CSE176 Introduction to Machine Learning
Fall semester 2025

Lab: nonparametric methods
Miguel Á. Carreira-Perpiñán

The objective of this lab is for you to explore the behavior of several nonparametric methods for density estimation, classification and regression in Matlab and apply them to some datasets. The TA will first demonstrate the results of the algorithms on a toy dataset. Then, you will replicate those results, and further explore other datasets.

We provide you with the following:

- The scripts `lab05_knn2.m` and `lab05_kde1.m` set up the problem: a k-nearest-neighbor classifier for 2D data and a Gaussian KDE for 1D data, respectively.
- `knn.m` (k-nearest-neighbor classifier) and various functions from the GM tools (Gaussian kernel density estimates).

I Datasets

Construct your own toy datasets to visualize the result easily. Take the input instances $\{\mathbf{x}_n\}_{n=1}^N$ in $\mathbb{R}$ and the labels $\{y_n\}_{n=1}^N$ in $\{1,\dots,K\}$ (classification) or $\mathbb{R}$ (regression). You can also take $\mathbf{x} \in \mathbb{R}^2$ and use surface or contour plots.

II Using nonparametric methods

k-nearest-neighbor (KNN) classifier In file `lab03_knn2.m` we apply the KNN classifier to datasets in 2D and plot its results. Explore its behavior in various settings (see suggestions at the end of the file). Note: this is the k-nearest-neighbor *classifier*, not the k-nearest-neighbor *density estimate*.

Kernel density estimate (KDE) In file `lab03_kde1.m` we have implemented the following methods for datasets in 1D (i.e., feature vectors $x_n \in \mathbb{R}$):

- *Histogram* with origin $x_0 \in \mathbb{R}$ and bin width $h > 0$, for density estimation. We plot the resulting density estimate $p(x)$ using a bar chart.
- *Gaussian kernel density estimate* with bin width $h > 0$:

$$p(\mathbf{x}) = \frac{1}{Nh^D} \sum_{n=1}^N K\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \qquad \text{Gaussian kernel: } K\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) = (2\pi)^{-D/2} e^{-\frac{1}{2}\left\|\frac{\mathbf{x}-\mathbf{x}_n}{h}\right\|^2},\ \mathbf{x} \in \mathbb{R}^D. \tag{1}$$

  We plot:

  - Density estimation: the resulting density estimate $p(x)$ as a continuous curve in $\mathbb{R}$.
  - Classification: the resulting posterior distribution estimate $p(k|x)$ for each class $k = 1,\dots,K$ as a continuous curve in $\mathbb{R}$, using a different color for each class; and the data points colored according to the predicted label $\arg\max_{k=1,\dots,K} p(k|x)$.
  - Regression: the resulting regression function $g(x)$ as a continuous curve in $\mathbb{R}$.

Using this code, explore histograms and KDEs in various settings (see suggestions at the end of file `lab03_kde1.m`). Consider the following questions:

- How does the histogram change if you change x_0? How does it change if you change h?
- How does the result change if you change h? How does the estimated density $p(x)$ and the regression function $g(x)$ behave for $h \to 0$ and for $h \to \infty$?
- How well does the estimated density $p(x)$ or regression function $g(x)$ approximate the true one?
- How does the regression function $g(x)$ behave near discontinuities in the true function $f(x)$, or in regions $x \in \mathbb{R}$ that have no data points?
- For classification and for density estimation, how do Gaussian KDEs compare with Gaussian classifiers?

Further things to do:

- Extend the code to work with 2D datasets. Use the plots in `lab02.m` as a guideline.
- Extend the code to work with 1D datasets but use kernels other than the Gaussian, specifically use the following two kernels in eq. (1):

$$\text{Uniform: } K\left(\frac{x-x_n}{h}\right) = \begin{cases} \frac{1}{2}, & \left|\frac{x-x_n}{h}\right| \leq 1 \\ 0, & \text{otherwise} \end{cases} \qquad \text{Epanechnikov: } K\left(\frac{x-x_n}{h}\right) = \begin{cases} \frac{3}{4}\left(1-\left(\frac{x-x_n}{h}\right)^2\right), & \left|\frac{x-x_n}{h}\right| \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

The following Matlab functions will be useful (among others): `hist bar randn rand find linspace scatter mode`.