Algorithms Associated with Factorization Machines

Yanyu Liang
Xin Lu
Xupeng Tong

Carnegie Mellon University

{yanyul,xlu2,xtong}@andrew.cmu.edu

December 12, 2016

Overview

Factorization Machines

- Factorization machines [1] models $y \in \mathbb{R}$ given $x \in \mathbb{R}^p$ using the following expression:

$$\begin{aligned}\hat{y} &= w_0 + w^T x + \sum_{i,j} (V_i x_i)^T (V_j x_j) \\ &= w_0 + w^T x + x^T V^T V x\end{aligned}$$

  , where V_i is $k \times 1$ vector.

- FMs is widely used in recommendation system, since it implicitly regularizes the model complexity by setting a k which is much smaller than p.

Other Formulation

- $\sum_{i,j}(V_i x_i)^T(V_j x_j)$ makes FMs nonconvex, and researcher has proposed convex FMs [2] by replacing low rank constraint by trace norm (replace $V^T V$ with W).
- Also, by replacing one V in $V^T V$ with U, [3] proposed generalized FMs with an online learning algorithm

Our Goal

- Propose ADMM method to sovle convexFMs with element-wise l_1 constraint
- Overview optimization algorithm to solve classic FMs problems

ConvexFMs with l_1 constraint

- l_1 penalty is widely used and it is potential helpful in many applications beyond recommendation systems.
- Consider the regression problem in convexFMs with l_1 penalty:

$$\min_{w_0,w,W} \sum_{i=1}^{n} \underbrace{(y_i - w_0 - w^T x_i - x_i^T W x_i)^2}_{f(w_0,w,W)} + \lambda_1 \|W\|_{\mathrm{tr}} + \lambda_2 \|W\|_1 + \lambda_3 \|w\|_2^2$$

ADMM Formulation

- By introduing axilluary variable U, it can be fit into ADMM framework and the augmented Lagrangian is:

$$\mathcal{L}(w_0, w, W) = f(w_0, w, W) + \lambda_1 \|U\|_{\text{tr}} + \lambda_2 \|W\|_1 + \lambda_3 \|w\|_2^2 \\ + \langle W - U, u \rangle + m\|W - U\|_2^2$$

- Then the ADMM loop is:
  1. Update w_0, w, W:

$$w_0^k, w^k, W^k = \arg \min_{w0, w, W} f(w_0, w, W) + \lambda_2 \|W\|_1 \\ + \frac{\rho}{2}\|W - U^{k-1} + u^{k-1}\|_2^2$$

  2. Update U:

$$U^k = \arg\min_U \lambda_1 \|U\|_{\text{tr}} + \frac{\rho}{2}\|W^k - U + u^{k-1}\|_2^2$$

  3. Update u:

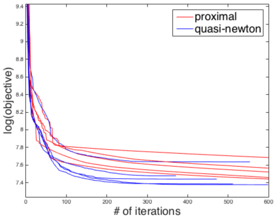
$$u^k = u^{k-1} + W^k - U^k$$

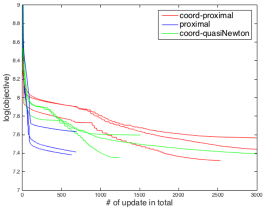
ADMM Subproblem

- The second step can be solve exactly with proximal operator of trace norm.
- We explored two approaches to solve the first subprobelm:
  - proximal graident descent on w_0, w, W (proximal)
  - blockwise coordinate descent on w_0, w and W respectively, where we applied coordinate descent on the first block and tried proximal gradient (coor-proximal) and Quasi-Newton (coor-newton) method on the second block

ADMM Results

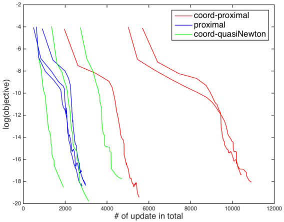
- Quasi-Newton method performs better than proximal graident descent in solving $\arg\min_W g(W)$ subproblem (as we expected)
- proximal gradient descent performs better and stable than blockwise coordinate descent

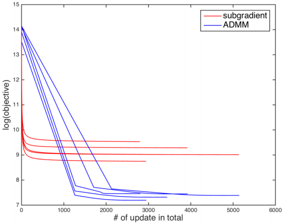




ADMM Results (con'd)

- ADMM converges and it achieves feasiblity along the path
- with the same number of updates (consider inner loop), ADMM outperforms sub-gradient method





References I

S. Rendle, "Factorization machines," in 2010 IEEE International Conference on Data Mining. IEEE, 2010, pp. 995–1000.

M. Blondel, A. Fujino, and N. Ueda, "Convex factorization machines," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2015, pp. 19–35.

M. Lin and J. Ye, "A non-convex one-pass framework for generalized factorization machine and rank-one matrix sensing," in Advances In Neural Information Processing Systems, 2016, pp. 1633–1641.

The End