Machine learning in fMRI

Validation

Alexandre Savio, Maite Termenón, Manuel Graña

[1] Computational Intelligence Group, University of the Basque Country

December, 2010

Outline

Outline

The purpose of validation I

- We have:
  - a model with one or more unknown parameters
  - a data set to which the model can be fit (the training data set).
- The fitting process optimizes the model parameters to make the model fit the training data as well as possible.

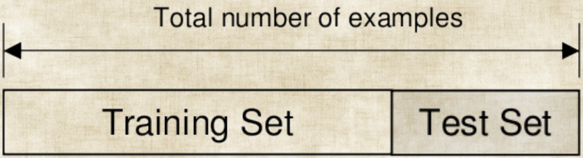
Overfitting I

- If we then take an independent sample of validation data from the same population as the training data, it will generally turn out that the model does not fit the validation data as well as it fits the training data.
- This is particularly likely to happen when:
  - the size of the training data set is small,
  - or when the number of parameters in the model is large.

Outline

Hold-Out

- Split dataset into two groups:
    - Training set: used to train the classifier
    - Test set: used to estimate the error rate of the trained classifier

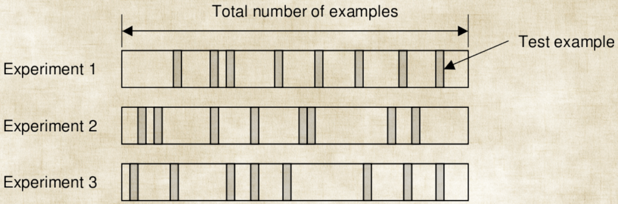


Cross-validation

- Cross-validation is:
  - a way to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available,
    - using computation in place of mathematical analysis.

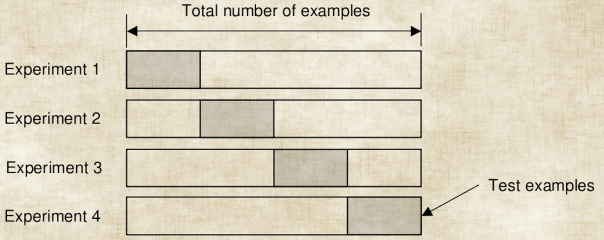
Repeated random sub-sampling

- Each split randomly selects a (fixed) number of examples without replacement.
- For each data split we retrain the classifier from scratch with the training examples and estimate E_i with the test examples.



K-fold

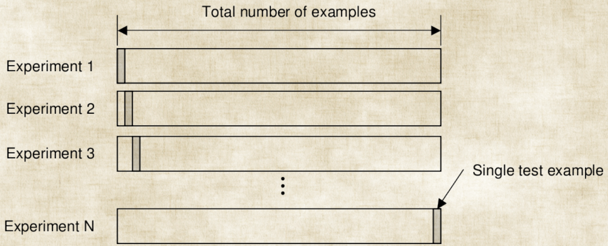
- For each of K experiments, use K-1 folds for training and the remaining one for testing.



- The advantage of K-Fold cross-validation is that all the examples in the dataset are eventually used for both training and testing.
- Common choice for K-Fold cross-validation is K=10.

Leave-one-out

- Leave-one-out is the degenerate case of K-Fold cross Validation, where K is chosen as the total number of examples.
- For each experiment use N-1 examples for training and the remaining example for testing, then N experiments will be performed.



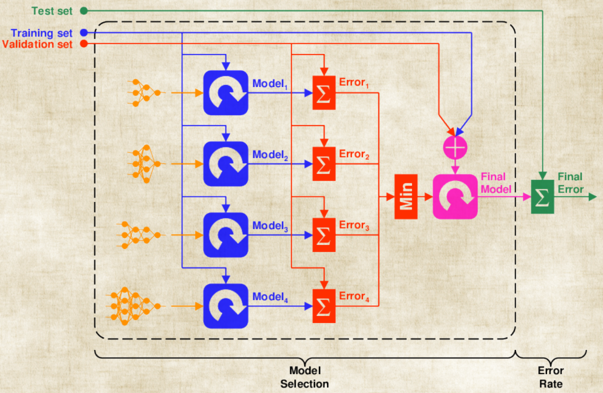
How many folds are needed?

- With a large number of folds
  - [+] The bias of the true error rate estimator will be small (the estimator will be very accurate).
  - [-] The variance of the true error rate estimator will be large.
  - [-] The computational time will be very large as well (many experiments).

- With a small number of folds
  - [+] The number of experiments and, therefore, computation time are reduced.
  - [+] The variance of the estimator will be small.
  - [-] The bias of the estimator will be large (conservative or higher than the true error rate).

- In practice, the choice of the number of folds depends on the size of the dataset.

Three-way data splits



Confusion matrix

		actual value		
		p	***n***	**total**
prediction outcome	***p'***	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

Measures of fit I

- $TP + FP + FN + TN = N$
- Accuracy: $Acc = \frac{TP+TN}{N}$
- True Positive Rate or Sensitivity or Recall: $TPR = \frac{TP}{TP+FN} = sensitivity$
- False Positive Rate: $FPR = \frac{FP}{FP+TN} = 1 - specificity$

Measures of fit II

- Precision
  - The proportion of the predicted positive cases that were correct, as calculated using the equation:
  - $Precision = \frac{TP}{TP+FP}$
- F-score: $F - score = Precision * Recall$
- Misclassification Ratio: $MCR = \frac{FP+FN}{N}$
- Jaccard Similarity: $Jaccard = \frac{TP}{TP+FN+FP}$
- Dice's Coefficient: $DC = \frac{2*TP}{2*TP+FN+FP}$

Measures of fit III

- Accurate Segmentation Ratio (ASR)
  - The *ASR* is only meaningful when we are dealing with multiple class problems.
  - Then we will have specific contigency matrices, one for each class of the problem.
  - $ASR = \frac{sum_c(TP_c)}{sum_c(TP_c+FP_c)}$ where sum_c means the sum over all different classes.

Summary

- Cross-validation is a way to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available.
- K-fold cross-validation and leave-one-out are the most used approaches in fMRI machine learning applications.
- There are many coefficients of fit measure which can be chosen.
  - Accuracy, Sensitivity and Specificity are the most used in medical literature.