Non-Paramatric Statistics Exercise 4

Osman Ceylan, Jiahui Wang, Zhuoyao Zeng

27. November 2020

Exercise 2.3

Implement the histogram rule $h_{D,s}$ in an algorithm that only uses $O(n)$ spaces, where n is the number of samples. Visualize the effect of different widths on the data sets of Exercise 2.2.

Solution:

Here we describe our implementation:

Input is the data set D of n samples, a given point as origin for cell generation (x_0, y_0) and the width of cells s.

Our algorithm identifies each cubic cell A with its center c_A and uses a dictionary to store c_A as keys and the respective histogram values of each cell (as values of the dictionary). This data structure enables storage complexity to stay within $O(n)$.

Step 1: For each cell A, the algorithm calculates $|\{i \in \mathbb{N} : x_i \in A\}|$.

It means, that for each point $d \in D$ our algorithm determines $A(x)$ by a simple calculation and sees whether $c_{A(x)}$ is already a key in the dictionary. If $c_{A(x)}$ already exists in the dictionary, the value of $c_{A(x)}$ will increase 1; else the key $c_{A(x)}$ will be created and receives the value 1.

Step 2: For each key $c_{A(x)}$, the algorithm devides its value $|\{i \in \mathbb{N} : x_i \in A\}|$ by $n * s^2$, so that the histogram values are generated.

Step 3: The algorithm plots $c_{A(x)}$ as scatters and uses colours to represent different histogram values. The module *matplotlib.cm* is deployed for the colour scheme.

Now we present our graphical results for Exercise 2.2 i):

We draw 10,000 samples from the distribution $\mathbf{P}$:

$N(\left(\begin{smallmatrix} 0 \\ e \end{smallmatrix}\right), \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$