Computer Lab: Quantile Regression

Convex analysis, monotone operators and optimization
Olivier Fercoq olivier.fercoq@telecom-paristech.fr
12 December 2019

You can choose any programming language and work either alone or in pairs. Please send your code and answers to the questions to `olivier.fercoq@telecom-paristech.fr` before Wednesday, December 18th.

1 Data

We will be using the census dataset for this computer lab. Please download the dataset and helper file on
`https://perso.telecom-paristech.fr/ofercoq/tp_qr/`.

2 Quantile regression with linear kernels

For $\tau \in (0,1)$, let us consider the pinball loss defined as $L_\tau(v) = \max\{-(1-\tau)v, \tau v\}$.

Question 2.1
Calculate L_τ^*, $\mathrm{prox}_{\gamma L_\tau^*}(v)$ and $\mathrm{prox}_{\gamma L_\tau}(v)$ for $\gamma > 0$.

The quantile regression problem consists in estimating conditional quantiles. Given a pair of random variables $(X,Y) \in \mathbb{R}^d \times \mathbb{R}$ and a number $\tau \in [0,1]$, our goal is to estimate the conditional quantile function

$$\mu_\tau(x) = \inf\{\mu \in \mathbb{R} \,:\, \mathbb{P}(Y \leq \mu \mid X = x) \geq \tau\} \ .$$

Given a training set $\{(x_{i,:}, y_i)\}_{0 \leq i \leq n-1}$, we estimate the conditional quantile using the solution of the following optimization problem:

$$\min_{w \in \mathbb{R}^d, w_0 \in \mathbb{R}} \frac{\alpha}{2} \sum_{j=1}^d w_j^2 + \sum_{i=1}^n L_\tau\big(y_i - \sum_{j=1}^d x_{i,j} w_j - w_0\big) \tag{1}$$

where $\alpha > 0$ is a regularization constant and set $\hat{\mu}_\tau(x) = \sum_{j=1}^d x_j w_j - w_0$. In the rest of the lab, we shall take $\alpha = 1$.

Question 2.2
Define $g : (w, w_0) \mapsto \frac{\alpha}{2}\|w\|^2$. Calculate $\mathrm{prox}_{\gamma g}((w, w_0))$.

Question 2.3
For $z \in \mathbb{R}^n$, denote $\mathbf{L}_\tau(z) = \sum_{i=1}^n L_\tau(z_i)$ and $e = (1, \dots, 1)$. Show that

$$\begin{aligned}
\min_{w \in \mathbb{R}^d, w_0 \in \mathbb{R}} \frac{\alpha}{2}\|w\|^2 + \mathbf{L}_\tau(y - xw - w_0 e) &= \min_{w \in \mathbb{R}^d, w_0 \in \mathbb{R}} \max_{z \in \mathbb{R}^n} \frac{\alpha}{2}\|w\|^2 - \mathbf{L}_\tau^*(z) + z^\top(y - xw - w_0 e) \\
&= \max_{z \in \mathbb{R}^n} y^\top z - \frac{1}{2\alpha}\|x^\top z\|^2 - \mathbf{L}_\tau^*(z) - \iota_{\{0\}}(e^\top z) \\
&= \max_{z \in \mathbb{R}^n} \min_{u \in \mathbb{R}} y^\top z - \frac{1}{2\alpha}\|x^\top z\|^2 - \mathbf{L}_\tau^*(z) - u e^\top z
\end{aligned}$$

3 Implementation

Question 3.1
Implement at least two algorithms for the resolution of the quantile regression problem. You may choose `test_size = 0.99` in order to test your algorithm on small data.

Question 3.2
Define a stopping criterion. Why did you choose it?

Question 3.3
Compare the performance of the algorithms you implemented on the census dataset with $\tau = 0.7$ and `test_size=0.33`.