INF264 - Homework 3

Pierre Gillot Natacha Galmiche[*]

Week 36 - 2021

1 Model selection for regression

In this first exercise we want to compare the regular model selection using a simple validation set versus more involved model selection methods. In particular, we are interested in the so-called "KFold cross-validation" procedure. This is a follow-up to the exercise from the file 'polynomial_regression_2D.ipynb', thus you can use your code from this assignment as a starting point.

Consider once again the Boston Housing dataset: `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`. Remember that in this dataset, each sample corresponds to a house, whose target price is to be inferred from 13 features contained in the dataset. You will implement a code that answers the following questions (you can help yourself with the provided template code):

1. Load the Boston dataset: store the 13 features in a matrix X and the target price in a vector Y.
2. Investigate the Boston dataset. It is obviously not possible to visualize all the 14 dimensions at the same time, but it may be a good idea to represent the target price as a function of each of the 13 features individually. **Hint:** To load the dataset, use the 'load_boston()' from 'sklearn.datasets'. You can plot using the 'scatter' function from 'matplotlib.pyplot'.
3. Identify (at least) 2 features that are continuous and appear to be correlated to the target price. Extract these features to obtain a simpler features matrix (with less columns than the full features matrix). **Note:** Ideally, the extracted features should be (pairwise) as decorrelated as possible.

We want to build a polynomial regression model that learns the target prices from the extracted features. We are not sure about which polynomial order we should use to obtain the best model. Moreover, we were told that adding a regularization term can be useful to prevent overfitting, but once again we do

[*] Not a TA for this course this year.

not know how to choose this hyper-parameter. A solution is to perform model selection with cross-validation on the hyper-parameters.

4. Finish the implementation of a KFold cross-validation procedure in order to perform model selection of a Ridge model with respect to its hyper-parameters (polynomial order and regularization value), using the MSE metric. More specifically:

   i Split the whole dataset into a train and a test set, then set aside the test set.

   ii Split the train set into 5 (train,validation) folds using the KFold class from sklearn.

   iii Loop over each hyper-parameters instance you cross-validate on (outer loop).

   iv For current instance of hyper-parameters, loop over each (train,validation) fold (inner loop).

   v In the inner loop, assess a "surrogate" model with current instance of hyper-parameters on the current (train,validation) fold, i.e fit your surrogate model on the train fold and evaluate it using the MSE metric on the validation fold.

   vi In the outer loop (after the inner loop finishes), compute the mean validation MSE over each (train,validation) fold.

   vii Select the model with the smallest mean validation MSE.

   viii Train the selected model on the whole train set, then evaluate it on the test set which was set aside at the end of step i.

   Cross-validation is to be done on the polynomial order ranging in $\{1, 2, 3\}$ and on the regularization value ranging in $\{0, 0.001, 0.01, 0.1\}$ for a total of 12 hyper-parameters combinations. Indicate which hyper-parameters combination obtained the best results with respect to the MSE metric during the KFold cross-validation procedure.

5. When fitting a Ridge model of 3rd degree, you should have encountered the warning "Singular matrix in solving dual problem. Using least-squares solution instead.". Can you explain why this warning occured ? How much do you need to increase the regularization hyper-parameter in order to get rid of this warning ?

6. Perform a regular model selection using a simple validation set to select the best Ridge model with respect to the polynomial order and the regularization value. Comment on the differences with the KFold cross-validation procedure.

2 Model selection for classification

In this second exercise, we will first illustrate how misleading the accuracy metric can be when assessing a classifier on unbalanced data. Finally, we will cross-validate different classifiers on an unbalanced dataset with a relevant metric (you can help yourself with the provided template code).

1. Create randomly generated binary datasets, with a 1st class ratio ranging in $\{0.6, 0.75, 0.9, 0.95, 0.98, 0.99\}$. For each dataset generated this way, train a K-NN classifier with $K = 10$, then evaluate it on its corresponding test set with respect to the accuracy metric, the $F1$-score metric and the confusion matrix metric. Plot all of those results in a single figure (using subplots). Does the accuracy metric appear to assess the quality of your model in an appropriate way ?

2. Load the custom randomly generated binary dataset contained in the file 'custom_unbalanced_dataset.pickle'. Visualize this dataset. How unbalanced is it ? Inspiring yourself from the previous exercise, perform model selection on three different classification models: a K-NN classifier, a logistic regression classifier and a decision tree classifier. In order to do this, use Kfold cross-validation with the number of folds set to $k > 5$. Indicate and justify which metric you decided to use in order to cross-validate the different models. Finally, train and evaluate the best model on the whole dataset (evaluate on the test set with the $F1$-score and the confusion matrix).