Evaluation of Classification Quality

Prof. Wai Lam

Reference:
Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations, by I. Witten and E. Frank, Morgan Kaufmann
Testing and Error

- Error rate: proportion of errors made over the whole set of instances.
- Test set: 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

• What shall we do if the amount of data is limited?
• 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, $P_1, \ldots, P_k$

<table>
<thead>
<tr>
<th>Training set</th>
<th>Testing set</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_2, \ldots, P_k$</td>
<td>$P_1$</td>
<td>$A_1$</td>
</tr>
<tr>
<td>$P_1, P_3, \ldots, P_k$</td>
<td>$P_2$</td>
<td>$A_2$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$P_1, P_2, \ldots, P_{k-1}$</td>
<td>$P_k$</td>
<td>$A_k$</td>
</tr>
</tbody>
</table>

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

The contingency table:

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicated Class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>YES</td>
<td>True Positive</td>
</tr>
<tr>
<td></td>
<td>NO</td>
<td>False Negative</td>
</tr>
<tr>
<td>NO</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
A Sample 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

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicated Class</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YES</td>
<td>100</td>
<td>900</td>
<td>1,000</td>
</tr>
<tr>
<td>YES</td>
<td>NO</td>
<td>99,900</td>
<td>899,100</td>
<td>999,000</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>100,000</td>
<td>900,000</td>
<td>1,000,000</td>
</tr>
</tbody>
</table>

Accuracy = 0.8992  
(Error = 0.1008)

## A Data Mining Method

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicated Class</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YES</td>
<td>400</td>
<td>600</td>
<td>1,000</td>
</tr>
<tr>
<td>YES</td>
<td>NO</td>
<td>99,600</td>
<td>899,400</td>
<td>999,000</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>100,000</td>
<td>900,000</td>
<td>1,000,000</td>
</tr>
</tbody>
</table>

Accuracy = 0.8998  
(Error = 0.1002)
Lift Factor

• The random response rate is 0.1% (i.e., 100 respondents).
• The response rate of a certain data mining method is 0.4% (400 respondents).
• The increase in response factor, is known as the lift factor (a factor of 4 in this case).
Parameter tuning

- It is important that the test data is not used in any way to create the classifier
- Some learning schemes operate in two stages:
  - Stage 1: builds the basic structure
  - Stage 2: optimizes parameter settings
- The test data can’t be used for parameter tuning!
- Proper procedure uses three sets: training data, validation data, and test data
  - Validation data is used to optimize parameters