Optimization for Big Data - Logistic regression and few other things

Summary

The goal of this homework is to study from a practical point of view several datasets with machine learning tools.

1) You are expected to produce a report that follows linearly the tutorial. You are asked to answer the theoretical and practical questions with a pdf file. Don't forget to provide somme graphical illustrations.
2) You are also asked to call your file : NAME-SURNAME.pdf.
3) You can use R or Python for the first dataset, create a Notebook. In any case, providing a commented code is mandatory !

Deadline : 14th of april 2019
Individual work

First dataset : cleaning and classification

We study a first dataset that presents wheter a patient is sick or not. This information is stored in the variable "outcome" of the csv file. A value 1 indicates a sick patient.

The other variables gather several biological or physiological informations measured on the patients :

- Pregnancies : Number of times pregnant
- Glucose : Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure : Diastolic blood pressure (mm Hg)
- SkinThickness : Triceps skin fold thickness (mm)
- Insulin : 2-Hour serum insulin (mu U/ml)
- BMI : Body mass index
- DiabetesPedigreeFunction : Diabetes pedigree function
- Age : Age (years)

Question 1 : Produce a statistical descriptive analysis (histograms, PCA, etc). What do you think about some strange 0 values ?

Question 2 : Is there any correlation between 0's and the outcome variable ?

Question 3 : Clean the dataset with a method of your choice.

Question 4 : Produce an unsupervised classification of the dataset without the outcome variable. Then, represents the value of the outcome variable associated with the obtained clustering.

Question 5 : Use the best algorithm you can find (in terms of accuracy) for supervised classification of the outcome variable.

Question 6 : Provide an estimation of the misclassification rate of the whole process, from the cleaning step to the final prediction.

Question 7 : What are the meaningful variables for predicting the outcome variable ?

Second dataset : high dimensional dataset

This work is interested in a problem of certification of a production chain of cookies. It is necessary to control the amount of each ingredient in the preparation before baking in an industrial hoven. The purpose is to guarantee a good proportion rate of lipids, sugar, floor, water so that the preparation is guaranteed to be close to the nominal recommandation for the recipe. It is also important to detect as soon as possible the presence of a gap between the preparation and the nominal recommandation. The measures and their analyses are performed in a chemistry laboratory and are generally long and costly.

It is however possible to perform a spectral analysis with infra-red signals in the cookie-firm. Hence, a "simplification" of the NIR analysis is needed : the selection of a few amount of spectral frequencies and a robust model of regression can bypass a whole chemistry analysis This kind of problem is classical in food-processing industry and corresponds to a "chemistrymetry" calibration

Data

The dataset can be loaded with R through the use of the package *ppls*.

Measures are performed on two samples, the sizes of the training and test sets are respectively 40 and 32 samples. For each of the 72 cookies, the amount of lipids, sugar, floor and water are measured with a classical chemistry approach, although in the same time a spectral infra-red analysis is made from the frequency 1100 nm to the frequency 2498 nm, regularly sampled with a stepsize of $2nm$. We hence have in our hands 700 observed values for each cookie. All of these values may be informative (or not) to explain the cookie preparation. This study is thus a typical example of a high dimensional problem with $p >> n$. The main objective is to answer the following question : is it possible to infer from this spectrum the composition of the cookie preparation ? In case of a favorable answer, the time and money saved would be important. The practical session focuses on the modelization of sugar and how to build a good prediction of this sugar rate.

You can load the dataset using this list of commands.

```
library(ppls)
data(cookie)
# Extraction of the sugar rate and spectrum
cook = data.frame(cookie[,702],cookie[,1:700])
names(cook)= c("sucre",paste("X",1:700,sep=""))
```

You can use the caret package to produce easily understandable R commands

Question 8 : Warning : one observation seems to be an outlier (as pointed in the litterature). Why ?

Question 9 : Create a training set and a test set to learn your algorithm and measure the efficiency of your method.

Question 10 : Use a ridge regression to learn the sugar rate, optimize the ridge regression with a suitable choice of the penalty parameter. Discuss on the quality of the model on the learning set and provide a prediction on the test set and a computation of the error. (precise its definition). Note the error of the method so that we can compare it with other methods.

Question 11 : Investigate the PLS regression on the www, explain the model, and its resolution.

Question 12 : Optimize the use of the PLS regression (optimal number of components) and estimate the model with this number of components. Predict on the test set and compute the error of the method.

Question 13 : Check that on this dataset, the aggregation models with trees are not very strong. (Explain the meaning of an "agregation" method).

Question 14 : Explain the Bayesian exponential weighted aggregation and program it with R. You are assumed to read carefully the paper " Sparse Regression Learning by Aggregation and Langevin Monte-Carlo " of Dalalyan and Tsybakov.

Question 15 : Use the Lasso estimator with the lars algorithm.

Question 16 : What is a deep learning neural network method ? Explain the theoretical difficulties and the practical ones. Use a neural network method with several hidden layers.

Question 17 : Provide a summary of the results, what are the good methods, what are the useful variables ?