Assignment 2. Linear Regression, Basis Expansion and Feature Selection

Due Date: February 10, 2022, 11:30 pm
Assessment: 6% of the total course mark

Description:

In this assignment you are required to experiment with linear regression, basis expansion and feature selection. You will use the Boston housing data set (available in scikit-learn; use "load_boston"). Feature selection refers to selecting a subset of the features to be used in the prediction.

In some cases, selecting just a subset of features might improve the prediction. Another situation when we seek to use only a subset of features is when we want to determine a smaller subset of features that affect the prediction the most.

One approach for feature selection is "forward-stepwise" selection [1]. This is a greedy algorithm that starts with an empty set and grows it gradually by selecting at each stage the feature that increases the performance of the predictor the most. In other words, you will start with S_0, which is empty. At stage 1, you select one feature to add to S_0 and obtain the set S_1. At stage 2 you select another feature to add to S_1 in order to obtain S_2. You continue in this manner generating the sets S_3 with 3 features, S_4 with 4 features, and so on up to S_{13}, which includes all 13 features there are available for the Boston housing data set.

In order to decide what feature to select at each stage, you will use K-fold cross-validation (see Topic 5). Say, you are at the beginning of stage n. Thus, you have the set S_{n-1} with the $n-1$ features selected before. You want to choose another feature F_0 to include in this set. For this, for each feature F which was not already selected, you have to assess the performance of the linear model that uses F in addition to the $n-1$ features in S_{n-1} (thus n features in all) to make the prediction. For this, you will use the K-fold cross-validation error as a measure of performance. This means that after dividing the training set into K folds, you keep the first fold aside and train a linear model with the n features specified above on the remaining $K-1$ folds. Then compute the test error using the first fold as the test set. You repeat the process $K-1$ times, each time putting aside a different fold as a test set. At the end, compute the average of the K test errors to obtain the K-fold cross-validation error. This value multiplied by -1 will be considered as the score assigned to the feature F. After computing the score for each feature which is not in S_{n-1}, you select the feature F_0 with the highest score, i.e., with the lowest K-fold cross-validation error, and add it to S_{n-1} to obtain S_n[1]. After that, train the model that uses the features in S_n for prediction, and compute the test error.

Next, you have to use basis expansion on the whole set of 13 features , i.e., augment or replace the set of features with some functions of them (see the slides at the end of Topic

[1] For more details and refinements of this technique and for other subset selection methods, refer to [1, Section 3.3].

2). Consider at least two models with basis expansion and select the best one among these two and the model without basis expansion using K-fold cross-validation. Then train the selected model and compute the test error. The choice of the basis functions to use is yours. The goal is to obtain an improvement in performance. You will have to try more basis functions, if necessary, until at least one of your models with basis expansion achieves a cross-validation error smaller than the model without basis expansion.

At the beginning you have to split the data into the training and test sets. **Whenever you use randomization in your code, use a number formed with the last 4 digits of your student ID (in any order) as a seed for the pseudo number generator or as a random state.**

You have to write a report to present your results and their discussion. You have to specify the value of K that you choose. For each $n, 1 \leq n \leq 13$, you have to specify the set S_n. For the S_{13} model, specify the basis expansions that you have tried as well. Include in your report, the cross-validation errors for all models you consider. Then train the selected model and compute the test error. Organize well these results.

For the 14 models that you have selected The report should contain a plot of the cross-validation errors and of test errors for the 14 models that you have selected (corresponding to $S_1, \cdots, S_{13}$ and S_{13} plus basis expansion), versus n (the size of the subset of features). Discuss the relation between the cross-validation error and the test error. (For each model, which one is larger? Is this relation consistent for all models? etc.). Among all models without basis expansion, which one has the smallest cross-validation error? Does the same model have the smallest test error? Include any other observation that you think are interesting.

Also specify what guided you in choosing the basis functions in your trials. Additionally, specify all the sources that you have used for inspiration.

Besides the report, you have to submit your numpy code. The code has to be modular. Write a function for each of the main tasks. Also, write a function for each task that is executed multiple times (e.g, to compute the average error, to compute the K-fold cross-validation error for a model, etc). The code should include instructive comments. You are allowed to use from scikit-learn only functions to split the data into the training and test sets and the functions to split the training set into K folds.

Submission Instructions:

- Submit the report in pdf format, the python file (with extension ".py") containing your code, and a short demo video. The video should be 2 minutes or less. In the video, you should scroll down your code, show that it runs and that it outputs the results for each part of the assignment. Submit the files in the Assignments Box on Avenue.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd Ed., Springer, 2009 (ISBN 9780387848570), available for free download at https://web.stanford.edu/ hastie/ElemStatLearn/