

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 ($X_{\text{train}}, Y_{\text{train}}$).
- 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 ($X_{\text{test}}, Y_{\text{test}}$) 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 13),
- 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 13a) ? 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 .

Submission instructions

- We accept both handwritten and electronic submissions. So you can choose what is more convenient for you. In any case, you should specify full names and immatriculation IDs of all team members. Obviously, programming tasks you can submit only electronically.
- Handwritten submissions should be submitted in the lecture hall of Monday's lecture (before the lecture starts).
- Electronic submissions should be zipped, containing the m-files (**Basis** etc.), your plots (png files) and the matlab data files (.mat) and emailed to the corresponding tutor:
 - a. Apratim Bhattacharyya (Wednesday 8-10): `abhattach@mpi-inf.mpg.de`
 - b. Maksym Andriushchenko (Thursday 8-10): `s8mmandr@stud.uni-saarland.de`
 - c. Max Losch (Friday 16-18): `mlosch@mpi-inf.mpg.de`

If not all 3 students belong to the same tutorial group, then you should email your submission to **only** one tutor (e.g. to the tutor of the first author of your homework), so please do not put other tutors in copy of the email.

The email subject must have the following form: "[ML18/19 Exercise] Sheet X", where X is the number of the current exercise sheet. Then please specify in the email full names and immatriculation IDs of all team members. Then please attach all your files as a single zip archive, which consists of your immatriculation IDs, e.g. "2561234_2561235_2561236.zip".

- Reminder: you should submit in groups of 3. Otherwise, we will later on merge the groups smaller than 3 students.