IS580 – Knowledge, Discovery and Mining

Assignment 4

Due Date: 29 May 2018

In this assignment, you are expected to fit a generalized linear model via penalized maximum likelihood (i.e. elastic net), fit a time series model, and make a model selection. For the elastic net task, you are provided with training and test datasets, namely "train.csv" and "test.csv". For the time series analysis, you are provided with the data set named "timeSeries.txt".

You can fit the elastic net model with the *glmnet* package. You may refer to the references below for fitting the model:

- https://cran.r-project.org/web/packages/glmnet/glmnet.pdf
- http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

For the time series analysis, you will use the *tseries* and *forecast* packages.

For the last part (performance analysis), you may refer to the references below for the paired t-test and Wilcoxon signed rank test:

- https://www.r-bloggers.com/paired-students-t-test/
- https://www.r-bloggers.com/wilcoxon-signed-rank-test/

Deliverables:

1. RMarkdown file named as "yourID_Assignment4.Rmd including R chunks, stats, results of the analysis, graphs and comments.
2. Data file named as "yourID_Assignment4.Rdata" including elastic net model and predictions of the model in Question A and time series, additive model for the time series and forecasts in the Question B.

Questions

A. Consider the data sets "train.csv" and "test.csv"
   1. Generate a multinomial elastic net model with cross-validation option on the "train.csv" data set (**Hint:** Use *cvglmnet* function in *glmnet* package. *cvglmnet* only accepts *matrix* data type, so be sure to convert your data set to matrix. Also, pay attention to the *family* parameter of the function.).
   2. Make predictions for the "test.csv" data set by using the model.
   3. Compare the predictions of the model with the actual classes. Comment on the performance of the model.
   4. Save your model and predictions in the data file "yourID_Assignment4.RData"

B. Consider the data set "timeSeries.txt"
   1. Is the time series stationary? Why? Support your reasons with ACF, PACF plots and/or statistical tests. If it is not stationary, convert it to a stationary series. Explain your steps in detail while converting it.

2. Fit an additive model upon the time series. Explain the model fit and why you select it, and show the model is appropriate for the time series.
3. Make a forecast for the next 12 observations. Plot your forecast.
4. Save the time series (manipulated if you need), model fit and forecast in the data file "yourID_assignment4.RData"

C. Consider the model performance table below. The table shows the AUC (area under curve) values for two classification models on 15 data sets. Which model performed better on the data sets overall? Use paired t-test or Wilcoxon signed rank test in order to show the performance differences of the two models (**Hint**: Consider the distribution of the values before deciding the statistical test). Comment on the results and explain in detail your model selection.

	Model 1	*Model 2*
Dataset 1	0.018	0.578
Dataset 2	0.022	0.990
Dataset 3	0.455	0.124
Dataset 4	0.153	0.187
Dataset 5	0.373	0.390
Dataset 6	0.555	0.662
Dataset 7	0.158	0.545
Dataset 8	0.777	0.734
Dataset 9	0.308	0.587
Dataset 10	0.081	0.124
Dataset 11	0.369	0.864
Dataset 12	0.816	0.213
Dataset 13	0.247	0.491
Dataset 14	0.185	0.961
Dataset 15	0.312	0.996

Data set information:

"train.csv" and "test.csv":

Datasets contain the features extracted from a database of colonoscopic videos showing gastrointestinal lesions. There are 3 types of lesion: hyperplasic, adenoma and serrated adenoma. The last column of the dataset denoted as **Y** corresponds to the lesion name: 3 for adenoma, 1 for hyperplasic, and 2 for serrated. All other columns (denoted as **X1, X2, …, X698**) are the raw features (without any kind of pre-processing):

422 2D TEXTURAL FEATURES

- First 166 features: AHT: Autocorrelation Homogeneous Texture (Invariant Gabor Texture)

- Next 256: Rotational Invariant LBP

76 2D COLOR FEATURES

- 16 Color Naming

- 13 Discriminative Color

- 7 Hue

- 7 Opponent

- 33 color gray-level co-occurrence matrix

200 3D SHAPE FEATURES

- 100 shapeDNA

- 100 KPCA

"timeSeries.txt":

This dataset contains 373 instances of daily averaged sensor response for the PT08.S1 (tin oxide) levels in an Air Quality Chemical Multi-sensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city.