Loss estimation

Ivan Nikolov

2024-03-21

Setup

A proxy for true risk

Q: How did I determine that 100000 is enough to reduce the error to the 3rd decimal digit?

```
sample_set <- toy_data(100, 0)
df_dgp_1 <- toy_data(100000, 1)
h_test <- glm(y~ ., data=sample_set, family = binomial())

loss_estimate <- log_loss(df_dgp_1$y, get_preds(h_test, df_dgp_1))

sd_risk <- sd(loss_estimate) / sqrt(1e5)
sd_risk
```

```
## [1] 0.002162163
```

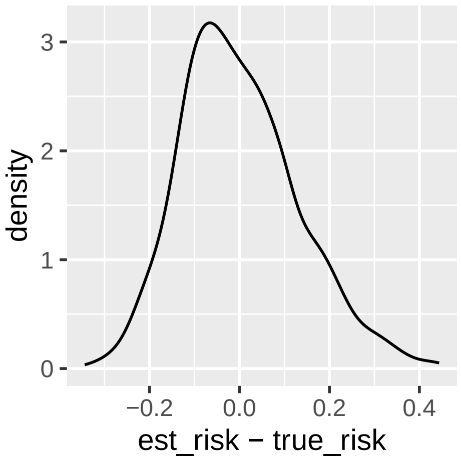
```
(mean(loss_estimate) + 1.96 * sd_risk) - (mean(loss_estimate) - 1.96 * sd_risk)
```

```
## [1] 0.008475681
```

By the strong law of large numbers the sample average of a risk from a given dataset will converge to the true value. By the central limit theorem the error of the risk (our loss function) will be normally distributed with standard deviation $\sqrt{n}$ smaller than the standard deviation of the sample. If we calculate the difference between the 2.5th percentile and 97.5th percentile, we can see that the difference is only at the third decimal.

Holdout estimation

Model loss estimator variability due to test data variability



```
## [1] "## True risk proxy: 0.5759"
```

```
## [1] "## Mean difference: 0.0007"
```

```
## [1] "## 0.5-0.5 baseline true risk: 0.6931"
```

```
## [1] "## Median standard error: 0.1268"
```

```
## [1] "## Percentage of 95CI that contain the true risk proxy: 92.9"
```

When using holdout estimation, in average the estimator is unbiased which can be noticed from the small bias value (`est_risk - true_risk`). From the pdf estimate, we can see that the errors are skewed to the right, which means that the risk is underestimated. In practice, this means that our estimates will largely depend on our test set (especially if the test set is small). In our case the model performs better that the baseline, however the median standard error is also large. With increasing training set size, the estimated model will get close to the optimal attainable h, and the risk decrease. Smaller train set can easily be impacted by outliers and yield a suboptimal model. A smaller test set would introduce a larger variance in the estimations, while with a larger data set we would decrease the standard error.

Overestimation of the deployed model's risk

```
## [1] "Summary of true risk h1 - true risk h2:"
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -0.02558  0.02763  0.12924  0.18009  0.22342  0.94565
```

We overestimate risk by using holdout estimation because the model only uses a portion of the data for training. With larger dataset set size, these differences will become smaller because we have more data to learn and test on.

Loss estimator variability due to split variability

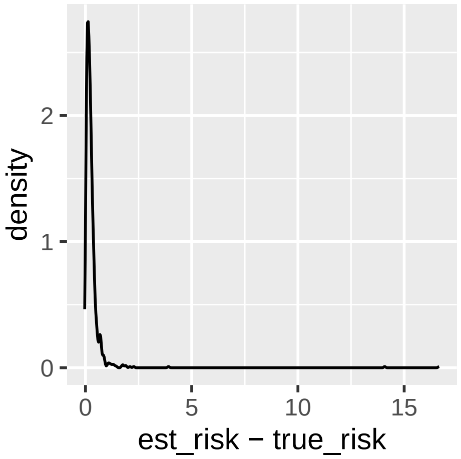
```
## [1] 0.4857711
```

```
## [1] "True risk proxy: 0.4858"
```

```
## [1] "Mean difference: 0.2848"
```

```
## [1] "Median standard error: 0.1259"
```

```
## [1] "Percentage of 95CI that contain the true risk proxy: 74.70"
```



From the statistics and the density estimation we can see that there is a large positive bias (skewness of the risk difference distribution). This is because the model only uses half of the data for training. In addition, the dataset is small and different splits introduce variance in the model learning and loss estimation procedure. Both variance and bias would decrease if the dataset was larger. Smaller dataset will increase the bias and variance. Larger training data will decrease bias, however the variance when estimating risk on the test set will be larger. Larger testing data will decrease variance.

Cross validation

```
## [1] "2-fold"
```

```
## [1] "Mean difference: 0.4395"
```

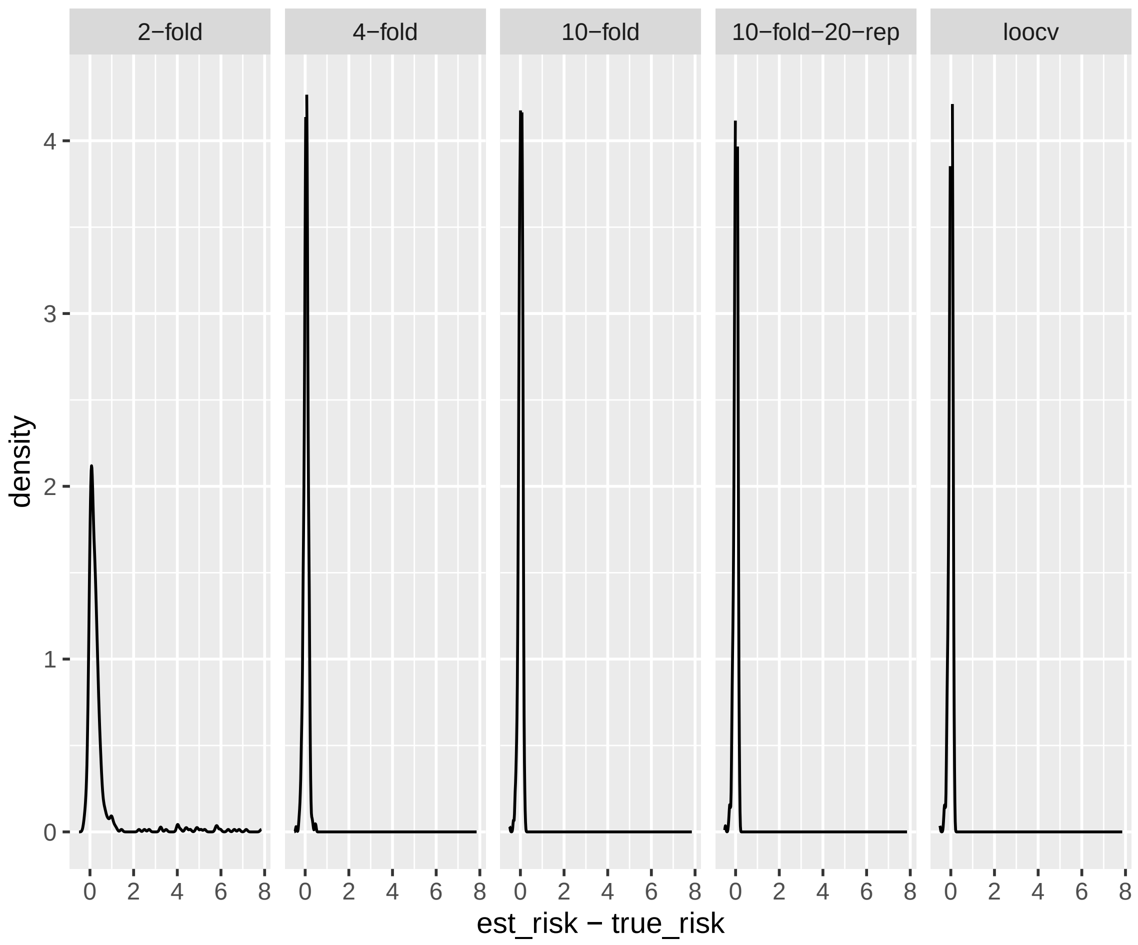
```
## [1] "Median standard error: 0.1137"
```

```
## [1] "Percentage of 95CI that contain the true risk proxy: 65.0"
```

```
## [1] "----------"
```

```
## [1] "4-fold"
```

```
## [1] "Mean difference: 0.0397"
## [1] "Median standard error: 0.0841"

## [1] "Percentage of 95CI that contain the true risk proxy: 89.6"

## [1] "----------"

## [1] "10-fold"

## [1] "Mean difference: 0.0100"

## [1] "Median standard error: 0.0775"

## [1] "Percentage of 95CI that contain the true risk proxy: 93.2"

## [1] "----------"

## [1] "20 repeated 10 fold"

## [1] "Mean difference: 0.0098"

## [1] "Median standard error: 0.0765"

## [1] "Percentage of 95CI that contain the true risk proxy: 93.2"

## [1] "----------"

## [1] "LOO"

## [1] "Mean difference: -0.0013"

## [1] "Median standard error: 0.0747"

## [1] "Percentage of 95CI that contain the true risk proxy: 92.0"

## [1] "----------"
```



From the results that we got, using CV with bigger fold number will give us better estimates with lower variance. We achieved the best results with 10-fold CV, 20 times repeated 10-fold CV, and LOOCV. LOOCV in theory should give us the best results, however the main drawback is the long computation time. In case the model is unstable (very sensitive) to training data, the variance of the k-fold CV with also increase with k. It is important to note that because the dataset is small, the training sets that are small are also small can introduce outliers in the model risk estimates.