $M = 3$, elseif $A+B+D > 4.5$ then $M = 1$, else $M = 2$

Problem 2

Suppose that your data is as pictured below. The predictive attributes are the x and y coordinates of the dots. The classification attribute is the color (red or blue). The colored points are the training data. The empty dots represent test points.

A. How does 1-nearest neighbors classify the points A, B, C, and D? How does 3-nearest neighbors classify them?

B. Suppose that you use the optimal linear separator. (Don't run a program; just experiment with a straight edge on the picture.) What is the maximum accuracy that you can obtain on the training set? How will it classify the points A, B, C, and D?