Statistical and Machine Learning **Exercise 4**
Grün, Bajons **WS 2025/26**

4 CART

Sub-task 1:

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.

Sub-task 2:

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 predictor 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.

Sub-task 3:

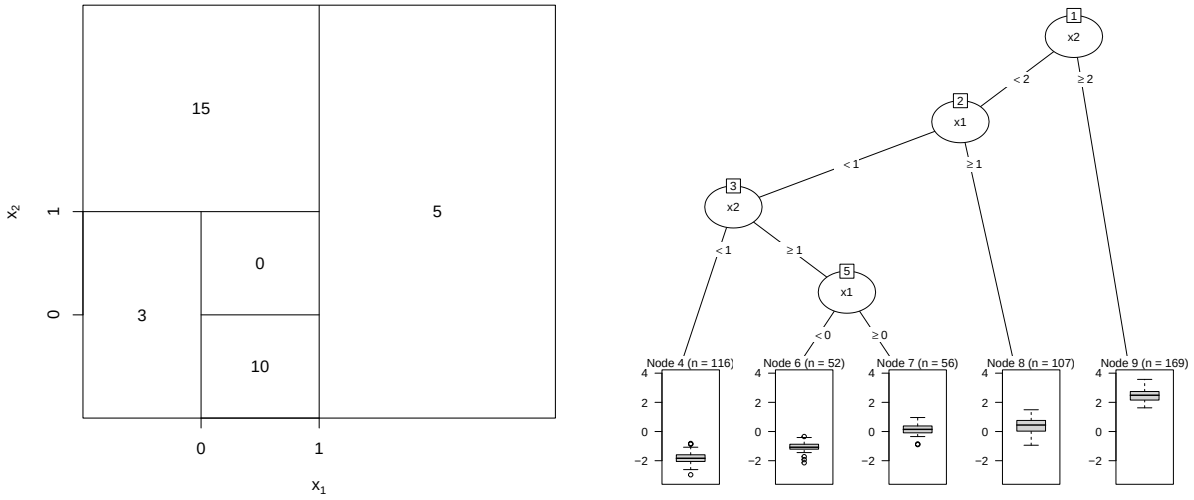
Assume the following data generating process

$$\begin{aligned} \Pr(Y = 1|X) &= \pi(X), \\ \mathrm{logit}(\pi(X)) &= X, \end{aligned}$$

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

- Determine the Bayes error for the 0-1 loss for this classification problem.
- Determine the expected prediction error for the 0-1 loss for the fitted model which predicts always 1.
- Determine the expected prediction error for the 0-1 loss for the fitted model which predicts 1 for positive X and 0 otherwise.

Sub-task 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 *
```

Sub-task 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?

Sub-task 6:

- 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?

Sub-task 7:

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 expected prediction error based on squared error loss.
  - Fit a regression tree using cost-complexity pruning to select a suitable tree. Determine the tree size and the expected prediction error based on squared error loss.
- 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 the expected 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`.

Sub-task 8:

The `spam` dataset is available in package **ElemStatLearn** to learn a classifier for e-mails being spam or not.

Before fitting a model, split the dataset into a training and testing set in the following way:

```
> data("spam", package = "ElemStatLearn")
> set.seed(1234)
> test_index <- sample(nrow(spam), 1000)
> spam_train <- spam[-test_index,]
> spam_test <- spam[test_index,]
```

- Fit a generalized linear model with all other variables as additive linear covariates and using a binomial likelihood with logit link.
- Fit a classification tree where you use cross-validation in order to determine the optimal level of tree complexity.
- Assess the performance of the two fitted models on the test data. Determine the accuracy as well as 95% confidence intervals for the accuracy estimates.
- The performance of the two fitted models is evaluated using the same test data. The error rates are hence correlated. Determine the confusion matrix of the predicted values for the two models and compare the performance using a McNemar test.
- Interpret the results and discuss if the results allow to conclude that one model provides a better fit.