1 Introduction

This document discusses how to solve feature-wise split L2-loss SVM problems using ADMM. Most contents are directly copied from the supplementary materials of Zhuang et al. (2015).

2 Details of ADMM for Logistic Regression

2.1 Feature-wise Splitting

Given J machines, the data matrix X is decomposed to J blocks, each of which contains several feature columns.

$$X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l]^T = [X_{fw,1}, \ldots, X_{fw,J}].$$

Feature-wise ADMM solves

$$\begin{aligned} \min_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_J,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_J} \quad & \frac{1}{2}\sum_{j=1}^{J} \|\boldsymbol{w}_j\|^2 + C\sum_{i=1}^{l} \max(0, 1 - \sum_{j=1}^{J} (\boldsymbol{z}_j)_i)^2 \\ \text{subject to} \quad & YX_{fw,j}\boldsymbol{w}_j = \boldsymbol{z}_j, \quad j = 1, \ldots, J, \end{aligned} \tag{1}$$

where $\boldsymbol{w}_j$ a sub-vector of $\boldsymbol{w}$ corresponding to features stored in the jth machine, $\boldsymbol{z}_j \in \mathbb{R}^{l\times 1}$, $(\boldsymbol{z}_j)_i$ refers to the ith dimension of $\boldsymbol{z}_j$, and $Y \in \mathbb{R}^{l\times l}$ is a diagonal matrix with $Y_{ii} = y_i$. In the kth iteration with the use of feature-wise splitting, ADMM sequentially performs

$$\begin{aligned} \boldsymbol{w}_j^{k+1} &= \arg\min_{\boldsymbol{w}_j} \frac{1}{2}\|\boldsymbol{w}_j\|^2 + \frac{\rho}{2}\|YX_{fw,j}\boldsymbol{w}_j - YX_{fw,j}\boldsymbol{w}_j^k - \bar{\boldsymbol{z}}^k + \frac{1}{J}\sum_{p=1}^{J} YX_{fw,p}\boldsymbol{w}_p^k + \frac{1}{\rho}\boldsymbol{\mu}^k\|^2, \\ \bar{\boldsymbol{z}}^{k+1} &= \arg\min_{\bar{\boldsymbol{z}}} C\sum_{i=1}^{l} \max(0, 1 - J\bar{z}_i)^2 + \frac{\rho J}{2}\|\bar{\boldsymbol{z}} - \frac{1}{J}\sum_{p=1}^{J} YX_{fw,p}\boldsymbol{w}_p^{k+1} - \frac{1}{\rho}\boldsymbol{\mu}^k\|^2, \\ \boldsymbol{\mu}^{k+1} &= \boldsymbol{\mu}^k + \rho(\frac{1}{J}\sum_{p=1}^{J} YX_{fw,j}\boldsymbol{w}_p^{k+1} - \bar{\boldsymbol{z}}^{k+1}). \end{aligned}$$

If we use $\boldsymbol{\mu}^k$ to denote $\boldsymbol{\mu}^k/\rho$ to remove redundant arithmetical multiplications, the updating rules can be transformed into

$$\boldsymbol{w}_j^{k+1} = \arg\min_{\boldsymbol{w}_j} \frac{1}{2}\|\boldsymbol{w}_j\|^2 + \frac{\rho}{2}\|YX_{fw,j}\boldsymbol{w}_j - YX_{fw,j}\boldsymbol{w}_j^k - \bar{\boldsymbol{z}}^k + \frac{1}{J}\sum_{p=1}^{J} YX_{fw,p}\boldsymbol{w}_p^k + \boldsymbol{\mu}^k\|^2, \tag{2}$$

$$\bar{\boldsymbol{z}}^{k+1} = \arg\min_{\bar{\boldsymbol{z}}} C\sum_{i=1}^{l} \max(0, 1 - J\bar{z}_i)^2 + \frac{\rho J}{2}\|\bar{\boldsymbol{z}} - \frac{1}{J}\sum_{p=1}^{J} YX_{fw,p}\boldsymbol{w}_p^{k+1} - \boldsymbol{\mu}^k\|^2, \tag{3}$$

$$\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k + \frac{1}{J}\sum_{p=1}^{J} YX_{fw,p}\boldsymbol{w}_p^{k+1} - \bar{\boldsymbol{z}}^{k+1}. \tag{4}$$

The optimum of (2) occurs while its gradient is zero. Therefore, the optimal condition of (2) is equivalent to a linear system

$$A_j\boldsymbol{w}_j + \boldsymbol{v}_j = 0, \tag{5}$$

where

$$A_j = I + \frac{\rho}{2}X_{fw,j}^T X_{fw,j}$$
$$\boldsymbol{v}_j = \frac{\rho}{2}X_{fw,j}^T Y(YX_{fw,j}\boldsymbol{w}_j^k - \bar{z}^k + \frac{1}{J}\sum_{p=1}^{J} YX_{fw,p}\boldsymbol{w}_p^k + \boldsymbol{\mu}^k).$$

We use standard conjugate gradient method to solve (5). That is the trcg procedure of tron without the trust region part. Note that we use $\xi = 10^{-3}$, choosing by an ad hoc, as the parameter of the stopping criterion during CG.

On the other hand, (3) is composed of l independent single-variable problems which seperatedly minimize

$$f(\bar{z}_i) = C\max(0, 1 - J\bar{z}_i)^2 + \frac{\rho J}{2}(\bar{z}_i - b_i)^2 \quad \forall i \in 1, \ldots, l,$$

where b_i is the ith component of

$$\frac{1}{J}\sum_{j=1}^{J} YX_{fw,j}\boldsymbol{w}_j^{k+1} + \boldsymbol{\mu}^k. \tag{6}$$

This decomposition implies that these l subproblems can be solved in parallel. We note that (6) is a quadratic convex function so we can set its derivative to zero to obtain the optimal solution.

$$0 = -2JC\max(0, 1 - J\bar{z}_i) + \rho J(\bar{z}_i - b_i) \Rightarrow \rho\bar{z}_i = \rho b_i + 2C\max(0, 1 - J\bar{z}_i).$$

With some simple calculations, we have

$$\bar{z}_i = \begin{cases} b_i & \text{if } Jb_i > 1, \\ \frac{2C+\rho b_i}{2CJ+\rho} & \text{otherwise.} \end{cases}$$

Following Zhuang et al. (2015), the stopping condition is set to be until

$$|f'(\bar{z}_i)| \leq 10^{-3} |f'(\bar{z}_i^0)|.$$

References

Y. Zhuang, W.-S. Chin, Y.-C. Juan, and C.-J. Lin, "Distributed Newton method for regularized logistic regression," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2015.