CS795/895: Topics in Data Mining and Security

Summer 2011

Homework #3: Due: June 3, 2011

Problem 1. Given the following 30 instances of a dataset of which 15 are yes, along with the predicted probability of a **yes** response, determine (i) Lift chart (ii) ROC curve

Instance#	**A1 {a,b,c}**	**A2 {x,y,z}**	**A3 {p,q,r,s}**	**Actual Response {yes, no}**	**Predicted Probability of yes**
1	a	x	p	yes	0.99
2	b	z	q	yes	0.94
3	b	y	q	no	0.97
4	a	z	q	yes	0.80
5	a	x	q	yes	0.75
6	b	x	p	no	0.40
7	b	y	p	no	0.90
8	a	y	q	yes	0.70
9	b	z	p	yes	0.87
10	b	x	q	no	0.55
11	a	y	p	yes	0.77
12	b	z	q	Yes	0.94
13	a	z	p	no	0.60
14	b	y	p	yes	0.65
15	a	x	p	no	0.70
16	a	y	r	yes	0.65
17	b	z	r	no	0.95
18	a	z	r	no	0.87
19	b	x	r	yes	0.90
20	a	x	r	no	0.95
21	c	x	p	yes	0.97
22	c	y	q	yes	0.6
23	c	z	r	no	0.87
24	c	x	s	no	0.60
25	c	y	p	yes	0.75
26	c	z	q	yes	0.95
27	c	x	r	no	0.65
28	c	y	r	yes	0.99
29	c	x	q	yes	0.70
30	c	y	s	no	0.45

Problem 2. Consider the above data ignoring the predicted probability. Manually (i) Build a decision tree (DT). Use pruning where possible. (ii) Build classification rules (CR)

In both cases (DT and CR), clearly show your work: how you decided on a particular attribute first (in the case of decision tree), what computations were used (e.g., information gain, etc.), pruning, errors computed, etc.

Clearly, explain the final output for each.