Machine Learning	Prof. Dr. Bernt Schiele
	Dr. Paul Swoboda & Dr. Gerard Pons-Moll
Exercise Sheet 7 - 03.12.2018	due: 10.12.2018, 14:15

Exercise 13 - Cross Validation in Feature Selection

We are given a dataset from some biological experiment - the features represent some physiological parameters of some person and the label is 1 if the person is ill and -1 otherwise. The biologists claim that using least squares together with feature selection they can solve the problem with around 85% accuracy and only a small number of features.

Your task is to check if the biologists have done a good job in the data analysis. Check their claim by implementing an exhaustive search for the best feature subset. As a classifier use least squares, that is the weight vector is given as

$$w = (X^T X)^{-1} X^T Y,$$

where X is the design matrix and Y the label vector.
Classification of a new test instance x is done via $f(x) = \text{sign}(\langle w, x \rangle)$. As error measure use the classification error,

$$L(Y_i, f(X_i)) = \frac{1}{2}|Y_i - \text{sign}(\langle w, X_i \rangle)|.$$

a. **(6 Points)** There are in total 15 features in the training data. Use 5-fold cross-validation (use the ordering of the data as it is provided) for the linear least squares classifier in order to determine the best feature subset among all possible $2^{15} - 1 = 32767$ possible feature subsets by minimizing the 5-fold cross validation error on the training data `(Xtrain,Ytrain)`.

- Report the best feature subset(s) and its/their 5-fold cross-validation error (**written in a separate document (not within code)**).
- Use the whole training data of the best feature subset(s) obtained in a) to learn the final classifier. In the meantime the biologists have obtained new data. Use this data, given as `(Xtest,Ytest)` to evaluate the performance of the classifier(s) and report it. Do you have an idea why the cross-validation error obtained in a) and the just computed test error are so different. Has this to do with the classifier or the feature selection ? What will you tell the biologists ?

For the second part you have to submit a written answer and submit the code (see below).

Hint:

a. Be aware that in your cross-validation routine you use for every set of features the **same** partition of the data into the five folds (otherwise the results are not comparable !)

b. The following matlab code generates together with the function `allsets.m` a cell array of all possible subsets of 15 variables.

```
Subsets = cell(1,2^15-1);
counter=1;
BinCodes = allsets(15);
numbers=1:15;
for i=2:size(BinCodes,1)
  Subsets{i-1}=numbers(BinCodes(i,:)==1);
end
```

Exercise 14 - Permutation test

In the meantime you might have become suspicious what kind of data the biologists have provided. The natural question is if we could have found out that something is wrong without getting the new data (the test set).

a. **(6 Points)** For performance reasons restrict yourself now to the best feature subset selection among the first 6 features (discard features 7 to 15), that means there are only $2^6 - 1 = 63$ possible feature subsets.
Perform a permutation test, where

   - the null hypothesis is that features X_i and labels Y_i are independent,
   - the test statistic is the best 5-fold cross-validation error over all possible subsets of the 6 features (learning method and evaluation as in exercise 16),
   - significance level is $\alpha = 0.05$,
   - Define a rejection region for this test. What cross-validation error do you expect under the null hypothesis ?
   - Restrict yourself to 1000 samples of permutations of the labels and compute with the obtained distribution of values of the test statistic the p-value (use

     `rand('state',1)`

     to initialize the random number generator before drawing your data.)

   Report your computed p-value (report 4 digits), your rejection region, your decision (reject/not reject the null hypothesis) (written on paper). What does the result of the test imply for the result obtained in 20a) ? Generate a histogram of the test statistic (use `hist(T,20)` where `T` is the vector of computed test statistics).

Hint:

a. The function `randperm(n)` generates a random permutation of the integers from 1 to n.

Solution

Exercise 13 - Cross Validation in Feature Selection

a. One usually permutes the data first before one does the partitioning for cross validation. For my randomly chosen permutation (which is fixed for all trials with different feature subsets) the best 5-fold cross validation error among all possible feature subsets is 0.175 and is achieved by 2 sets of parameters.

   1. features: $4, 5, 7, 9, 10, 14$ - Test error: 0.474
   2. features: $3, 4, 5, 10, 14$ - Test error: 0.496

   Performing the experiment without initial permutation of the data which corresponds to the setting chosen by most of you. In this case the best cross validation error is again 0.175 but is achieved by a single feature subset:

   1. features: $4, 10, 11$ - Test error: 0.503

   Note, that the chosen feature subsets heavily depends on the partitioning which should not be the case if the problem had an underlying structure. Here, we see an example that wrapper methods can overfit. The data has been no biological dataset (sorry for that :)), the class variable Y was generated completely random given the features, that is we have a Bayes error of 0.5 for all possible feature subsets. Why do we then see a cross-validation error of 0.175? Well, we have 32767 possible feature subsets and only 40 training samples. Thus for

this large set of possible subsets there exists just by random fluctuations a feature subset which also performs well on the excluded folds in cross validation. Given a larger number of training samples the possibility that this happens gets smaller since we have more data to evaluate the error (e.g. for 500 training points generated in this way the minimal cross validation error over all possible feature subsets has been in 3 trials larger than 0.4). The reason for this effect of overfitting lies in this case not in the classifier. A linear classifier is so simple that hardly any overfitting can occur given that one has more samples than features. Nevertheless, we still see that overfitting with a linear classifier is possible if one tries all possible feature subsets. The cleanest way to detect such overfitting is to have a separate validation subset but one can always perform a permutation test in order to detect if the class variable Y is independent of the features.

Exercise 14 - Permutation test

a. The permutation test checks how unique the original cross validation error is given that we inject random label noise and repeat the experiment. If the original result is not due to heavy overfitting then by randomly permuting the labels the result should be close to 50% error. Thus the rejection region should be of the form $[0, c)$ for $c < 0.5$. If we do no prior permutation before the partitioning for cross validation, then the best cross validation error among the 6 features is 0.325 (average over $40/5 = 8$ test examples) or 2.6 (no division by 8). We perform 1000 permutations of the labels, the (approximate) p-value is then

$$p = \frac{1}{1000} \sum_{i=1}^{1000} \mathbb{1}_{CV(i)<0.325} = 0.266,$$

and since $p \geq \alpha$ we do not reject the null hypothesis. Thus we cannot refute the claim that the labels Y are independent of the features. Note that the p-value of 0.266 depends on the random permutation (and will be different without ***rand(state,1)***). Thus this test tells us that despite we have found this quite low cross validation error in the first part of the exercise, this does not imply that we have really found any meaningful structure in the data. In particular, using this test we could have refuted the claim of the biologists without even having some additional new test data. In figure 1 the histogram of all 1000 cross validation errors is shown. For the permuted data the original cross-validation error is 27.5% and the p-value is now

$$p = \frac{1}{1000} \sum_{i=1}^{1000} \mathbb{1}_{CV(i)<0.275} = 0.058,$$

so that we still do not reject the null hypothesis, but the p-value is now much smaller than before. Note however, that if we compute the p-value with a smaller or equal instead, we get

$$p = \frac{1}{1000} \sum_{i=1}^{1000} \mathbb{1}_{CV(i)\leq 0.275} = 0.144.$$

This shows again the problem of the p-value for a discrete distribution. The histogram for this case is plotted in Figure 2.

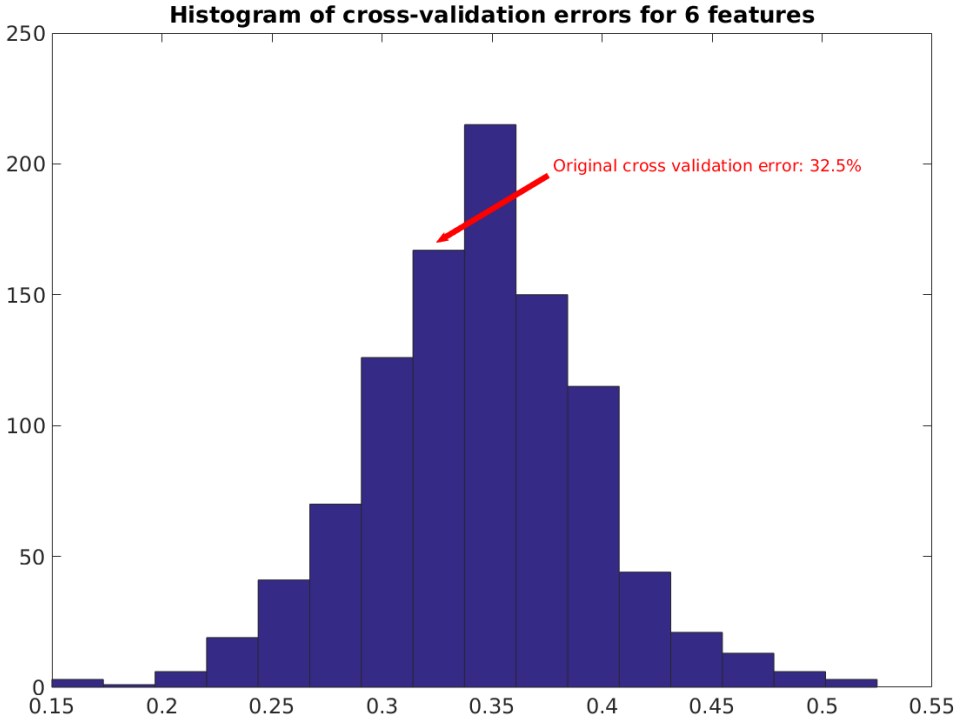

Figure 1: Figure 1: The histogram for the 1000 cross validation errors obtained by permuting 1000 times the original labels and repeating the feature selection procedure. We see that the original cross validation error is nothing special and thus we get the very high p-value of 0.266.

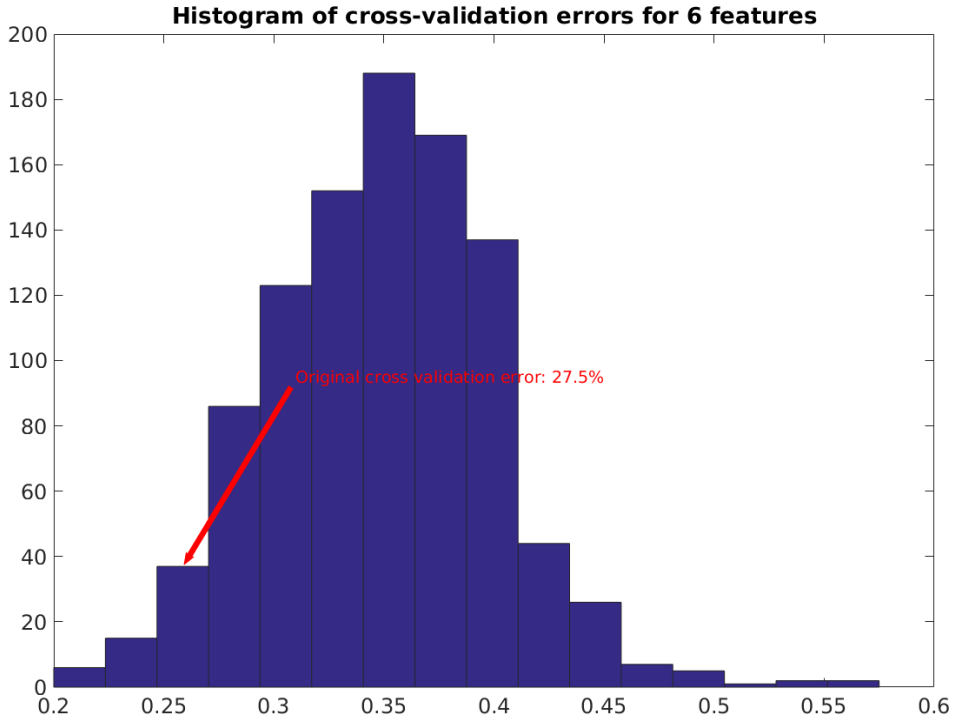

Figure 2: Figure 2: The histogram for the 1000 cross validation errors obtained by permuting 1000 times the original labels (with the randomly chosen partition) and repeating the feature selection procedure. In this case the p-value is of 0.058, so that we still do not reject the null hypothesis. Note, however that the smaller the estimated p-value is, the more permutations should be done in order to decrease the variance in the estimate of the p-value.