MSDS 534 **Homework 2** Due November 03 2021

Notes. **NO** late submission will be accepted. Computer generated output without detailed explanations and remarks will not receive any credit. You may type out your answers, but make sure to use different fonts to distinguish your own words from computer output. Only hard copies are accepted. For the simulation and data analysis problems, keep the code you develop as you may be asked to present your work in class.

1. Exercise 1 of ISL, P. 368.

2. Exercise 2 of ISL, P. 368.

3. Exercise 3 of ISL, P. 368-369.

4. Exercise 8 of ISL, P. 371-372.

5. This problem uses the MNIST dataset of handwritten digits. You may want to read `http://yann.lecun.com/exdb/mnist/` to get more information. For this problem, install the `R` package `dslabs`, and use the following code to read the MNIST dataset (you need Internet to download the data):

```
mnist=read_mnist()
```

Now the `mnist` is a list with two components: train and test. Each of these is a list with two components: images and labels. The images component is a matrix with each column representing one of the 28*28 = 784 pixels. The values are integers between 0 and 255 representing grey scale. The labels components is a vector representing the digit shown in the image. For example, you can use the following code to print the 5-th image in the training set.

```
i <- 5
image(1:28, 1:28, matrix(mnist$test$images[i,], nrow=28)[ , 28:1],
                  col = gray(seq(0, 1, 0.05)), xlab = "", ylab="")
## the label for this image is:
mnist$test$labels[i]
```

Now create the training and test set for this problem as follows, each of size 800.

- Select the first 400 images of "3" in `mnist$test$images`, and the first 400 images of "5" in `mnist$test$images`, as the training set. Create the corresponding label vector, which has length 800.
- Select the next 400 images of "3" in `mnist$test$images`, and the next 400 images of "5" in `mnist$test$images`, as the test set. Create the corresponding label vector.

Answer the following questions.

(a) Perform logistic regression on the training set, and use it to predict the labels of the test set. Report the training and testing mis-classification rates.

(b) For the logistic regression, the size of the training set is $N = 800$, and the number of features is $p = 784$, which is almost the same as N. Now try to run the logistic regression using the `glmnet()` function in the `glmnet` package. This function adds a Lasso type penalty to the logistic regression. Use the tuning parameter `lambda=.1` and `family="binomial"` in the `glmnet()` function (you don't need to specify any other parameters). Report the training and testing mis-classification rates.

(c) Try some other values of `lambda`, and report the smallest testing mis-classification rate you obtain, with the corresponding value of `lambda`.

(d) Build a support vector classifier using the training set, and use it to predict the labels of the test set. Report the training and testing mis-classification rates. [Hint. You can use `cost=1`, and add `scale=FALSE` in the `svm()` function.]

(e) From now on only use the 400 images of "3" in the training set. Plot the average image of them.

(f) Perform the PCA, and plot the images given by the first three principal directions. [Hint. You can use `svd()` as I did in the lecture, but you need to center the data first by yourself. Or you can use the function `prcomp()`, which does the centering automatically. See the book ISL for more details on the function `prcomp()`.]

6. Consider the kernel PCA. Suppose $\boldsymbol{x}_0$ is a new subject, of which we want to calculate the principal component vector $\boldsymbol{z}_0$. Recall that $\boldsymbol{K}$ is the un-centered kernel matrix, and $\tilde{\boldsymbol{K}} = (\boldsymbol{I} - \boldsymbol{J})\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{J})$ is the centered kernel matrix, where $\boldsymbol{J} = \mathbf{1}\mathbf{1}'/N$. Let $\tilde{\boldsymbol{K}} = \boldsymbol{U}\Lambda\boldsymbol{U}'$ be the eigenvalue decomposition of $\tilde{\boldsymbol{K}}$. Let

$$\boldsymbol{k}_0 = [K(\boldsymbol{x}_0, \boldsymbol{x}_1), K(\boldsymbol{x}_0, \boldsymbol{x}_2), \ldots, K(\boldsymbol{x}_0, \boldsymbol{x}_N)]'.$$

(a) Show that

$$\boldsymbol{z}_0 = \Lambda^{-1/2}\boldsymbol{U}'(\boldsymbol{I} - \boldsymbol{J})(\boldsymbol{k}_0 - \boldsymbol{K}\mathbf{1}/N),$$

where $\Lambda^{-1/2}$ is the diagonal matrix with diagonal elements $\{\lambda_1^{-1/2}, \ldots, \lambda_r^{-1/2}\}$.

(b) Use part (a) to verify that the m-th principal component is given by

$$z_{0m} = \frac{1}{\sqrt{\lambda_m}} \sum_{i=1}^{N} u_{im} \tilde{K}(\boldsymbol{x}_0, \boldsymbol{x}_i),$$

where

$$\tilde{K}(\boldsymbol{x}_0, \boldsymbol{x}_i) = \left[K(\boldsymbol{x}_0, \boldsymbol{x}_i) - \frac{1}{N}\sum_{j=1}^{N} K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \frac{1}{N}\sum_{j=1}^{N} K(\boldsymbol{x}_0, \boldsymbol{x}_j) + \frac{1}{N^2}\mathbf{1}'\boldsymbol{K}\mathbf{1} \right].$$