Technische Universität München
Lehrstuhl für Informatik V
Univ.-Prof. Dr. M. Bader
M.Sc. M. Wille

SS 2022
Worksheet 5
27.05.2022

Algorithms for Scientific Computing (Algorithmen des Wissenschaftlichen Rechnens)

1D Classification

This week we focus on the question how to calculate the binary clustering of a 1D dataset S sampled from an unknown distribution f(x). We try to approximate f(x) by using the function

$$f_N(x) = \sum_{j=1}^{N} v_j \varphi_j(x) \approx f(x)$$

where N is the number of used basis functions.

For the following exercises we will use the training dataset:

$$S = [(0.1, 1), (0.2, 1), (0.3, -1), (0.35, 1), (0.4, 1), (0.55, -1), (0.6, -1), (0.65, -1), (0.7, -1), (0.8, 1)]$$

where the first element of our tuple is the feature $x_i \in [0, 1]$ und the second is the label $y_i \in \{-1, 1\}$ of the i-th datapoint.

Hint: If not specified otherwise, the domain considered from now on is the unit interval $\Omega = [0, 1]$.

Exercise 1: Interpolation-like classification

For the first task, we construct our function $f_N(x)$ by using one basis function $\varphi_j(x)$ per data point. We therefore adapt the standard hat functions from the lecture to fit the variable distance between our data points.

a) Draw the hat functions $\varphi_j(x)$ for our dataset in the interval.

   Hint: The formal definition of the hat function is:

$$\varphi_k(x) := \begin{cases} \frac{1}{h_{k-1}}(x - x_{k-1}) & x_{k-1} < x < x_k \\ \frac{1}{h_k}(x_{k+1} - x) & x_k < x < x_{k+1} \\ 0 & \text{otherwise} \end{cases}$$

   with $h_k := x_{k+1} - x_k$

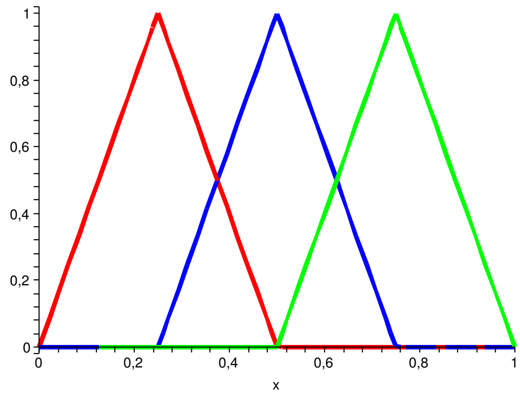
b) Determine the v_j for this classification and draw the resulting function $f_N(x)$ including the scaled hat functions $(v_j \cdot \varphi(x))$.

c) Evaluate $f_N(0.5)$. What is the result of the classification for this point?

d) What is the problem of this approach?

Exercise 2: Equidistant nodal basis

In this task we use the hat functions on equidistant intervals which are not aligned with the data points. Similar to the lecture, we consider 3 basis functions resulting in 4 equidistant intervals in the domain $\Omega = [0,1]$:



As there are more data points than basis functions, we cannot fit the function perfectly to our training set. We will therefore use the least-squares approach from the lecture to calculate the best fit of our function $\tilde{f}_N(x)$ to the data. This is equal to solving

$$\underset{v}{\operatorname{argmin}}(||Gv - y||_2^2).$$

where $G_{ij} = \tilde{\varphi}_j(x_i)$, v is the vector containing the v_k and y_i is the label of x_i. This can be calculated by solving the system of linear equations for the unknown coefficients v which results from the normal equation:

$$G^T G v = G^T y.$$

a) Construct G for the training data.

b) Calculate v and draw the resulting function.

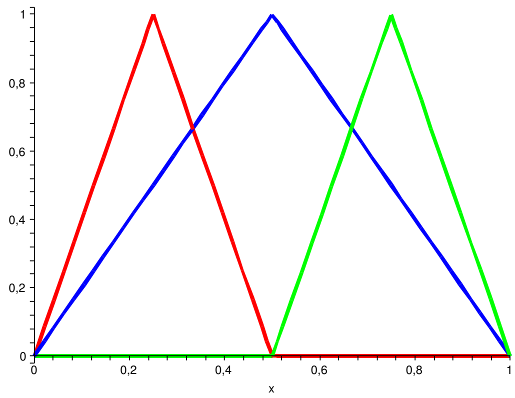
c) Evaluate $\tilde{f}_N(0.5)$. What is the result of the classification for this point?

Exercise 3: Hierarchical classification

In this task we use the hierarchical construction of the hat functions. Similar to exercise 2 we have to solve:

$$\underset{v}{\operatorname{argmin}}(||Gv - y||_2).$$

where $G_{ij} = \varphi_j^*(x_i)$ and y_i is the label of x_i. However, the hierarchical basis for constructing our approximation $f_n^*(x)$ functions ϕ are constructed in the following way:



a) Construct G for the training data.

b) Calculate v and draw the resulting function.

c) Evaluate $f_N^*(0.5)$. What is the result of the classification for this point?

d) Does $\tilde{f}_N(x)$ differ from $f_n^*(x)$?

e) What is an advantage of the hierarchical basis compared to the standard nodal basis?

f) Experiment with other training sets and other numbers of basis functions and repeat exercise 1-3 (solve with python code).

What happens if there are no data points between 0.25 and 0.75?