Weekly Exercise - Week 4

Julian Karch

For this week's exercises, we will use a real dataset: The Breast Cancer Wisconsin (Diagnostic) Dataset. You can find information on this dataset here: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. Briefly, the goal is to predict whether a breast tumor is malignant or benign (`diagnosis` column) based on data from a sample of breast tissue (remaining columns but `id` and last column). You can find the data set itself on Brightspace (data.csv). In this assignment, you will essentially select the best performing model and estimate its performance. You will do this using the cross-validation + test set approach discussed during the lecture.

- Split the data into training and test sets, putting 80% in the training and 20% in the test set. Put the test set away and do not look at it.
- Pick three candidate models. A method is defined as method + tuning parameter. kNN with $k = 3$ is thus one method. Ideally, this should be done by considering the training set, but you can also just make a random choice. Hints: Among others, you could consider the number of predictors, the size of the training set, which assumptions are likely to be met, and methods that complement each other (e.g., a low variance method and a low bias method).
- Select the most accurate model by applying 10-fold cross-validation on the training set. Hints: Have a look at the classroom exercises / solutions.
- Estimate the accuracy, specificity, and sensitivity of the selected model on the test set. Hint: Have a look at the classroom exercises / solutions.