Machine Learning For Data Science I, Homework 02

Maj Gaberšček, 27212075

March 2023

Code description and instructions

This homework was coded in programming language `Python`, version `3.11`, although it should work in some older versions as well. The only packages used for this homework were

- `numpy` and `pandas` for data processing,
- `sklearn` for logistic regression model,
- `seaborn` and `matplotlib` for plotting.

The code is organized so that every part of this report is in a separated script (which is accordingly named) and all auxiliary functions are imported from file `aux_functions.py`.

1 Setup: a proxy for true risk

First we generated the toy dataset of 100.000 rows. As we were coding this homework in `Python`, we had to write the `toy_data` and the `log_loss` functions from scratch.

We generated 100.000 rows of data and stated, that it is enough, to safely reduce the error to the first decimal digit. We can see this by estimating the margin of error or standard error of the sample mean. **Margin of error** can be calculated as

$$ME = \frac{z \cdot s}{\sqrt{n}},$$

where s is the sample standard deviation, n is the sample size and z is the z-score (1.96 for 95% confidence interval). We want to have $ME <= 0.0005$ (for accuracy to the 3rd decimal digit with 95% confidence). We can also assume a large value for s, such as 0.5. If we insert the numbers, we can see, that $n = 100.000$ is enough to have $ME <= 0.0005$.

2 Holdout estimation

2.1 Model loss estimator variability due to test data variability

In this part of the homework, we wanted to test, how true risk of a model differs to estimated risk, calculated on test data. The output of my code is printed below and the plot of estimated risk differences is drawn in Figure 1:

```
True risk proxy:              0.4853
Mean difference:              0.0006
0.5-0.5 baseline true risk:  0.6399
Median standard error:        0.0592
Percentage of 95CI that contain the true risk proxy: 93.1
```

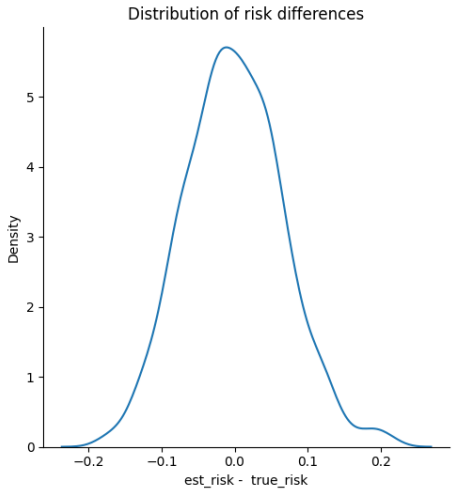


Figure 1: Density estimate of the differences between estimates and the true risk proxy

From the results, we can conclude, that sometimes holdout estimation overestimates the risk and sometimes it underestimates it. The mean difference is very close to 0, from which we can see, that holdout estimation does indeed return expected risk same as the true one (if we performed estimation enough times on different data, we could get as close as needed to true risk). So, in practice, we could estimate true risk proxy as accurately as needed by repeating this procedure more and more times.

If the training set would be bigger, mean difference would drop even more (because larger training set would learn the model more accurately). The median standard error of estimations would also decrease. Number of times, that the confidence interval contains the true risk proxy would also increase with larger testing set (as the training set would now represent the population even

better). Similarly, with a smaller training set, the effect would be the other way around.

2.2 Overestimation of the deployed model's risk

Here, we wanted to see, how the true risk calculated on model, which learned on 100 observations, differs to true risk of a model, which learned on a subset of this 100 observations of size 50. Results of true risk differences (model on smaller dataset minus model on bigger dataset), repeated 50 times, is copied below:

```
Minimum:        -0.0349
1st Quantile:    0.0133
Median:          0.0322
Mean:            0.0381
3rd Quantile:    0.0666
Maximum:         0.1265
```

We can see, that almost every time, we achieved better predictions (smaller risk), if training on additional 50 data occurrences. If both datasets would be bigger (and difference in size would remain the same), the difference of risks would start to get smaller, as additional samples would not change the prediction that much (as population would already be well represented). Similarly, if both datasets were smaller, each individual training sample would contribute more towards final prediction, and differences in models' risks would be bigger.

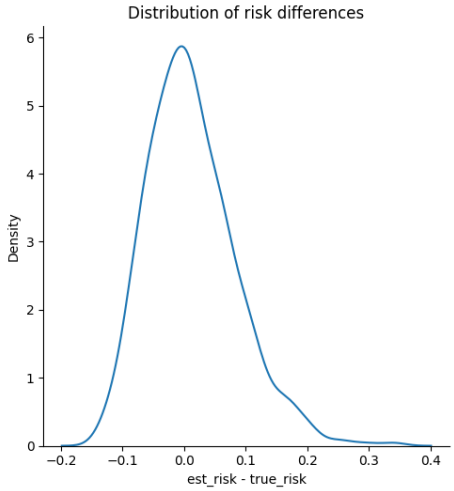


Figure 2: Density estimate of the differences between estimates and the true risk proxy

2.3 Loss estimator variability due to split variability

Here, we generated a toy dataset of size 100 and calculated the true risk proxy. Then we split the dataset into training and test and then trained another model on training data. We estimated original model's risk (model trained on data of size 100) by testing on test data. Code output was:

```
True risk proxy:         0.4933
Mean difference:         0.0123
Median standard error:   0.0986
Percentage of 95CI that contain the true risk proxy: 97.8
```

We plotted difference between estimated risk and true risk in Figure 2.

We can conclude, that if the data would be bigger, estimated risk would also get closer to true risk. Mean difference and median standard error would decrease.

3 Cross-validation

In this section, we tested risk estimations with different kinds of cross-validations. Results of this experiment:

```
----------------------------------------
Estimator: 2-fold cross validation
Mean difference:                                0.0438
Median standard error:                          0.0037
Confidence interval contains true risk proxy:   82.4%
----------------------------------------
Estimator: leave-one-out cross validation
Mean difference:                                0.0017
Median standard error:                          0.0027
Confidence interval contains true risk proxy:   91.2%
----------------------------------------
Estimator: 10-fold cross validation
Mean difference:                                0.0068
Median standard error:                          0.0028
Confidence interval contains true risk proxy:   90.0%
----------------------------------------
Estimator: 4-fold cross validation
Mean difference:                                0.0174
Median standard error:                          0.0030
Confidence interval contains true risk proxy:   87.6%
----------------------------------------
Estimator: 10-fold cross validation repeated 20 times
Mean difference:                                0.0061
Median standard error:                          0.0027
Confidence interval contains true risk proxy:   32.0%
```

If we compare different estimators, we can conclude, that 2-fold cross validation generated biggest mean difference in comparison to other estimators, as

well as the biggest median standard error. Best model in terms of mean difference and also median standard error is leave-one-out, which was well expected. However, for practical purposes, it is important to point out, that leave-one-out model is by far the most time consuming one.

If we compare 10-fold cross validation and 10-fold cross validation repeated 20 times, we can see almost no difference in term of median standard error or mean difference.

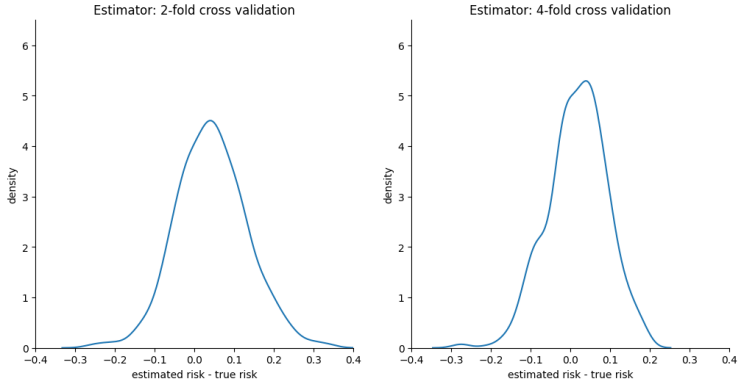
Plots of risk differences can be seen in Figure 3.

3.1 A different scenario

I did not code this part of the exercise. I did, however, find an article [1], which states, that leave-one-out technique may provide a poor estimate of model performance for dataset with very high variance (because of overfitting).

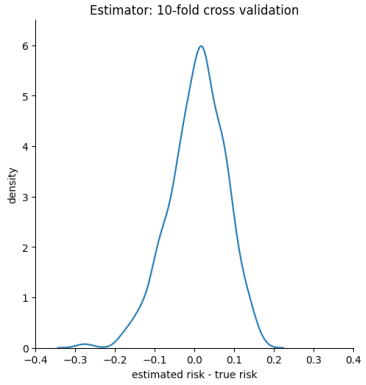
References

[1] Ron Kohavi and Roger Longbotham. "The Relationship between Leave-One-Out Cross-Validation and k-Fold Cross-Validation". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*. Montreal, Canada, 1995.

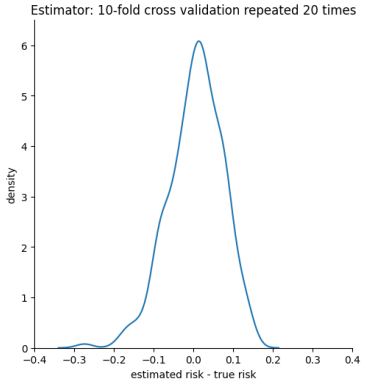


(a) 2-fold

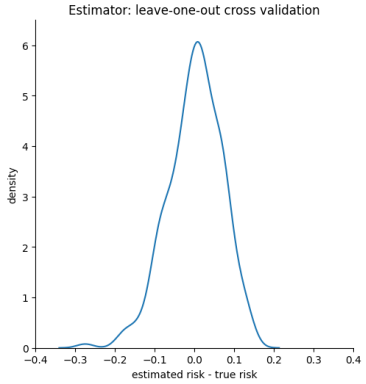
(b) 4-fold



(c) 10-fold



(d) 20 times 10-fold



(e) leave-one-out

Figure 3: Density estimate of the differences between estimates and the true risk proxy