CS 363D/ SSC 358 Statistical Learning and Data Mining Spring 2014

Homework 1

Lecturer: Prof. Pradeep Ravikumar Date Due: Feb 12, 2014

Keywords: *Classification*

Note: Hard copy of the solutions is due beginning of class, Wed Feb 12, 2014.

1. (2 points) Consider two binary vectors u and v. Suppose the total number of ones in both the binary vectors together is n; and that dot product of the two vectors is d. What is the Jaccard distance between u and v?

2. (6 points) Consider the binary classification data set in Table 1, with two attributes A and B.

   (a) Calculate the gain in Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

   (b) Repeat part (a) with information gain (i.e. gain in the entropy measure).

A	B	Class Label
1	0	-
1	1	-
1	1	-
1	0	+
1	1	-
0	0	+
0	0	+
0	0	+
1	1	+
1	0	+

Table 1: Data set 1.

3. (7 points) Compute a two-level decision tree for the training data in Table 2 using the greedy approach discussed in the class, with gain in Gini index as the criterion for splitting. What is the overall error rate on the training data of the induced tree?

X	Y	Z	No. of Class 1 examples	No. of Class 2 examples
0	0	0	5	35
0	0	1	0	15
0	1	0	10	5
0	1	1	40	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

Table 2: Data set 2.