Evaluating What's Been Learned

Cross-Validation

- Foundation is a simple idea – "***holdout***" – holds out a certain amount for testing and uses rest for training
- Separation should NOT be "convenience",
  - Should at least be random
  - Better – "***stratified***" random – division preserves relative proportion of classes in both training and test data
- 10-fold cross validation has become standard
- This is improved if the folds are chosen in a "stratified" random way

For Small Datasets

- Leave One Out
- Bootstrapping

Leave One Out

- Train on all but one instance, test on that one (pct correct always equals 100% or 0%)
- Repeat until have tested on all instances, average results
- Really equivalent to N-fold cross validation where N = number of instances available
- Plusses:
  - Always trains on maximum possible training data (without cheating)
  - No stratification, no random sampling necessary
- Minuses
  - Guarantees a non-stratified sample – the correct class will always be at least a little bit under-represented in the training data
  - Statistical tests are not appropriate

Bootstrapping

- Sampling done ***with replacement*** to form a training dataset
- Particular approach – 0.632 bootstrap
  - Dataset of n instances is sampled n times
  - Some instances will be included multiple times
  - Those not picked will be used as test data
  - On large enough dataset, .632 of the data instances will end up in the training dataset, rest will be in test

- This is a bit of a pessimistic estimate of performance, since only using 63% of data for training (vs 90% in 10-fold cross validation)
- This procedure can be repeated any number of times, allowing statistical tests

Counting the Cost

- Some mistakes are more costly to make than others
- Giving a loan to a defaulter is more costly than denying somebody who would be a good customer
- Sending mail solicitation to somebody who won't buy is less costly than missing somebody who would buy (opportunity cost)
- Looking at a confusion matrix, each position could have an associated cost (or benefit from correct positions)

Information Retrieval (IR) Measures

- E.g., Given a WWW search, a search engine produces a list of hits supposedly relevant
- Which is better?
  - Retrieving 100, of which 40 are actually relevant
  - Retrieving 400, of which 80 are actually relevant
  - Really depends on the costs

Information Retrieval (IR) Measures

- IR community has developed 3 measures:
  - Recall = $\frac{\text{number of documents retrieved that are relevant}}{\text{total number of documents that are relevant}}$
  - Precision = $\frac{\text{number of documents retrieved that are relevant}}{\text{total number of documents that are retrieved}}$
  - F-measure = $\frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}}$

Confusion Matrix

Results of experiments by cross-validation method: client_A

Actual		Predicted						
		Promo...	Passive	Detrac...	MISSI...	No. of obj	Accuracy	Coverage
	Promoter	152.3	0.9	0.1	0	168.5	0.993	0.91
	Passive	1	7.9	0	0	19.1	0.868	0.474
	Detractor	0.7	0.1	3.4	0	10.2	0.849	0.496
	MISSING	0	0	0	0	1.2	0	0
	True positive rate	0.99	0.92	0.98	0			

```
Total number of tested objects: 199
Total accuracy: 0.983
Total coverage: 0.836
```