Density Estimation & Estimator Bias

Exercise 10.1: Kernel Density Estimation & Validation (7 points)

Data: Each pixel inside the image `testimg.jpg` contains an 8-bit grayscale intensity value which can be stored using a type such as `unsigned char` (or `uint8`, or similar). Therefore, $X \in [0, 255]^{m \times n}$, where X is the image matrix and m, n are the height and width of that image, respectively. Treat the pixel values as *independent and identically distributed* (i.i.d.) samples drawn from the same unknown distribution P. This assumes that there are no statistical dependencies between neighboring elements in the matrix. This independence assumption is actually not valid when it comes to neighboring image pixels but we will ignore this throughout this problem sheet.

1. Load the data into a vector and scale the values to the [0,1] range. The location of the pixels are no longer relevant, only the pixel values.
2. Create *two new datasets* by adding to the data Gaussian noise with zero mean and standard deviation $\sigma_1 = 0.05$ for the first Gaussian and $\sigma_2 = 0.1$ for the second Gaussian.
3. Create a figure that combines the histograms of all 3 datasets (i.e. the original scaled to [0,1] without any noise + the 2 sets of the data with the added noise). Use enough bins that gives you a high enough resolution to differentiate between the datasets. Use the same bins for all 3 datasets.
4. Create a figure that shows the empirical cumulative distribution functions (ECDF) of the 3 datasets in one plot. Visualizing the ECDF may provide a less cluttered comparison between the distriubtions that may be clearer than the histograms overlayed in the same figure.
5. For each dataset:
   a) Data split:
   Take a **random** subset of $p = 100$ observations to form a "training-set". **Randomly** select another 5000 points to form a "test-set". Ensure that the two sets are disjoint (i.e. no point is shared between training and test sets).
   b) KDE:
   Estimate the probability density $\widehat{P}(x; h)$ of intensities using a *rectangular*[1] kernel ("gliding window"/"top-hat") that is parametrized by the window width h.
   c) Visualization:
   Plot $\widehat{P}$ resulting from (e.g. 10) different random training-sets[2] with the same size p. Overlay them onto the same figure.

[1] You will be asked to repeat everything using another kernel. Just a heads up to modularize your implementation accordingly.
[2] You do not have to explicitly ensure/enforce that the different training sets are disjoint.

d) Validation:
Calculate the negative log-likelihood per datapoint of your estimator using only the 5000 points from the test-set which were **not** used in computing $\widehat{P}$. Average the negative log-likelihood over the 5000 test points.

e) Average negative log-likelihood vs. h:
Repeat the above steps (specifically, estimating $\widehat{P}(x;h)$ and computing the average negative log-likelihood for the test points) for a set of kernel widths h. From what range should you select the values of h? This should give you for a fixed data set size p, the relation between kernel-width and mean validated likelihood. You do not need to plot $\widehat{P}$ for all the new h values anymore.

6. Analysis:
   (a) Plot mean validated likelihoods (y-axis) vs. kernel width h (x-axis) for all datasets in the same figure (i.e. one line for each dataset).
   (b) Repeat the above to show mean validated likelihoods vs. kernel width h for training-sets of size $p = 500$.
   (c) Repeat the previous two steps (a) & (b) for the Gaussian kernel with $\sigma = h$.

Interpretation: Which kernel width h yields the minimal validated negative log-likelihood hence the minimal generalization error, for each combination? How do you interpret this result?

Exercise 10.2: Maximum Likelihood (3 points)

Suppose we are given a data set $x^{(1)}, \ldots, x^{(p)}$ representing p *i.i.d.* observations of the scalar random variable x, which follows a Gaussian distribution:

$$P(x) \sim \mathcal{N}(\mu, \sigma^2)$$

(a) Find the *maximum likelihood* estimates $\widehat{\mu}_{ML}$ and $\widehat{\sigma}^2_{ML}$ of the distribution parameters.

(b) Show that $\widehat{\mu}_{ML}$ is *unbiased* and $\widehat{\sigma}^2_{ML}$ is *biased*.

(c) Replace $\widehat{\mu}_{ML}$ with the true value μ^* in your expression of the maximum likelihood estimator $\widehat{\sigma}^2_{ML}$ and verify that the resulting estimator $\widehat{\sigma}^2_{ML}$ becomes *unbiased*.

Total 10 points.