

768 Assignment 2*

Your Name and Student ID

2024-10-07

Instructions: *You are to use Quarto Markdown for generating your assignment output file. You begin with the Quarto Markdown script downloaded from A2L, and need to pay attention to information provided via introductory material posted to A2L on working with R, and Quarto Markdown. Having downloaded and added your answers to the Quarto Markdown script, you then are to generate your output file using the “Render” button in the RStudio IDE and, when complete, upload both your Quarto Markdown file and your PDF file to the appropriate folder on A2L.*

Continuous Density and Cumulative Distribution Functions Questions

1. When $X \in \mathcal{D}$, we defined the sample proportion as $p_n(x) = n^{-1} \sum \mathbf{1}(X_i = x)$, and when $X \in \mathbb{R}$, we defined the histogram as $f_H(x) = (nh)^{-1} \sum \mathbf{1}(X_i \text{ is in the same bin as } x)$. The former is a ratio of the sum of an indicator function to the sample size n , while the latter is a ratio of the sum of an indicator function to nh , where h is the bin width. Demonstrate that each is a proper estimate of a probability mass function and a probability density function, respectively.
2. Demonstrate that the naïve density estimator $f_N(x)$ is proper.
3. Suppose that someone who is presented with an i.i.d. sample $\{X_1, X_2, \dots, X_n\}$ would like to estimate the arithmetic mean of a continuous random variable via

$$\hat{\mu} = \int_{-\infty}^{\infty} x \hat{f}(x) dx,$$

where $\hat{f}(x)$ is the Rosenblatt-Parzen kernel density estimator (you may assume that a second-order symmetric kernel function is being used).

*Keiths-MacBook-Pro-2.local, x86_64, root

- i. Simplify this estimator via change of variable arguments (hint - since the kernel function is symmetric, you can interchange $K((x - X_i)/h)$ and $K((X_i - x)/h)$).
 - ii. Derive the bias of the simplified estimator.
 - iii. Derive the variance of the simplified estimator.
 - iv. What is the rate of convergence of the simplified estimator?
4. The expectation of a continuous random variable is defined as $\mu = \int_{-\infty}^{\infty} xf(x) dx$, while the variance of a continuous random variable is defined as $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$. For $X \in \mathbb{R}^1$, simplify the expression for σ^2 when $f(x)$ is replaced by the univariate Rosenblatt-Parzen kernel density estimator that uses a symmetric second-order kernel function. In particular, determine whether or not this estimator of σ^2 coincides with the maximum likelihood estimator of the variance, which you recall is given by

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \hat{\mu})^2 = n^{-1} \sum_{i=1}^n X_i^2 - \left(n^{-1} \sum_{i=1}^n X_i \right)^2,$$

where $\hat{\mu} = n^{-1} \sum_{i=1}^n X_i$.

5. Consider the Rosenblatt-Parzen kernel estimator of a univariate density function, which is given by

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

where the X_i are i.i.d.

- i. Define a *higher-order kernel function* (e.g., a kernel function with the property that the order ν of its first nonzero moment is greater than 2).
- ii. What is the purpose of using a higher-order kernel function?
- iii. How does using a higher-order kernel function affect the variance?
- iv. How does using a higher-order kernel function affect the bias?
- v. Derive the IMSE and the optimal bandwidth for an eighth-order ($\nu = 8$) mean-zero symmetric kernel function.
- vi. Can you generalize your IMSE result for the eighth-order kernel function so that it applies to an arbitrary ν th-order kernel function, where $\nu > 1$ is even. That is, can you express the IMSE and the resulting optimal bandwidth in terms of an even ν ?
- vii. Obtain the pointwise rate of convergence of the kernel estimator $\hat{f}(x)$ that uses a ν th-order kernel function (hint - use your optimal bandwidth in the formula for the root MSE and observe the power to which n is raised in your solution).

6. Consider a mean zero, symmetric, fourth-order kernel function written in polynomial form as

$$K(u) = (a + bu^2)\Phi(u),$$

where a and b are two constants that satisfy the requirements of a fourth-order kernel function (since the kernel function is symmetric, the only two restrictions that we need to consider are $\int K(u) du = 1$ and $\int u^2 K(u) du = 0$). Let $\Phi(u)$ be the standard normal kernel function given by $\Phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$. Hence $\int \Phi(u) du = 1$, $\int u\Phi(u) du = 0$, $\int u^2\Phi(u) du = 1$, $\int u^3\Phi(u) du = 0$, $\int u^4\Phi(u) du = 3$, $\int u^5\Phi(u) du = 0$, $\int u^6\Phi(u) du = 15$, and so on. Demonstrate that this fourth-order kernel function has a solution that is given by

$$K(u) = \left(\frac{3}{2} - \frac{1}{2}u^2\right)\Phi(u).$$

What is $\kappa_4 = \int u^4 K(u) du$ for this kernel function?