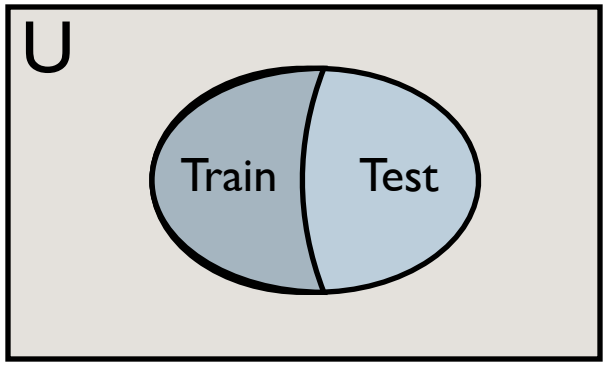
Evaluation in Predictive Analytics

Evaluation Structure

Classification (Prediction)

- Typical Question:
  - Which is better, Classifier A or Classifier B?
- Interested in ***Generalisation Accuracy***
- ***Hold back*** some training data to use for testing
  - Use performance on Test data as a ***proxy*** for performance on unseen data (i.e. Generalization).

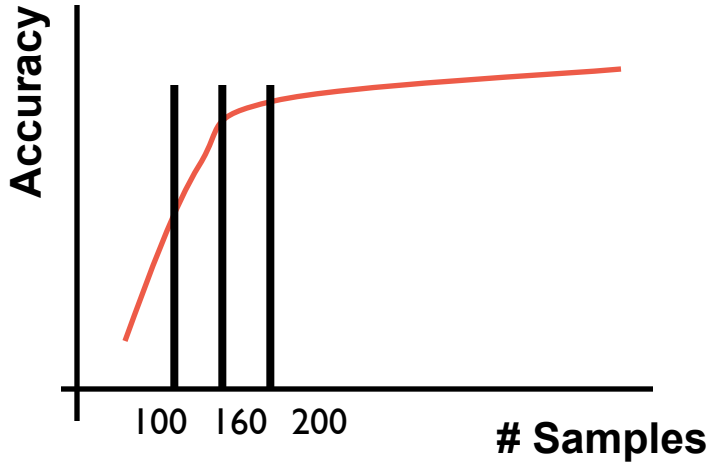


Insight

Problems with '*Hold-out' Validation*

Imagine 200 samples are available for training:

- 50:50 split underestimates generalisation acc.
- 80:20 estimate based on a small sample (40)
  - Different hold-out sets - different results



Insight

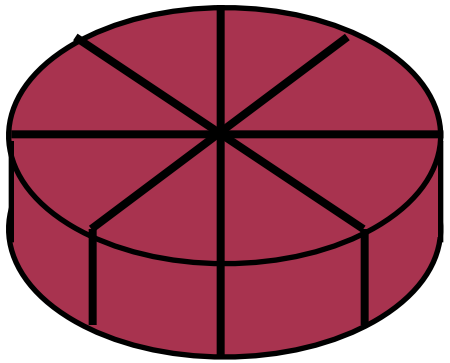
k-Fold Cross Validation

Having your cake and eating it too…

Divide data into *k* folds

For each fold in turn

- Use that fold for testing and
- Use the remainder of the data for training



Insight

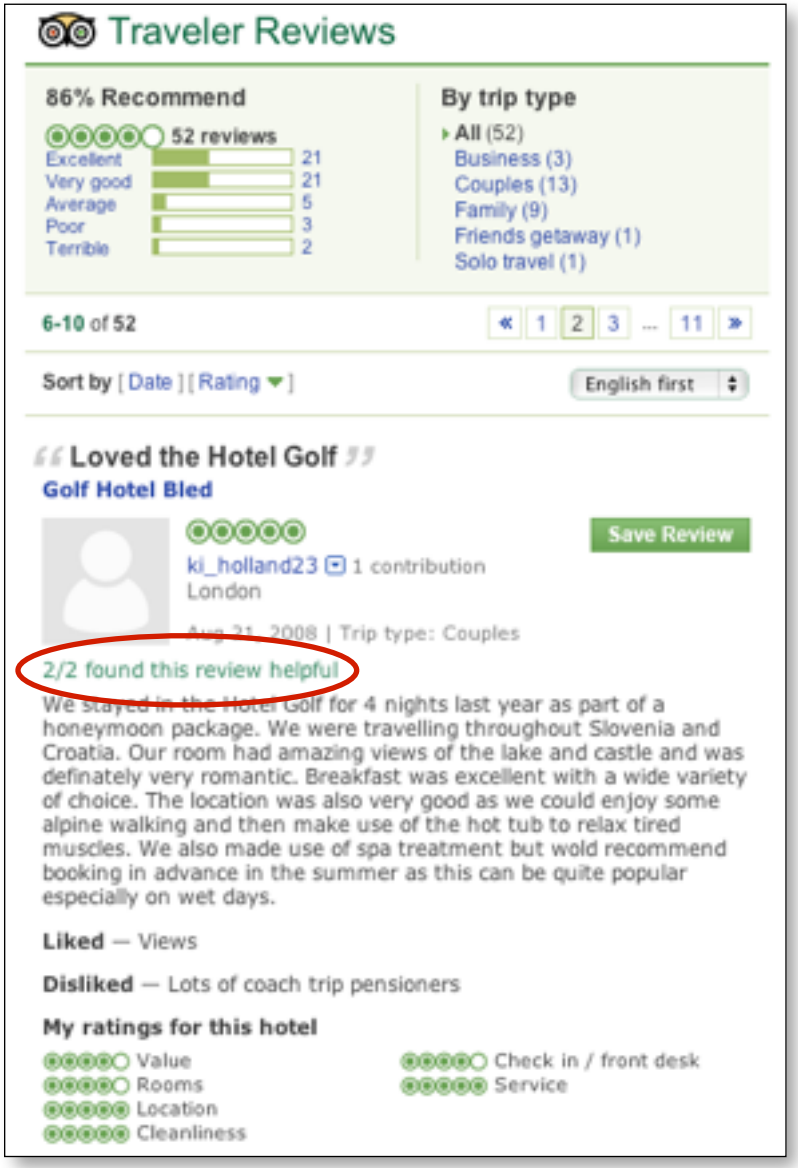
Comparing Two Classifiers

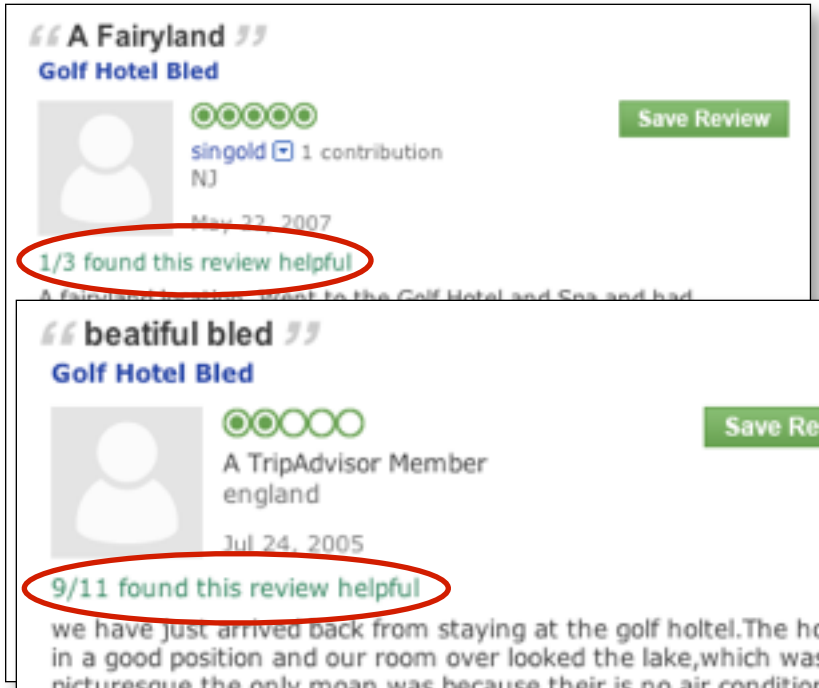
1. Divide dataset S into k folds (say 10)
2. For each of the k folds
   1. Create training and test sets R & T
   2. Divide R into sets R1 and V
   3. For each of the classifiers
      1. Use V to tune parameters on a model trained with R1
      2. Use these 'good' parameters to train a model with R
      3. Measure Accuracy on T
      4. Record 0-1 loss results for each classifier
3. Assess statistical significance of results

Tuning is explicit

Note: It is common to run x-val multiple times, e.g. 10 x 10-fold x-val.

Example: Review 'helpfulness'





Insight

Loss Functions

Hotel Reviews dataset

Predict if users will find a review helpful

O'Mahony & Smyth (2009) 3rd ACM RecSys conference

- "Learning to Recommend Helpful Hotel Reviews"

Subset of dataset

- 24 features + class label good/bad
- all reviews have received ≥ 5 ratings
- review is classified as helpful (good) if ≥ 75% of ratings helpful
- 486 cases: 308 good, 178 bad
- Available here

  - http://www.csi.ucd.ie/content/ecml-pkdd-2009-workshop-evaluation

Compare performance of Naive Bayes and SVM (SVM)

- Using Weka http://www.cs.waikato.ac.nz/ml/weka/

SVM v's Naive Bayes

Hold-out validation - 33% holdout set

SVM

```
Correctly Classified Instances         117               70.9091 %
Incorrectly Classified Instances        48               29.0909 %
Kappa statistic                          0.3071
Mean absolute error                      0.2909
Root mean squared error                  0.5394
Relative absolute error                 62.6804 %
Root relative squared error            112.1168 %
Total Number of Instances              165

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.617      0.718     0.895     0.797      0.639    good
                 0.383     0.105      0.676     0.383     0.489      0.639    bad
Weighted Avg.    0.709     0.431      0.703     0.709     0.685      0.639

=== Confusion Matrix ===

  a  b   <-- classified as
 94 11 |  a = good
 37 23 |  b = bad
```

Naive Bayes

```
Correctly Classified Instances         103               62.4242 %
Incorrectly Classified Instances        62               37.5758 %
Kappa statistic                          0.1995
Mean absolute error                      0.3793
Root mean squared error                  0.5316
Relative absolute error                 81.7353 %
Root relative squared error            110.5048 %
Total Number of Instances              165

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.686     0.483      0.713     0.686     0.699      0.674    good
                 0.517     0.314      0.484     0.517     0.5        0.674    bad
Weighted Avg.    0.624     0.422      0.63      0.624     0.627      0.674

=== Confusion Matrix ===

  a  b   <-- classified as
 72 33 |  a = good
 29 31 |  b = bad
```


Insight

Naive Bayes v's SVM

Test set: 105 good, 60 bad

NB Accuracy 62.4%

SVM Accuracy 70.1%

SVM

Classified as

good	bad	
94	**11**	good
37	**23**	bad

Act. Class

SVM biased toward majority class

Naive Bayes

Classified as

good	bad	
72	**33**	good
29	**31**	bad

Act. Class

What if this is important?

Insight

Other Scores from Weka

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.895     0.617      0.718     0.895      0.797      0.639    good
                   0.383     0.105      0.676     0.383      0.489      0.639    bad
Weighted Avg.      0.709     0.431      0.703     0.709      0.685      0.639

=== Confusion Matrix ===

  a  b   <-- classified as
 94 11 |  a = good
 37 23 |  b = bad
```

Other Measures

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N}$$

$$TPRate = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Classified as

	Pos	Neg	
P	**TP**	**FN**	Pos
N	**FP**	**TN**	Neg

Act. Class

$$FPRate = \frac{FP}{FP + TN} = \frac{FP}{N}$$

Classified as

	Pos	Neg	
P	**TP**	**FN**	Pos
N	**FP**	**TN**	Neg

Act. Class

Popular in medical research

$$Sensitivity = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$Specificity = \frac{TN}{FP + TN} = \frac{TN}{N} = 1 - FPRate$$

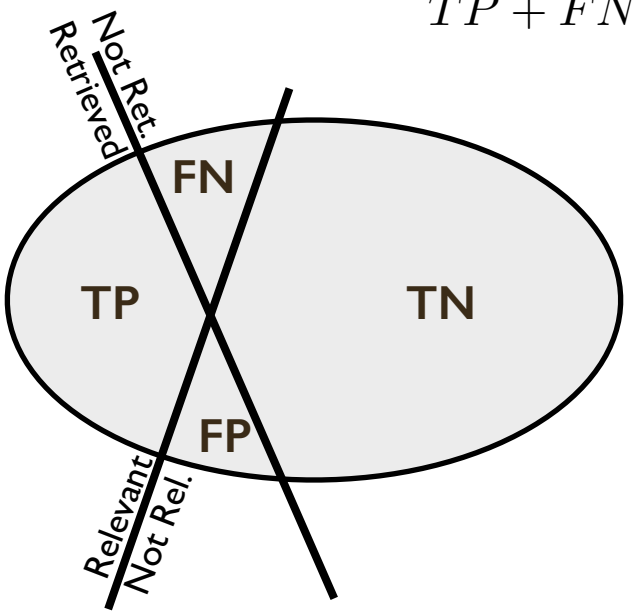
Information Retrieval Perspective

"proportion of retrieved docs that are relevant"

$$Precision = \frac{TP}{TP + FP}$$

"proportion of relevant docs that are retrieved"

$$Recall = \frac{TP}{TP + FN} = Sensitivity$$



Classified as

		Pos	Neg	
P		**TP**	**FN**	Pos
N		**FP**	**TN**	Neg

Act. Class

Classified as

		Pos	Neg	
P		**TP**	**FN**	Pos
N		**FP**	**TN**	Neg

Act. Class

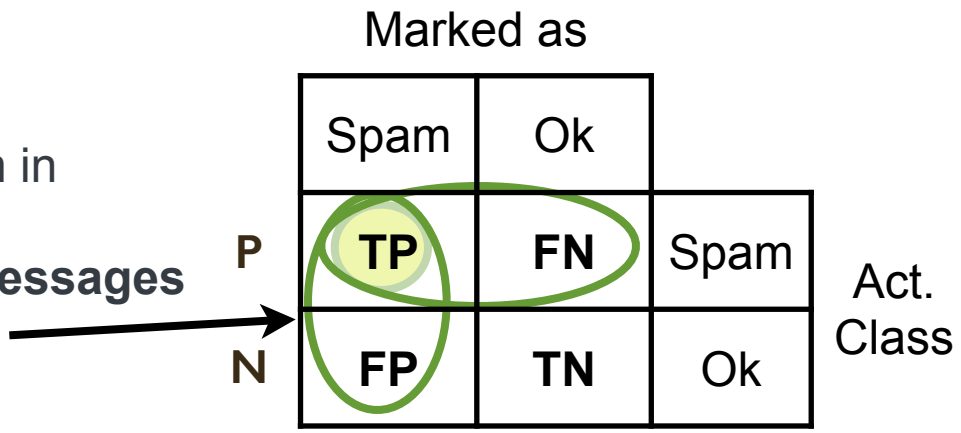
Insight

Precision & Recall

Precision and Recall can be used in 'non-IR' situations:

Spam Filtering

- ***Precision:*** proportion of spam in **messages marked as spam**
- ***Recall:*** proportion of **spam messages** marked as spam



Review Helpfulness

- ***Precision:*** Proportion of reviews classified as helpful (good) that are actually helpful.

Insight

Comparing Classifiers

Using the Hotel Review Helpfulness dataset:

- Which of the following classifiers is most accurate using the training set for testing?
  - Logistic Regression
  - Naive Bayes
  - Support Vector Machine (SMO)
- Which of the three is most accurate using a 33% holdout set?
- Which wins using 10-fold cross validation?
- Which has best recall on the minority class?
  - Which is best for weeding out bad reviews?