Data Mining

Practical Machine Learning Tools and Techniques

Slides for Section 5.7

Counting the cost

- In practice, different types of classification errors often incur different costs
- Examples:
  - Loan decisions
  - Oil-slick detection
  - Fault diagnosis
  - Promotional mailing

Counting the cost

- The *confusion matrix*:

		Predicted class	
		Yes	No
Actual class	Yes	True positive	False negative
	No	False positive	True negative

There are many other types of cost!

- E.g.: cost of collecting training data

Aside: 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*	

- Number of successes: sum of entries in diagonal (D)
- *Kappa* statistic: $\frac{D_{observed} - D_{random}}{D_{perfect} - D_{random}}$

measures relative improvement over random predictor

Classification with costs

- Two cost matrices:

		Predicted class	
		yes	*no*
Actual class	*yes*	0	1
	no	1	0

		Predicted class		
		a	*b*	*c*
Actual class	*a*	0	1	1
	b	1	0	1
	c	1	1	0

- Success rate is replaced by average cost per prediction
  - Cost is given by appropriate entry in the cost matrix

Cost-sensitive classification

- Can take costs into account when making predictions
  - Basic idea: only predict high-cost class when very confident about prediction
- Given: predicted class probabilities
  - Normally we just predict the most likely class
  - Here, we should make the prediction that minimizes the expected cost
    - Expected cost: dot product of vector of class probabilities and appropriate column in cost matrix
    - Choose column (class) that minimizes expected cost

Cost-sensitive learning

- So far we haven't taken costs into account at training time
- Most learning schemes do not perform cost-sensitive learning
  - They generate the same classifier no matter what costs are assigned to the different classes
  - Example: standard decision tree learner
- Simple methods for cost-sensitive learning:
  - Resampling of instances according to costs
  - Weighting of instances according to costs
- Some schemes can take costs into account by varying a parameter, e.g. naïve Bayes

Lift charts

- In practice, costs are rarely known
- Decisions are usually made by comparing possible scenarios
- 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

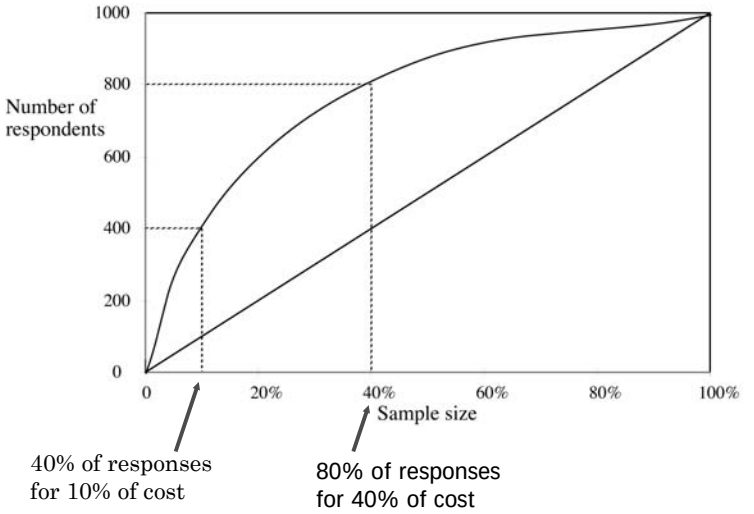
Generating a lift chart

- 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

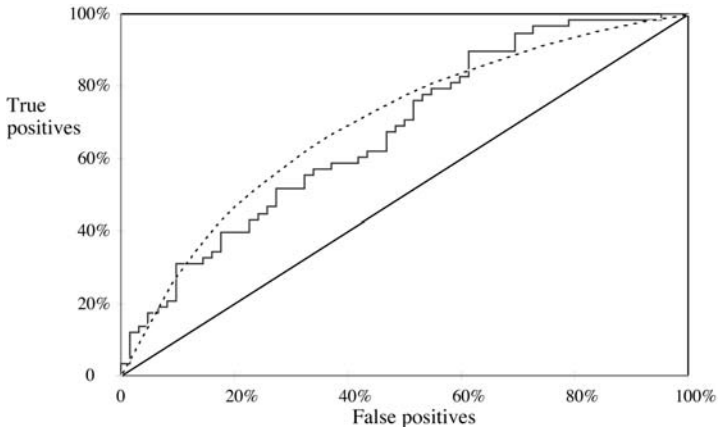
A hypothetical lift chart



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

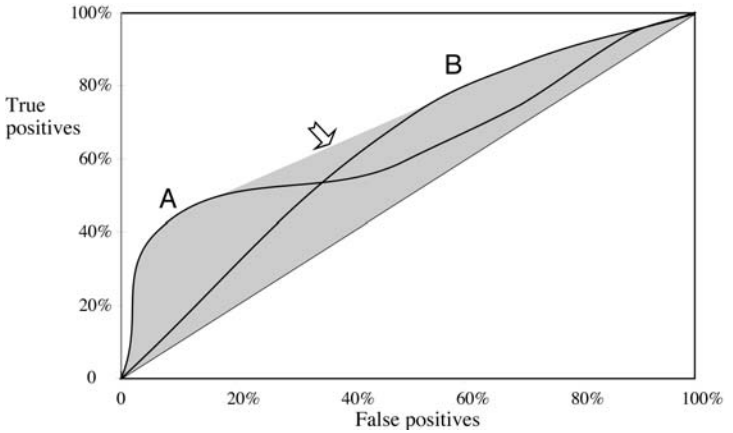


- Jagged curve—one set of test data
- Smooth curve—use cross-validation

Cross-validation and ROC curves

- Simple method of getting a ROC curve using cross-validation:
  - Collect probabilities for instances in test folds
  - Sort instances according to probabilities
- This method is implemented in WEKA
- However, this is just one possibility
  - Another possibility is to generate an ROC curve for each fold and average them

ROC curves for two schemes



- For a small, focused sample, use method A
- For a larger one, use method B
- In between, choose between A and B with appropriate probabilities

The convex hull

- Given two learning schemes we can achieve any point on the convex hull!
- TP and FP rates for scheme 1: t_1 and f_1
- TP and FP rates for scheme 2: t_2 and f_2
- If scheme 1 is used to predict $100 \times q$ % of the cases and scheme 2 for the rest, then
  - TP rate for combined scheme: $q \times t_1 + (1-q) \times t_2$
  - FP rate for combined scheme: $q \times f_1 + (1-q) \times f_2$

More measures...

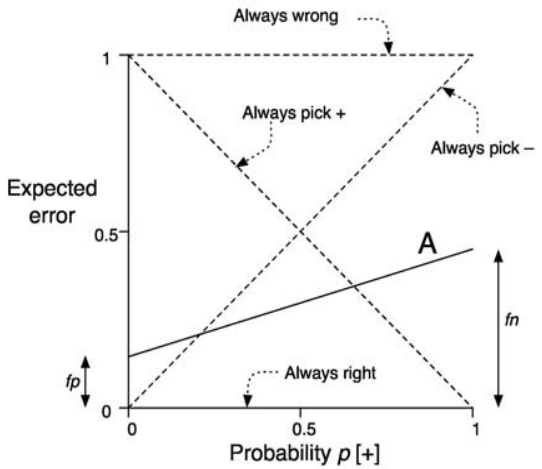
- Percentage of retrieved documents that are relevant: *precision*=TP/(TP+FP)
- Percentage of relevant documents that are returned: *recall* =TP/(TP+FN)
- Precision/recall curves have hyperbolic shape
- Summary measures: average precision at 20%, 50% and 80% recall (*three-point average recall*)
- *F-measure*=(2 × recall × precision)/(recall+precision)
- *sensitivity* × *specificity* = (TP / (TP + FN)) × (TN / (FP + TN))
- Area under the ROC curve (*AUC*): probability that randomly chosen positive instance is ranked above randomly chosen negative one

Summary of some measures

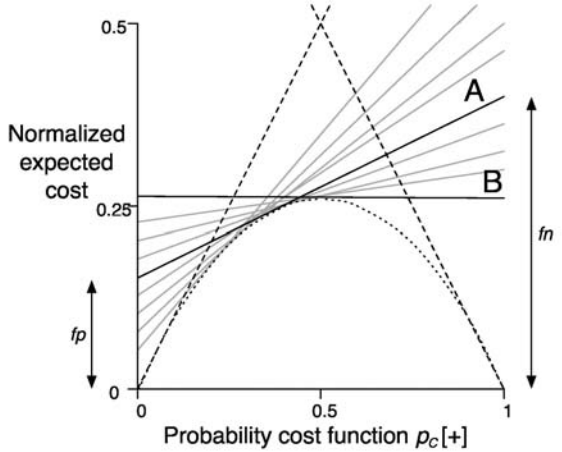
	Domain	Plot	Explanation
Lift chart	Marketing	TP Subset size	TP (TP+FP)/(TP+FP+TN+FN)
ROC curve	Communications	TP rate FP rate	TP/(TP+FN) FP/(FP+TN)
Recall-precision curve	Information retrieval	Recall Precision	TP/(TP+FN) TP/(TP+FP)

Cost curves

- *Cost curves* plot expected costs directly
- Example for case with uniform costs (i.e. error):



Cost curves: example with costs



$$\text{Probability cost function } p_c[+] = \frac{p[+]C[+|-]}{p[+]C[+|-]+p[-]C[-|+]}$$

$$\text{Normalized expected cost} = \text{fn} \times p_c[+] + \text{fp} \times (1 - p_c[+])$$