

Homework 4

Due Thursday, April 28, end of day. Please follow the following format for the filename of your submission: Lastname_Firstname_HWx.zip, where x needs to be substituted by the homework #.

Maximum possible points to earn: 15; 100% = 15 points, rest counts as bonus.

Feature selection and cross-validation

The goal of this homework is to make you aware of potential pitfalls in the use cross-validation for model selection.

You are given a data set from some biological experiment - the features represent physiological parameters of a group of individuals and the label is 1 if the individual has a certain disease and -1 otherwise. The biologists who have generated the data claim that they can predict the labels with around 80% accuracy, using least squares together with feature selection.

Your task is to check if the biologists have done a good job in their 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

$$\hat{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

where \mathbf{X} is the feature matrix and \mathbf{y} the vector of labels.

Classification of a new test instance x is done via $f(x) = \text{sign}(\langle \hat{w}, x \rangle)$. As error measure use the classification error,

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

There are $d = 15$ 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 $2^{15} - 1 = 32,767$ possible feature subsets** by minimizing the 5-fold cross validation error on the training data ($\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}}$).

- Report the best feature subset(s) and its/their 5-fold cross-validation error (**write them down in your report**).
- 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?

Hints:

* the indices for the five folds need to be taken as $\{1, 2, \dots, n/5\}, \{n/5 + 1, \dots, 2n/5\}, \dots, \{4n/5 + 1, \dots, n\}$.

** Use existing code to generate all possible feature combinations (unless you really want to do it yourself, but then make sure your code works). In Python, I recommend using `itertools.combinations`. In R, use `library(data.table)`
`allp=do.call(CJ, replicate(n, 0:1, FALSE))`

- c) Explain how randomly permuting the labels Y_{train} could have been used to detect the issue in b) without the need of an additional test set.
- d) Implement the idea in c) for 20 random permutations***. For each of those random permutations, re-run a) with Y_{train} being replaced by its permuted version, and save the minimum cross-validation error over all feature subsets. How do the minimum errors obtained in this way compare to the one in a)?

*** Depending on your machine, this may take between five and ten minutes of time.

Point distribution:

a)	b)	c)	d)
6	4	2	3