Homework Assignment N°4

BML36
Thibault Douzon
Rajavarman Mathivanan

September 26th, 2018

Contents

1 Exercise 1: Decision Trees

1.1 Part a

0.51996

1.2 Part 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

1.3 Part 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

1.4 Part d

best split is a1 (information gain is 0.229)

1.5 Part e

Error rate:

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

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

1.6 Part f