HW4

1.1 – Exercise 2. Consider the training examples shown in Table 3.5 for a binary classification problem.

a) Compute the Gini index for the overall collection of training examples.

Answer:

Gini Index = $1 - \sum_{i=1}^{n}(pi)^2$

The table consists of 20 rows .

P(C1) = 10/20, P(C2) = 10/20

The Gini index for overall collection of training example is:

1-$\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2$ = 0.5

b) Compute the Gini index for the Customer ID attribute.

Answer:

The customer Id is unique identifier in the table because the Gini index will be 0. The columns in the table is dependent on the Id, which is used to discriminate between individual rows. So the Gini index for the customer id will be 0 and also for every row in Gini index.

c) Compute the Gini index for the attribute.

Answer:

Calculating Gini index for Gender:

P(Gender=M): 10/20

P(Gender=F): 10/20

If(Gender= M & Class=C0)

= $1 - \left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2$= 0.48

If (Gender = F & class= C1)

=$1 - \left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2$= 0.48

Weighted average of Gini index can be calculated for the given split:

$\text{Gini}_{Split}\ (Gender)$ = $\left(\frac{10}{20}\right) * 0.48 + \left(\frac{10}{20}\right) * 0.48$ =0.48.

The Gini index for the Customer ID IS 0.48

d) Compute the Gini index for the Car Type attribute using multiway split.

Answer

Calculating Gini index for Car type:

P(Car type=Family)= 4/20

P(car type= Sports)= 8/20

P(Car type= Luxary)=8/20

Gini index for Family is:

P(Car type=Family)

$1-\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2$ = 0.375

Gini index for Sports:

$1-\left(\frac{0}{8}\right)^2+\left(\frac{8}{8}\right)^2$ = 0

Gini index for Luxury:

$1-\left(\frac{1}{8}\right)^2+\left(\frac{7}{8}\right)^2$ = 0.2656

Weighted avg of Car Type Gini indexes is:

$\mathrm{Gini}_{Split}\ (Car\ Type)$ = $\left(\frac{4}{20}\right)*0.375+\left(\frac{8}{20}\right)*0.2656+\left(\frac{8}{20}\right)*0$ = 0.1625

The Gini index for car type is 0.1625

e) Compute the Gini index for the Shirt Size attribute using multiway split.

Answer:

P(Shirt size= small)= 5/20

P(Shirt size=medium)= 7/20

P(shirt size= Large)= 4/20

P(Shirt size = extra large)=4/20

Gini Index for Small:

$1-\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2$ = 0.48

Gini Index for Medium:

$1 - \left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2$ = 0.4897

Gini Index for Large:

$1 - \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2$ = 0.5

Gini Index for Extra Large:

$1 - \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2$ = 0.5

Weighted avg of these Gini indexes is:

$\text{Gini}_{Split}\ (Shirt\ Size)$ = $\left(\frac{5}{20}\right) * 0.48 + \left(\frac{7}{20}\right) * 0.4898 + \left(\frac{4}{20}\right) * 0.5 + \left(\frac{4}{20}\right) * 0.5$

= 0.49143

Gini Index for Shirt Size Attribute is 0.4914

f) Which attribute is better, Gender, Car Type, or Shirt Size?

Answer:

Gini Index for Gender is = 0.48

Gini Index for Car Type is = 0.1625

Gini Index for Shirt Size is = 0.4919

The Gini index of the Car Type is lowest among the others and it is better, so it will be chosen as the parent node.

g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

Answer:

The Gini Index for the Customer Id attribute is zero. Because each partition will only contain one record, there won't be any purity benefit even if we increase the number of records in the table.

1.1 – Exercise 3. Consider the training examples shown in Table 3.6 for a binary classification problem.

a) What is the entropy of this collection of training examples with respect to the class attribute?

Answer:

Given

The no of positives are =4/9

The no of negatives are =5/9

Entropy = $\sum_{i=1}^{c} -pi \ \log_2 pi$

$= -\left[\left(\frac{4}{9}\right) * \log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) * \log_2\left(\frac{5}{9}\right)\right]$

= -[-0.52 -0.4711]

entropy = 0.9911

The Entropy is 0.9911

b) What are the information gains of a1 and a2 relative to these training examples?

Answer:

Entropy(a1) = $\frac{4}{9} * \left[-\left(\frac{3}{4}\right) * \log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right) * \log_2\left(\frac{1}{4}\right)\right] + \frac{5}{9} * \left[-\left(\frac{1}{5}\right) * \log_2\left(\frac{1}{5}\right) - \left(\frac{4}{5}\right) * \log_2\left(\frac{4}{5}\right)\right]$

=0.44*[0.3112-(-0.5)]+0.55*[0.464-(-0.257)]

=0.44*(0.811)+0.55*(0.721)

=0.36+0.3970

=0.7540

Entropy(a2) = $\frac{5}{9} * \left[-\left(\frac{2}{5}\right) * \log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right) * \log_2\left(\frac{3}{5}\right)\right] + \frac{4}{9} * \left[-\left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right)\right]$

=0.55*[0.528+0.44]+0.44*[0.5+0.5]

=0.534+0.44

=0.97405

Information Gain(S, A)=Entropy(S)-Entropy(S,A)

Gain(a1) =0.9911-0.7540

= 0.2371

Gain(a2)=0.9911-0.98388

= 0.0072

c) For a3, which is a continuous attribute, compute the information gain for every possible split.

Answer:

Total Entropy is 0.9911

We must first sort the values in the A3 attribute. After dividing, we must then identify the midpoint between two adjacent numbers.

A3	Target Class
1	+
3	-
4	+
5	-
5	-
6	+
7	-
7	+
8	-

Here the split point is =0+1/2=0.5

	<=0.5	>0.5
+	0	4
-	0	5

The entropy is 0.9911

Information Gain = Entropy(s)-Entropy(s,a)

=0.9911-0.9911

=0

The split is 0.5, there are no values less than split 0.5 ,s so we considered same as the total entropy of the dataset.

The Split here is = 1+3/2= 2

$$\text{Entropy}=\frac{1}{9}\left[-\frac{1}{1}\log_2\frac{1}{1}-\frac{0}{1}\log_2\frac{0}{1}\right]+\frac{8}{9}\left[-\frac{3}{8}\log_2\frac{3}{8}-\frac{5}{8}\log_2\frac{5}{8}\right]$$

$$\text{Entropy} = 0+\frac{8}{9}\left[-\frac{3}{8}\log_2\frac{3}{8}-\frac{5}{8}\log_2\frac{5}{8}\right]$$

=0.8478

	<=2	>2
+	1	3
-	0	5

Information Gain = 0.9911 – 0.8487

= 0.14326

The split is =3+4/2= 3.5

$$\text{Entropy} = \frac{2}{9}\left(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}\right)+\frac{7}{9}\left[-\frac{3}{7}\log_2\frac{3}{7}-\frac{4}{7}\log_2\frac{4}{7}\right]$$

	<=3.5	>3.5
+	1	3
-	1	4

Entropy = $\frac{2}{9}(1) + \frac{7}{9}\left[-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right]$

= 0.988533

Information Gain = 0.9911 – 0. 988533

= 0.0026

The split is 4+5/2=4.5

Entropy = $\frac{5}{9}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right] + \frac{4}{9}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$

= 0.9838

Information Gain = 0.9911 – 0.9838

= 0.0072

	<=4.5	>4.5
+	2	2
-	1	4

The split is =5+6/2=5.5

Entropy = $\frac{5}{9}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right] + \frac{4}{9}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$

= 0.9838

Information Gain = 0.9911 – 0.9838

= 0.0072

	<=5.5	>5.5
+	2	2
-	3	2

The split is =6+7/2=6.5

Entropy = $\frac{6}{9}\left[-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}\right] + \frac{3}{9}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]$

= 0.9727

Information Gain = 0.9911 – 0.9727

= 0.0183

	<=6.5	>6.5
+	3	1
-	3	2

The split is 7+8/2=7.5

Entropy = $\frac{8}{9}\left[-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}\right] + \frac{1}{9}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right]$

= 0.8888

Information Gain = 0.9911 – 0.8888

= 0.1022

	<=7.5	>7.5
+	4	0
-	4	1

The Split point 8+9/2=8.5

Entropy = $\frac{9}{9}\left[-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9}\right]$

	<=8.5	>8.5

= 0.9911

+	4	0
-	5	0

Information Gain = 0.9911 – 0.9911

= 0

Therefore, Split point ‘2’ provides the largest information gain value of 0.14326 after calculating the information gain for each split position.

d) What is the best split (among A1, A2 and A3) according to the information gain?

Answer:

These are the IG’s of A1,A2,A3

Information Gain (A1): 0.2371

Information Gain (A2): 0.0170

Information Gain (A3): 0.1432

Maximum of the three information gain values mentioned above would be the best split. The value for information gain that is highest is in A1. So the optimal split would be A1.

(e) What is the best split (between a1 and a2) according to the misclassification error rate?

A)

The misclassification error rate is a performance metric that indicates how many predictions are incorrect.

Error(k) = $\left(\frac{|Sv|}{|S|}\right) * 1 - \max(p_t^k)$

P(A1)	+	-
T	$\frac{3}{4}$	$\frac{1}{4}$
F	$\frac{1}{5}$	$\frac{4}{5}$

Error(A1) = $\frac{4}{9}\left(1-\frac{3}{4}\right)+\frac{5}{9}\left(1-\frac{4}{5}\right)$

$=\frac{4}{9}*\frac{1}{4}+\frac{5}{9}*\frac{1}{5}$

$=\frac{2}{9}$

P(A2)	+	-
T	$\frac{2}{5}$	$\frac{3}{5}$
F	$\frac{2}{4}$	$\frac{2}{4}$

Error(A2) = $\frac{5}{9}\left(1-\frac{3}{5}\right)+\frac{4}{9}\left(1-\frac{2}{4}\right)$

$=\frac{5}{9}*\frac{2}{5}+\frac{4}{9}*\frac{2}{4}$

$=\frac{2}{9}+\frac{2}{9}$

$=\frac{4}{9}$

A1 produces the best split because its error rate is lower than that of A2.

f) What is the best split (between a1 and a2) according to the Gini index?

Answer:

Gini Index = $1 - \sum_{i=1}^{n}(pi)^2$

Gini Index for A1 Attribute:

Total Records for A1 is T = $\frac{4}{9}$

Total Records for A2 is F = $\frac{5}{9}$

Gini Index = $\frac{4}{9}\left(1-\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)+\frac{5}{9}\left(1-\left(\frac{1}{5}\right)^2+\left(\frac{4}{5}\right)^2\right)$

= $\frac{4}{9}(1-0.625)+\frac{5}{9}(1-0.68)$

=$\frac{4}{9}(0.375)+\frac{5}{9}(0.32)$

=$0.1666+0.1777$

=0.3443

Gini Index for A2 Attribute:

Total Records for A2 is T = $\frac{5}{9}$

Total Records for A2 is F = $\frac{4}{9}$

Gini Index = $\frac{5}{9}\left(1-\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right)+\frac{4}{9}\left(1-\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)$ = 0.488

= $\frac{5}{9}(1-0.52)+\frac{4}{9}(1-0.5)$

= $\frac{5}{9}(0.48)+\frac{4}{9}(0.5)$

=$\frac{5}{9}(0.48)+\frac{4}{9}(0.5)$

=0.266+0.222

=0.4889

Gini Index for Attribute A1 is lower. So, A1 would be the best split.

1.1-Exercise 5. Consider the following data set for a binary class problem.

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

Answer:

entropy= $\sum_{i=1}^{c} -pi\ \log_2 pi$

= $\left[-\left(\frac{4}{10}\right) * \log_2\left(\frac{4}{10}\right) - \left(\frac{6}{10}\right) * \log_2\left(\frac{6}{10}\right)\right]$

=[-(0.4)* -1.321- (0.6)- 0.767]

= 0.971

entropy(A) $= \frac{7}{10} * \left[-\left(\frac{4}{7}\right) * \log_2\left(\frac{4}{7}\right) - \left(\frac{3}{7}\right) * \log_2\left(\frac{3}{7}\right)\right] + \frac{3}{10} * \left[-\left(\frac{3}{3}\right) * \log_2\left(\frac{3}{3}\right) - \left(\frac{0}{3}\right) * \log_2\left(\frac{0}{3}\right)\right]$

entropy(A) = 0.689

Information gain= E(S) –E(S,A)

= 0.971-0.68968

IG = 0.28132

Entropy(B) $= \frac{4}{10} * \left[-\left(\frac{3}{4}\right) * \log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right) * \log_2\left(\frac{1}{4}\right)\right] + \frac{6}{10} * \left[-\left(\frac{1}{6}\right) * \log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right) * \log_2\left(\frac{5}{6}\right)\right]$

$= \frac{4}{10} * 0.811 + \frac{6}{10} * 0.650$

$= 0.3244\ +\ 0.390$

= 0.7144

Information Gain = 0.971 – 0.714

= 0.257

The best split would be the maximum of the above two information gain values. A has the highest information gain value. As a result, A would be the best split.

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

Answer

Gini Index for overall collection of training examples is: $1-\left(\frac{4}{10}\right)^2+\left(\frac{6}{10}\right)^2=0.48$

Gini Index for A Attribute:

Total Records for A is T = 7/10

Total Records for A is F = 3/10

Gini Index = $\frac{7}{10}\left(1-\left(\frac{4}{7}\right)^2+\left(\frac{3}{7}\right)^2\right)+\frac{3}{10}\left(1-\left(\frac{3}{3}\right)^2+\left(\frac{0}{3}\right)^2\right)$

$$=0.342$$

Gini Index for B Attribute:

Total Records for B is T = 4/10

Total Records for B is F = 6/10

Gini Index = $\frac{4}{10}\left(1-\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)+\frac{6}{10}\left(1-\left(\frac{1}{6}\right)^2+\left(\frac{5}{6}\right)^2\right)$

$$=0.3166$$

Gain(A) = Entropy(S) – Entropy (S, A) = 0.48 – 0.342

= 0.138

Gain(B) = 0.48 – 0.3166

= 0.16333

Information Gain from B is higher than from A. So, we need to choose B.

c) Figure 3.11 shows that entropy and the Gini index are both monotonically increasing on the range [0, 0.5] and they are both monotonically decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.

Answer:

Yes, the measures have the same range as 0 to 1and the same monotonously actions, these are having same ranges but their respective gains , Δ these are scaled on the different in measures, and they are having different behavior even they are scaled in same range. The gain from the information gain and Gini index will favor different attributes.

1.2-Exercise 18. Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains instances from two classes, "+ and "-". Half of the data set is used for training while the remaining half is used for testing.

a) Suppose there are an equal number of positive and negative instances in the data and the decision tree classifier predicts every test instance to be positive. What is the expected error rate of the classifier on the test data?

Answer:

A	B	Target
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	+
T	T	-
T	F	-
T	F	+
F	T	-

Assume the data is evenly distributed, as stated in the question. Consider the following example from the table. There are ten rows in total, with positive and negative values distributed equally, implying that half of the dataset belongs to the positive class and the rest to the negative class.

If the classifier predicts that every test instance will be positive, the accuracy is 50%.

We can compute the accuracy score using the confusion matrix.

$$Accuracy\ Score = \frac{TP + TN}{TP + FP + TN + FN}$$

$$= \frac{6+0}{6+6+0+0} = \frac{1}{2}$$

Accuracy score = 0.5

	Actual		
Predicted		+	-
	+	6	6
	-	0	0

Error Rate in % = (1 – Accuracy Score) x 100

= 1 – 0.5 =0.5

=0.5*100

Error Rate = 50%

b) Repeat the previous analysis assuming that the classifier predicts each test instance to be positive class with probability 0.8 and negative class with probability 0.2.

Answer:

The probability on classifier of predicting True positives os 80%, so the

TP =0.8

The probability of True Negative the probability is 20%

TN= 0.2

The TP and TN in the confusion matrix above

TP= 6 and TN=0

Accuracy== $\frac{0.8(6)+0.2(6)}{12}$= 0.5

Error rate percentage=(1 – Accuracy Score) x 100

= 50%

c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test instance to be positive?

Answer:

two-thirds of the data belong to the positive class then the TP will be 2/3 and the

remaining one third belongs to negative class means TN is 1/3 , if this is the case then the accuracy will be

$Accuracy\ Score = \frac{\left(\frac{2}{3}\right)*(12)+0}{12}$ = 0.666

Error rate in % = (1 – Accuracy Score) x 100

=(1-0.666)*100

=0.33*100

Error rate = 33

=33%

The error rate is 33 percentage

d) Repeat the previous analysis assuming that the classifier predicts each test instance to be positive class with probability 2/3 and negative class with probability 1/3.

Answer:

The probability of classifier predicting positive class is 2/3 and the classifier predicting Negative class is 1/3 as given, then the accuracy will be

Accuracy= $\frac{\left(\frac{2}{3}\right)*\left(\frac{2}{3}\right)*(12)+\left(\frac{1}{3}\right)*\left(\frac{1}{3}\right)*(12)}{12}$

Accuracy= $\frac{6.6666}{12}$

Accuracy=0.55555

Error rate=(1-Accuracy score)*100

=(1-0.55555)*100

=0.44445*100

=44

The error rate is 44%.

1.3. Multiclass Classification

Using the confusion matrix from multiclass.Rmd notebook (from Lecture 7), create a binary-class confusion matrix using the "one-vs-many" strategy for each class. Then, for each class, compute the sensitivity, specificity and precision to two decimal places. Show all work, including the binary class confusion matrices.

Answer

		Actual		
		Setosa	Versicolor	Virginica
Predicted	Setosa	10	0	0
	Versicolor	0	10	1
	Virginica	0	0	9

To generate a binary class confusion matrix, we must first manually compute all of the TP, TN, FP, and FN for each class.

Binary-class Confusion matrix for "Setosa" Class

True Positive(TP): The actual and predicted values should match. The value should be obtained from the corresponding cells. In our case, we need to get the value from cell 1 for class setosa.

TP= 10

False Positive: The total of the values in cells 2 and 3 that correspond to the relevant rows but do not include the TP value

FP= 0+0 =0

False Negative: The sum of all comparable column values, represented by cells 4 and 7, but excluding the TP value.s

FN= 0+0 =0

True Negative: The values of cells 5, 6, 8, and 9 that are in classes other than the one for which we are performing the calculation.

TN = 10 + 1 + 0 + 9 = 20

Confusion Matrix for “Setosa” class:

	Actual	
Predicted	10	0
	0	10

Sensitivity: $\frac{TP}{TP+FN} = \frac{10}{10+0} = 1$

Specificity: $\frac{TN}{TN+FP} = \frac{20}{20+0} = 1$

For class “Setosa”,

Sensitivity = 1

Specificity = 1

Binary-class Confusion matrix for “Versicolor” Class:

TP = 10, FP = 1, FN = 0, TN = 19

Confusion Matrix for “Versicolor” class:

	Actual	
Predicted	10	1
	0	20

Sensitivity: $\frac{TP}{TP+FN} = \frac{10}{10+0} = 1$

Specificity: $\frac{TN}{TN+FP} = \frac{19}{19+1} = 0.95$

For class “Versicolor”,

Sensitivity = 1

Specificity = 0.95

Binary-class Confusion matrix for “Virginica” Class:

TP = 9, FP = 0, FN = 1, TN = 20

Confusion Matrix for Virginica class:

	Actual	
Predicted	9	0
	1	20

Sensitivity: $\frac{TP}{TP+FN} = \frac{9}{9+1} = 0.9$

Specificity: $\frac{TN}{TN+FP} = \frac{20}{20+0} = 1$

For class “Virginica”,

Sensitivity = 0.9

Specificity = 1`