Statistical and Machine Learning **Homework assignment 4**
Grün **WS 2023/24**

4 CART & ensemble methods: bagging

Exercise 1:

A binary dependent variable is generated by

$$\Pr(Y = 1|X) = q + (1 - 2q) \cdot 1\left[\sum_{j=1}^{J} X_j > J/2\right],$$

where 1[] is the indicator function, $X \sim U[0,1]^p$, $0 \leq q \leq 1/2$, and $J \leq p$ is some predefined (even) number.

Describe this probability surface, and give the Bayes error rate.

Exercise 2:

Suppose x_i, $i = 1, \ldots, N$ are i.i.d. with mean μ and variance σ^2.

Let $\bar{x}_1^*$ and $\bar{x}_2^*$ be two bootstrap realizations of the sample mean.

Show that the sampling correlation

$$\mathrm{Cor}(\bar{x}_1^*, \bar{x}_2^*) = \frac{N}{2N-1} \approx 50\%.$$

Along the way, derive $\mathrm{Var}(\bar{x}_1^*)$ and the variance of the bagged mean $\bar{x}_{\mathrm{bag}}$.

Note: $\bar{x}$ is a linear statistic; bagging produces no reduction in variance for linear statistics.

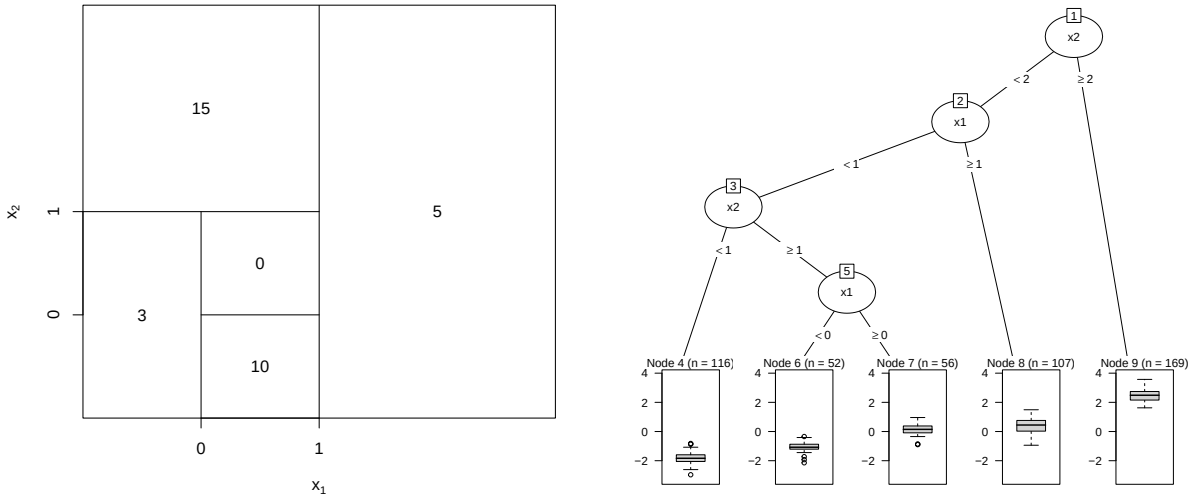
Exercise 3:

Suppose we fit a linear regression model to N observations with response y_i and predictors $x_{i1}, \ldots, x_{ip}$. Assume that all variables are standardized such that for example for $\boldsymbol{y}$ it holds $\boldsymbol{y}^\top \mathbf{1} = 0$ and $\frac{1}{N}\boldsymbol{y}^\top \boldsymbol{y} = 1$. Let RSS be the mean-squared residuals on the training data, and $\hat{\beta}$ the estimated OLS coefficient. Denote by ${RSS_j}^*$ the mean-squared residuals on the training data using the same $\hat{\beta}$, but with the N values for the jth variable randomly permuted before the predictions are calculated. Show that

$$\mathrm{E}_P[{RSS_j}^* - RSS] = 2\hat{\beta}_j^2,$$

where E_P denotes expectation with respect to the permutation distribution.

Exercise 4:



- Sketch the tree corresponding to the partition of the predictor space illustrated on the left of the figure. The numbers inside the boxes indicate the mean of Y within each region.
- Create a diagram similar to the plot on the left in the figure, using the tree illustrated on the right of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region. Determine also the fitted function.

  Additional information on the fitted tree is summarized below:

```
n= 500

node), split, n, deviance, yval
      * denotes terminal node

 1) root 500 1500.0  0.41
   2) x2< 2 331  370.0 -0.65
     4) x1< 1 224  170.0 -1.20
       8) x2< 1 116   16.0 -1.80 *
       9) x2>=1 108   51.0 -0.44
        18) x1< 0.0003 52    6.0 -1.10 *
        19) x1>=0.0003 56    7.7  0.12 *
     5) x1>=1 107   23.0  0.40 *
   3) x2>=2 169   26.0  2.50 *
```

Exercise 5:

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 6:

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.

Fit a classification tree to the data without pre-processing:

- Use very loose stopping criteria such that the tree might be overfitting.
- Inspect the fitted tree and describe it.
- Use pruning to select a tree with a suitable size. Determine this smaller tree and inspect and describe it.

Note: Omit the variable `id` from the set of potential predictors.

Exercise 7:

- Draw 100 observations from four independent variables $X_1, \ldots, X_4$ 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 8:

Assume the following data generation process:

$$y = x + \epsilon,$$

where $x \sim N(0, 1)$ and $\epsilon \sim N(0, 0.1)$ independently.

- Repeat 100 times:
  - Draw 100 observations from the data generation process.
  - Fit a linear regression and determine the prediction error.
  - Fit a regression tree using cost-complexity pruning to select a suitable tree. Determine the tree size and the prediction error.
- Visualize one data set together with the fitted predictions using the linear model as well as the tree.

- Summarize the results across the 100 repetitions regarding prediction error of the linear model and the fitted tree as well as the tree size.

  Do the results indicate that regression trees might have problems to capture linear relationships?

Note: $\epsilon \sim N(0, 0.1)$ means that ϵ has mean zero and a variance of 0.1, i.e., a standard deviation of $\sqrt{0.1}$. The R function `*dnorm` has as arguments for the parameters `mean` and `sd`.

Exercise 9:

We assume data with the following data generation process:

$$x = y + \epsilon,$$

where y is a categorical variable with values 1, 2, 3, which occur with equal probability and $\epsilon \sim N(0, 0.2)$ independent.

- Draw 100 data sets of size 100.
- Determine the sum of the misclassification rates, Gini indices and deviance criteria weighted with the number of observations in each subgroup for the subgroups obtained when splitting the observations using x with thresholds 1.5, 2, and 2.5 and y as dependent variable in the classification problem.
- Calculate the best threshold according to each of the three impurity measures for each of the 100 data sets. Summarize and interpret the results.

Note: $\epsilon \sim N(0, 0.2)$ means that ϵ has mean zero and a variance of 0.2, i.e., a standard deviation of $\sqrt{0.2}$. The R function `*dnorm` has as arguments for the parameters `mean` and `sd`.

Exercise 10:

Assume that $y \sim N(0, 3)$.

- Draw a sample of size 10 from the data generating process.
- Determine the mean estimate from the sample.
- Use bootstrapping with $B = 1000$ to determine a bootstrap-based estimate for the mean.
- Compare the results. Which value would the bootstrap estimate have if not Monte Carlo sampling would be used to approximate the estimate, but if the estimate based on the true bootstrap distribution would be determined?

Note: $y \sim N(0, 3)$ means that y has mean zero and a variance of 3, i.e., a standard deviation of $\sqrt{3}$. The R function `*dnorm` has as arguments for the parameters `mean` and `sd`.