HW1: Classification trees, random forests

Marko Ivanovski

Decision tree misclassification rates

Table 1 shows the misclassification rate on the decision tree model evaluated on the train and test dataset. To understand how I evaluated the standard error, we first must understand that the misclassification rate itself is an average over all **wrong** predictions. Therefore, we can assume that the train or test dataset are two samples with instances on which we can just use the sample variance formula and divide the result with the square root of the number of instances to get the variance of the mean:

$$s_\mu = \frac{s}{\sqrt{n}}, s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}, \bar{x} = \frac{\sum_{i=1}^{n}(x_i)^2}{n} \quad (1)$$

	Misclassification rate	Standard error
Train	0.77%	0.76%
Test	20.69%	5.32%

Table 1. Misclassification rates and their standard errors on the decision tree model evaluated on two sets of data.

Random Forest misclassification rates

Table 2 shows the scores for the Random Forest model. In it, 100 decision trees were trained on randomly sampled data.

	Misclassification rate	Standard error
Train	0.00%	0.00%
Test	1.73%	1.71%

Table 2. Misclassification rates and their standard errors on the Random Forest model evaluated on two sets of data.

Misclassification rates vs the number of trees

Figure 1 illustrates how the misclassification rate changes by increasing the number of trees in the forest.

Variable importance in context with the trees root variables

Figure 2 shows the variable importance. In it, the red dots are the variables that appeared in the root node when 100 DT were built on randomly sampled data. The number next to them is how many times that particular variable was used as a root variable. The numbers coincide with the feature importance algorithm output in a way that more important features were tree roots more times.

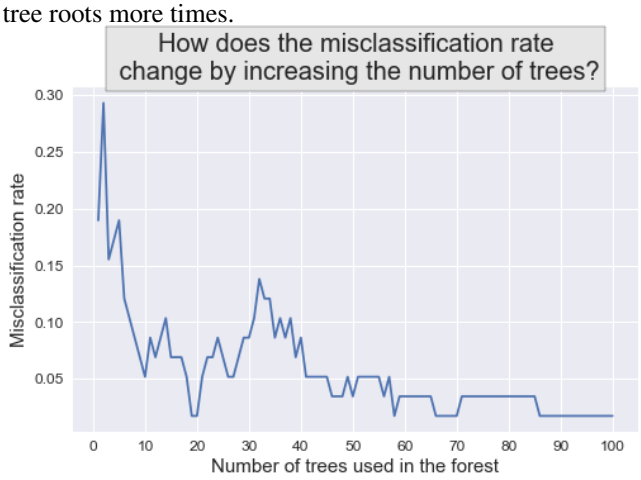

Figure 1. Number of trees vs MCR scores. A visualization of how the MCR rates change on the testing data based on the number of trees used in the RF model.

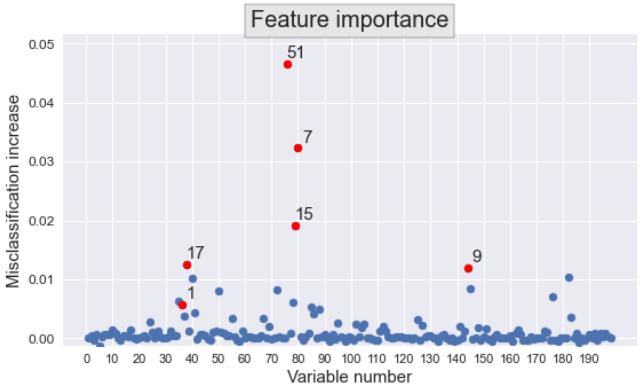

Figure 2. Variable importance. The x-axis contains all variables with their misclassification rate increase (when shuffled) on the y-axis. The red dots are the variables present in the tree roots when 100 DT were built on sampled data. The numbers next to them correspond to how many times that variable was used in the root of the 100 trees.