Exercises for
Pattern Analysis Programming
Lina Felsner, Mathias Seuret, Dalia Rodriguez-Salas
Task 2: April 29 – May 6

Pattern Recognition Lab

Density Estimation

Our First Empirical Distribution: Sampled from a Non-parametric Ground Truth Distribution

A central challenge in statistical machine learning is the discrepancy between a true, ideal data distribution and the actual, observed distribution. The purpose of statistical machine learning is to deduce the ideal data distribution from the observed distribution. We will write some first code here to better understand the relationship between both distributions.

In this task, we create a ground truth distribution, and draw an empirical distribution from it. We can "misuse" a gray-scale image for this task, namely if we imagine that the x- and y-coordinates of a pixel are feature dimensions x_1 and x_2, and the grayscale intensity represents the relative density at its respective (x, y)-location.

Exercise 1

Use `scipy.misc.face(gray=true)` to load a gray-scale image of a raccoon face. Apply a Gaussian filter (e.g., with $\sigma = 3$) to slightly smooth the image. Use `matplotlib` to visualize the image.

Implement a function that draws new samples from this density, using the approach with the cumulative density function from the lecture. Maybe you can re-use some code from the warmup exercise? To visualize the drawn samples, it is most convenient to create a new (empty) image with the same dimensions as the raccoon, and to draw a point at every location that is sampled from the density. An example is shown in Fig. 1.

Draw 10.000, 50.000, 100.000, 200.000, and 400.000 samples. What do you observe?

Exercise 2

Use the Parzen Window Estimator to "reconstruct" the image. Thus, use the sampled density from the previous task as input, and output a density from the Parzen window estimator.

More in detail, Implement a Parzen window estimator with a box kernel. Use our empirical raccoon densities from the previous task as an input. Vary the window size of the Parzen estimator. What do you observe?

A First Taste of the Model Selection Problem

Which kernel size is best suited for our raccoon density? Implement a cross-validation to automatically determine the best kernel size from a reasonably large list of candidate kernel sizes (with at least 3 candidate sizes). The objective function shall be chosen analogously to the lecture.

If you are familiar with cross-validation, then please just go ahead and implement it. Otherwise, you may follow these steps: to perform k-fold cross-validation, split the samples $\mathcal{S}$ into a test set $\mathcal{S}_j$ of size $|\mathcal{S}_j| = N/k$ and a training set $\mathcal{T}_j = \mathcal{S}$

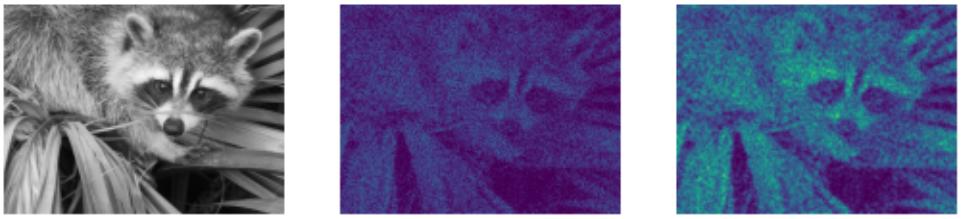

Figure 1: Ground truth distribution (raccoon) as well as sampled density (100000 samples) and a reconstruction with a Parzen Window kernel of width 9.

$\mathcal{S}_j$. Build a density $p(\mathbf{x})_\theta^j$ from $\mathcal{T}_j$ where θ denotes the candidate kernel size. Choose the best θ with an ML estimation across all folds, by choosing $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum \log p(\mathbf{x})_\theta^j$ on all $\mathbf{x} \in \mathcal{T}_j$ across all folds j.

Play with it! Which observations are new or surprising to you? In other words, what do you experience in the experiments that was not obvious to you from the theoretical derivation of the model selection problem?

Please create a figure with the raccoon or any other image of your choice as well as a sampling and a reconstruction with the best kernel size (example in Fig. 1). Please post it to the forum and add a short text about one observation you made.

Comments:
We ask for only one figure per group. Please also state your group number. Bring your code to the joint meeting on Mai 6 or Mai 7 for a potential little extra experiment.