Statistical Machine Learning for Data Science Assignment 1

Due Date: Saturday, January 12, 2024 - By 11:59pm

Exercise 1: Practical Machine Learning - Digit Recognition (50 points)

This exercise features the analysis of the USPS digit recognition dataset using kNN with various neighborhood sizes.

```
library(dslabs)        # Package contributed by Yann LeCun to provide the MNIST data
mnist <- read_mnist()  # Read in the MNIST data

xtrain <- mnist$train$images
ytrain <- mnist$train$labels
ytrain <- as.factor(ytrain)
```

For all your random samples throughout this exercise, you must use the seed 19671210, that is

```
set.seed(19671210)
```

Also you will find the provided RMarkDown file MNIST-Classification-PCA-1.Rmd *useful. It is crucial to note that this file does not give answers, but instead contains fragments that should help you answer the questions of this exercise.* Recall that for a kNN classifier, the predicted label of any given $\mathbf{x} \in \mathscr{X}$ is given by

$$\widehat{f}_{\texttt{kNN}}(\mathbf{x}) = \underset{g \in \mathscr{Y}}{\texttt{argmax}} \left\{ \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}(y_i = g) \mathbb{1}(\mathbf{x}_i \in \mathcal{V}_k(\mathbf{x})) \right\}$$

Consider classifying digit '1' against digit '7', with '1' representing positive and '7' representing negative. Store in memory your training set and your test set. Of course you must show the command that extracts only '1' and '7' from both the training and the test sets. The learning machines you will be using here at just `1NN`, `9NN`, `18NN` and `27NN`.

1. Choose n a training set size and m a test set size, and write a piece of code for sampling a fragment from the large dataset. Explain why you choose the numbers you chose.
2. Display both the training confusion matrix and the test confusion matrix for each of the four learning machines under consideration.
3. Display the comparative ROC curves of the four learning machines, and do so for both the training set and the test set.
4. Identify two false positives and two false negatives at the test phase, and in each case, plot the true image against its falsely predicted counterpart.
5. Comment in greater details on any pattern that might have emerged.

Exercise 2: Investigating the Bayes risk $R^\star$ is Regression (34 points)

Let X and Y be two continuous random variables with joint probability density function

$$p(\mathbf{x},\mathbf{y}) = \frac{1}{2\pi\sqrt{2\pi\frac{9}{\pi^2}}}\exp\left\{-\frac{\pi^2}{18}\left[\mathbf{y}-\frac{\pi}{2}\mathbf{x}-\frac{3\pi}{4}\cos\left(\frac{\pi}{2}(1+\mathbf{x})\right)\right]^2\right\}$$

It can be shown that the conditional density of Y given X is given by

$$p_1(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{2\pi\frac{9}{\pi^2}}}\exp\left\{-\frac{\pi^2}{18}\left[\mathbf{y}-\frac{\pi}{2}\mathbf{x}-\frac{3\pi}{4}\cos\left(\frac{\pi}{2}(1+\mathbf{x})\right)\right]^2\right\}, \quad -\infty<\mathbf{x}<\infty$$

while the marginal density of X is given by

$$p_2(\mathbf{x}) = \frac{1}{2\pi}, \quad 0\le\mathbf{x}<2\pi.$$

1. Find and write down the expression of $\mathbb{E}[Y|X]$.
2. Let $n = 99$. Generate $\{(X_i,Y_i),\ i=1,\cdots,n\}$, an iid sample of size n where each (X_i,Y_i) as density $p(\mathbf{x},\mathbf{y})$. [Hint: Draw each X_i from the marginal of X, and then draw the corresponding Y_i from the conditional distribution of Y given X.] .
3. Draw the scatterplot from $\{(X_i,Y_i),\ i=1,\cdots,n\}$.
4. Consider now using regression analysis learning machines to learn the true function underlying your created dataset. We are herein learning under the squared error loss, namely $\ell(Y,f(X)) = (Y-f(X))^2$, with corresponding risk functional

$$R(f) = \mathbb{E}[\ell(Y,f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(\mathrm{y},f(\mathrm{x}))p_{XY}(\mathrm{x},\mathrm{y})d\mathrm{x}d\mathrm{y}.$$

   1. Find and write down the expression of

$$f^*(X) = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\ell(Y,f(X))]$$

   2. Find and write down the expression

$$R^* = R(f^*) = \min_f R(f)$$

   3. Consider a $60\% - 40\%$ Training-Test Set split and from it plot the comparative boxplots of the test errors based on 100 replications, for (a) kNN regression (b) Linear regression (c) Polynomial Regression (d) Regression tree learner
   4. Comment on how each average test error compares with R^*.

Exercise 3: Discovering the concept of ultra high dimensionality

You will then need to load the following data sets in order to answer the questions contained in this exercise:

```
data(DNA)    % From library(mlbench)    ### Binary predictor variables
data(BreastCancer) % From library(mlbench)
spam<-read.csv('spam-classification-1.csv')    # Spam detection data set
leukemia<-read.csv('leukemia-data-1.csv')      # DNA Microarray Gene Expression
prostate <- read.csv('prostate-cancer-1.csv')  # DNA Microarray Gene Expression
colon <- read.csv('colon-cancer-1.csv')        # DNA Microarray Gene Expression
```

For each of the above datasets, perform the following tasks:

1. Download the dataset and open it.
2. Provide a succinct description of the dimensionality of the data, sample size and homogeneity, featuring both type-homogeneity and scale homogeneity.
3. Comment on the quintessentially important index $\kappa := n/p$.
4. Generate the basic graphical summaries (plots) of 9 of the variables in the data. (You may select the 9 randomly the case of very large p)
5. Discuss correlations and multicollinearity wherever possible.