Statistical and Machine Learning **Exercise 5**
Grün, Bajons **WS 2025/26**

5 Ensemble methods; Deep learning

Sub-task 1:

Show that the solution to

$$(\beta_m, G_m) = \arg\min_{\beta, G} \sum_{i=1}^{N} w_i^{(m)} \exp[-\beta y_i G(x_i)],$$

with $w_i^{(m)} = \exp(-y_i f_{m-1}(x_i))$ can be obtained in two steps:

1. For any value $\beta > 0$ the solution for $G_m(x)$ is

$$G_m = \arg\min_{G} \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i)),$$

   which is the classifier minimizing the weighted error rate in predicting y.

2. Plugging this G_m into the criterion and solving for β one obtains

$$\beta_m = \frac{1}{2} \log \frac{1 - \text{err}_m}{\text{err}_m},$$

   where err_m is the minimized weighted error rate

$$\text{err}_m = \frac{\sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i^{(m)}}.$$

Sub-task 2:

Assume $Y \in \{1, -1\}$.

- Show that the population minimizer for the exponential loss is given by

$$f^*(x) = \arg\min_{f(x)} \text{E}_{Y|x}(e^{-Yf(x)}) = \frac{1}{2} \log \left(\frac{\Pr(Y = 1|x)}{\Pr(Y = -1|x)} \right).$$

- Show that the population minimizer for the deviance loss is given by

$$p^*(x) = \arg\min_{p(x)} \text{E}_{Y|x} \left(\frac{Y+1}{2} \log p(x) + \frac{1-Y}{2} \log(1 - p(x)) \right) = \Pr(Y = 1|x).$$

- Show that the population minimizer for squared error loss is given by

$$f^*(x) = \arg\min_{f(x)} \text{E}_{Y|x}(Y - f(x))^2 = \text{E}(Y|x) = 2\Pr(Y = 1|x) - 1.$$

Sub-task 3:

Let the predictive function be given by

$$f(x_1, x_2) = x_1$$

and assume that (X_1, X_2) are bivariate Gaussian with mean zero and variance-covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Determine the partial dependence function $\bar{f}(x_2)$ capturing the average effect of x_2 on the predictions and the marginal effect $\tilde{f}(x_2)$ of x_2 on the predictions:

$$\begin{aligned}
\bar{f}(x_2) &= \mathrm{E}_{X_1}(f(X_1, x_2)) \\
\tilde{f}(x_2) &= \mathrm{E}_{X_1}(f(X_1, X_2)|X_2 = x_2).
\end{aligned}$$

Sub-task 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.

Sub-task 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:

$$X_1 \sim N(0,1), \qquad X_2 \sim U(0,1),$$
$$X_3 \sim M(1,(0.5,0.5)), \qquad X_4 \sim M(1,(0.2,0.2,0.2,0.2,0.2)).$$

  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.

Sub-task 6:

In the following the influence of the ratio of relevant to irrelevant predictor variables on the performance is assessed for random forests with $m = \sqrt{p}$.

- 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 $X \sim U[0,1]^p$, $0 \leq q \leq 1/2$.

- Two predictor variables are assumed to be informative, i.e., $J = 2$. The number of noise predictor variables is varied between $\{5, 25, 50, 100, 150\}$. The value for q is set to 0.1 to obtain a Bayes error rate of 0.1.
- The training sample size is $N = 300$ and the test sample size is 500.
- Fit random forests with $m = \sqrt{p}$ and visualize the test misclassification rates obtained for 50 repetitions. Interpret the results.

Sub-task 7:

In the following we will develop a predictive model for the South African heart disease data available as data object `SAheart` in package **ElemStatLearn**.

- Split the data set into a training and a test dataset such that 75% of observations are in the training dataset.
- Fit a logistic regression model with backward-stepwise regression using only linear effects for the covariates (`glm` and `step`) to the training dataset.
- Fit a boosted logistic regression model with generalized additive effects (`gamboost` from package **mboost**) to the training dataset.
- Compare the predictive performance of the two fitted models on the test dataset.

Hint: Use function `gamboost` from package **mboost** with `family = Binomial()`.

Sub-task 8:

Consider the `IMDb` dataset from the **keras3** package to perform document classification. Restrict the vocabulary to the most frequently-used words and tokens.

- Fit a fully-connected neural network with two hidden layers, each with 16 units and ReLU activation to the data with dictionary size 1000.
- Consider the effects of varying the dictionary size. Try the values 500, 1000, 3000, 5000, and 10,000, and compare the results.

Hint: See James et al. (2023, Chapter 10).