Optimization for Big Data - High dimensional regression

Summary

The goal of this third session is to solve some regression problems involved in big data nowadays problems : the high dimensional regression problem and the sequential logistic regression with Python.

You are expected to produce a report that follows linearly the tutorial. You are asked to :

- *Deal with Section 1, "the high dimensional regression". In this part, you have to choose one option among two :*
  1. *Option 1 : Subsection 1.2.1*
  2. *Option 2 : Subsection 1.2.2*
- *Deal with Section 2, "the sequential logistic regression"*

Deadline : 15th of april 2018

Work with a team of 2

Expected length : less than 20 pages (without any Python program in an appendix section)

1 High dimensional regression

We observe i.i.d. realizations $(X_i, Y_i)_{1\leq i\leq n}$ of the standard linear model

$$Y_i < X_i, \theta > +\epsilon_i, \tag{1}$$

and we assume that each observation X_i belongs to $\mathbb{R}^p$ with $p >> n$. We also assume that θ is s-sparse, with s small, *i.e.* much smaller than n.

1.1 Lasso

Question 1 : Create a sample of size $n = 100$ with $p = 10^3$ and $s = 10$. Each X_i is sampled with a centered Gaussian distribution and $\epsilon_i \sim \mathcal{N}(0,1)$. θ may be chosen arbitrarily and Y is given by (1)

Question 2 : We are interested in using the Lasso, with the minimization of

$$f_\lambda(\theta) = \|Y - X^t\theta\|_2^2 + \lambda\|\theta\|_1,$$

where λ is a positive parameter. Recall the main step of the minimization algorithm of f_λ.

Question 3 : Program the Iterative Soft Thresholding method !

Question 4 : Compare your result with the $sklearn.linear\_model$ package

```
import scipy.io
import numpy as np
import pylab as pl
from sklearn.linear_model import lasso_path, LassoCV
```

Help yourself with some ressources you may found on the web.

1.2 Option 1 : application

1.2.1 Real database

Question 7-a : Use a real dataset and test your program and the package.

1.2.2 Comparison with the ridge regression

Question 7-b : Use your lecture notes to check what is ridge regression. Explain briefly the pros and cons associated to this method, especially when handling high dimensional datasets.

Question 7-c : Compare the results (from a statistical and numerical point of view) between the Lasso and the Ridge regression.

1.3 Option 2 : improving the computational time

Question 8-a : What is the computational time associated to the first order gradient descent scheme ? associated to the iterative soft thresholding algorithm for the Lasso ?

Question 8-b : Investigate on the Nesterov Accelerated Gradient Descent method on www. Explain the method. What are the improvements brought by NAGD, in comparison to AGD ? We ddo not ask to provide some proofs !

Question 8-c : Implement in python the FISTA (Fast Iterative Soft Thresholding algorithm) associated to the Lasso problem.

2 Sequential regression problem

We consider a logistic regression model : a pair of variables of random variables $(X,Y) \in \mathbb{R}^p \times \{\pm 1\}$ such that X is uniformly distributed in $[-1,1]^p$:

$$X \sim \mathcal{U}([-1,1]^p),$$

and $Y|X$ is a binary random variable :

$$\mathbb{P}[Y=1|X] = \frac{e^{<\theta^\star, X>}}{1+e^{<\theta^\star, X>}},$$

where $\theta^\star$ is an unknown parameter to be recovered.

2.1 Theory

Question 9 : Compute the log-likelihood of the model, denoted by $\ell_n(\theta)$, based on a set of n observations $(X_1,Y_1),\ldots,(X_n,Y_n)$. Recall the properties of the M.L.E and define

$$\ell(\theta) := \frac{\ell_n(\theta)}{n}.$$

Question 10 : Recall the properties of ℓ, as a function of $\theta \in \mathbb{R}^p$.

Question 11 : Compute the gradient of ℓ. Check that

$$\theta^\star = \arg\min_{\theta \in \mathbb{R}^p} \ell(\theta).$$

Question 12 : Define a sequential stochastic gradient algorithm for the maximization of ℓ :

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \nabla \ell(\theta_n) + \gamma_{n+1}\epsilon_{n+1}.$$

Question 13 : Give an admissible step-size sequence $(\gamma_n)_{n\geq 1}$ for the convergence of $(\theta_n)_{n\geq 1}$ towards $\theta^\star$. Using the lecture notes, describe the theoretical properties with a convergence rate.

Question 14 : Program the sequential logistic regression algorithm and check its good behaviour for reasonable size of p ($p = 5, 10, 20$).

Question 15 : Illustrate the convergence rate. Is it better then expected (from a theoretical point of view).