Homework 2 : Loss Estimation

Stefanela Stevanović, 63220492

Introduction

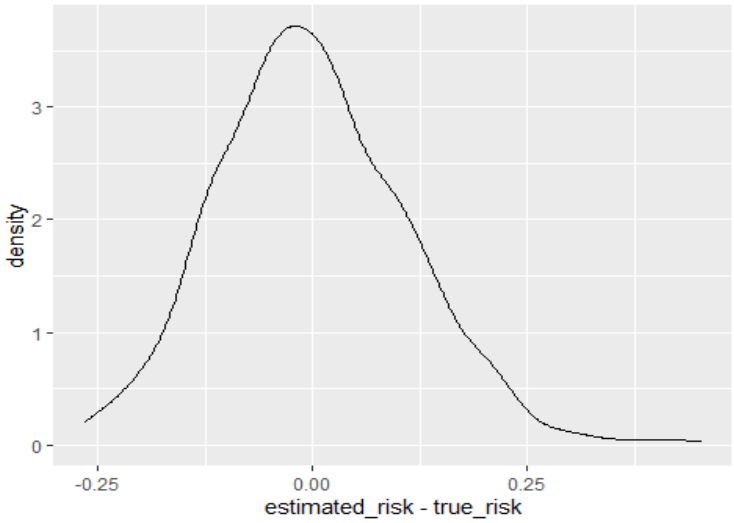
The true risk is an important concept in statistical learning theory because it provides a measure of how well our model is likely to perform on new, unseen data. In practice, we cannot directly calculate the true risk because we do not know the DGP. Instead, we estimate the risk using the training data and various statistical techniques. However, the model's risk on the test data is just an estimated of the model's true risk, and may not reflect the real situation. The goal of this homework is to prepare the data generator, so that we would be familiar with DGP and use holdout estimation to estimate a model's risk and investigate the sources of variability and bias that contribute to the difference between the estimated model's risk and the true risk. We will be using logistic regression learner and log-loss function, a classification metric based on probabilities, for evaluating the performance of binary classification models. The smaller the value of the log-loss, the better the model's performance.

Setup

We firstly prepared the generator for our toy binary classification data, which consists of 8 independent variables, with 3 of them having no relationship with the target variable. We generated a dataset of 100000 samples as a proxy for the data generating process and determining the ground-truth true risk. This dataset is considered large enough to reduce the error between the model's risk on this dataset and the true risk to the 3rd decimal digit. According to the law of large numbers, as the sample size increases, the sample mean approaches the true population mean. Or, in the case of this proxy dataset, as the number of observations in the proxy dataset increases, the model's risk on this dataset will approach the true risk. Due to Central limit theorem, the rate at which the error decreases is proportional to the square root of the sample size. The square root of 100000 is 316.23, which means that we can expect the difference between the model's risk on this dataset and the true risk to be at most 1/316.23, or approximately 0.003.

Model loss estimator variability due to test data variability

In this experiment we inspected how the test data risk estimate varies with the test data. We trained a model h on a toy dataset with 50 observations and computed true risk proxy using the huge dataset with 100000 observations. Then, we iteratively (1000 times) generated new toy dataset with 50 observations and estimated the risk of model h on this dataset. The obrained results are present below.




```
## True risk proxy: 0.5755

## Mean difference: 2e-04

##  0.5-0.5 baseline true risk: 0.6931

## Median standard error: 0.1094

## Percentage of 95CI that contain the true risk proxy: 93
```

From the results of the analysis, we see that the true risk proxy is 0.5755, which indicates that on average, the model is performing better than randomly guessing (baseline true risk of 0.6931). This was expected, since we know that model was trained on the data containing 5 independent variables that influence the target variable. However, median standard error of 0.1094 indicates high variability of estimates for different test sets. The mean difference between the estimated risk and the true risk proxy is very small (0.0002), indicating that the model's performance on the test set is very close to its performance on the huge dataset. We can see from the plot that median difference between estimated risk and true risk proxy lies a little below 0. This tells us that in most cases the risk was a little underestimated. Also, a high percentage of 95% confidence intervals that contain the true risk proxy (93%) suggests that the estimated risk is likely to be a good approximation of the true risk on new data. From the obtained results, we can conclude that the model doesn't generalize well due to high variability across different test sets.

If the training set was larger, we would expect the model to perform better on the testing set, as it has more data to learn from. More accurate predictions would cause less variability (lower standard error) and thus result in the lower median standard error of the estimates across all iterations. Larger training set would also cause lower true risk proxy and lower estimated risk on the testing set, but we can't tell if the bias would change, or how it would change. From the larger testing set, in general, we expect an estimate of the risk to be closer to the true risk proxy, resulting in lower bias and higher percentage of 95%CI that contain true risk proxy. Fewer observations in the testing set could lead to incorrect risk estimations, making significant underestimation or overestimation possible. Also, with a larger testing set, the estimates of model performance will be based on a larger number of observations, which can help to reduce the effects of random variation in the data. This may result in lower standard error of the estimates.

Overestimation of the deployed model's risk

In this experiment we trained model h1 on the dataset with 50 observations and model h2 on the dataset with 100 observations, in order to inspect how the size of the data used during the training process influences the true risk of the model.

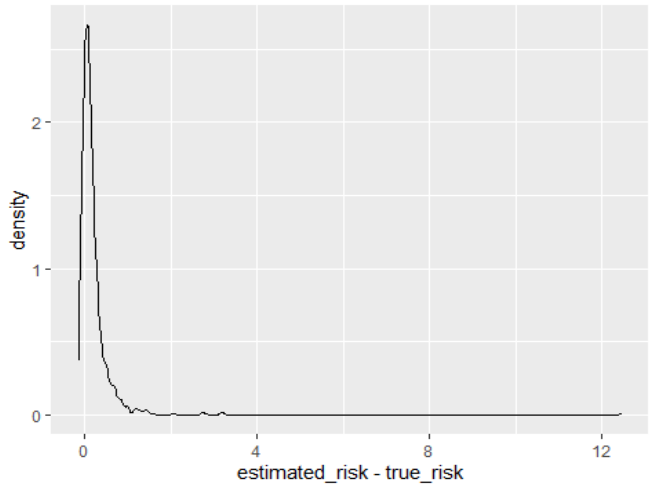
```
## Summary of true risk h1 - true risk h2:

##          Min.    1st  Qu.      Median          Mean    3rd  Qu.         Max.
## -0.03188  0.05576  0.11605  0.63321  0.24404  8.66872
```

From the results, we can see that the true risk of the model trained on a smaller dataset (h1) is consistently higher than the true risk of the model trained on a larger dataset (h2). This demonstrates that using more data typically leads to lower true risk and better performance of the deployed model. If we start increasing the size of the data used for training the model, we would likely see a similar trend where more data leads to lower true risk, until we encounter the problem of over-fitting. Over-fitting happens when the model is simply memorizing the training data instead of learning the general patterns, causing the true risk to start increasing, as model starts performing poorly on unseen data. On contrary, using smaller data for training the model would increase the true risk. If we decrease the sizes of both datasets by the same factor, the mean (and median) difference between their true risks will increase. This happens because the learner is not provided with enough data, and each additional data point is of great importance for the learner. If we increase the sizes of both datasets the mean difference between their true risks will decrease, because the learner has already captured some important patterns in the data and additional data may not introduce new information.

The implications for practical use are that it is important to use as much data as possible when training machine learning models (while also being mindful of overfitting), as this can significantly improve their performance on new data. However, it is also important to balance the cost of collecting and processing large datasets with the expected gains in model performance, because the performance improvements become smaller and smaller as the dataset size increases.

Loss estimator variability due to split variability

It this experiment we inspect the loss estimator variability due to train-test split variability, caused by different observations ending up in the training and testing sets. Firstly, we trained a model h0 on a toy dataset with 100 observations and computed the true risk proxy using the huge dataset with 100000 observations. Then, we iteratively (1000 times) split this toy dataset into train and test sets at random, with each having 50 observations in total. In each iteration we trained the model h on the the training set and used testing set for estimating the model's risk.




```
## True risk proxy: 0.5255

## Mean difference: 0.2025

## Median standard error: 0.1256

## Percentage of 95CI that contain the true risk proxy: 86
```

Firstly, we notice that the true risk proxy calculated in this experiment is lower than in the first experiment, because the model was trained on twice as many data observations. Also, we notice that the mean difference between estimated risk and true risk (0.2025) is far higher when compared to the first experiment (0.0002). This high bias comes from the fact that the tested models are trained on less data than the model we are interested in. This also causes the percentage of 95%CI that contains the true risk proxy to drop in comparison to the first experiment. From the plot we can see that most differences between estimated risk and true risk are positive, indicating that we overestimated the risk. Median standard error of 0.1256 indicates high variance due to split variability.

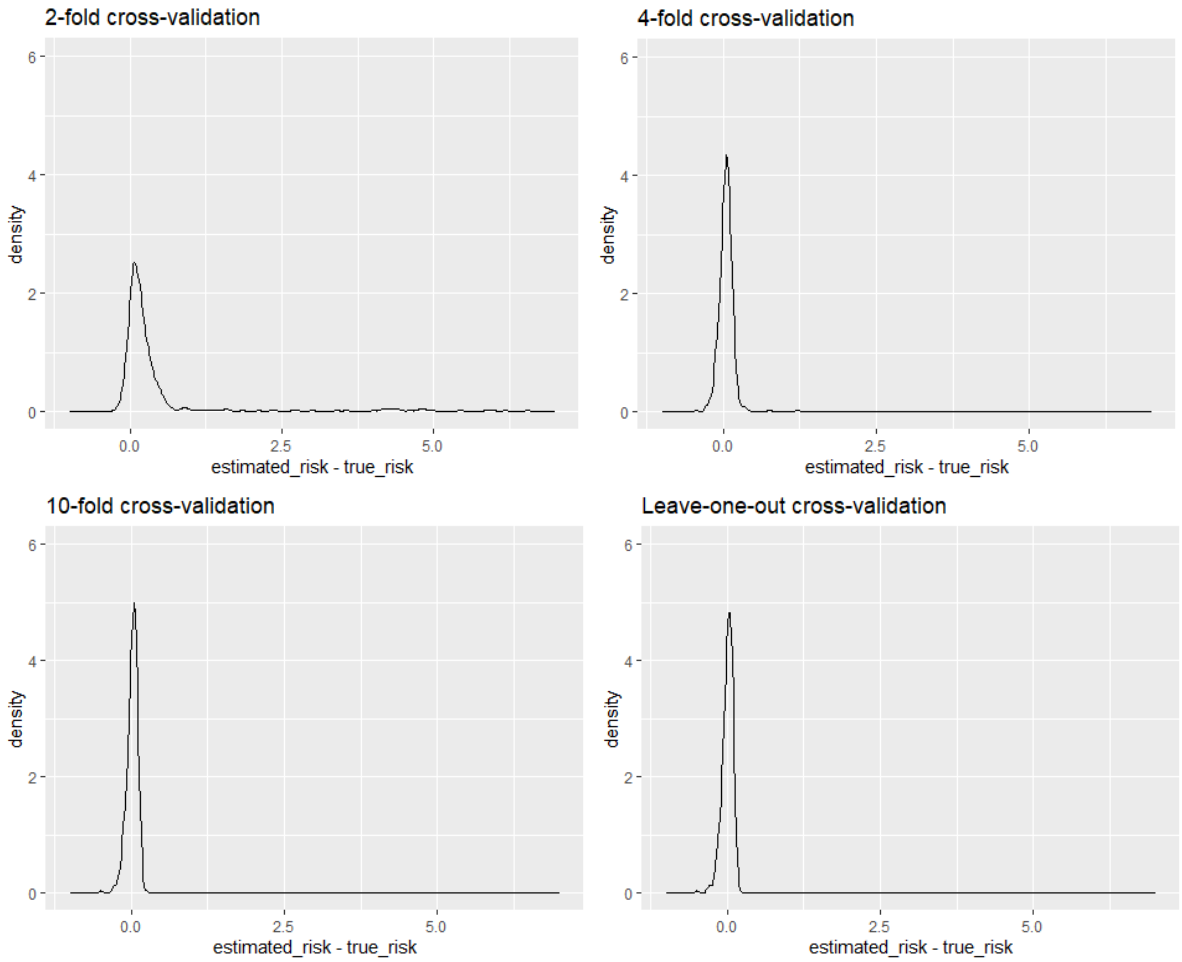
If the dataset was larger, we would expect the bias and median standard error of the estimates to decrease. This is because a larger dataset would provide more data for training and testing the model. On the other hand, using the smaller dataset would further increase the bias and variability of the estimates. However, we would not want to use a dataset that is larger than the dataset we are interested in, as this may result in underestimation of the risk and increase in bias.

If we used a smaller proportion of the training data, we would expect model to have poorer performance on the test set, resulting in higher median standard error and higher estimated risk. Since the model already had high bias in 50:50 split, due to using less data in training tested models than the model we are interested in, this would cause even higher increase in bias and overestimation of the risk. This would also decrease the percentage of 95%CI that contains the true risk proxy. On the other hand if we used a higher proportion of the training data, we would expect lower median standard error and lower estimated risk, which consequently leads to reducing the overestimation of the risk and smaller bias.

Cross-validation

In this experiment we inspect how different cross-validation methods affect the loss estimator variability. Through 500 iterations, we generated a toy dataset with 100 observations, trained a model ho and computed true risk on the huge dataset. Then, we performed cross-validation to estimate ho model's risk using 4 estimators: 2-fold cross-validation, 4-fold cross-validation, 10-fold cross-validation and leave-one-out cross-validation.

```
## ESTIMATOR: 2-fold

## Mean difference: 0.425

## Median standard error: 0.1092

## Percentage of 95CI that contain the true risk proxy: 68.8

## ESTIMATOR: 4-fold

## Mean difference: 0.0387

## Median standard error: 0.083

## Percentage of 95CI that contain the true risk proxy: 90.2

## ESTIMATOR: 10-fold

## Mean difference: 0.0107

## Median standard error: 0.0777

## Percentage of 95CI that contain the true risk proxy: 93.4

## ESTIMATOR: LOOCV

## Mean difference: -3e-04

## Median standard error: 0.0752

## Percentage of 95CI that contain the true risk proxy: 92
```



As the number of folds increases, the mean difference between the estimated and true risk proxy and the median standard error also decrease. This suggests that using a higher number of folds leads to more accurate estimates of the model's risk. Since with the higher number of folds, more data is used in the training, reduction in the overestimation of the risk was expected. Furthermore, the percentage of 95CI that contain the true risk proxy also increases as the number of folds increases. This indicates with using a higher number of folds, the uncertainty around the estimates is better captured by the confidence intervals. The results also show that using leave-one-out cross-validation (LOOCV) leads to the lowest median standard error and bias, but the difference in performance between LOOCV and 10-fold cross-validation is relatively small. This suggests that LOOCV may be a good choice when the dataset is small, and the computational cost is not a concern. However, if the dataset is large, 10-fold cross-validation may be a more practical choice due to its lower computational cost.

In this experiment, we didn't implement 20 times repeated 10-fold cross-validation, as we would not expect results to significantly change compared to 10-fold cross-validation. This is because we are using 500 iterations to generate new dataset and perform the cross-validation. We would expect that the effect that potential imbalanced splits on some datasets have on the estimated risk is made less significant through 500 iterations. However, in practice, when we would not be able to just generate new dataset 500 times, it would be a

good idea to use a repetition in cross-validation, as it can reduce the impact of random variations in the data splits and provide a more robust estimate of the model's risk.

A different scenario

By choosing a different learner, DGP or dataset size it is possible to get the results that don't agree with the previous experiment. We experimented with using 2-fold cross-validation, 4-fold cross-validation and 10-fold cross-validation estimators on a significantly larger dataset of 100000 observations. Because of the huge datasets, we reduced the numbers of iterations from 500 to 200 iterations, in order to speed up the computation.

```
## ESTIMATOR: 2-fold

## Mean difference: 4e-04

## Median standard error: 0.0054

## Percentage of 95CI that contain the true risk proxy: 95

## ESTIMATOR: 4-fold

## Mean difference: 0

## Median standard error: 0.0054

## Percentage of 95CI that contain the true risk proxy: 95

## ESTIMATOR: 10-fold

## Mean difference: -1e-04

## Median standard error: 0.0054

## Percentage of 95CI that contain the true risk proxy: 96
```

Firstly, we notice that with larger datasets risk estimates are more accurate than with smaller datasets, and the effects of increasing the number of folds in cross-validation are less visible. This is because with large datasets enough data is available for training, even when using a smaller number of folds. Also, with larger datasets, computational cost becomes more of a concern, so using a cross-validation with smaller number of folds may be a better option.