

CS795/895: Topics in Data Mining and Security

Summer 2009

Homework #3: Due: June 3, 2009 (Wednesday)

Problem 1. Given the following 15 instances of a dataset of which 9 are yes, along with the predicted probability of a **yes** response, determine (i) lift chart (ii) ROC curve

Instance#	A1 {a,b}	A2 {x,y,z}	A3 {p,q}	Actual Response {yes, no}	Predicted Probability
1	a	x	p	yes	0.95
2	b	z	q	no	0.94
3	b	y	q	yes	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

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.