Statistical Pattern Recognition
Assignment 4

Ali Gholami

Department of Computer Engineering & Information Technology
Amirkabir University of Technology

https://aligholamee.github.io
aligholami7596@gmail.com

Abstract

Keywords. *KNN Classifier, Kernel Density Estimation.*

1 Kernel Density Estimation & KNN Estimation

Given dataset $X = 2, 3, 4, 4, 4, 5, 5, 10, 11, 11, 11, 12, 14, 16$ use Parzen windows to estimate density $p(x)$ at $x = 5$ and $x = 12$; assuming $h = 4$ in the following conditions.

(a) If you use standard kernel function

$$K(u) = \begin{cases} 1 & |u| \leq \frac{1}{2} \\ 0 & o.w. \end{cases}$$

(b) If you use Gaussian Kernel, $N(0, 10)$.

(c) For the same dataset and the same sample points, i.e. $x = 5$ and $x = 12$, estimate the density using KNN approach. Take $K = 5$ in your estimations.

Solution

(a) We'll use (1.1) to estimate the density using different kernels.

$$\hat{p}_{(x)} = \frac{1}{n * h^d} \sum_{i=1}^{k} \Phi(\frac{x - x_i}{h}) \tag{1.1}$$

In this case, we'll compute the distance of each of the given points from our dataset. For every distance less than 2 we'll consider the effect of that point in our estimation.

$$x = 5 \quad \rightarrow \quad \hat{p}_{(x)} = \frac{1}{14 * 4} * (6) = \frac{3}{28} = 0.107$$

Doing the same for the point $x = 12$ and we'll have the following results:

$$x = 12 \quad \rightarrow \quad \hat{p}_{(x)} = \frac{1}{14 * 4} * (6) = \frac{3}{28} = 0.107$$

(b) In this case, the value of items in the $\sum$ are no longer 1. They will have different outputs, specially on the center of the curve. This can lead to an smoother and more realistic estimation of the density. The Gaussian Kernel equation is given in (1.2).

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-u^2}{2}) \tag{1.2}$$

$$x = 5 \quad \rightarrow \quad \hat{p}_{(x)} = \frac{1}{14 * 4} * \frac{1}{\sqrt{2\pi}}(\exp(-2) + 3\exp(\frac{-1}{2}) + 2)) = 0.028$$

$$x = 12 \quad \rightarrow \quad \hat{p}_{(x)} = \frac{1}{14 * 4} * \frac{1}{\sqrt{2\pi}}(2\exp(-2) + 3\exp(\frac{-1}{2}) + 1)) = 0.021$$

(c) In this case, we have to consider the window size as a dynamic variable. In case $h = 2$ then $x = 4$, $x = 4$, $x = 4$ and $x = 5$ will be in our window of estimation as well as the centered point $x = 5$ (K = 5). To estimate the density we'll have the following equation:

$$x = 5 \quad \rightarrow \quad \hat{p}_{(x)} = \frac{1}{14 * 2} * (5) = 0.178$$

Note that we have considered the kernel to be the same as part a. In the second case, choosing the $h = 2$ yields 5 points.

$$x = 12 \quad \rightarrow \quad \hat{p}_{(x)} = \frac{1}{14 * 2} * (5) = 0.178$$

2 Parzen Windows & KNN Estimation

Consider the following training set drawn from an unknown density $f(x)$:

$$X = 0.01, 0.12, 0.19, 0.32, 0.41, 0.48$$

(a) Let $\phi(x) = N(1, 0)$. Find and sketch the Parzen Windows estimate for the values of h_n of 0.1 and 1.0.

$$\hat{f}_{n(x)} = \frac{1}{n * h_n} \sum_{i=1}^{k} \Phi(\frac{x - x_i}{h_n}) \tag{2.1}$$

(b) Find and sketch the 3-nearest neighbor estimate of $f(x)$.

Solution

(a) For each point in our dataset, we'll center the Gaussian Kernel and add the values of these kernels. Here are the computations in case that $h_n = 0.1$.

$$\hat{f}_{n(x=0.01)} = \frac{1}{6 * 0.1} \sum_{i=1}^{k} \Phi(0) = 0$$

$$\hat{f}_{n(x=0.12)} = \frac{1}{6*0.1}\sum_{i=1}^{k}\Phi(0) = 0$$

$$\hat{f}_{n(x=0.19)} = \frac{1}{6*0.1}\sum_{i=1}^{k}\Phi(0) = 0$$

$$\hat{f}_{n(x=0.32)} = \frac{1}{6*0.1}\sum_{i=1}^{k}\Phi(0) = 0$$

$$\hat{f}_{n(x=0.41)} = \frac{1}{6*0.1}\sum_{i=1}^{k}\Phi(0) = 0$$

$$\hat{f}_{n(x=0.48)} = \frac{1}{6*0.1}\sum_{i=1}^{k}\Phi(0) = 0$$

As denoted above, estimation will give 0 for all of the points in the dataset with $h_n = 0.1$. We can understand that, the window size has been chosen too small that no point is being considered by the estimator.

(b)