Statistical Learning **Homework assignment 3**
Grün **SS 2024**

3 Regression and classification trees, random forests

Exercise 1:

The data set `Carseats` from package **ISLR2** is used to predict `Sales` using regression trees, treating the response as a quantitative variable.

- Split the data set into a training set and a test set.
- Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?
- Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

Exercise 2:

- Draw 100 observations from four independent variables $X_1, \dots, X_4$ where
  - X_1 follows a uniform distribution,
  - X_2 follows a standard normal distribution,
  - X_3 follows a Bernoulli distribution with success probability $\pi = 0.5$,
  - X_4 follows a Bernoulli distribution with success probability $\pi = 0.1$.
- Repeat 1000 times the following:
  - Draw a dependent variable y from a standard normal distribution which is independent of the four independent variables.
  - Fit a tree stump, i.e., a tree which contains only one split.
  - Determine which variable was used for splitting.
- Create the table of relative frequencies how often each of the variables was selected for splitting. Given that all independent variables are not associated with the dependent variable, is the probability of including them as a split variable the same? If not, why would they differ?

Exercise 3:

Assume the following data generating process

$$Y = X + \epsilon,$$

with $X \sim N(0, 1)$ and $\epsilon \sim N(0, 1)$ independent.

In addition 20 covariates $Z_1, \dots, Z_{20}$ are given with

$$Z_i \sim \sqrt{0.9}X + \epsilon_{Z_i},$$

where $\epsilon_{Z_i} \sim N(0, 0.1)$.

- Draw a training data with 30 observations and a test data with 10,000 observations from the data generating process including the additional covariates $\boldsymbol{Z}$.
- Sample 100 bootstrap samples of size 30 from the training data of the previous example by drawing with replacement.
- Fit to each bootstrap sample:
  (a) a regression tree,
  (b) the null model with predicted value equal to the observed empirical mean of Y,
  (c) a linear model including linear effects for X and all Z variables and
  (d) a linear model potentially including linear effects for X and all Z variables, but using model selection with the AIC to select a suitable model starting from the null model.
- Determine the predicted values on the test data for the bagged model estimator by calculating the average predictions over the 100 trees fitted to the bootstrap samples, the 100 null models, the 100 linear models including all linear effects and the 100 linear models based on model selection.
- Determine the mean squared error of the four bagged model estimators on the test sample of size 10,000.

Exercise 4:

The dataset `icu` in package **aplore3** contains information on patients who were admitted to an adult intensive care unit (ICU). The aim is to develop a predictive model for the probability of survival to hospital discharge of these patients. Use random forests to fit a predictive model to the data.

- Select a suitable number of bootstrap iterations.
- Assess the influence of varying the hyperparameter m on the out-of-bag error obtained and select a suitable value.
- Inspect the variable importance measures. Compare the mean decrease Gini and the mean decrease accuracy measures and assess if the observed differences in relative importance assigned might be related to the predictor variable being numeric or not.

Exercise 5:

In the following we analyze the performance of the variable importance measures for random forests using a simulation study.

- Assume that there are four predictor variables which have the following distributions:

$$\begin{aligned} X_1 &\sim N(0,1), & X_2 &\sim U(0,1), \\ X_3 &\sim M(1,(0.5,0.5)), & X_4 &\sim M(1,(0.2,0.2,0.2,0.2,0.2)). \end{aligned}$$

  This means we have two continuous variables which follow either a standard normal or a standard uniform distribution ($U(0,1)$ and two categorical variables with balanced categories with either 2 or 5 categories, i.e., $M(N,\pi)$ is the multinomial distribution for N trials and success probability vector π.
- The dependent variable y is assumed to be a binary categorical variable with equal-sized classes.
- Set the sample size to $N = 200$.
- Generate 100 datasets for each setting and fit a random forest to each dataset and determine the mean decrease Gini and mean decrease accuracy values for each of the predictor variables. Suitably visualize the results and interpret them.