École Nationale de la Statistique et de l'Administration Économique
Mastère Spécialisé Data Science

SUJET 2 : Implémentation de CoCoA (Communication-efficient distributed dual Coordinate Ascent)

Charline Fraioli
Kévin Martel
Louis Millot

Professeur :
Marco Cuturi

Année universitaire 2016-2017

Introduction

Nous avons choisi comme sujet l'implémentation de l'algorithme de distribution de calculs COCOA pour la résolution d'une régression logistique.

On cherche à minimiser une fonction de perte convexe avec un terme de régularisation convexe :

$$\min_{w \in \mathbb{R}^d} \quad \left[P(w) := \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell_i(w^T x_i) \right]$$

Où :
$x_i \in R^d$: vecteur contenant les données d'apprentissage
l_i : la fonction de perte logistique qui dépend des labels y_i
$\lambda > 0$: le paramètre de régularisation

Le problème dual correspondant :

$$\max_{\alpha \in \mathbb{R}^n} \quad \left[D(\alpha) := -\frac{\lambda}{2}\|A\alpha\|^2 - \frac{1}{n}\sum_{i=1}^{n} \ell_i^*(-\alpha_i) \right]$$

Où :
l_i^* : la conjuguée de la fonction de perte logistique qui dépend des labels y_i
$A \in R^{dxn}$: matrice contenant en colonne les vecteurs x_i, $A_i = \frac{1}{\lambda n} x_i$, avec $w(\alpha) = A\alpha$ la correspondance primal-dual

Pour résoudre ce problème, nous utiliserons la méthode SDCA *Stochastic Dual Coordinate Ascent* et nous distribuerons les calculs sur plusieurs processeurs via la méthode CoCoa.

Ce papier a pour but de résumer notre notebook, vous pourrez retrouver tous les détails et les résultats au sein de celui-ci.

Programmation

Nous avons programmé l'algorithme comme défini dans l'article « *Communication-Efficient Distributed Dual Coordinate Ascent* » de Virginia Smith and al.

La fonction « obj_log_los » calcule la fonction de coût qui permet d'étudier la convergence de l'algorithme. Elle correspond à la fonction objective de la régression logistique.

La fonction « SDCA » permet de résoudre le problème.

La fonction « COCOA » réalise la distribution des calculs.

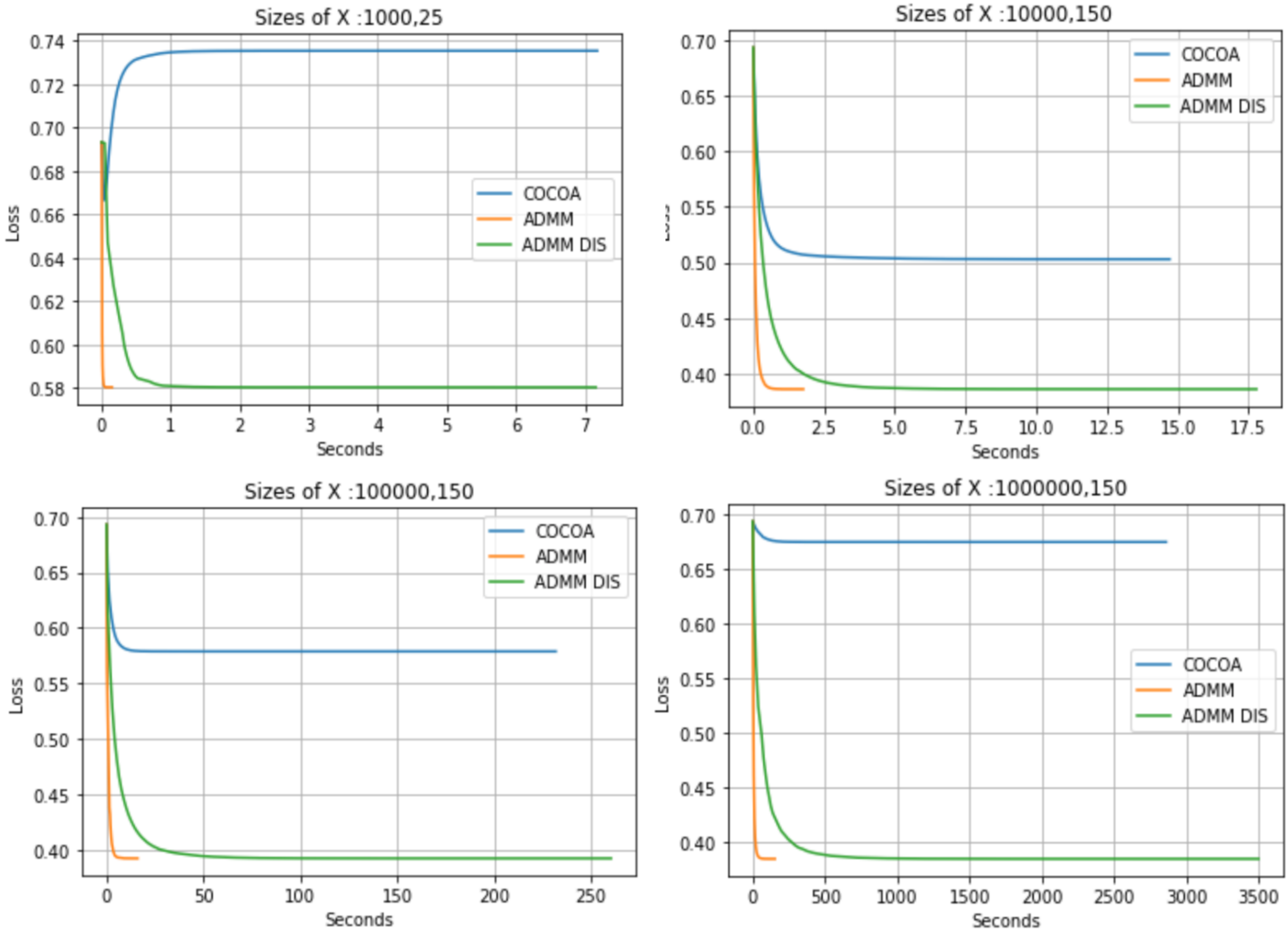
Nous avons aussi programmé deux méthodes (ADMM distribuée et non distribuée) afin de pouvoir comparer la performance des algorithme de distributions.

Vous trouverez une explication plus détaillée du programme dans le notebook.

Résultats

L'objectif de ce projet fut de montrer la performance de l'algorithme CoCoA sur un jeu de données très grand.
Nous avons donc fait évoluer la taille de notre jeu de données afin d'étudier le seuil à partir duquel il est intéressant de distribuer avec CoCoA.

Voici la comparaison des convergences des fonctions objectives pour 4 tailles de X différentes :



En termes de calcul distribué la méthode CoCoA converge plus rapidement que l'algorithme ADMM distribué à partir de 10 000 observations.
Cependant le calcul non distribué reste intéressant même au seuil du million d'observations.

Il est possible que la performance des calculs distribués soit améliorée lorsque l'on distribue sur plusieurs machines.
De plus le nombre de features contenus dans X joue aussi sur la convergence et la distribution s'imposerait si ce nombre était plus élevé.

Conclusion

Avant de proposer des solutions de distributions de calcul, il faut bien connaître son jeu de données. En effet les coûts de communications sont si élevés, qu'une distribution sur un faible échantillon sera plus couteuse en termes de temps.
La distribution est donc un outil à utiliser avec beaucoup de précaution. Cependant lorsque celle-ci est nécessaire l'algorithme CoCoA surperforme une distribution plus naïve des calculs.