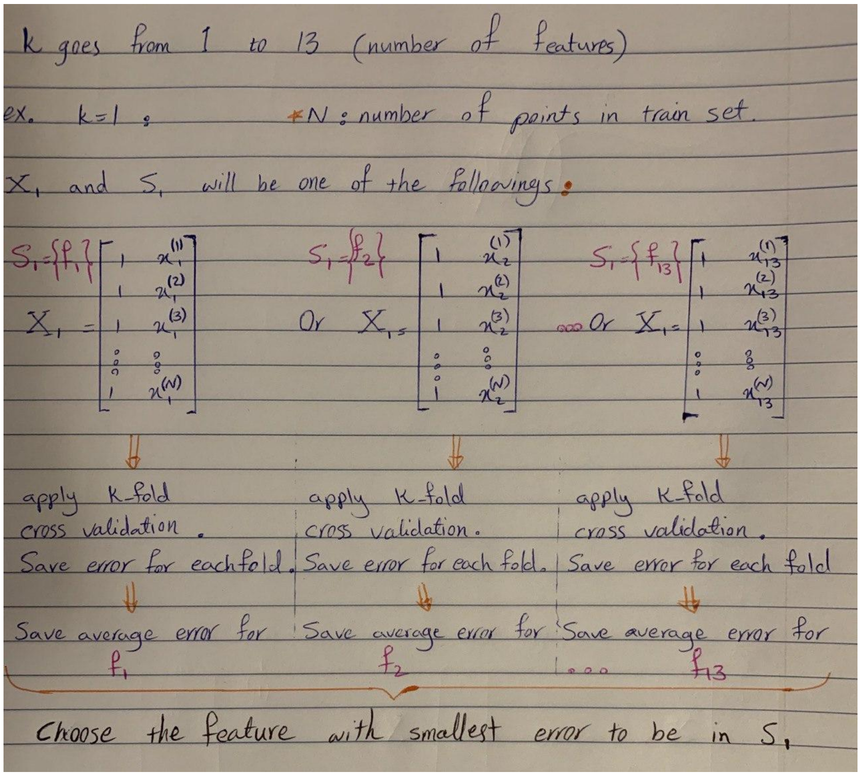
Notes for A1P2; How to use Kfold cross validation

Sara Zendehboodi

I. What to do in the first part of Assignment 1 part 2

In this part, you should select the most effective features among the 13 features. Consider k to be the numbet of steps (or the number of features is set S) in each iteration. Then k goes from 1 to 13. At the begining, split the whole data to test and train (use the instructions in the assignment). Then set the test part aside for future use. Then use the train part to train the models and find the best features.

For example, in $k = 1$, you need to find the best feature that gives the smallest error. Therefore, $|S_{(1)}| = 1$ and $|X_{train}| = N_{train} \times 1$. Note that you can add the dummy 1 to make the computations easier.



k goes from 1 to 13 (number of features)

ex. k=1 : *N : number of points in train set.

X_1 and S_1 will be one of the followings:

$S_1=\{f_1\}$ $X_1 = \begin{bmatrix} 1 & x_1^{(1)} \\ 1 & x_1^{(2)} \\ 1 & x_1^{(3)} \\ \vdots & \vdots \\ 1 & x_1^{(N)} \end{bmatrix}$ $S_1=\{f_2\}$ Or $X_1 = \begin{bmatrix} 1 & x_2^{(1)} \\ 1 & x_2^{(2)} \\ 1 & x_2^{(3)} \\ \vdots & \vdots \\ 1 & x_2^{(N)} \end{bmatrix}$ $S_1=\{f_{13}\}$ ... Or $X_1 = \begin{bmatrix} 1 & x_{13}^{(1)} \\ 1 & x_{13}^{(2)} \\ 1 & x_{13}^{(3)} \\ \vdots & \vdots \\ 1 & x_{13}^{(N)} \end{bmatrix}$

apply k-fold cross validation. Save error for each fold. | apply k-fold cross validation. Save error for each fold. | apply k-fold cross validation. Save error for each fold

Save average error for f_1 | Save average error for f_2 | Save average error for ... f_{13}

Choose the feature with smallest error to be in S_1

Then, in the next iteration, $k = 2$, you should find the next effective feature among the remaining 12. As an example, lets say $f_{(3)}$ was chosen in the previous iteration. Therefore, $S_{(1)} = \{f_{(3)}\}$. Now $S_{(2)}$ will have $f_{(3)}$ in it, plus a new feature that you are going to find in this iteration. So you should try $S_{(2)} = \{f_{(3)}, f_{(1)}\}$, $S_{(2)} = \{f_{(3)}, f_{(2)}\}$, $S_{(2)} = \{f_{(3)}, f_{(4)}\}$, $S_{(2)} = \{f_{(3)}, f_{(5)}\}$, up to $S_{(2)} = \{f_{(3)}, f_{(13)}\}$, to find the one with smallest error and save it in $S_{(2)}$.

You do the same process until you get to $S_{(13)}$ which is the complete set of features.

Whenever you want to measure the performance of a set S, you should apply cross validation to that set. To find the smallest error, you should use K-fold cross validation, which I will discuss in the next section.

Note that after finding the best features in each set (for each k), you have to train the model using train data only using features in that set. Then you can compute the predictions and the test error (in each k) for the test data you put aside at the begining.

II. Using K-fold cross validation in your assignment

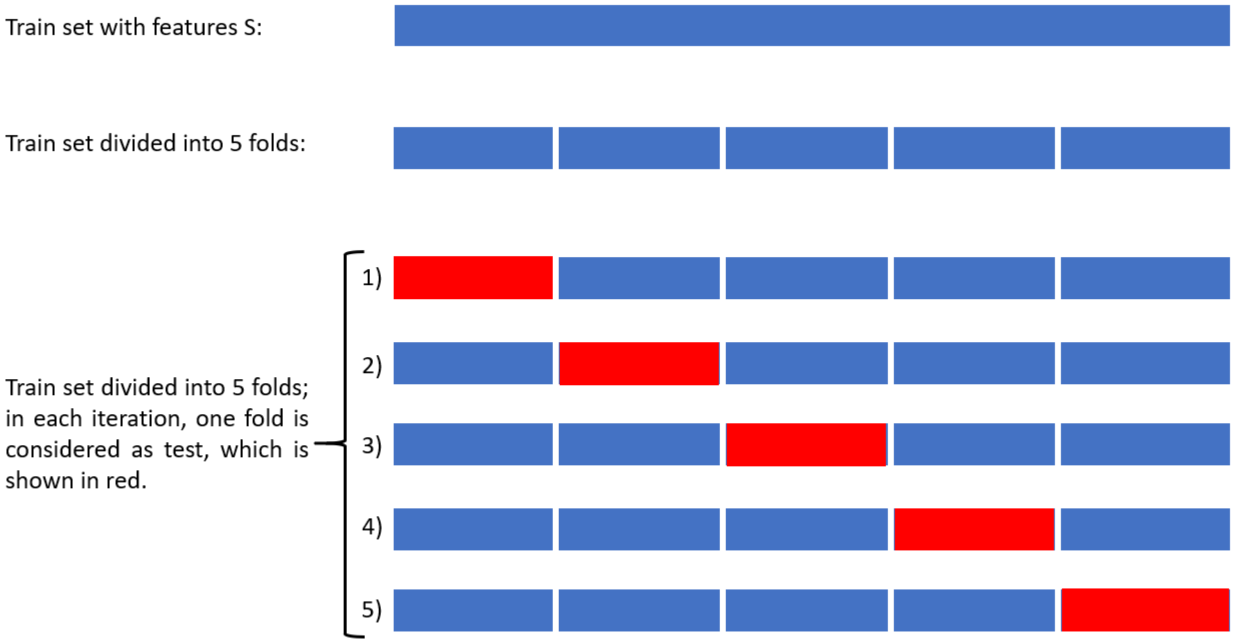
To use K-fold cross validation, follow the instruction given in Topic 3.5 in course notes. Note that α here is the feature to be found for each k.

You must give the train data to the K-fold function. For value of K, number of splits, the default is usually 5. Using ”sklearn.model_selection.KFold()”, you can split the train data into K folds. The way this function gives you the splits, is as follows:

1) Set $n.splits$ as the number of folds you want. Lets say 5.

2) Then give the whole train data (with only the features that you should consider in each S, each iteration inside k) as X, and the targets as Y. Then function KFold gives you 5 models for the split of the data into train/test. Note that you will get the indexes corresponding to train and test in 5 lines. The test set here, works as the validation set. This is equivalent to spliting the data into 5 folds, and each time puting 1 fold aside as the test and using the rest as the train.

3) Then for each line of index, use the train part to train the model and use the test part to test your model and find the test error. This is the error you should save and after K-fold process was finished for the 5 lines of index, you should take the average of 5 errors you found, and set it as the error of using the specific features.


Train set with features S:
Train set divided into 5 folds:
Train set divided into 5 folds; in each iteration, one fold is considered as test, which is shown in red.
1)
2)
3)
4)
5)

If you have questions about the assignments, please attend the TA office hours using teams. If you could not attend, please leave your questions in the channels for TAs and we will respond as soon as we can.