IE 582 Statistical Learning for Data Mining
Homework 3, due November 12th, 2018

Instructions: Please solve the following exercises using R (http://www.r-project.org/). Also, you are requested to use certain R packages for this particular homework. You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals (as you did for Homework 0).

Do not share your code (except the one in your progress journals)! As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Task 1 (60 points)

The conference paper by Liu et al. (2009) starts with the following statement: "Gestures have recently become attractive for spontaneous interaction with consumer electronics and mobile devices in the context of pervasive computing". The aim is to provide efficient personalized gesture recognition on wide range of devices.

To achieve this, Liu et al. (2009) uses a single three-axis accelerometer to collect data from eight users to characterize eight gesture patterns. The library, uWaveGestureLibrary, consists over 4000 instances each of which has the accelerometer readings in three dimensions (i.e. *x*, *y* and *z*). Eight gestures are illustrated in Figure 1.

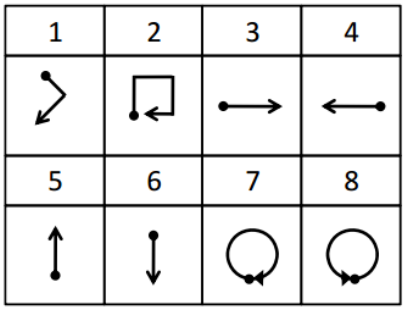

Figure 1: Gesture vocabulary considered by Liu et al. (2009). The dot denotes the start and the arrow the end

a) The dataset is provided in the following link:
https://drive.google.com/drive/folders/13553neknux7U8why55KM1WrjgkA9IJKm?usp=sharing

Note that there are separate files for each axis and each row corresponds to one gesture in the files. First column has the class information. The information between second and last column is the time ordered observations in the corresponding axis (provided in the file name as X, Y or Z).

Read the data and visualize one instance (all axes) from each class and try to relate the shape (time series) you see with the gestures shown in Figure 1 (this is just for fun, sometimes it is good to start with data visualization to understand what is going on). A 3D scatter plot would be interesting. Note that this is an acceleration information. You can transform this information to a velocity vector by computing the cumulative sum of acceleration over time.

b) Suppose we decided to apply a nearest-neighbor (NN) classifier to find the labels of test instances based on acceleration information. Combine the information of X, Y and Z coordinates columnwise (i.e. cbind) to obtain a single vector (i.e. time series of X, Y and Z together) for each gesture. Propose two distance measures for computing similarity between two time series. For each distance measure alternative, use the training data to identify the optimal value of *k* which minimizes the error of a 10-fold cross-validation.

c) Using the value of *k* (identified for each distance measure) in part (b) and evaluate your final performance on the test data and present your results in a (8-by-8) confusion matrix, showing the counts for actual and predicted labels. In addition, quote the runtime and accuracy for your results.

Task 2 (40 points)

Background

We have covered two penalized regression approaches in class, namely ridge and lasso regression/classification with linear models. There are also alternative penalization approaches for data with a specific structure such as time-series. Suppose we would like to perform regression given a time series data set. In other words, we have time series observations and there is a continuous response associated with this time series observations. A penalized regression approach with fused penalties (Tibshirani et al., 2005) minimizes the following loss function:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2|\boldsymbol{\beta}|^2 + \lambda_1 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$$

The first part is the sum of squared errors, second part is the ridge penalty over the coefficients and the last part is the fused lasso penalties. As you can see from the equation, the coefficients of the variables that are temporally close (i.e. coefficient for an observation at time *t* and time *t+1*) are motivated to be close to each other. A simple example and codes can be found on http://www.mustafabaydogan.com/blog/11-codes/76-time-series-classification-with-fused-lasso-using-lqa-package.html (this example benefits from "lqa" package in R). Note that the first part of the loss function is different in classification setting (it is logistic loss which is not our concern for now).

You are given a nearest-neighbor classification problem and code on Moodle (titled as "NN Classification ECG" under week named as "6 October - 12 October"). The zip file contains a sample NN code (you can also use this code as an example for the first task) and ECG data. The details of the dataset are explained in the class.

a) Train a logistic regression model on the training data using fused lasso penalties. You may want to check "penalized" or "lqa" package for training). Use the learned model to predict the class for test data. Present your results in a (2-by-2) confusion matrix.

Do not forget to learn the parameters of fused lasso through cross-validation on training data. If you prefer to use the "penalized" or "lqa" packages, cross-validation is already implemented in the functions. You just need to provide the necessary parameters as arguments to the corresponding function.

b) Comment on the regression coefficients. Is there any interesting information? Try to interpret the model.

c) Suppose you take the difference between consecutive time series observations to transform the time series. Let x_t be the observation at time point *t* and you are asked to create a new time series which is $y_t = x_t - x_{t-1}$ for $t = 2,3,\dots,T$ where *T* is the length of the time series. Perform the same operation in part (a) for this newly created dataset and compare the results.

d) Comment on the regression coefficients for the model trained on the new dataset. Is there any interesting information? Try to interpret the model.

References

J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan. uWave: Accelerometer-based personalized gesture recognition and its applications. Pervasive Computing and Communications, IEEE International Conference on, 0:1-9, 2009.
(link: http://www.ruf.rice.edu/~mobile/publications/liu09percom.pdf)

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. "Sparsity and smoothness via the fused lasso." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, no. 1 (2005): 91-108.