Universität Stuttgart
Fachbereich Mathematik

Prof. I. Steinwart

Blatt 1 **Nichtparametrische Statistik** **02.11.2020**

Votieraufgabe 1. For $X := [0,1]^d$ and $d = 1,\ldots,5$ fix an arbitrary but somewhat interesting probability measure P of your choice, which is absolutely continuous with respect to the Lebesgue measure. For each of these five distributions, generate a data set of length $n = 10,000$ and save it to a file `<your-last-name>-<d>.train.csv` of the format

$$
\begin{array}{lll}
x_{1,1}, & \ldots, & x_{1,d} \\
\ldots & & \\
x_{n,1}, & \ldots, & x_{n,d}
\end{array}
$$

where between each numerical entry there is first a comma and then a `\space`. In addition, each line should end with `\n` (Linux end-of-line marker LF).

Furthermore, write an ASCII file `<your-last-name>-<d>.txt` containing a brief description of the chosen probability measure and its density.

Aufgabe 2. (schriftlich) Let $X = [0,1)^2$, $Y = \{-1,1\}$, $m \in \mathbb{N}$ and $b \in [0,0.5]$. Consider the distribution P on $X \times Y$, that is defined by the following conditions:

i) The marginal distribution P_X of P on X equals the uniform distribution on $[0,1)^2$.

ii) The conditional distribution at $x \in X$ is

$$
P(y=1|x) := \frac{1}{2} + \left(\frac{1}{2} - b\right) \sum_{i,j=0}^{m-1} (-1)^{i+j} \mathbf{1}_{[i/m,(i+1)/m)\times[j/m,(j+1)/m)}(x)\,.
$$

Write a program that generates a data set $D \sim P^n$ of length n. The result needs to be written to a file named as `chess-<m>-<b>-<n>.csv` . The format of the file should be

$$
\begin{array}{lll}
y_1, & x_{1,1}, & x_{1,2} \\
\ldots & & \\
y_n, & x_{n,1}, & x_{n,2}
\end{array}
$$

with the additional formatting requirements described in Exercise 1. Finally, create a graphical illustration of the data set for $m = 4$, $n = 1000$ und $b = 0.1$.

Votieraufgabe 3. Install the program `ggobi`, see `http://www.ggobi.org/` . Get acquainted with the program with the help of the data sets produced in Exercises 1 and 2, and be prepared to demonstrate it in the tutorial.

Aufgabe 4. (schriftlich) Find an implementation of the k-nearest neighbor classification rule in the internet and apply it to the data set constructed in Exercise 2. Illustrate the behavior for different k.