Stat/CS 5525 Homework 1

(Due on September 14th, 2023)

1. Please show that when the loss function $L(Y, f(X))$ is the L_1 loss function, i.e., $L(Y, f(X)) = E|Y - f(X)|$, the solution to the optimization problem

$$\min_c E_{Y|X}(|Y - c| \mid X = x)$$

is $median(Y|X = x)$.

2. Ex. 2.7(a), (b), (c) from the textbook. The text book can be founded from

`https://hastie.su.domains/Papers/ESLII.pdf`

3. Program: Code up a k-nearest-neighbor classifier. Your codes should include two files. One file is the kNN classifier function, which has four input arguments:

   - new: input variable values for the test data set.
   - x: input variable values for the training data set.
   - y: output labels for the training data set.
   - k: neighborhood size.

   The output of the kNN classifier function should be the prediction results of the test data set. The other code file should be the "main" file, which includes

   - loading/reading the training and test data sets.
   - calling the kNN classifier function and return the prediction results. Specify three k values: $k = 5, 10, 15$.

   The training data set is shown in Figure 1. What you should submit:
   1. Three plots of the test data set with the predicted results colored as in Figure 1.
   2. Two files of codes.

   The training and test data sets can be downloaded. In the training data, the first two columns (from left to right) are input variables, the last column is the output label. In the test data, all the two columns are input variables.

4. Compare the classification performance of linear regression and kNN classification on the data in Problem 3. Show both the training and testing errors.

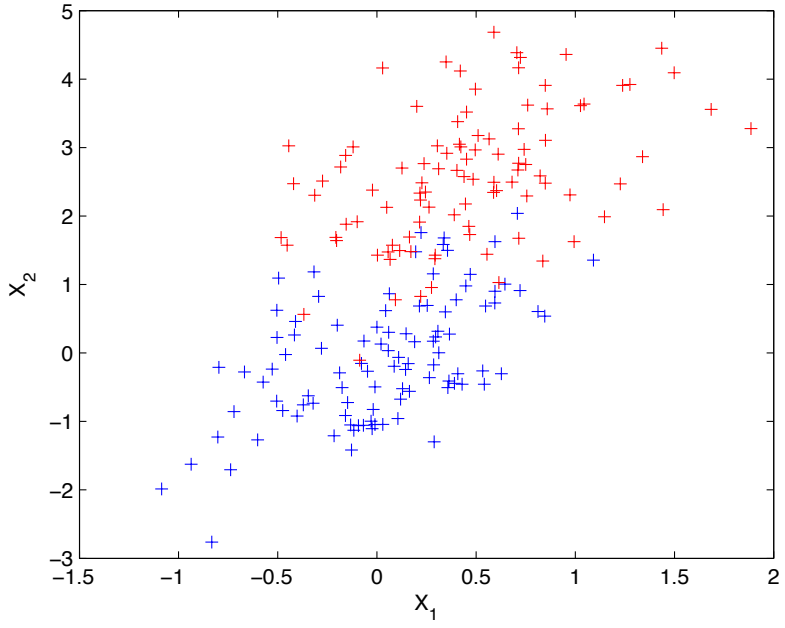

Figure 1: Training data set.