Machine Learning and Physics
Winter Semester 2024/2025

Exercise Sheet #2

Due: 04.11.24 14h00

Peter Lippmann

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group).
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes please make sure that they are legible and not too blurry.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at `https://github.com/sciai-lab/mlph_w24`. Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter. Please hand in your notebook (`.ipynb`), as well as an exported pdf-version of it.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of two. Specify all names of your group in the submission.

1 Kernel Density Estimation

(a) Implement a Quartic (biweight) kernel

$$k\left(x-\mu; w\right) = \frac{15}{16w}\left(1-\left(\frac{x-\mu}{w}\right)^2\right)^2 \quad \text{with support } x \in [\mu - w, \mu + w]$$

and plot it for $\mu = 0$ and $w = 1$ over the range $[-1, 1]$. (2 pts)

(b) Take the first $N = 50$ data points from `samples.npy`, compute and plot the kernel density estimate over the range $[-10, 20]$ for a set of different bandwidths (e.g. $w \in \{0.1, 0.5, 1, 3, 5\}$). Discuss the results and the influence of the bandwidth. Which bandwidth is optimal in your opinion? Explore what happens as you increase the number of samples N. (5 pts)

2 Bonus: Average shifted Histograms and KDE

Average shifted histograms do what their name implies: they compute a number h of histograms with random offsets and average their results.

Prove that, for $h \to \infty$, average shifted histograms converge to a kernel density estimate. What does the shape of the kernel look like for

(a) 1D histograms with uniform bin width (2 pts)

(b) 2D histograms with axis-aligned rectangular bins (1 pt)

(c) 2D histograms made from any regular tiling (covering of the plane using a single shape, without gaps or overlaps, and without rotating the shape) (1 pt)

3 Mean-Shift

(a) Gradient ascent on the KDE with the Epanechnikov kernel corresponds to the update step

$$x_j^{t+1} = x_j^t + \alpha_j^t \frac{2}{n} \sum_{i:||x_i - x_j^t||<1} (x_i - x_j^t)$$

For which choice of the adaptive learning rate α_j^t is this equivalent of updates to the local mean? Why is this a sensible choice of learning rate? (3 pts)

(b) Implement the updates to the local mean in python. Apply your implementation to the 1D dataset from exercise 1 and visualize how the points move over time, by plotting a line of x over t for every data point. Repeat the same for "blurring" meanshift, where the current x_i^t are used instead of the x_i in the computation of the local mean. How does the convergence compare between the variants? Are the resulting clusters the same? (5 pts)

4 Bonus: On KDE Bandwidth and Modes

The RBF ("Radial Basis Function") kernel is defined by

$$k(x; w) = \frac{1}{w\sqrt{2\pi}} \exp\left(-\frac{\|x\|^2}{2w^2}\right).$$

Disprove by counterexample or otherwise the following false statement:
For every set of points $x_i \in \mathbb{R}^n, i = 1, \ldots, N$, the number of modes of their KDE with the RBF kernel decreases monotonously as the bandwidth w increases. (5 pts)

5 Linear Regression: Heteroscedastic Noise

The standard formulation of linear regression is of homoscedastic noise, i.e. the variances of the observation noise is independent of $\mathbf{x}$. A generalization is to have a data point dependent variance on the observation noise, i.e. we have

$$y_n = \boldsymbol{\beta}^T \mathbf{x}_n + \varepsilon_n, \tag{1}$$

with $\mathbb{E}[\varepsilon_n] = 0$ and known $\mathrm{var}[\varepsilon_n] = \sigma_n^2$, so called *heteroscedastic noise*. Give the sum-of-squares problem in that case, interpret it and derive mean and covariance structure of the $\hat{\boldsymbol{\beta}}$. (4 pts)