ECLT 5810/SEEM5750
Evaluation of Classification Quality

Reference:
Data Science for Business
by F. Provost and T. Fawcett, O'Reilly
Chapter 5

Testing and Error

- Error rate: proportion of errors made over the whole set of instances.
- Test set (Holdout data): set of independent instances that have played no part in formation of classifier
  - Assumption: both training data and test data are representative samples of the underlying problem

Holdout estimation

- The *holdout* method reserves a certain amount for testing and uses the remainder for training
  - Usually: one third for testing, the rest for training
- Problem: the samples might not be representative
  - Example: class might be missing in the test data
- Advanced version uses *stratification*
  - Ensures that each class is represented with approximately equal proportions in both subsets

Repeated holdout method

- Holdout estimate can be made more reliable by repeating the process with different subsamples
  - In each iteration, a certain proportion is randomly selected for training (possibly with stratification)
  - The error rates on the different iterations are averaged to yield an overall error rate
- This is called the *repeated holdout* method
- Still not optimum: the different test sets overlap
  - Can we prevent overlapping?

Cross-validation

- Cross-validation avoids overlapping test sets
  - First step: data is split into k subsets of equal size
  - Second step: each subset in turn is used for testing and the remainder for training
- This is called *k-fold cross-validation*
- Often the subsets are stratified before the cross-validation is performed
- The error estimates are averaged to yield an overall error estimate

Cross-validation

- Split the available data set into k equal partitions, namely, $P_1, \ldots P_k$

Training set	Testing set	Accuracy
$P_2, \ldots, P_k$	P_1	A_1
$P_1, P_3, \ldots, P_k$	P_2	A_2
:	:	
$P_1, P_2, \ldots, P_{k-1}$	P_k	A_k
	Average Accuracy	A

More on cross-validation

- Standard method for evaluation: stratified ten-fold cross-validation
- Why ten? Extensive experiments have shown that this is the best choice to get an accurate estimate
  - There is also some theoretical evidence for this
- Stratification reduces the estimate's variance
- Even better: repeated stratified cross-validation
  - e.g. ten-fold cross-validation is repeated ten times and results are averaged (reduces the variance)

Binary Classification

- For each testing instances, there are only four possible situations:
  - predicted: yes, actual: yes
  - predicted: yes, actual: no
  - predicted: no, actual: yes
  - predicted: no, actual: no
- The contingency table records the total number of testing instances for each situation

		Predicated Class	
		YES	NO
Actual Class	YES	True Positive	False Negative
	NO	False Positive	True Negative

Binary Classification

		Predicated Class	
		YES	NO
Actual Class	YES	True Positive (TP)	False Negative (FN)
	NO	False Positive (FP)	True Negative (TN)

Error rate = $$\frac{FP + FN}{TP + FP + FN + TN}$$

Accuracy rate = 1 – Error rate

A Marketing Application Scenario

- In a direct mailing business, a mass mailout of a promotional offer to a million households (1,000,000).
- Let the response rate is 0.1% (i.e., 1,000 respondents).
- Suppose a random selection of a subset of 100,000 households for mailing.
  - The number of respondent is 100.
- Suppose a data mining method is used and the response rate is 0.4% (400 respondents)

Undesirable Effect of Accuracy

Random Prediction

		Predicated Class		
		YES	NO	total
Actual Class	YES	100	900	1,000
	NO	99,900	899,100	999,000
	total	100,000	900,000	1,000,000

Accuracy = 0.8992
(Error = 0.1008)

A Data Mining Method

		Predicated Class		
		YES	NO	total
Actual Class	YES	400	600	1,000
	NO	99,600	899,400	999,000
	total	100,000	900,000	1,000,000

Accuracy = 0.8998
(Error = 0.1002)

Lift Factor

- The random response rate is 0.1% (due to 100 respondents out of 100,000).
- The response rate of a certain data mining method is 0.4% (due to 400 respondents out of 100,000)
- The increase in response factor, is known as the *lift* factor
  - In the previous example, the lift factor is:

$$\frac{0.4}{0.1} = 4$$

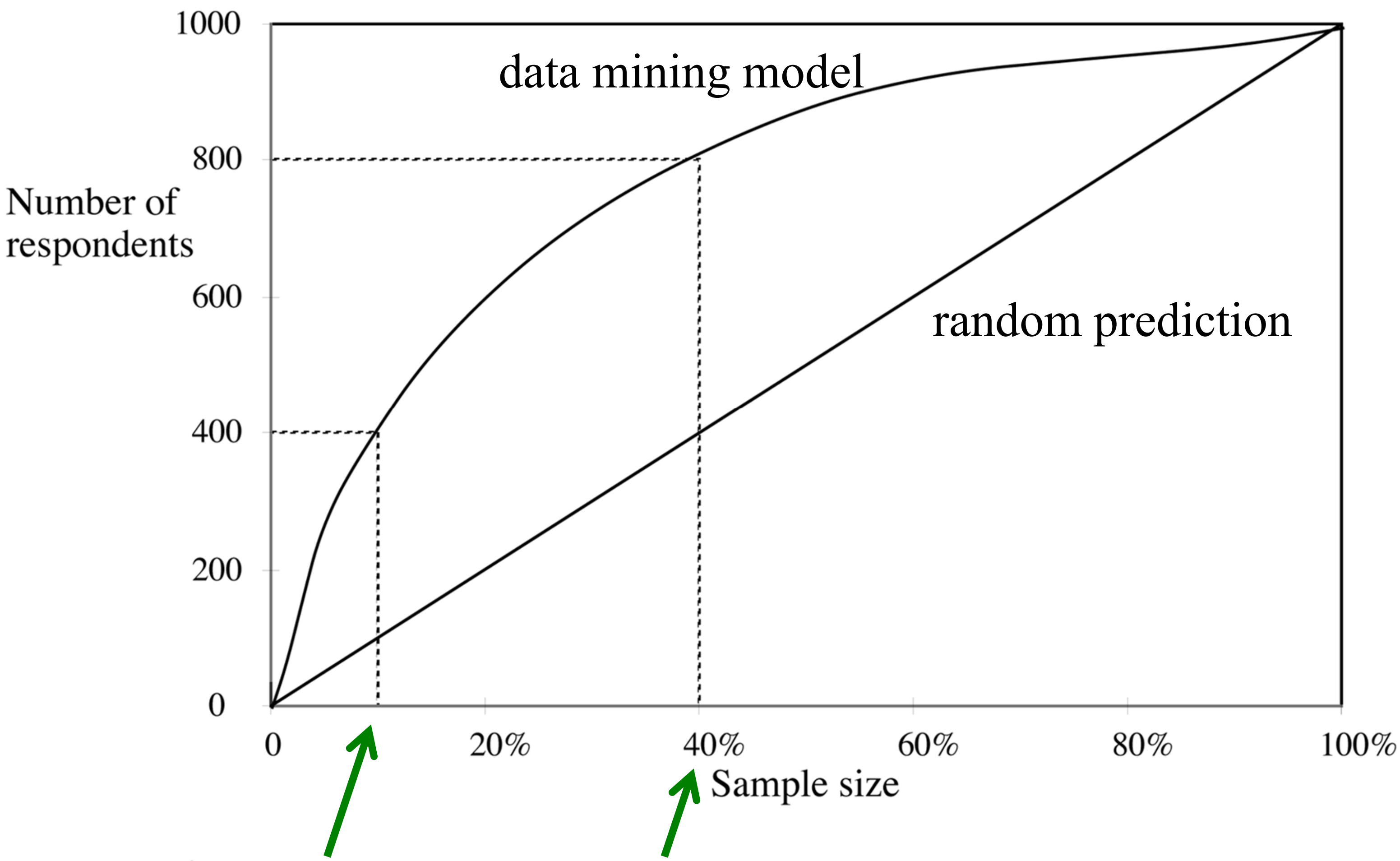
Generating a Lift Chart

- Assume that the classifier can output a predicted probability of being positive
- Sort instances according to predicted probability of being positive:

	Predicted probability	Actual class
1	0.95	Yes
2	0.93	Yes
3	0.93	No
4	0.88	Yes
…	…	…

x axis is sample size
y axis is number of true positives

A Sample Lift Chart



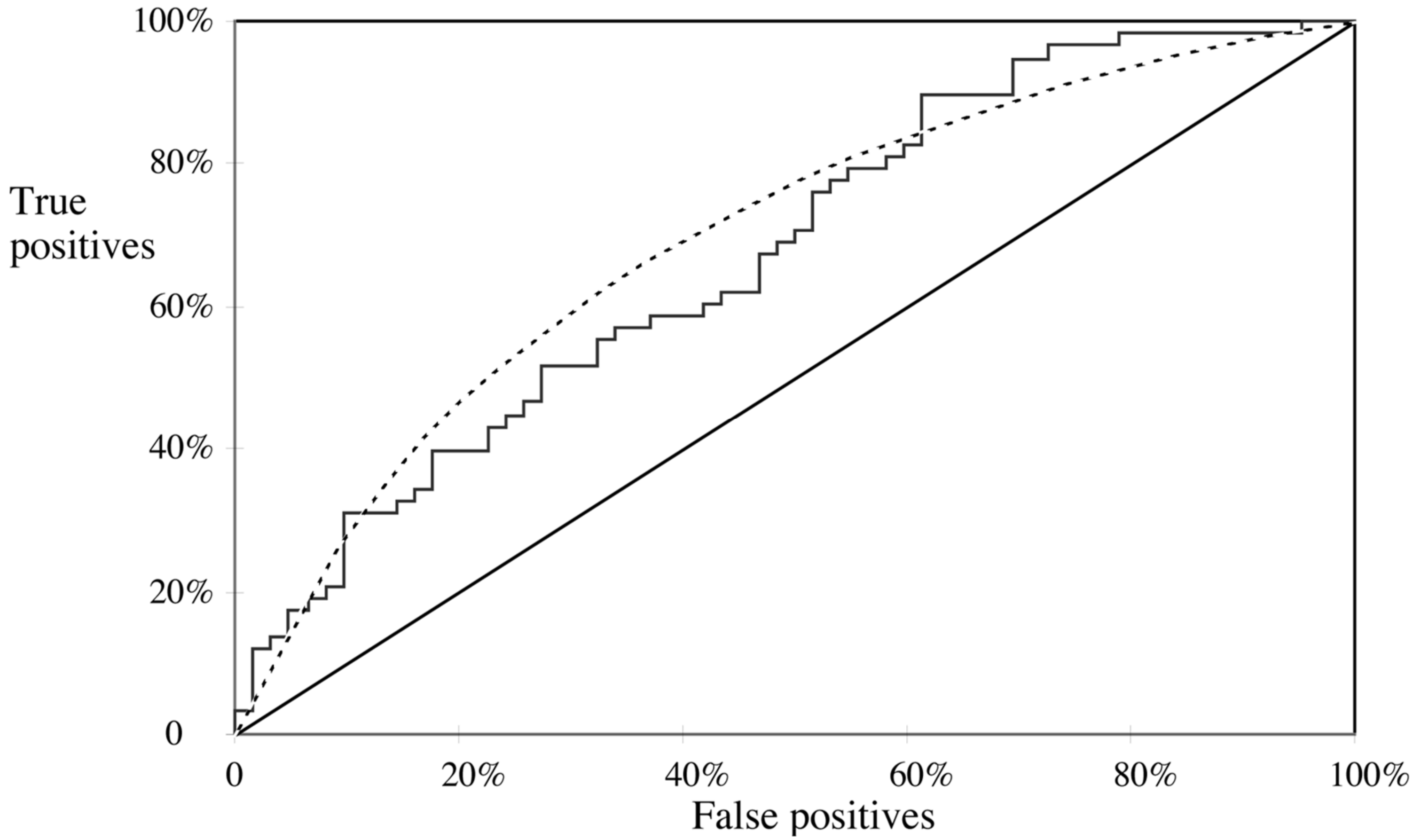
40% of responses
for 10% of resources

80% of responses
for 40% of resources

ROC Curves

- ROC curves are similar to lift charts
  - Stands for “receiver operating characteristic”
  - Used in signal detection to show tradeoff between hit rate and false alarm rate over noisy channel
- Differences to lift chart:
  - y axis shows percentage of true positives in sample rather than absolute number
  - x axis shows percentage of false positives in sample rather than sample size

A Sample ROC Curve



Jagged curve—one set of test data

Smooth curve—use cross-validation

Considering Cost

- In practice, different types of correct/incorrect prediction incur different costs
- 0-1 loss (for each data instance):
  - correct prediction – loss is 0
  - incorrect prediction – loss is 1
- Loss Matrix for 0-1 loss:

		Predicted class	
		yes	no
Target (actual) class	yes	0	1
	no	1	0

- Note that this table captures loss / cost, which is *different* from the previous contingency table.
- The loss matrix is to be considered during learning and classification

Considering Cost

- Minimizing the loss is equivalent to minimizing the error rate
- Extending 0-1 loss via using different costs in the loss matrix

		Predicted class	
		yes	no
Target (actual) class	yes	-1	10
	no	5	0