Homework 4

(Canvas online submission, **due time 11:59pm EST, Wednesday, April 6, 2022**)

1 (**Factor Analysis vs PCA, Ex 14.15 in ELS**) Generate 200 observations of $Y = (Y_1, Y_2, Y_3)^\top$ according to the model

$$Y_{3\times 1} = \Lambda^*_{3\times 3} X_{3\times 1}$$

where

$$\Lambda^* = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0.001 & 0 \\ 0 & 0 & 10 \end{pmatrix}_{3\times 3} \quad and \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}_{3\times 1} \sim N_3(0, I_3).$$

Compute the leading PC component and factor analysis directions (i.e., assuming 1 component/factor). Show that the leading PC component aligns itself in the direction of Y_3 while the leading factor essentially picks up the correlated component $Y_1 + Y_2$. Explain why.

2 **Spam classification.** Consider the email spam data set given on Canvas. This consists of 4601 email messages, from which 57 features have been extracted. These are as follows:

- 48 features giving the percentage of certain words (e.g., ”business”, ”free”, ”george”) in a given message
- 6 features giving the percentage of certain characters (; ([! \$ #)
- feature 55: the average length of an uninterrupted sequence of capital letters
- feature 56: the length of the longest uninterrupted sequence of capital letters
- feature 57: the sum of the lengths of uninterrupted sequences of captial letters

The data set contains a training set of size 3065, and a test set of size 1536. One can imagine performing several kinds of preprocessing to this data. Try each of the following separately: (i) Standardize the columns so they all have mean 0 and unit variance; (ii) Transform the features using $\log(x_{ij} + 1)$; (iii) Discretize each features using $\mathbb{I}(x_{ij} > 0)$.

(a) For each version of the data, write your code to fit a penalized logistic regression model, which seeks to minimize

$$-\text{log conditional likelihood} + \lambda\theta^T\theta.$$

The parameter λ can be chosen by cross-validation. Report the mean error rate on both the training and test sets.

(b) Write your LDA code to fit the standardized data and the log transformed data. Report your error rates.

(c) Write your code to implement a Naive Bayes classifier on the discretized data. Report your error rates.

(d) Perform Kernel logistic regression to fit the standardized data and the log transformed data (*You may use a existing package or write your own code*). Please consider Gaussian and polynomial kernels to produce the corresponding classifiers. Try a range of tuning parameters and show how such choices affect the behavior of the classifier obtained and the misclassification error on the test set.

Compare error rates using different methods and different preprocessed data in a table, and comment on the different performances.

3 **Comparison of different methods.** The Canvas website contains a data set "classification_dat" of (x_n, y_n) pairs, where the x_n are 2-dimensional vectors and y_n is a binary label. *(You may use existing packages, or you may implement the programs yourself)*

(a) Plot the data, using O's and X's for the two classes. The plots in the following parts should be plotted on top of this plot.

(b) Fit LDA model. Calculate the posterior probability of class 1, and plot the line where this probability is equal to 0.5.

(c) Fit a logistic regression model. Plot the line where the logistic function is equal to 0.5.

(e) Fit a linear regression to the problem, treating the class labels as real values 0 and 1. Plot the line where the linear regression function is equal to 0.5.

(f) Fit the additive logistic regression model and plot the curve where the logistic function is equal to 0.5.

(g) Perform Kernel logistic regression with Gaussian kernel and plot the curve where the logistic function is equal to 0.5.

(h) The data set "classification_test" is a separate data set generated from the same source. Test your fits from parts (b)–(g) on these data and compare the results.