Evaluation – Counting the Costs

CSCI 347, Data Mining

Confusion Matrix for Binary Classification

		Predicted	
		Yes	**No**
Actual	**Positive**	TP	FN
	Negative	FP	TN

On diagonal:

TP – true positive

TN – true negative

False results in corners

The Kappa Statistic

Kappa statistic - measures relative improvement over random predictor

$$\frac{D_{observed} - D_{random}}{D_{perfect} - D_{random}}$$

Two confusion matrices for a 3-class problem: actual predictor (left) vs. random predictor (right)

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	88	10	2	*100*
	b	14	40	6	*60*
	c	18	10	12	*40*
	total	*120*	*60*	*20*	

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	60	30	10	*100*
	b	36	18	6	*60*
	c	24	12	4	*40*
	total	*120*	*60*	*20*	

The Kappa Statistic

Two confusion matrices for a 3-class problem: actual predictor (left) vs. random predictor (right)

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	88	10	2	*100*
	b	14	40	6	*60*
	c	18	10	12	*40*
	total	*120*	*60*	*20*	

		Predicted class			
		a	*b*	*c*	*total*
Actual class	*a*	60	30	10	*100*
	b	36	18	6	*60*
	c	24	12	4	*40*
	total	*120*	*60*	*20*	

$$\frac{D_{observed} - D_{random}}{D_{perfect} - D_{random}}$$

$$\frac{140 - 82}{200 - 82} = \frac{58}{118} = 0.49$$

Cost-Sensitive Learning

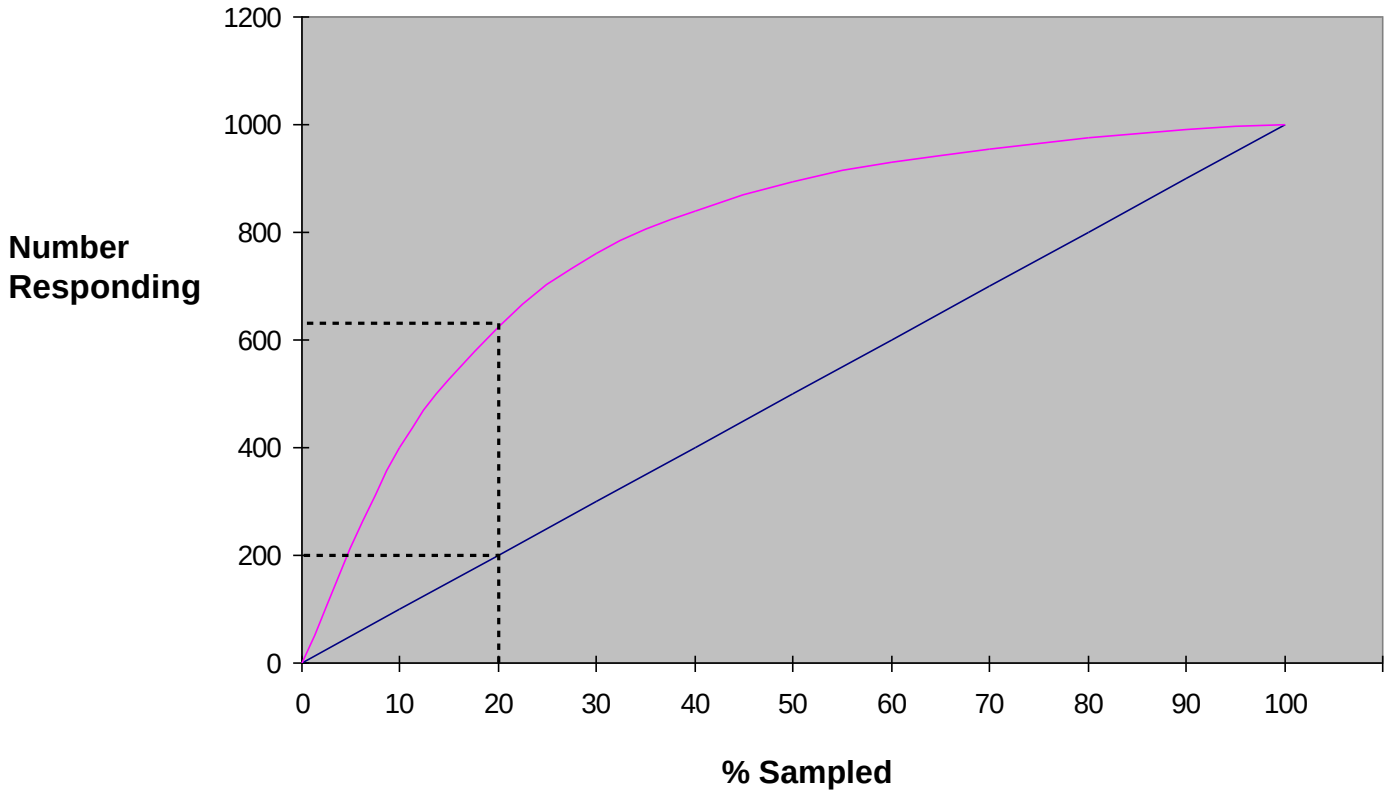
Cost-sensitive learning –

- Can skew the data according to the costs
- Can weight instances according to the costs ,for example, give more weight to an instance that costs less (some versions of Naïve Bayes allow weighting the instances)

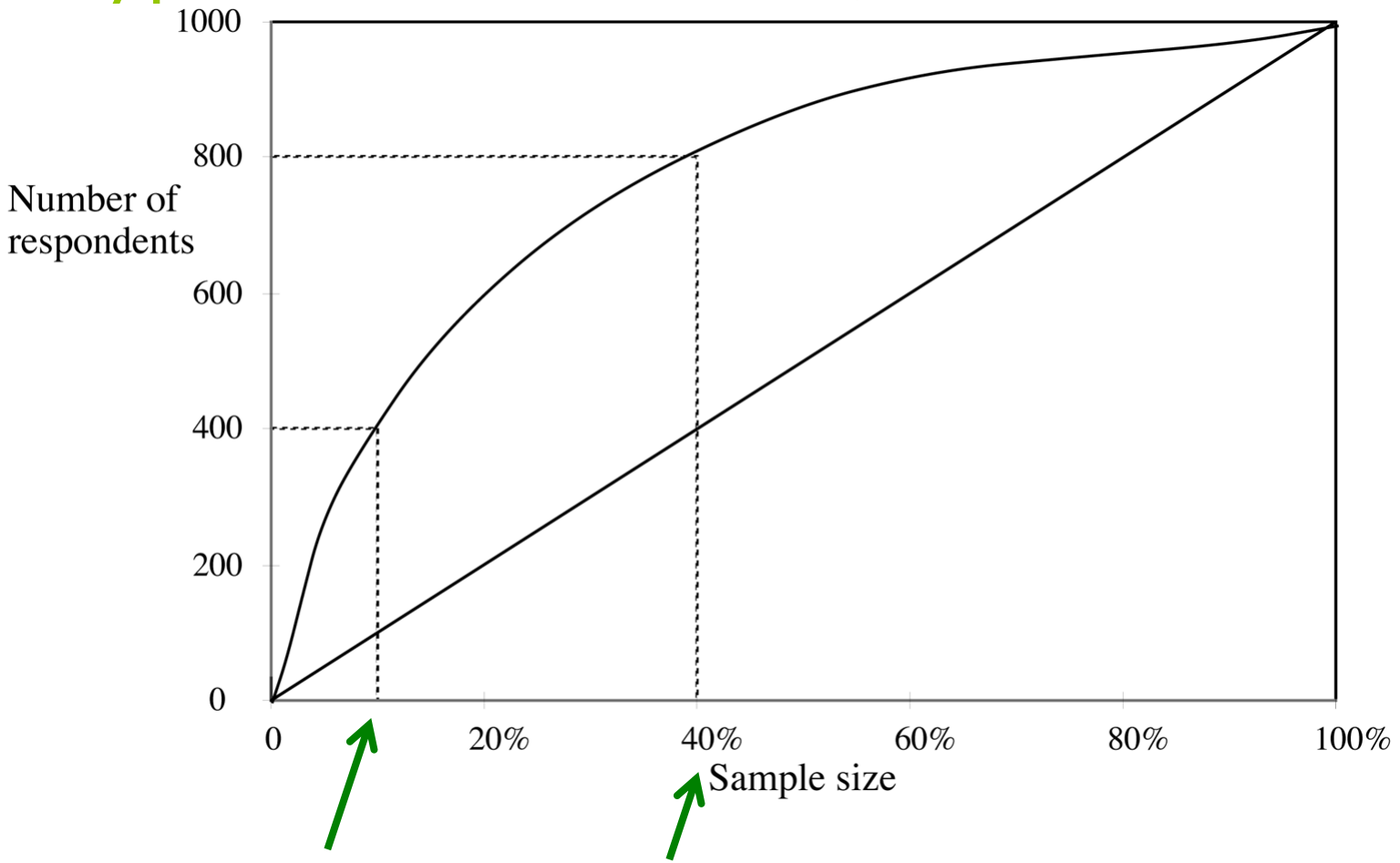
Lift Charts

- Example: promotional mailout to 1,000,000 households
  - Mail to all; 0.1% respond (1000)
  - Data mining tool identifies subset of 100,000 most promising, 0.4% of these respond (400)
    40% of responses for 10% of cost may pay off
  - Identify subset of 400,000 most promising, 0.2% respond (800)

A *lift chart* allows a visual comparison


Number Responding
1200
1000
800
600
400
200
0
0
10
20
30
40
50
60
70
80
90
100
% Sampled

A Hypothetical Lift Chart



40% of responses for 10% of cost

80% of responses for 40% of cost

More Measures…

Precision=TP/(TP+FP)

Test specificity

When the test predicts the person has the disease, how often is it true?

Recall =TP/(TP+FN)

Test sensitivity

When a person has the disease, how often does the test find it?

Summary measure:

F-measure=(2 × recall × precision)/(recall+precision)