Topics in Business Intelligence

Lecture 3: Model validation

Tommi Tervonen

Econometric Institute, Erasmus University Rotterdam

Erasmus Universiteit Rotterdam

Model validation

- In practice we always have multiple methods to choose from, and even within a single method we often need to choose parameter values

$\rightarrow$ need for model validation

$\rightarrow$ need for accuracy measures

Accuracy measures (classification)

- Probability of making a misclassification error
- We should perform better than the "Naive classification rule": classify everything to the most prevalent class

Confusion matrix

		Predicted	
		0	1
Actual	0	True Negative	False Positive
	1	False Negative	True Positive

Overall error rate

		Predicted 0	Predicted 1
Actual	0	True Negative	False Positive
	1	False Negative	True Positive

- Overall error rate $= \frac{FN+FP}{n}$
- If n is reasonably large, the estimation of error rate is good (e.g. misclassification rate 0.05, 99% confidence $\rightarrow$ 3152 cases)

Overall accuracy

		Predicted	
		0	1
Actual	0	True Negative	False Positive
	1	False Negative	True Positive

- Overall accuracy $= \frac{TN+TP}{n}$

Classification cutoff

- Many algorithms use a cutoff for classification probability in deciding the predicted class
- Cutoff value of 0.5 provides the optimal overall accuracy and error rate
- However, sometimes false negatives are more expensive than false positives (or *vice versa*), and the asymmetric costs should be taken into account (e.g. direct mailing)
- Suppose it is more important to predict membership in class 1 than 0

Sensitivity

		Predicted	
		0	1
Actual	0	True Negative	False Positive
	1	False Negative	True Positive

- Sensitivity $= \frac{TP}{FN+TP}$
- Ability of the classifier to detect the important class 1 members correctly

Specificity

		Predicted 0	Predicted 1
Actual	0	True Negative	False Positive
Actual	1	False Negative	True Positive

- Specificity $= \frac{TN}{FP+TN}$
- Ability to rule out class 0 members correctly

False positive rate

		Predicted	
		0	1
Actual	0	True Negative	False Positive
	1	False Negative	True Positive

- False positive rate $= \frac{FP}{FP+TP}$

False negative rate

		Predicted	
		0	1
Actual	0	True Negative	False Positive
	1	False Negative	True Positive

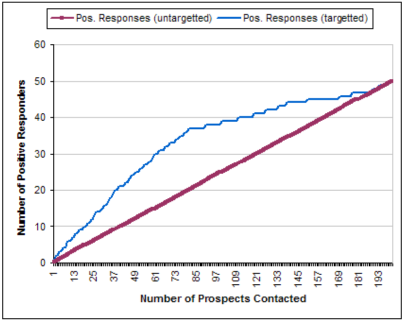
- False negative rate $= \frac{FN}{FN+TN}$

False negative rate

		Predicted 0	Predicted 1
Actual	0	True Negative	False Positive
Actual	1	False Negative	True Positive

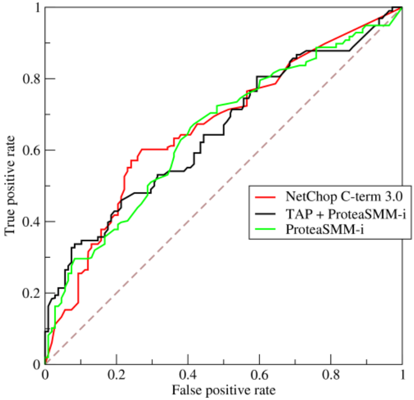
- False negative rate $= \frac{FN}{FN+TN}$
- Accuracy measures can be plotted against cutoff values to find a value that balances the measure

Lift charts



- Lift chart visualize the cumulative lift (or gain) curve
- x-axis: cumulative number of cases in decreasing probability
- y-axis: cumulative number of true positives (the important class 1)
- Example: construction of a lift chart

ROC Curves



- True positive rate vs false positive rate

Asymmetric misclassification costs

- Suppose our direct mail offer is accepted by 1% of the receivers
- A naive classifier would classify all as nonresponders, and have only 1% error rate (but be useless)
- A classifier that would classify 30% of nonresponders as responders and 2% of responders as nonresponders would probably be better
- $\rightarrow$ asymmetric misclassification costs between classes

Asymmetric misclassification costs

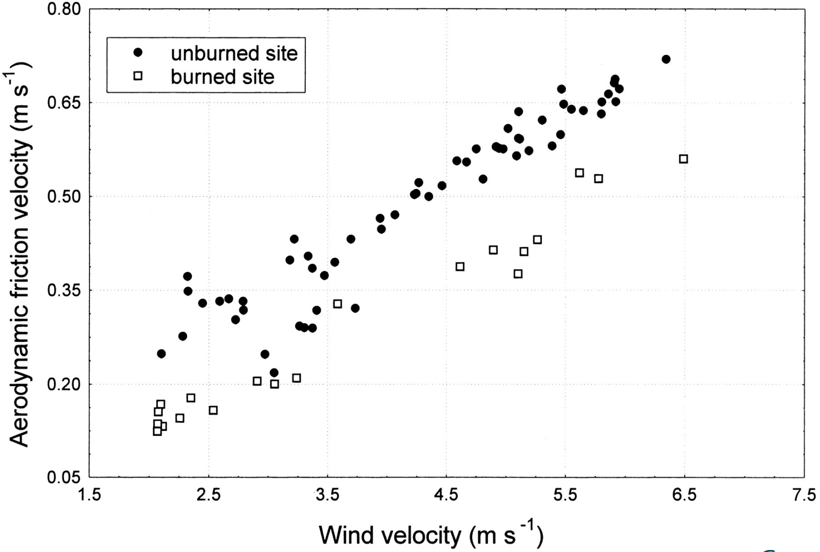
	Predict class 1	predict class 0
Actual 1	8	2
Actual 0	20	970

- 2.2% overall error rate
- Now, suppose sending an offer costs 1e, and profit from response is 10e (after sending costs)
  - Send to all $\rightarrow$ loss of 692 euros
  - Naive classifier $\rightarrow$ 0 euros
  - Use classifier above, send to 28 people $\rightarrow$ profit of 60e

Oversampling for asymmetric costs

- Stratified sampling is used to oversample rare cases
- Similarly, we can oversample (sample multiple times, with or without replacement) to affect the classification errors
- Consequently the costs are indirectly taken into account

Oversampling



Oversampling - model validation

For validating the model with oversampled training, we can:

1. Score the model to a validation set that has been selected without oversampling
2. Score the model to an oversampled validation set, and reweight the results to remove the effects of oversampling

The first option is always preferred, but not might be feasible due to lack of data

Reweighing oversampled validation set

- Assume 2% response rate, oversampling 25x $\rightarrow$ response of 50%
- Assume confusion matrix:

	Actual 1	Actual 0	Total
Predicted 1	420	110	530
Predicted 0	80	390	470
Total	500	500	1000

- Overall misclassification rate $= (80 + 110)/1000 = 19\%$, and model ends up classifying 53% of the records as 1's

Reweighing oversampled validation set

- To reweight to account to the actual number of 0's and 1's in the validation set, we need to add enough 0's to get the original balance $(1 : 50)$, that is

$$500 + 0.98x = x$$

Reweighing oversampled validation set

- To reweight to account to the actual number of 0's and 1's in the validation set, we need to add enough 0's to get the original balance (1 : 50), that is

$$500 + 0.98x = x$$

- which yields $x = 25000$. Now we augment # of actual nonresponders, and get:

	Actual 1	Actual 0	Total
Predicted 1	420	5 390	5 810
Predicted 0	80	19 110	19 190
Total	500	24 500	25000

Reweighing oversampled validation set

- To reweight to account to the actual number of 0's and 1's in the validation set, we need to add enough 0's to get the original balance $(1 : 50)$, that is

$$500 + 0.98x = x$$

- which yields $x = 25000$. Now we augment $\#$ of actual nonresponders, and get:

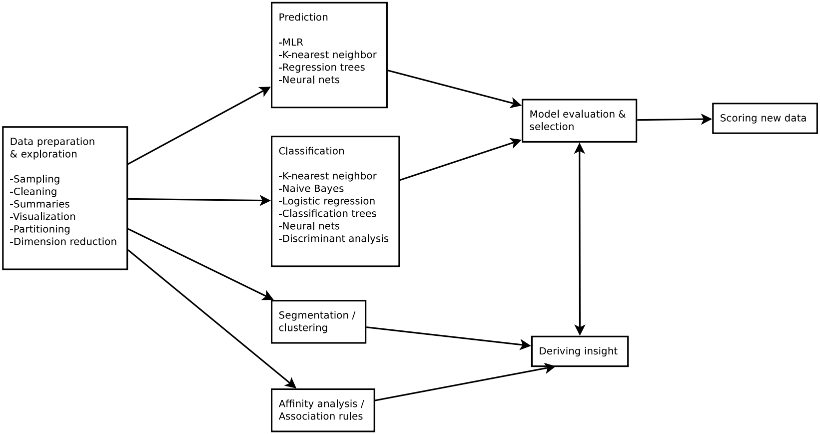
	Actual 1	Actual 0	Total
Predicted 1	420	5 390	5 810
Predicted 0	80	19 110	19 190
Total	500	24 500	25000

- $\rightarrow$ adjusted misclassification rate $(80 + 5390)/25000 = 21.9\%$
- Model classifies 21.4% of records as 1's.

Evaluating predictive performance

- Mean absolute error/deviation
- Average error
- Mean absolute percentage error
- Root mean-squared error
- Total sum of squared errors

Data mining process



Groups and topics

Group	Topic	Week
Yiwei et al	k-NN and Naive Bayes'	4
Stamenova et al	Classification trees	5
Zaghainov et al	Neural nets	6
Merkle et al	Logistic regression	7

Note! These 4 lectures have mandatory attendance