Lab II - Feature Selection

module II, Data Collection/Understanding/Preparation

DT4031, *Applied Machine Learning*
Jens Lundström

Introduction

As you have been studying from *module II* the process of dimensionality reduction is crucial to create efficient, robust and understandable machine learning models. One of the two main methods is to select a subset of the available features, a process called *Feature Selection.* This lab will be about implementing and testing *Sequential Backward Selection*. You will study if the two previously mentioned methods for reducing the features propose the same set of features after elimination and to study how many features that are suggested (to give the best generalization error).

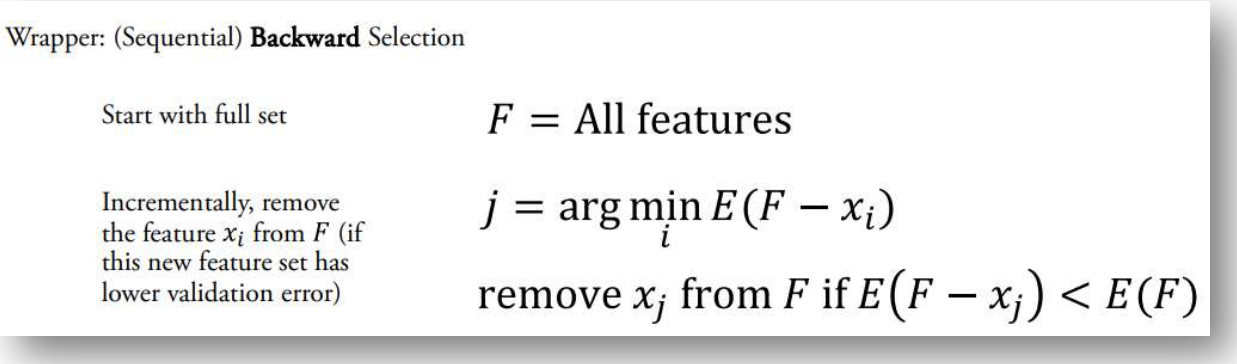
For this lab you will use an artificially created dataset by functions contained in scikit-learn itself. The machine learning algorithms (for *regression*) that will be used are *k*-NN regression and linear regression. The size of the dataset is 1000x100 (*NxD*) and your task is to reduce *D* to improve the prediction accuracy (decrease the generalization error). The features are already scaled by their standard deviation and centered around their mean which means that you do not have to apply any *feature transformation* in this lab. Keep in mind that the ML task to be done here is a *regression* task (prediction of a numerical value) and not *classification* (which we have seen examples of before in the course). Therefore, the model performance is not measured in the same way as classification accuracy. Instead, the two models are assessed using the *coefficient of determination*, called R^2. This measure (ranging from 0 to 1) corresponds to the fraction of how much of the variance of the target variable that is accounted for (how good the fit is). For further details please read *section 4.6* (and especially page 82) in the course literature, *Introduction to Machine Learning*.

Lab assignment

Open the notebook attached to lab II (*Lab_II_Feature_Selection_STUDENT_version.ipynb*). Study the code, run the notebook and try to understand the resulting numbers and figures. Answer the questions:

- What is the training (empirical) error for the two models (expressed in R^2)?
- What is the test (generalization) error for the two models (expressed in R^2)?
- Which of the 2 models and their feature selection seems to be most suitable for the regression task?

Examine the function *find_feature_index_which_gives_smallest_validation_error_if_removed(…)* and try to understand the inner workings of it. Your task is now to use the function to implement backward selection. Remember the equations/pseudo-code for doing so:



Wrapper: (Sequential) **Backward** Selection

Start with full set $\quad F = \text{All features}$

Incrementally, remove the feature x_i from F (if this new feature set has lower validation error) $\quad j = \arg\min_i E(F - x_i)$

remove x_j from F if $E(F - x_j) < E(F)$

Figure 1 - "Pseudo-code" for Sequential Backward Selection

After implementing backward selection (which should be approximately 10-20 lines of code), please answer the following questions:

Update for students taking the course during 2023: Try first to implement the SBS function by yourself. If you need assistance the code will be provided by Kunru.

- To how many features is the feature set reduced to?
- Is the backward selection processing resulting in the same features for the two models?
- Which model now gives the best generalization error?
- Is the experiment requiring a validation dataset?
- Which conclusions could you make from the above experiment?

Lab report instructions

The lab solution shall be documented in the form of a written report. As a data scientist it is your task to explain what you have done in a detail such that the solution is reproducible, the reasoning behind the decisions, the outcome (results) and the conclusions you have drawn from such results. It is suggested to have the following lab report structure: Introduction & Motivation; Methods & Solutions; Results; Conclusions & Discussion; Appendix. Keep the number of pages in the lab report limited, no more than five pages (excluding *Appendix* which should contain Notebook code and figures not directly linked to the result).

Lab assistance

Feel free to reach out to Kunru or Jens if you have questions regarding the lab.