Problem 3: Parametric and Non-Parametric Methods

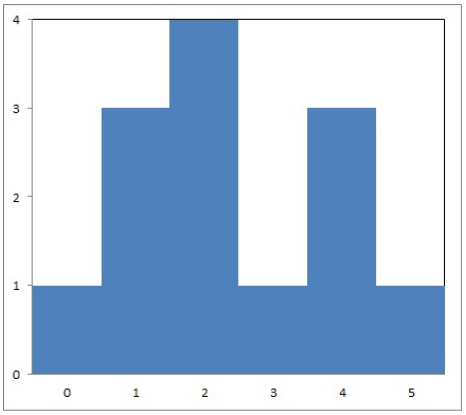

Figure 1: Histogram

1. The histogram is shown in Figure 1.
2. The kernel density function is defined as

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})$$

3. Since bandwidth is 2, $V_n = h^d = h^1 = 2$, the kernel function will be

$$K(x-x_i) = (1-|\frac{x-x_i}{h}|)\delta(|x-x_i| \leq 1)$$

Thus the estimated density for a given x will be

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{x-x_i}{h_n}) = \frac{1}{13}\sum_{i=1}^{n}\frac{1}{2}(1-|\frac{x-x_i}{h}|)\delta(|x-x_i| \leq 1)$$

with $x_0, x_1, ..., x_n = D = 0, 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 4, 5$.
Thus, we can get

$$p_n(0) = \frac{5}{52} \approx 0.096; p_n(1) = \frac{11}{52} \approx 0.21; p_n(2) = \frac{12}{52} \approx 0.23; p_n(3) = \frac{9}{52} \approx 0.17; p_n(4) = \frac{8}{52} \approx 0.15; p_n(5) = \frac{}{}$$

4. Parzen window specifies the size of the windows as some function of n such as $V_n = 1/\sqrt{n}$, while k-nearest-neighbor specifies the number of samples k_n as some function of n such as $k_n = \sqrt{n}$. Both of them converge to $p(x)$ as $n \to +\infty$. For parzen window method, the choice of V_n has an important effect on the estimated $p_n(x)$: if V_n is too small, the estimation will depends mostly on closer points and will have too much variability based

on a limited number of training samples (over-training); if V_n is too large, the estimation will be an average over a large range of nearby samples, and will loss some details of $p(x)$. By specifying the number of samples, kNN methods circumvent this problem by making the window size a function of the actual training data. If the density is higher around a particular x, the corresponding V_n will be smaller. This means if we have more samples around x, we will use smaller window size to capture more details around x. If the density is lower around x, the corresponding V_n will be larger, which means less samples can only give us an estimation of a larger scale and cannot recover many details.

5. If assume the density is a Gaussian, the maximum likelihood estimate of the Gaussian parameters μ, σ^2 is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx 2.38$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \approx 2.08$$

The unbiased result is

$$\hat{\sigma}^2_{unbias} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \approx 2.26$$

6. Histogram captures the density in a discrete way and can have large errors around boundaries of bins. The triangle-kernel better captured the density with a smoother representation. Although Gaussian estimation is also smooth, it cannot capture the data samples very well. That's because it assumed Gaussian distribution of the samples, while the given sample data doesn't fit into a Gaussian distribution. The triangular-kernel best captured the data and I would choose this one to estimate the distribution of this particular data set.