universität innsbruck
Department of Computer Science

Intelligent and Interactive Systems (IIS)	MACHINE LEARNING
Digital Science Center (DiSC)	Exercises 9: Nonparamteric density estimation
Assist. Prof. Matteo Saveriano	

Exercise 1

Given a set of 5 data points $\{x_1 = 2, x_2 = 2.5, x_3 = 3, x_4 = 1, x_5 = 6\}$, find the pdf estimate at $x = 3$ using the Gaussian kernel function with variance $\sigma = 1$ as a window function.

Solution Exercise 1

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n^D}\varphi(\frac{x - x_i}{h_n})$$

The previous equation equation suggests a general approach to estimate density by substituting the Parzen window function by another window function. This equation can be considered as the average of a set of window functions

If we consider a Gaussian window function:

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(x_i - x)^2}{2\sigma^2})$$

Then we have that:

x1: $\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(x^{(1)}-x)^2}{2\sigma^2}) = \frac{1}{\sqrt{2\pi}}exp(-\frac{(2-3)^2}{2}) = 0.2420$

x2: $\frac{1}{\sqrt{2\pi}}exp(-\frac{(2.5-3)^2}{2}) = 0.3521$

Similarly you can compute:

x3: ... $= 0.3989$
x4: ... $= 0.054$
x5: ... $= 0.0044$

Thus:

$$p_n(3) = \frac{0.2420 + 0.3521 + 0.3989 + 0.054 + 0.0044}{5} = 0.2103$$

Exercise 2

The following table shows 4 training samples from a survey. Two attributes (x_1, x_2) have been selected to classify data samples as good or bad.

x_1	x_2	$y(classification)$
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

An incoming sample is the $(x_1 = 3, x_2 = 7)$. Classify the sample by using K-nearest neighbor method. Use $K = 3$.

Solution Exercise 2

Step 1: Define number of neighbors. Suppose $K = 3$.

Step 2: Calculate distance between query pointy and training samples:

$(3, 7)$ from $(7, 7) : (7 - 3)^2 + (7 - 7)^2 = 16$
$(3, 7)$ from $(7, 4) : (7 - 3)^2 + (4 - 7)^2 = 25$
$(3, 7)$ from $(3, 4) : (3 - 3)^2 + (4 - 7)^2 = 9$
$(3, 7)$ from $(1, 4) : (1 - 3)^2 + (4 - 7)^2 = 13$

Step 3: Determine nearest neighbors based on k-th minimum distance.
The 3 nearest neighbors are $(3, 4), (1, 4), (7, 7)$. 2 Good , 1 Bad. Thus, query sample is classified as 'Good'.