The objective of this lab is for you to program in Matlab several nonparametric methods for density estimation, classification and regression, apply them to some datasets and observe their behavior. The TA will first demonstrate the results of the algorithms on a toy dataset, and then you will program them, replicate those results, and further explore the datasets with the algorithms. You can use the textbook, lecture notes and your own notes.

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 Implementing and using nonparametric methods

k-nearest-neighbor (KNN) classifier In file `lab05_knn2.m` we apply the KNN classifier to datasets in 2D and plot its results. You have to program the function `knn.m` that implements the KNN classifier. Then, explore its behavior in various settings (see suggestions at the end of file `lab05_knn2.m`).

Note: this is the k-nearest-neighbor *classifier*, not the k-nearest-neighbor *density estimate*.

Kernel density estimate (KDE) In file `lab05_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 $h > 0$:

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

  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 `lab05_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 (ch. 4–5)?

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):

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

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

III What you have to submit

Follow these instructions strictly. Email the TA the following packed into a **single** file (`lab05.tar.gz` or `lab05.zip`) and with email subject `[CSE176] lab05`:

- Matlab code for:

  1. The function `knn.m`, using the template provided. Read carefully the template to understand what the function should do; it should work when called from `lab05_knn2.m`.
  2. The following scripts, which are variations of `lab05_kde1.m` (Gaussian KDE for 1D datasets):
     (a) `lab05_kde2.m`: Gaussian KDE but for 2D datasets.
     (b) `lab05_kde1u.m`: 1D datasets but using the uniform kernel.
     (c) `lab05_kde1e.m`: 1D datasets but using the Epanechnikov kernel.

  Note: you are not allowed to use any functions from the Matlab Toolboxes (in particular, the Statistics and Machine Learning Toolbox, or the Neural Network Toolbox). You can only use basic Matlab functions, and the GM tools.

- A brief report (2 pages) in PDF format describing your experience with the algorithms. The more extensive and insighful your exploration, the higher the grade. Be concise. Don't include code or figures, we can recreate them by running your functions. Indicate the part that each member of the group did.