

Data Mining 2

CAT 2 - 2019/2020

Name _____ Surname _____ ID: _____

Test id. AUTO

Answers

Q1: $y_i = 1 \text{ if } w\dot{x}_i + b \geq 1 - \xi_i \text{ else } -1 \text{ if } w\dot{x}_i + b \leq -1 + \xi_i$

Q2: The trick is to transform the data from its original space x into a new space $\Phi(x)$ so that a linear decision boundary can be used.

Q3: Cross-Industry Standard Process for Data Mining.

Q4: 2, 4

Q5: 1, 2 ¶

Q6:

id	X_1	X_2	Y
a	1	1	1
b	0	2	1
c	1	0	-1

Lambda = 0.4
f = sign

test

id	X_1	X_2	Y
1	0	0	-1
2	-1	1	1
3	2	0	-1

it	W_0	W_1	W_2	$X.W$	$f(X.W)$	error	δ_{a_0}	δ_{a_1}	δ_{a_2}
1	-1	0	0	-1	-1	2	0,8	0,8	0,8
2	-0,2	0,8	0,8	1,4	1	0	0	0	0
3	-0,2	0,8	0,8	0,6	1	-2	-0,8	-0,8	0
4	-1	0	0,8	-0,2	-1	2	0,8	0,8	0,8
5	-0,2	0,8	1,6	3	1	0	0	0	0
6	-0,2	0,8	1,6	0,6	1	-2	-0,8	-0,8	0
7	-1	0	1,6	0,6	1	0	0	0	0
8	-1	0	1,6	2,2	1	0	0	0	0
9	-1	0	1,6	-1	-1	0	0	0	0
10									
11									
12									

Q7: 3

Q8:

<u>Plan</u>	Y	N	<u>Sex</u>	Y	N
travel	3/5	2/5	M	2/4	2/4
normal	3/4	1/4	F	2/5	3/5

 Delta Gain = $5/9 * 2/5 + 4/9 * 1/4 = 3/9$ Delta Gain = $5/9 * 2/5 + 4/9 * 2/4 = 4/9$

plan	sex	minutes	churn	weight	new weight	norm weight
travel	F	90	N	1/9	0.06	0.07
travel	F	130	Y	1/9	0.06	0.07
travel	M	70	N	1/9	0.06	0.07
travel	M	80	N	1/9	0.06	0.07
normal	M	90	Y	1/9	0.21	0.255
normal	M	120	Y	1/9	0.06	0.07
normal	F	100	Y	1/9	0.06	0.07
normal	F	110	N	1/9	0.21	0.255
travel	F	100	N	1/9	0.06	0.07

Gain Function = Misclassification Error $Z = 0.84$

70	80	90	90	100	100	110	120	130
N	N	N	Y	Y	N	N	Y	Y

<u>M<90</u>	Y	N	<u>M<100</u>	Y	N
<90	5/7	2/7	<100	1/4	3/4
>=90	0	2/2	>=100	3/5	2/5

 Delta Gain = $7/9 * 2/7 = 2/9$ Delta Gain = $4/9 * 1/4 + 5/9 * 2/5 = 3/9$

<u>M<120</u>	Y	N	<u>M<110</u>	Y	N
<120	2/2	0	<110	2/6	4/6
>=120	2/7	5/7	>=110	1/3	2/3

 Delta Gain = $7/9 * 2/7 = 2/9$ Delta Gain = $6/9 * 2/6 + 3/9 * 1/3 = 3/9$

Split by: Minutes

Error = 2/9

Alpha = $1/2 * \ln((1-2/9)/(2/9)) = 1/2 \ln(7/2) = 0.626$

$w_i^{i+1} = 1/9 * e^{0.626} = 0.21$
misclassified

$w_i^{i+1} = 1/9 * e^{-0.626} = 0.06$ classified correctly

List id questions: [3, 9, 10, 13, 17, 230, 27, 400]