Dual Formulations

October 30, 2016

1 Dual LP Boost-1

$$\begin{aligned}
&\arg\min_{\mu,\beta} \beta \\
\text{s.t} &\sum_{i=1}^{l} \Omega_i H_{ij} \leq \beta \ j = 1 \ldots m \\
&\sum_{i=1}^{l} \Omega_i = 1, \ \ 0 \leq \Omega_i \leq D \ i = 1 \ldots l
\end{aligned} \tag{1}$$

This version of the problem is amenable to distributed computing by writing the dual of the above problem by introduction of consensus variables. The consensus is done over β Consider we have K machines each catering a partition P_j of hypothesis $j = 1 \ldots K$

Now we add new variable for every machine β_j's and $\boldsymbol{\Omega_j}$'s[*] represents a vector of weights that a stored on the j^{th} machine , $j = 1 \ldots K$ and add constraints.

$$\begin{aligned}
\beta &= \beta_j \ , j = 1 \ldots K \\
\boldsymbol{\mu} &= \boldsymbol{\Omega_j} \ , j = 1 \ldots K
\end{aligned} \tag{2}$$

Now the new formulation is

$$\begin{aligned}
&\arg\min_{\beta,\mu} \beta + \sum_{k=1}^{K} \lambda_k(\beta_k - \beta) + \sum_{k=1}^{K} \boldsymbol{\psi_i}^T(\boldsymbol{\Omega_k} - \boldsymbol{\mu}) + \frac{\rho}{2}\sum_{k=1}^{K} ||\beta - \beta_k||^2 + \frac{\rho}{2}\sum_{k=1}^{K} ||\boldsymbol{\mu} - \boldsymbol{\Omega_k}||^2 \\
&\text{s.t}\forall j \in P_k \ \ \sum_{i}^{l} \Omega_{ki} y_i H_{ij} \leq \beta_k \ \ \sum_{i=1}^{l} \Omega_{ki} = 1, \ \ 0 \leq \Omega_{ki} \leq D \ i = 1 \ldots l
\end{aligned} \tag{3}$$

Update Equations for all the variables:

$$\begin{aligned}
\beta_{t+1} &= \frac{\sum_{k=1}^{K} (\lambda_k + \beta_{k,t})}{2} \\
\boldsymbol{\mu_{t+1}} &= \ldots
\end{aligned} \tag{4}$$

[*] $\boldsymbol{\mu}$ is a vector weighing all the sample points of length l. μ_i represents its i^{th} component of the consensus variable $\boldsymbol{\mu}$. And $\boldsymbol{\Omega_j}$ is the local copy of $\boldsymbol{\mu}$ at the j^{th} machine

Solve the LpBoost problem for rest of the partitions and update β_k and $\boldsymbol{\Omega_k}$ and 2 more constraints for β_k and Ω_k individually added to the optimization problem.
Issue: Do we still go for finding the max violator? or we need to considering some modifications?

$$\arg\min_{\beta_k,\boldsymbol{\Omega_k}} \beta + \lambda_k(\beta_k - \beta) + \boldsymbol{\psi_k}^T(\boldsymbol{\Omega_k} - \boldsymbol{\mu}) + \frac{\rho}{2}||\beta - \beta_k||^2 + \frac{\rho}{2}||\boldsymbol{\mu} - \boldsymbol{\Omega_k}||^2 \tag{5}$$

Update equations for the dual variables

$$\begin{aligned} \lambda_{t+1,k} &= \lambda_{t,k} + \rho(\beta_k - \beta) \\ \boldsymbol{\psi_{t+1}} &= \dots \end{aligned} \tag{6}$$

Solving the above optimization problems will require solving QP's using Column Generations. If that is done then we are good with the optimization problem.

1.1 Solving the local Linear Constrained Quadratic Objective

Consider the partition of hypothesis P_k

$$\begin{aligned} &\arg\min_{\beta_k,\boldsymbol{\Omega_k}} \beta_k + \lambda_k(\beta_k - \beta) + \boldsymbol{\psi_k}^T(\boldsymbol{\Omega_k} - \boldsymbol{\mu}) + \frac{\rho}{2}||\beta - \beta_k||^2 + \frac{\rho}{2}||\boldsymbol{\mu} - \boldsymbol{\Omega_k}||^2 \\ &\text{s.t}\forall j \in P_k \quad \sum_i^l \Omega_{ki} y_i H_{ij} \leq \beta_k \quad \sum_{i=1}^l \Omega_{ki} = 1, \ \ 0 \leq \Omega_i \leq D \ i = 1 \dots l \end{aligned} \tag{7}$$

The above problem is a good candidate for Analytic Center Cutting Plane Method (ACCPM). This will ensure we still have the idea of column generation in place.

The output of of ACCPM would be adding a constraint (a new hypothesis) to the of already existing set.

Steps for Cutting plane algorithm:

1. Worker Node receives β and $\boldsymbol{\mu}$ from the master.
2. H_k be the current hypothesis set under consideration at worker k
3. h $\notin H_0$ be the maximum violating constraint from the hypothesis partition(P_k) at worker k
4. Add h to the set H_0 and with constraints other sum and box constraints from 7
5. Feed the above problem with quadratic objective and minimize for β_k and Ω_k.
6. Continue going to step 3 until the maximum violating hypothesis is ϵ more else go to step 7
7. Update dual variables λ and $\boldsymbol{\psi_k}$ go to step 1

1.2 KKT Conditions

Formulate the dual of the above problem.

$$F_0 = \beta_k + \lambda_k(\beta_k - \beta) + \boldsymbol{\psi_k}^T(\boldsymbol{\Omega_k} - \boldsymbol{\mu}) + \frac{\rho}{2}||\beta - \beta_k||^2 + \frac{\rho}{2}||\boldsymbol{\mu} - \boldsymbol{\Omega_k}||^2$$

$$\text{s.t}\forall j \in P_k \;\; \sum_i^l \Omega_{ki} y_i H_{ij} \leq \beta_k \;\; \sum_{i=1}^l \Omega_{ki} = 1, \;\; 0 \leq \Omega_i \leq D \; i = 1 \ldots l$$

$$L(\beta_k, \boldsymbol{\omega_k}, \boldsymbol{\delta}, \alpha, \boldsymbol{\nu}, \boldsymbol{\eta}) = F_0 + \sum_{j \in P_k} \delta_j (\sum_i^l \Omega_{ki} y_i H_{ij} - \beta_k) + \alpha(\sum_{i=1}^l \Omega_{ki} - 1) + \sum_i^l \nu_i(-\Omega_{ki}) + \sum_i^l \eta_i(\Omega_{ki} - D) \tag{8}$$

Stationarity:

$$\nabla L_{\beta_k} = 1 + \lambda_k + \rho(\beta_k - \beta) - l \sum_{j \in P_k} \delta_j = 0 \tag{9}$$

$$\nabla L_{\Omega_{ki}} = \psi_{ki} + \rho(\omega_{ki} - u_i) + \alpha \sum_{j \in P_k} \delta_j y_i H_{ij} + \alpha - \nu_i + \eta_i = 0 \tag{10}$$

Complemantary Slackness

$$\forall j \in P_k \quad \delta_j (\sum_i^l \Omega_{ki} y_i H_{ij} - \beta_k) = 0 \tag{11}$$

$$\forall i \in 1 \ldots l \quad \nu_i(-\omega_{ki}) = 0 \tag{12}$$

$$\forall i \in 1 \ldots l \quad \eta_i(\Omega_{ki} - D) = 0 \tag{13}$$

Primal Feasibilty

$$\forall j \in P_k \quad \sum_i^l \Omega_{ki} y_i H_{ij} \leq \beta_k \tag{14}$$

$$\sum_{i=1}^l \Omega_{ki} = 1, \quad 0 \leq \Omega_i \leq D \quad i = 1 \ldots l \tag{15}$$

Dual Feasibility

$$\forall i \in 1 \ldots l \quad \delta_i \geq 0 \quad \nu_i \geq 0 \quad \eta_i \geq 0 \tag{16}$$

2 Primal Formulation

$$
\begin{aligned}
&\min \sum_{i=1}^{m} a_i + C \sum_{i=1}^{l} \eta_i \\
&\text{s.t } y_i H_i a + \eta_i \geq 1 \ , \ \eta_i \geq 0 \ \ , \ i = 1 \dots l \\
&\quad a_i \geq 0, \ \ , i = 1, \dots, m
\end{aligned} \tag{17}
$$

We are considering a model in which we have data partitioning in place.
Issue : The LpBoost problem is solved by invoking the dual of the problem and then trying to add hypothesis to it. And as we know that dual and primal only meet at the optimization point of the soln. So the problem here is that we need a consensus over the variable $\boldsymbol{a}$. The variable $\boldsymbol{a}$ is computed after solving the dual problem many number of times.

Now considering data is partitioned in different sub parts each partiotioned is labeled P_i where i $\in$ 1 to K. Consider solving the LpBoost problem form every partition and then getting a consensus on the a which is computed at the end. This is an ideal setting for ADMM where a set of machines ltry to solve a similiar problem. Inspired from SVM and ADMM. On similar lines we formulate our problem to solve LpBoost with data partitioning. Assume $\boldsymbol{w_i}$ is a local copy of $\boldsymbol{a}$ at every machine.

$$
\begin{aligned}
&\min \sum_{i=1}^{m} a_i + C \sum_{i=1}^{l} \eta_i + \frac{\rho}{2} \sum_{i=1}^{k} ||\boldsymbol{w_i} - \boldsymbol{a}||^2 \\
&\text{s.t } y_i H_i a + \eta_i \geq 1 \ , \ \eta_i \geq 0, \quad i = 1 \dots l \\
&\quad a_i \geq 0, \quad i = 1 \dots m \\
&\quad \boldsymbol{w_j} = \boldsymbol{a}, \quad \forall j \in 1 \dots Q
\end{aligned} \tag{18}
$$

Now adding the lagrangian to the the problem.

$$
L(w, a, \lambda) = \sum_{i=1}^{m} a_i + C \sum_{i=1}^{l} \eta_i + \sum_{i=1}^{k} \left(\frac{\rho}{2} ||\boldsymbol{w_i} - \boldsymbol{a}||^2 + \boldsymbol{\lambda_j^T}(\boldsymbol{a} - \boldsymbol{w_j})\right) \tag{19}
$$

ADMM consist of the following iterations

$$
\begin{aligned}
w^{k+1} &= \arg\min L(w, a^k, \lambda^k) \\
a^{k+1} &= \arg\min L(w^{k+1}, a, \lambda^k) \\
\lambda_j^{k+1} &= \lambda_j^k + \rho(\boldsymbol{w_j} - \boldsymbol{a^{k+1}})
\end{aligned} \tag{20}
$$

Solutions for solving a^{k+1} is closed form. Solving for w^{k+1} is an optimization problem that needs to be solved in a different way for every partition $P_i \in 1 \dots Q$

$$
\begin{aligned}
w_q^{k+1} &= \arg\min_{w} C \sum_{i \in P_q} \eta_i + \frac{\rho}{2} ||\boldsymbol{w_q} - \boldsymbol{a^k}||^2 + \boldsymbol{\lambda_q^T}(\boldsymbol{a^k} - \boldsymbol{w_q}) \\
&\text{s.t } y_i H_i w_q + \eta_i \geq 1 \ , \ \eta_i \geq 0, \quad i \in P_q
\end{aligned} \tag{21}
$$

This approach changes the objective function and but having the term with L2-Norm doesnot support sparsity of the primal space(the weights for the hypothesis)