Universität Salzburg
Ferndinand Ferber, BSc.
Univ.-Prof. Dr. Wolfgang Trutschnig

Salzburg, May 2024
SS 23/24

Exercise Sheet 09 @'Applied AI Using R'

We are working with decision (classification and regression) trees. As the previous sheets, exercises like these do prepare you for participating in projects and for internships in companies.

Exercise 38.
Consider the following small dataset that tracks a person's TV watching preferences. The columns `comedy`, `doctors`, `lawyers` and `guns` are the predictor variables, indicating if the TV show featured comedy, doctors, lawyers and/or guns. The column `likes` (target variable) specifies, if the person liked the TV show.

comedy	doctors	lawyers	guns	likes
FALSE	TRUE	FALSE	TRUE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	TRUE	TRUE
TRUE	TRUE	FALSE	FALSE	TRUE
FALSE	TRUE	FALSE	TRUE	FALSE
TRUE	TRUE	FALSE	TRUE	TRUE
TRUE	TRUE	FALSE	TRUE	TRUE
TRUE	TRUE	FALSE	TRUE	TRUE

Answer the following simple questions:

1. Compute the entropy for the `likes` variable, $H(\texttt{likes})$
2. Compute the conditional entropy $H(\texttt{likes}|\texttt{doctors} = 1)$
3. Compute the information gain $I(\texttt{likes}, \texttt{doctors} = 1)$
4. Compute the expected information gain $I(\texttt{likes}, \texttt{doctors})$
5. What is the optimal depth-0 decision tree (having only one leaf node and no tests)?
6. What is the optimal depth-1 decision tree (having one inner node and two leaves)?

Exercise 39.
We revisit the *indonesian nutrition dataset* from Aufgabe 28 (downloadable from `https://ufile.io/04mcf2kp`) and want to compare the performance of the linear model (studied in Aufgabe 28) and a simple regression tree by completing the following steps:

1. Randomly split the data into a training and a test dataset (70:30).
2. Fit a linear model (as described in Aufgabe 28) to the training data, predict the calories for the test data, and calculate the residuals.

3. Fit a regression tree (calories $\sim$ fat + proteins + carbohydrate) to the training data, predict the calories for the test data, and calculate the residuals. Work with `rpart` and set `method="anova"`.

4. Illustrate the obtained results by boxplots (one for the lm-residuals and one for the tree-residuals) in ggplot2 and include the mean absolute errors (MAE) in the boxplots.

5. Repeat the steps 1.-4. $R = 1.000$ times produce another boxplot summarizing the MAEs for lm and the tree models. Does one model outperform the other - which model would you chose?