Exercises for
Pattern Analysis Programming
Lina Felsner, Mathias Seuret, Dalia Rodriguez-Salas
Task 3: May 6 – May 13

Pattern Recognition Lab

Dealing with Highly Imbalanced Sample Distributions

In the previous task, we interpreted the intensities in our raccoon picture as relative frequencies in a raccoon PDF. However, we could also interpret these intensities as function values, and then aim to approximate these values via regression. This can also be done with a kernel, as shown, in Sec. 6.1 of our textbook by Hastie, Tibshirani, and Friedman.

The main advantage of regression via kernel smoothing is its simplicity. On the downside, this simplicity can be the source of systematic errors in the regression result.

The distribution of samples is always imbalanced at the boundary of the domain: here, samples can only be inside of the domain, but not outside. Thus, to regress a value x_0 close to the boundary, the kernel can only make use of samples on one side of the kernel window. Similar situations can also arise within the domain whenever the samples are very unevenly distributed within a local neighborhood.

In this exercise, we will investigate the special case at the boundary of the domain. This case is illustrated in Fig. 6.3 and Fig. 6.4 of Hastie, Tibshirani, Friedman (→ pages 195 and 196). The goal of this exercise is to reproduce this experiment, and to obtain a Figure that is similar to Fig. 6.3 and the left part of Fig. 6.4. An example how this could look link is shown in Fig. 1.

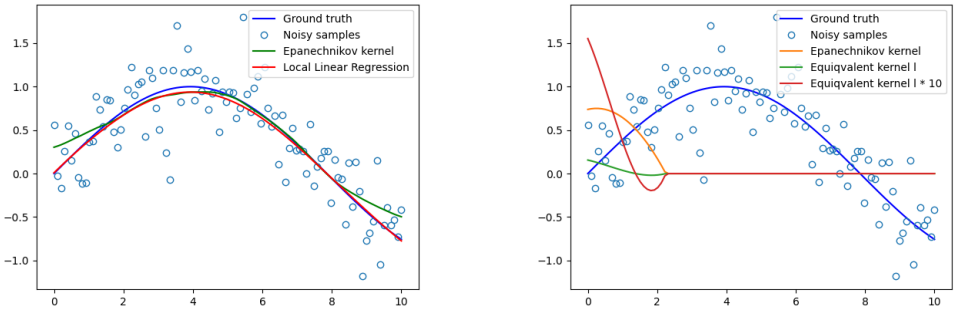

Figure 1: Left: Ground truth and noisy samples together with the regression with an Epanechnikov kernel and a a kernel-weighted local linear regression. Right: Kernels plotted for $x_0 = 0.25$. Note that the local linear equivalent kernel may become negative.

Exercise 1 Create the dataset. First, you will need a ground truth function (dark blue in Fig. 6.3). In our implementation, clamping the domain of $y = \sin(0.4x)$ looked qualitatively quite similar, but feel free to choose your own function if you like.

Draw noisy random samples along that function. In our implementation, we use additive Gaussian noise, which is a typical simple baseline.

Exercise 2 Perform kernel smoothing regression on the noisy samples to obtain the equivalent to the green curve in the left plot of Fig. 6.3. Use an Epanechnikov kernel, like Tibshirani, Hastie, Friedman, with a sufficiently large window size. Can you observe a similar error at the boundary? If not, wiggle on your function until a similar error occurs.

Exercise 3 Perform a kernel-weighted local linear regression, as described in Eqn. (6.8) and Eqn. (6.9), and plot this estimate. Hopefully, the boundary region is better extrapolated now? How well are the other areas of your function represented? Please also plot the Epanechnikov kernel and the local linear equivalent kernel (shown in Fig. 6.4). Do you observe a similar difference in both kernels as shown in the book?

Please note: Equations (6.8) and (6.9) are directly derived from a standard linear regression task. We find this derivation instructive, and we will tomorrow upload a short video on it to sec. A in studOn.

Comments:
We ask for only one figure per group. Please also state your group number. Bring your code to the joint meeting on May 13 or May 14 for a potential little extra experiment.