Homework #4
Non-parametric Density Estimation
Statistical Pattern Recognition

Due Azar 30th,1395
Format of homework file: Archive all files in a folder named as your student number and send it to mohammadhme@gmail.com. Send your emails with a subject of PR95F4_xxxxxx (replace xxxxxx by your student number)

---

1. You are a microcontroller with a sensor that measures the light level, x_i, in arbitrary units. Your job is to determine the weather: is today $\omega_1 = "sunny"$, or $\omega_2 = "cloudy"$? You have been programmed to compute Parzen window estimates of $p(x \mid \omega_1)$ and $p(x \mid \omega_2)$ using the rectangular window:

$$\hat{p}(x) = \frac{1}{nv} \sum_{i=1}^{n} \varphi\left(\frac{x - x_1}{h}\right)$$

$$\varphi\left(\frac{x}{h}\right) = \begin{cases} 1 & |x| < \frac{h}{2} \\ 0 & otherwise \end{cases}$$

You have five labeled training days. Three days were sunny, with light levels of $x_1 = 4$, $x_2 = 1$, and $x_3 = 5$ units, respectively. Two days were cloudy, with light levels of $x_4 = 3$ and $x_5 = 2$ units.

a. Plot the Parzen window estimated likelihood $\hat{p}(x \mid \omega_1)$ as a function of x, using h = 1.

b. Plot the Parzen window estimated likelihood $\hat{p}(x \mid \omega_2)$ as a function of x, using h = 1.

---

2. Consider the following training set drawn from an unknown density f(x):

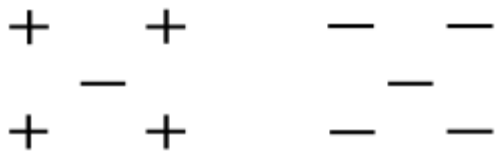
$$X = \{0.01, 0.12, 0.19, 0.32, 0.41, 0.48\}$$

a. Let $\varphi(x) \sim N(0,1)$. Find and sketch the Parzen window estimate of f(x) for values of h_n 0.1 and 0.01.

b. Find and sketch the 3-nearest neighbor estimate f(x).

---

3. Consider the following data set with two real-valued inputs x (i.e. the coordinates of the points) and one binary output y (taking values + or -). We want to use k-nearest neighbors (K-NN) with Euclidean distance to predict y from x.
    a. Calculate the leave-one-out cross-validation error of 1-NN on this data set.
    b. Calculate the leave-one-out cross-validation error of 3-NN on the same data set
    c. Describe how would you choose the number of neighbors K in K-NN in general?



---

4. [**computer project]** You are asked to build a k-Nearest Neighbor (kNN) classifier. The data set for evaluation is the heart data set. More information about the data can be found here:

   https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)

   The data set is included in heartstatlog.zip. In the heart data folder, there are three files: "trainSet.txt", "trainLabels.txt", and "test.txt". Each row of "trainSet.txt" corresponds to a data point whose class label is provided in the same row of "trainLabels.txt". Each row of "testSet.txt" corresponds to a data point whose class label needs to be predicted. You will train a classification model using "trainSet.txt" and "trainLabels.txt", and use it to predict the class labels for the data points in "testSet.txt".

   - Use the leave one out cross validation on the training data to select the best k among {1; 2;…; 10}.
   - Report the averaged leave-one-out error (averaged over all training data points) for each k={1; 2; … ; 10} and the best k used for predicting the class labels for test instances.
   - You should also report the predicted labels for the testSet.

---

5. **[Computer Project]** Consider arbitrary **one-dimensional** and **two-dimensional** distributions and draw some samples from each of these distributions. Use histogram, Parzen and KNN density estimation methods to estimate the distributions from these samples. Compare your results with the true real density. Evaluate the effect of number of samples and parameters of each method (number of cells for histogram, kernel type and bandwidth for Parzen and K for KNN method) in the accuracy of estimation.