NO 2018
Another ADMM for Logistic Regression

François Lauze, Oswin Krause

March 19, 2018

1 The Problem

In this part you will have to minimize the penalized Logistic regression problem

$$\underbrace{\sum_{i=1}^{n} \log\left(1 + e^{-y_i\theta^\top x_i}\right)}_{F(\theta)} + \mu\|\theta\|_1 \tag{1.1}$$

with $\theta = (\theta_0, \theta_1 \dots, \theta_d) \in \mathbb{R}^{d+1}$ and $\mu > 0$. In Absalon, you have one of the solvers described, it uses the fact that for $x \in \mathbb{R}$, $|x| = x^+ - x^-$ where $x^+ = \max(x, 0)$ and $x^- = -\min(x, 0)$, the positive and negative parts of x.

In this second programming assignment, we use a well known technique, adaptation of the Lasso algorithm [Tishbirani], so as to deal with the non smooth 1-norm $\|\theta\|_1$. Deliverables are the same as for the first programming assignment (up to the name of some variables...).

2 Augmented Lagrangian Formulation and ADMM

The idea is to modify the minimization of (1.1) to introduce a equality constraint around the 1-norm term in the problem. Minimizing (1.1) is clearly equivalent to

$$\begin{aligned} &\min_{\theta,\varphi} F(\theta) + \mu\|\varphi\|_1, \\ &\text{s.t. } \theta = \varphi \end{aligned} \tag{2.1}$$

We form the augmented Lagrangian for (2.1):

$$L_\rho(\theta, \varphi, \lambda) = F(\theta) + \mu\|\varphi\|_1 + \frac{\rho}{2}\|\theta - \varphi\|^2 + \lambda^\top(\theta - \varphi). \tag{2.2}$$

Applying ADMM to it, we obtain the following iterative algorithm

$$\theta^{k+1} = \arg.\min_\theta L_\rho(\theta, \varphi^k, \lambda^k) = \arg.\min_\theta F(\theta) + \frac{\rho}{2}\|\theta - \varphi\|^2 + \lambda^\top(\theta - \varphi) \tag{2.3}$$

$$\varphi^{k+1} = \arg.\min_\varphi L_\rho(\theta^{k+1}, \varphi, \lambda^k) = \arg.\min_\varphi \mu\|\varphi\|_1 + \frac{\rho}{2}\|\theta - \varphi\|^2 + \lambda^\top(\theta - \varphi) \tag{2.4}$$

$$\lambda^{k+1} = \lambda^k + \rho(\theta^{k+1} - \varphi^{k+1}). \tag{2.5}$$

2.1 Minimization in θ

A factorization trick is applied to (2.3), which will also be applied (in a slightly different way) to (2.4). For that, write

$$
\begin{aligned}
\frac{\rho}{2}\|\theta-\varphi\|^2+\lambda^\top(\theta-\varphi) &= \frac{\rho}{2}\left(\|\theta\|^2-2\theta^\top\varphi+\|\varphi\|^2-\frac{2\lambda^\top}{\rho}(\theta-\varphi)\right)\\
&= \frac{\rho}{2}\left(\|\theta\|^2-2\left(\varphi-\frac{\lambda}{\rho}\right)^\top\theta+L(\varphi)\right)\\
&= \frac{\rho}{2}\|\theta-\left(\varphi-\frac{\lambda}{\rho}\right)\|^2+M(\varphi).
\end{aligned}
$$

$L(\varphi)$ and $M(\varphi)$ regroup terms where θ does not appear, and thus do not affect minimization. So minimizing (2.3) is equivalent to compute

$$
\theta^{k+1}=\arg.\min_\theta F(\theta)+\frac{\rho}{2}\|\theta-\left(\varphi^k-\frac{\lambda^k}{\rho}\right)\|^2. \tag{2.6}
$$

In the case of logistic regression, F is $\mathcal{C}^2$, convex, positive semi-definite and we add to it a positive definite quadratic term making it positive definite. This strongly suggests a Newton method. This can become critical, as a too approximate solution of (2.6) may lead to unbounded Lagrange multiplier update ($\lambda^{k+1}=\lambda^k+(\theta^{k+1}-\varphi^{k+1})$).

2.2 Minimization in φ, soft shrinkage

With the same type of factorization as above, we can write

$$
\frac{\rho}{2}\|\theta-\varphi\|^2+\lambda^\top(\theta-\varphi)=\frac{\rho}{2}\|\varphi-\left(\theta+\frac{\lambda}{\rho}\right)\|^2+N(\theta).
$$

Solving (2.4) is clearly equivalent to

$$
\varphi^{k+1}=\arg\min_\varphi \mu\|\varphi\|_1+\frac{\rho}{2}\|\varphi-\left(\theta^{k+1}+\frac{\lambda^k}{\rho}\right)\|^2 \tag{2.7}
$$

Set $\pi^{k+1}=\theta^{k+1}+\frac{\lambda^k}{\rho}$ and note that

$$
\begin{aligned}
\mu\|\varphi\|+\frac{\rho}{2}\|\varphi_\pi^{k+1}\|^2 &= \mu\sum_{i=1}^{d+1}|\varphi_i|+\sum_{i=1}^{d+1}\frac{\rho}{2}(\varphi_i-\pi_i^{k+1})^2\\
&= \mu\sum_{i=1}^{d+1}\left(|\varphi_i|+\frac{\rho}{2\mu}(\varphi_i-\pi_i^{k+1})^2\right).
\end{aligned}
$$

The minimization can clearly be carried component by component. Its solution is given by *soft shrinkage*. What we need is to be able to solve a minimization of the form

$$
u^*=\arg\min_u |u|+\frac{\alpha}{2}(u-u_0)^2,\quad u\in\mathbb{R}. \tag{2.8}
$$

The solution is given by a sort of shrinkage and a partial proof is provided in the next section.

$$
u^*=\begin{cases} u_0+\frac{1}{\alpha} & \text{if } u_0<-\frac{1}{\alpha}\\ 0 & \text{if } u_0\in[-\frac{1}{\alpha},\frac{1}{\alpha}]\\ u_0-\frac{1}{\alpha} & \text{if } u_0>\frac{1}{\alpha}.\end{cases} \tag{2.9}
$$

3 1D Soft Shrinkage

The idea is also an idea of duality. The absolute value $|u|$ of u can indeed be written as

$$|u| = \max_{z\in[-1,1]} zu.$$

Inserting it in (2.8), we have the following minimization problem

$$\min_u \max_{z\in[-1,1]} zu + \frac{\alpha}{2}(u-u_0)^2. \tag{3.1}$$

A small "miracle" admitted here, is that a classical theorem, the *minimax* theorem [von Neuman][1], asserts that in that case (among others),

$$\min_u \max_{z\in[-1,1]} zu + \frac{\alpha}{2}(u-u_0)^2 = \max_{z\in[-1,1]} \min_u zu + \frac{\alpha}{2}(u-u_0)^2 \tag{3.2}$$

The minimizer u^* of the smooth, unconstrained expression $zu+\frac{\alpha}{2}(u-u_0)^2$ is simply $u^*(z) = u_0-\alpha z$ by for instance writing that the derivative in u is 0. By inserting the result in (3.2), one needs to solve

$$\max_{z\in[-1,1]} -\frac{z^2}{2\alpha} + zu_0 \tag{3.3}$$

Its unconstrained maximum is reached at $z = \alpha u_0$. We thus need to consider three cases here

$$z^* = \begin{cases} -1 & \text{if } \alpha u_0 < -1 \\ \alpha u_0 & \text{if } -1 \leq \alpha u_0 \leq 1 \\ 1 & \text{if } \alpha u_0 > 1 \end{cases} \tag{3.4}$$

And since $u^*(z^*) = u_0 - \alpha z^*$, we get finally the *soft shrinkage* as solution of (2.8)

$$u^* = \begin{cases} u_0 + \frac{1}{\alpha} & \text{if } u_0 < -\frac{1}{\alpha} \\ 0 & \text{if } u_0 \in [-\frac{1}{\alpha}, \frac{1}{\alpha}] \\ u_0 - \frac{1}{\alpha} & \text{if } u_0 > \frac{1}{\alpha}. \end{cases} \tag{3.5}$$

References

[Tishbirani] Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267 – 88

[von Neuman] von Neumann, John. (1928). "Zur Theorie der Gesellschaftsspiele". Math. Ann. 100: 295 – 320.

[1] This is actually not too complicated to prove in the 1D case of interest here.