The University of Texas
at Austin

CS363D Statistical Learning and Data Mining

Homework 01

Edited by LaTeX

Department of Computer Science

STUDENT

Jimmy Lin
xl5224

INSTRUCTOR

Pradeep Ravikumar

TASSISTANT

Adarsh Prasad

RELEASE DATE

Feb. 02 2014

DUE DATE

Feb. 08 2014

TIME SPENT

6 hours

February 8, 2014

Contents

List of Figures

1 Jaccard Distance

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**?

According to the Jaccard terminology for measuring the similarity of two vectors, M_{11} represents the number of dimensions where both vectors has value 1 and $M_{10} + M_{01}$ represents the number of dimensions where either vectors has value 1.

Since for given two binary vector **u** and **v**, total number of ones in both the binary vectors together is n, and dot product of the two vectors is d, we have

$$\begin{aligned} M_{11} &= d \\ M_{10} + M_{01} &= n - 2d \end{aligned}$$

Note that the dot product only count the dimensions where both vectors are 1.

By the definition of Jaccard Distance, we have

$$\begin{aligned} J &= \frac{M_{00}}{M_{01} + M_{10} + M_{11}} \\ &= \frac{d}{n - 2d + d} \\ &= \frac{d}{n - d} \end{aligned}$$

Thus, the Jaccard Distance between **u** and **v** is $\frac{d}{n-d}$.

2 Binary Classification Data Set

Consider the binary classification data set, with two attributes A and B.

Split on attribute A:

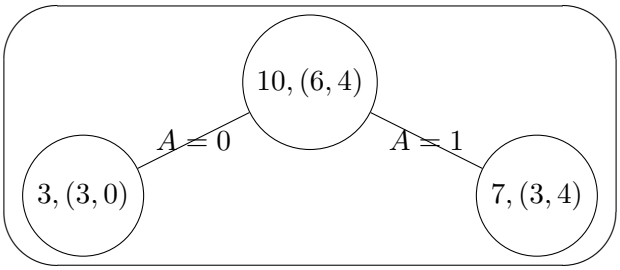

Figure 1: Decision Tree split on attribute A

Split on attribute B:

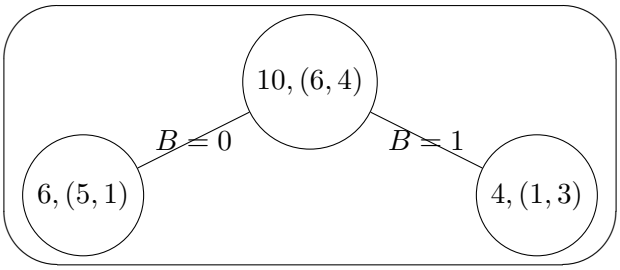

Figure 2: Decision Tree split on attribute B

Note that for $N, (A, B)$ in each node,

N represents number of records in that nodes
A represents number of records with class "+"
B represents number of records with class "-"

2.1 Gain in Gini index

Calculate the gain in Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose? Now let us look at the decision tree split by attribute A and attribute B.

The computation of Gini index split by attribute A is as follows:

$$Gini(root) = 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.48$$
$$Gini(leftChild) = 1 - (\frac{3}{3})^2 - (\frac{0}{3})^2 = 0$$
$$Gini(rightChild) = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.490$$

Thus, the gain of Gini index by spliting at attribute A $Gain_A$ is

$$\begin{aligned} Gain_A &= Gini(root) - \frac{3}{10}Gini(leftChild) - \frac{7}{10}Gini(rightChild) \\ &= 0.48 - \frac{7}{10} \times 0.49 = 0.137 \end{aligned}$$

The computation of Gini index split by attribute B is as follows:

$$Gini(root) = 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.48$$
$$Gini(leftChild) = 1 - (\frac{5}{6})^2 - (\frac{1}{6})^2 = 0.278$$
$$Gini(rightChild) = 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375$$

Thus, the gain of Gini index by spliting at attribute B $Gain_B$ is

$$\begin{aligned} Gain_B &= Gini(root) - Gini(leftChild) - Gini(rightChild) \\ &= 0.48 - \frac{6}{10} \times 0.278 + \frac{4}{10} \times 0.375 = 0.4632 \end{aligned}$$

Comparing the gain of Gini index $Gain_A$ and $Gain_B$, we have

$$Gain_A < Gain_B$$

Thus, **attribute B** is the one we should choose to split since it maximize the Gini gain.

2.2 Information Gain

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

The computation of Entropy split by attribute A is as follows:

$$\begin{aligned} Entropy(root) &= -\big(\frac{6}{10}log_2(\frac{6}{10}) + \frac{10-6}{10}log_2(\frac{10-6}{10})\big) = 0.971 \\ Entropy(leftChild) &= -\big(\frac{3}{3}log_2(\frac{3}{3}) + \frac{3-3}{3}log_2(\frac{3-3}{3})\big) = 0 \\ Entropy(rightChild) &= -\big(\frac{3}{7}log_2(\frac{3}{7}) + \frac{7-3}{7}log_2(\frac{7-3}{7})\big) = 0.985 \end{aligned}$$

Thus, the Information Gain by spliting at attribute A $Gain_A$ is

$$\begin{aligned} IGain^*_A &= Entropy(root) - \frac{3}{10}Entropy(leftChild) - \frac{7}{10}Entropy(rightChild) \\ &= 0.971 - 0 - \frac{7}{10} * 0.985 = 0.2815 \end{aligned}$$

The computation of Entropy split by attribute B is as follows:

$$\begin{aligned} Entropy(root) &= -\big(\frac{6}{10}log_2(\frac{6}{10}) + \frac{10-6}{10}log_2(\frac{10-6}{10})\big) = 0.971 \\ Entropy(leftChild) &= -\big(\frac{5}{6}log_2(\frac{5}{6}) + \frac{6-5}{6}log_2(\frac{6-5}{6})\big) = 0.65 \\ Entropy(rightChild) &= -\big(\frac{1}{4}log_2(\frac{1}{4}) + \frac{4-1}{4}log_2(\frac{4-1}{4})\big) = 0.811 \end{aligned}$$

Thus, the Information Gain by spliting at attribute B $Gain_B$ is

$$\begin{aligned} IGain^*_B &= Entropy(root) - \frac{6}{10}Entropy(leftChild) - \frac{4}{10}Entropy(rightChild) \\ &= 0.971 - \frac{6}{10} * 0.65 - \frac{4}{10} * 0.811 = 0.2566 \end{aligned}$$

Comparing the gain of Gini index $IGain_A$ and $IGain_B$, we have

$$IGain_A > IGain_B$$

Thus, **attribute A** is the one we should choose to split since it maximize the Information gain.

3 Decision Tree

Compute a two-level decision tree for the training data 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	class1	class2				
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				
		sum	95	95				
	First level division		class1	class2	#records within nodes	Gini Index	Gain	Split
	Divided by	x=0	55	55	110	0.5		
	x	x=1	40	40	80	0.5	0	
	Divided by	y=0	40	55	95	0.487534626		Attribute
	y	y=1	55	40	95	0.487534626	0.01246537	Z
	Divided by	z=0	30	65	95	0.432132964		
	z	z=1	65	30	95	0.432132964	0.06786704	
	Second Level division (z=0)				#records within nodes	Gini Index	Gain	Split
	Divided by	x=0	15	40	55	0.396694215		
	x	x=1	15	25	40	0.46875	0.07296651	
	Divided by	y=0	15	40	55	0.396694215		Attribute
	y	y=1	15	25	40	0.46875	0.07296651	X or Y
	Second Level division (z=1)							
	Divided by	x=0	40	15	55	0.396694215		
	x	x=1	25	15	40	0.46875	0.07296651	
	Divided by	y=0	25	15	40	0.46875		Attribute
	y	y=1	40	15	55	0.396694215	0.07296651	X or Y

Figure 3: Statistics about the decision tree production

Construction of decision tree is simple. The technique we employ here is to evaluate whether the class distribution derived by candidate division is close to uniform distribution. According to this principle, we expand over attribute z at the root node and afterwards we can either expand over attribute x or y on both branches of $z = 0$ and $z = 1$.

The resulted decision tree (just one possile solution) is as follows:

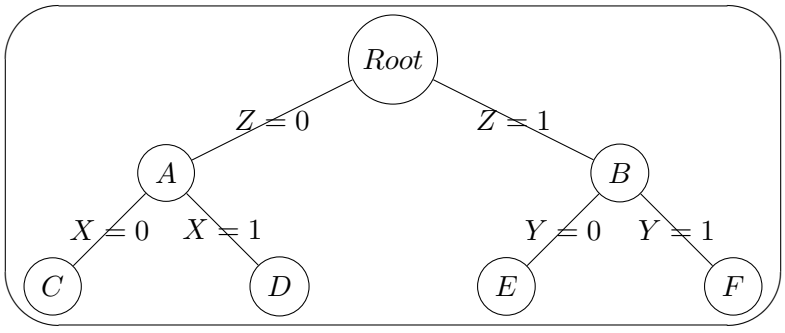

Figure 4: Constructed Decision Tree

For this instance, we classify all records in C and D as "class 1" object, and in the meanwhile assign all records in E and F as "class 2" object. Thus, the training error rate e would be

$$e = (30 + 30)/190 = 15.7\%$$