Homework Assignment N°4

BML36
Thibault Douzon
Rajavarman Mathivanan

September 26th, 2018

Contents

1 Exercise 1: Decision Trees

1.1 Part a

a 0.51996

b entropy of the dataset: 0.991

Feature a_1	+	-	p_+	p_-	entropy
T	3	1	$\frac{3}{4}$	$\frac{1}{4}$	0.811
F	1	4	$\frac{1}{5}$	$\frac{4}{5}$	0.722

new entropy for a_1: $0.811 \times \frac{4}{9} + 0.722 \times \frac{5}{9} = 0.762$
information gain of a1 = 0.229

Feature a_2	+	-	p_+	p_-	entropy
T	2	3	$\frac{2}{5}$	$\frac{3}{5}$	0.971
F	2	2	$\frac{2}{4}$	$\frac{2}{4}$	1

new entropy for a_2: $0.971 \times \frac{5}{9} + 1 \times \frac{4}{9} = 0.762$
information gain of a2 = 0.007

c Entropy for 0.5 split is: 0.9910760598382223, information gain: -1.1102230246251565e-16
Entropy for 1.5 split is: 0.8483857803777466, information gain: 0.14269027946047563
Entropy for 2.5 split is: 0.8483857803777466, information gain: 0.14269027946047563
Entropy for 3.5 split is: 0.9885107724710845, information gain: 0.002565287367137681
Entropy for 4.5 split is: 0.9182958340544896, information gain: 0.07278022578373267
Entropy for 5.5 split is: 0.9838614413637048, information gain: 0.007214618474517431
Entropy for 6.5 split is: 0.9727652780181631, information gain: 0.018310781820059074
Entropy for 7.5 split is: 0.8888888888888888, information gain: 0.10218717094933338
Entropy for 8.5 split is: 0.9910760598382223, information gain: -1.1102230246251565e-16

d best split is a1 (information gain is 0.229)

e Error rate:

$$\text{error}(t) = 1 - max_i[p(i|t)]$$

a1:
error on T node: $1 - 3/4$
error on F node: $1 - 4/5$
global classification error on a1 split: $(1 - 3/4) * 4/9 + (1 - 4/5) * 5/9 = 2/9$
a2:
error on T node: $1 - 3/5$
error on F node: $1 - 2/4$
global classification error on a1 split: $(1 - 3/5) * 5/9 + (1 - 2/4) * 4/9 = 4/9$
Best split is the one with fewer global classification error -> a1

f Gini :

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1}[p(i|t)]^2$$

a1:
Gini on node T: $1-((3/4)^2+(1/4)^2) = 0.375$
Gini on node F: $1-((4/5)^2+(1/5)^2) = 0.320$
Global Gini on a1 $= 0.375 * 4/9 + 0.320 * 5/9 = 0.344$
a2:
Gini on node T: $1-((3/5)^2+(2/5)^2) = 0.480$
Gini on node F: $1-((2/4)^2+(2/4)^2) = 0.5$
Global Gini on a1 $= 0.480 * 5/9 + 0.5 * 4/9 = 0.489$
Best split is the one with fewer Gini index -> a1

1.2 Part b

Feature a_1	low	high	p_{low}	p_{high}	Gini
bad	1	3	$\frac{1}{4}$	$\frac{3}{4}$	0.375
average	3	2	$\frac{3}{5}$	$\frac{2}{5}$	0.480
good	3	1	$\frac{3}{4}$	$\frac{1}{4}$	0.375

Overall Gini average index for the split is: $5/20 * 0.375 + 8/20 * 0.480 + 7/20 * 0.375 = 0.417$

1.3 Part c

confidense interval $= [0.8191; 0.9082]$

2 Exercise 2:Classification of 3 class confusion matrix

2.1 Part a

The accuracy of the classifier

Accuracy $= \frac{sumofalltruepositive}{sumofalltheresults} = \frac{110+130+120}{110+8+7+16+130+10+26+5+120} = 0.8333(83.3\%)$

2.2 Part b

The precision for class C2

Precision C2 $= \frac{truepositiveofC2}{sumofallpredictedpositiveofC2} = \frac{130}{130+8+5} = 0.909(90.9\%)$

2.3 Part c

The precision for class C3

Recall C3 $= \frac{Truepositive}{TotalActualPositive} = \frac{5}{26+5+120} = 0.033(3.3\%)$