21/9/21

DWDM

Tutorial-5

u18co021 - Sahil Bondre

Q1 Converting decesion tree to rules is preferred because:

a Converting to rules allows distinguishing among different contexts in which a decesion node is used

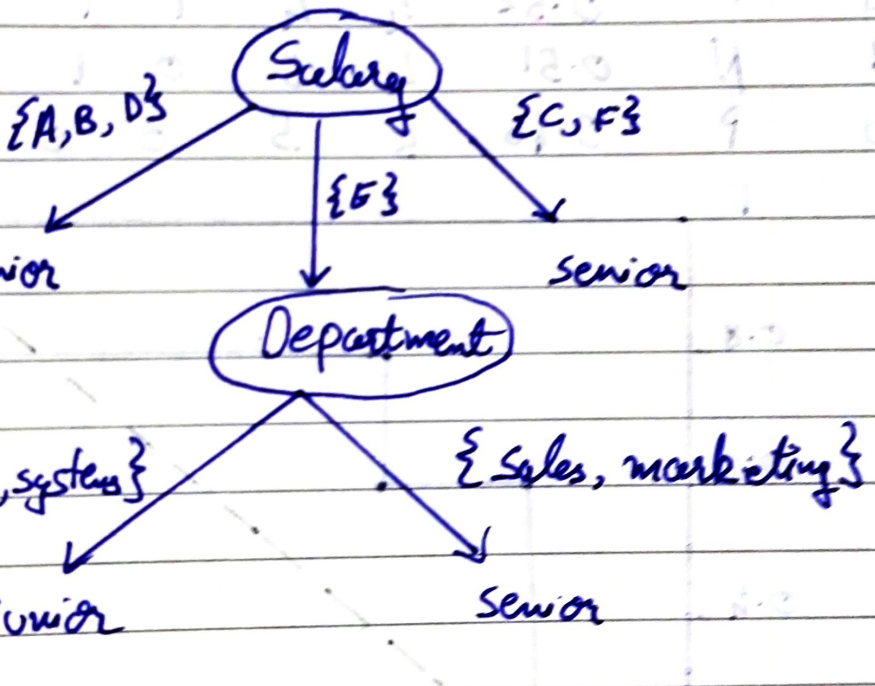
b It removes the distinction between attribute tests occuring at the root & leaves of the tree

c It improves readability

q2

a. To consider count of data tuple, it needs to be ~~under~~ included in the attribute selection measure (such as information gain)

b. Salary Ranges: $A = \{26-30k\}$, $B = \{31-35k\}$
$C = \{36-40k\}$, $D = \{41-45k\}$
$E = \{46-50k\}$, $F = \{66-70k\}$



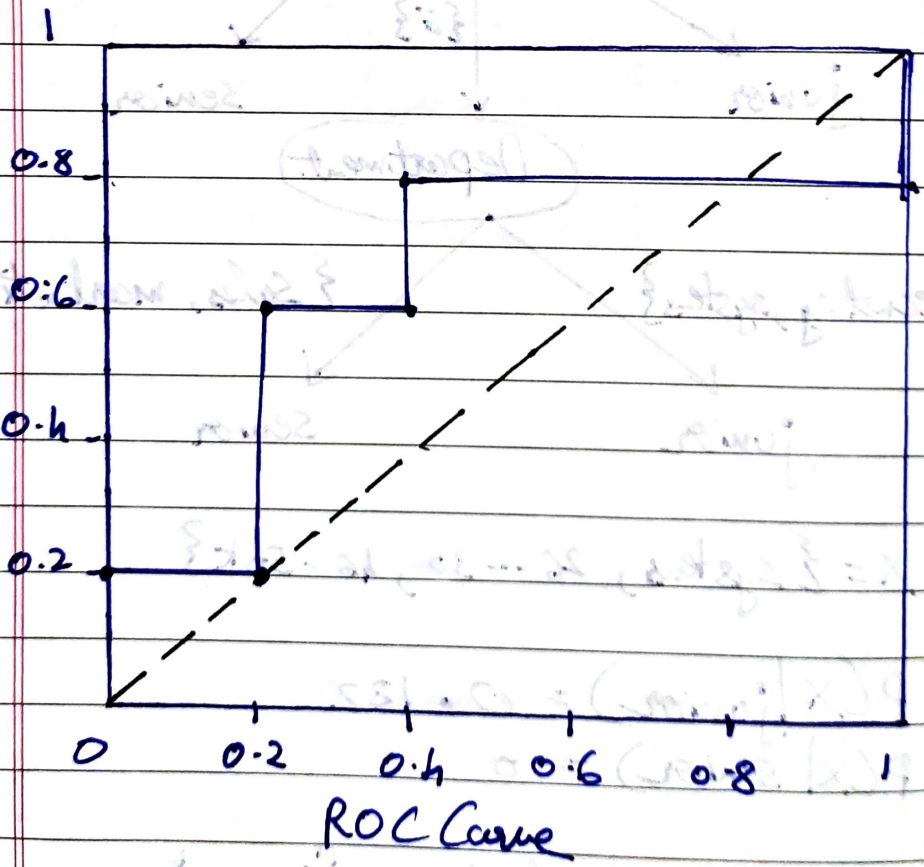
c. $X = \{systems, 26 \cdots 30, 46-50k\}$

$P(X|junior) = 0.182$
$P(X|senior) = 0$

$\therefore$ NB predicts "junior"

Q3

Tuple	Class	P	TP	FP	TN	FN	TPR	FPR
1	P	0.95	1	0	5	4	0.2	0
2	N	0.85	1	1	4	4	0.2	0.2
3	P	0.78	2	1	4	3	0.4	0.2
4	P	0.66	3	1	4	2	0.6	0.2
5	N	0.60	3	2	3	2	0.6	0.4
6	P	0.55	4	2	3	1	0.8	0.4
7	N	0.53	4	3	2	1	0.8	0.6
8	N	0.52	4	4	1	1	0.8	0.8
9	N	0.51	4	5	0	1	0.8	1
10	P	0.40	5	5	0	0	1	0



ROC Curve

classmate
Date
Page

Q4 Methods for class imbalance:

i Oversampling: Resample random tuples from minority class

ii Undersampling: Delete random tuples from majority class.

iii Threshold Moving: Shifting the threshold to reduce the costly false negative errors.

iv Cost-Modifying: Increasing the cost of mis-classifying the minority class

v Ensemble methods: Using boosting and random forests.

SVM & MLP would perform better on imbalanced datasets than tree based classifiers. SVM is not sensitive to imbalance as support vectors only look for boundries between clusters.

Oversampling perfors better for numeric datasets like ~~or~~ credit card fraud transaction. Cost Modifying would help too.