8STT108 – Big Data Statistics Analytical Tools

Lab #5

Kernel Density Estimation

Fall 2024 **Professor: Sara Séguin**

Python is required for the lab assignment. Use any relevant libraries such as: pandas, numpy, statistics, matplotlib.

The goal of the lab is for the students to experiment with the libraries, therefore, you should read any required documentation to understand the tools.

1. Implement the KDE algorithm from scratch. Follow these steps:
   - **Step 1:** Generate a synthetic dataset. For example, generate 100 random data points from a normal distribution with mean = 0 and standard deviation = 1.
   - **Step 2:** Define and implement a Gaussian kernel function. The Gaussian kernel formula is given by:

     $$K(x) = \frac{1}{\sqrt{(2\pi)}} e^{-\frac{x^2}{2}}$$

   - **Step 3:** Write a function to compute the KDE given a dataset, a kernel function, and a bandwidth. The KDE function is defined as:

     $$\hat{f}(x) = \frac{1}{nh} \sum_{i=1} K\left(\frac{x - x_i}{h}\right)$$

     where:

     - $\hat{f}(x)$is the estimated density at point x
     - n is the number of data points
     - h is the bandwidth
     - K is the kernel function

   - **Step 4:** Plot the estimated density using Matplotlib and compare it with a histogram of the data.

2. **Part 2: Exploring Different Kernel Functions**
   Implement the following kernel functions:
   - **Epanechnikov Kernel:**

$$K(x) = \frac{3}{4}(1 - x^2) for \ \mid x \mid \leq 1$$

   - **Uniform Kernel:**

$$K(x) = \frac{1}{2} for \ \mid x \mid \leq 1$$

   - **Triangular Kernel:**

$$K(x) = 1 - \mid x \mid for \ \mid x \mid \leq 1$$

   Apply each of these kernels to the synthetic dataset and plot the estimated densities. Discuss the differences in shape and smoothness between the kernels.

3. **Part 3: The Effect of Bandwidth on KDE**
   - **Step 1:** Fix the kernel function (e.g., Gaussian) and compute the KDE for different bandwidth values (e.g., 0.1, 0.5, 1, 2).
   - **Step 2:** Plot the resulting KDEs on the same graph and analyze how the choice of bandwidth affects the smoothness of the density estimate.
   - **Step 3:** Write a short analysis of the results, discussing the trade-off between bias and variance in KDE when choosing different bandwidths.

4. **Part 4: Application to a Real-World Dataset**
   - Download a real-world dataset (e.g., a dataset with a continuous variable like the “sepal width” from the Iris dataset).
   - Apply KDE using a suitable kernel and bandwidth to estimate the probability density function of the chosen variable.
   - Visualize and interpret the results. Discuss any insights you gained from the KDE plot.