General Information:

Regression Trees

Exercise 1 Based on a large number of patients, a study was published that showed a correlation between age, systolic blood pressure and the risk for stroke. The average risk according to age and systolic blood pressure of a patient is given as follows:
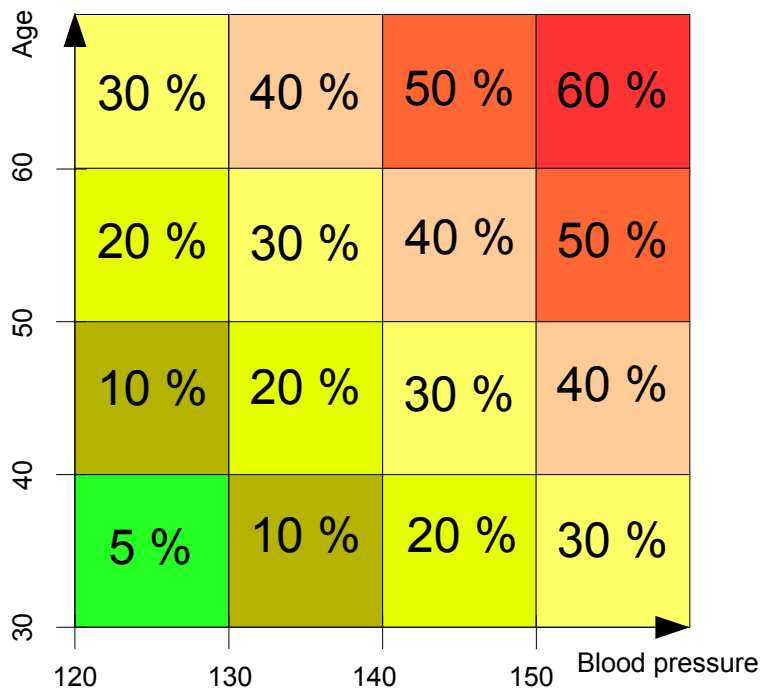
Design a regression tree for the risk of a stroke with

(a) four leaves.

(b) six leaves.

Exercise 2 Let $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set of n observations. For each observation $\boldsymbol{x}_i$, we can compute its occurrence in $\mathcal{S}$. A binary regression tree should be utilized to estimate a histogram based on the given set of observations.

(a) What are the benefits of histogram estimation based on regression trees compared to conventional binning methods (fixed bin sizes with fixed number of bins, adaptive bin sizes with fixed number of bins) discussed in the lecture?

(b) What is the major limitation of regression trees compared to the Parzen window approach in terms of probability density estimation?

(c) Write down the optimization problem to solve histogram estimation using regression trees. Explain the different terms involved in the objective function.

(d) In order to solve the optimization problem, we consider a simple Greedy strategy to determine the split position in the regression tree: Stop the splitting process if the approximation error achieved by the regression tree with respect to the relative frequencies of the samples (sum of squared differences) falls below a certain threshold.

Describe the Greedy algorithm to solve the optimization problem.

(e) Discuss the influence of the termination criterion of the Greedy algorithm to the estimated histograms. Explain the properties of the algorithm in terms of the bias-variance tradeoff.

(f) Does the proposed algorithm lead to optimal regression trees? Describe an alternative approach to train the regression tree.

Exercise 3   Matlab exercise

The concept of regression trees should be used to estimate histograms of 8-bit intensity images. Therefore, the set of observations is given by $\mathcal{S} = \{x_i, \ldots, x_n\}$ where $x_i \in [0; 255]$ is a gray value in the image. We apply the histogram estimation to retinal fundus images that were also used for Parzen window estimation.
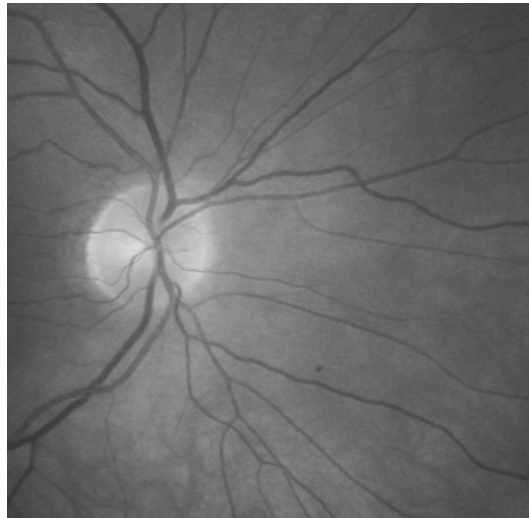
(a) Create a histogram with a fixed number of bins ($K = 256$) of equal size for the image `fundus.png`.

(b) Estimate the histogram using the Greedy algorithm discussed in the previous exercise and plot the result.

(c) Find a reasonable threshold t that can be used in the termination criterion of the algorithm. Compare the estimate with a the non-adaptive binning based on K bins.