In the Name of God

Statistical Pattern Recognition

Homework III

Non-Parametric Density Estimation

Assignment Date: **5 Azar** Submission Deadline: **14 Azar**

Contents

1 Objectives and Precautions

In this homework you will learn:

- How to implement non-parametric density estimation methods and how to visualize the resulted probability density functions.

Keep in mind that you can only use these three python libraries in your implementations:

- Pandas
- Numpy
- Matplotlib.pyplot

Also note that:

- You have 10 days to complete this homework.
- The home work instructions and datasets will be shared with you in your **Quera class**.
- Save your results and answers in any format you want, this will be your report.
- Save your code files and your report as a zip file and upload in Quera. (naming format is "your names.zip" for example "Achraf_Hakimi_Mohamed_Salah.zip")
- Late submission strategy is: 70 percent score for one day delay and 50 percent for 2 days delay. Submissions after 2 days will not be graded.
- Use only the python programming language.
- Feel free to ask your questions in the telegram channel.
- Do not copy other works, write your own code.

Thank you. Good Luck.

2 Understanding the data

You have 2 datasets for this homework:

2.1 1D_grades.csv

A 1-dimensional dataset which actually contains your midterm grades.

2.2 2D_synthetic_gaussians.csv

A 2-dimensional dataset which was synthetically generated from two Gaussian distributions.

There is not much of data preprocessing for this homework rather than:

1. Plot the datasets.
2. Normalize both datasets to have zero mean and unit variance.

3 Non-Parametric Density Estimation

Non-parametric density estimation is a statistical technique used to estimate the probability density function (PDF) of a random variable without assuming a specific parametric form for the underlying distribution.
There are several methods for non-parametric density estimation, and your task here is to implement 4 common ones on 2 dataset. Therefore please, follow the instructions.

3.1 Histograms

The simplest form of non-parametric density estimation involves dividing the range of the data into bins and counting the number of data points in each bin. The height of each bin then represents an estimate of the density. The resulted density function here is then referred to as histograms. Therefore:

- Implement histogram-based PDF estimation on both your datasets from scratch.
- Plot the resulted PDFs (refer to figures 1 and 2 for clarification).
- Do this for 4 different bandwidths of (0.1, 0.3, 0.5, 0.7).
- You can use equation 2.8 of your source book for more implementation details.

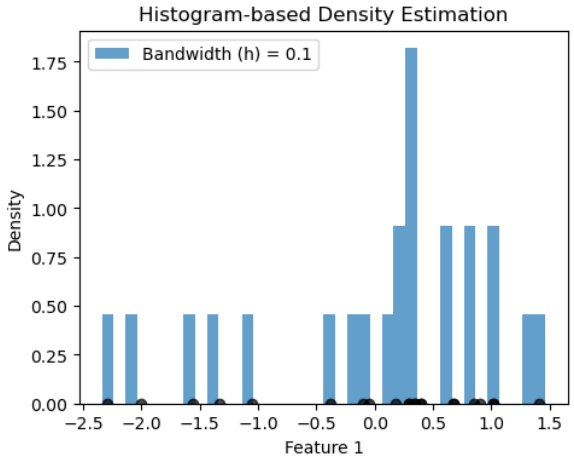

Figure 1: Plot the PDFs of your 1-D dataset in the 2D space along with the samples themselves. Here we provided a histogram of your 1-D dataset with bandwidth of 0.1 as an example

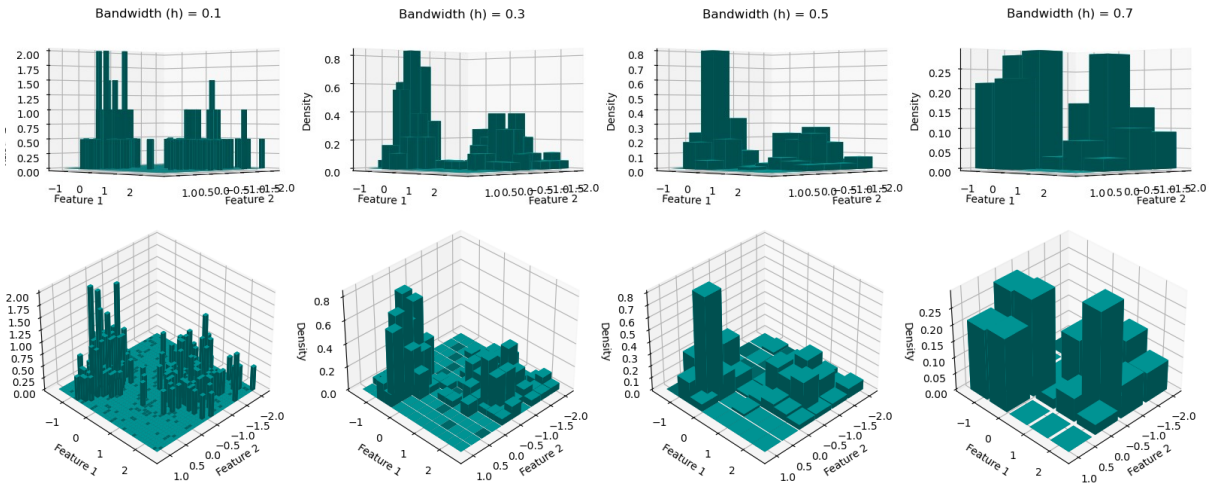

Figure 2: Plot the PDFs of your 2-D dataset from different viewpoints in the 3D space. Here we provided histograms as an example.

3.2 Kernels

KDE is a more sophisticated non-parametric method that involves placing a kernel at each data point and summing up these kernels to estimate the PDF. You are going to implement two kernels in this homework, Parzen window and the Gaussina function. Therefore:

3.2.1 Parzen Window as a Kernel

- Implement kernel density estimation using the Parzen window as the kernel, on both your datasets from scratch.
- Plot the resulted PDFs similar to the plots in figures 1 and 2.
- Do this for 4 different bandwidths of (0.1, 0.3, 0.5, 0.7).
- You can use equations 2.10, 2.11 and 2.12 of your source book for more implementation details.

3.2.2 Gaussian Function as a Kernel

- Implement KDE using the Gaussian kernel with zero mean and unit variance, on both your datasets from scratch.
- Plot the resulted PDFs similar to the plots in figures 1 and 2.
- Do this for 4 different bandwidths of (0.1, 0.3, 0.5, 0.7).
- You can use equations 2.13 and 2.14 of your source book for more implementation details.

3.3 K-Nearest Neighbours

KNN density estimation involves estimating the density at a point by considering the distances to its k nearest neighbors. The density is proportional to the inverse of the average distance to the k neighbors. In this part we are interested to use KNN as a density estimator. Therefore:

- implement KNN as a density estimator, on both your datasets from scratch.
- Plot the resulted PDFs similar to the plots in figures 1 and 2.
- Do this for 4 different k values. Use k=(1, 3, 5, 7) for your 1-D dataset, and k=(5, 50, 100, 200) for the 2-D one.
- You can use equations 2.24 and 2.25 of your source book for more implementation details.