A Dual and Primal Coordinate Descent Method for Linear SVM

Prathyusha Mallela*
MS in Computer Science (Specialization: Artificial Intelligence)
Oregon State University
Corvallis, Oregon 97331
mallelap@oregonstate.edu

December 9th, 2022

Abstract

This paper demonstrates the coordinate descent for the dual of Linear SVM. It applies a constant value for regularisation and the time taken for training is compared among the primal and the dual of SVM which uses a coordinate descent algorithm. The data used is the breast cancer, iris dataset which is imported from sklearn. This binary classification algorithm has been used to classify benign and malignant tumors, and versicolor and virginica respectively. It was observed that for small datasets which are not sparse, a constant regularisation factor works splendidly.

1 Introduction

SVM is one of the best known techniques for binary classification. SVMs find the maximal margin between the two projected convex hulls by solving quadratic programming problem (QPP) in the dual space. One of the disadvantages of the classical SVM is that it takes significant amount of time to train the dataset, and to locate the optimum parameter.Even though fast algorithms have been used to make primal SVMs to overcome the slow training, there are other methods that propel the training speed. One of them is through solving the dual of the QPP. The dual of QPP can be solved and made faster to train through successive overrelaxation algorithm and dual coordinate descent algorithm [1]. The study focuses on running a coordinate descent algorithm for both primal and dual forms on small datasets which are not sparse. The study wanted to extend to using L1 and L2 regularisation as mention in [2] and utilization of BFGS along with SVM. For a set of instance-label pairs $(\boldsymbol{x}_i, y_i), i = 1, \ldots, l, \boldsymbol{x}_i \in R^n, y_i \in \{-1, +1\}$, the primal form of SVM is given below:

$$\min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\xi\left(\boldsymbol{w};\boldsymbol{x}_i,y_i\right)$$

where $\xi\left(\boldsymbol{w};\boldsymbol{x}_i,y_i\right)$ is a loss function, and $C \geq 0$ is a penalty parameter. The two common loss functions which are used are Lasso and Ridge. Lasso adds absolute value of the magnitude as coefficient term to the loss function where as Ridge is the square magnitude of the coefficient term. A good scaling multiple for Ridge is important as this could determine in making the model overfit or underfit the data.

$$\max\left(1-y_i\boldsymbol{w}^T\boldsymbol{x}_i,0\right) \text{ and } \max\left(1-y_i\boldsymbol{w}^T\boldsymbol{x}_i,0\right)^2$$

The dual problem of SVM from [2] is:

$$\min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T\bar{Q}\boldsymbol{\alpha} - \boldsymbol{e}^T\boldsymbol{\alpha}$$
$$\text{subject to } 0 \leq \alpha_i \leq U, \forall i$$

The paper [2] discusses using dual coordinate descent on SVMs. Previous studies show that coordinated descent on the primal for of L2 SVM yielded in models which were trained at faster speeds. The dual problem here is twice differentiable for both L1 and L2 SVM whereas the primal problem is not twice differentiable. In [2] the authors Cho-Jui Hsieh et al., 2008, propose and successfully implement coordinate descent for the dual problem. They used the L1 and L2 regularisation techniques to demonstrate further that the coordinate descent on the dual form of SVM was faster while training and finding the optimum. My proposed experiment wanted to establish that even without the regularisation parameters and irrespective of the type of data, the dual form of the SVM along with the coordinate descent algorithm is more agile than the coordinate descent applied on the primal form. In [2] the authors perform the tests on sparse document classification data. I chose a dataset that isn't sparse, and didn't apply any regularisation as I didn't want to overfit or underfit data. I've used the packages available at the source [3], which contain coordinate descent for both primal and dual forms of SVM.

1.1 Problem: Dual Coordinate Descent Method

Here the coordinate descent method for L1 and L2 SVM is discussed. Suppose the optimization process starts at an initial point and generates a sequence of vectors, where each vector has 'l' elements in a sequence of 'k'. The above statement is put into mathematical notations from [2] as: $\boldsymbol{\alpha}^{k,i} \in R^l, i = 1,\ldots,l+1$, such that $\boldsymbol{\alpha}^{k,1} = \boldsymbol{\alpha}^k, \boldsymbol{\alpha}^{k,l+1} = \boldsymbol{\alpha}^{k+1}$, and

$$\boldsymbol{\alpha}^{k,i} = \left[\alpha_1^{k+1},\ldots,\alpha_{i-1}^{k+1},\alpha_i^k,\ldots,\alpha_l^k\right]^T, \forall i = 2,\ldots,l.$$

Given $\boldsymbol{\alpha}$ and the corresponding $\boldsymbol{w} = \sum_i y_i\alpha_i\boldsymbol{x}_i$ While $\boldsymbol{\alpha}$ is not optimal

For $i = 1, \ldots, l$
(1) $\bar{\alpha}_i \leftarrow \alpha_i$
(2) $G = y_i \boldsymbol{w}^T \boldsymbol{x}_i - 1 + D_{ii}\alpha_i$
(3)

$$ProjectedGradient = \begin{cases} \min(G, 0) & \text{if } \alpha_i = 0, \\ \max(G, 0) & \text{if } \alpha_i = U, \\ G & \text{if } 0 < \alpha_i < U \end{cases}$$

(4) If $|ProjectedGradient| \neq 0$

$$\alpha_i \leftarrow \min\left(\max\left(\alpha_i - G/\bar{Q}_{ii}, 0\right), U\right)$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \left(\alpha_i - \bar{\alpha}_i\right) y_i \boldsymbol{x}_i$$

The parameters that need to be computed are: $\bar{Q}_{ii}$ and $\nabla_i f\left(\boldsymbol{\alpha}^{k,i}\right)$. First, $\bar{Q}_{ii} = \boldsymbol{x}_i^T \boldsymbol{x}_i + D_{ii}$ They can be stored as global variables. Next to compute: $\nabla_i f\left(\boldsymbol{\alpha}^{k,i}\right)$, use

$$\nabla_i f(\boldsymbol{\alpha}) = (\bar{Q}\boldsymbol{\alpha})_i - 1 = \sum_{j=1}^{l} \bar{Q}_{ij}\alpha_j - 1$$

To check if alpha is optimal, re-computation of the whole gradient must be performed to check if it has the optimal solution. In order to override the gradient computation a method called shrinking implementation could be used. In [2] the authors also discuss the shrinking method which is combined with the dual coordinate descent with respective L1 and L2 regularisation. Here, at each iterative step of the gradient, the max of the gradient is iterated till the present value chosen is conditionally shrunken. This max value must be positive. To shrink it even further these shrinking values could be further scaled down by multiplying them with smaller ratios, which are lesser than 1. The shrinking value is kept under surveillance by holding a tolerance parameter. Here shrinking can be done without the reconstruction of gradients.

2 Hypothesis and experiment

My Hypothesis was to prove that L1 and L2 regularization needn't be applied to small datasets. I wanted to understand how the coordinate descent would work on small data sets that were not sparse. I wanted to use L1 and L2 regularisation, and the Quasi-Newton method to compare and check if they made a substantial impact on the training speed and mean average accuracy. The experiment was performed on breast cancer and iris data set that was available through the sklearn library. Breast cancer data set had 569 data points and from iris 105 data points were chosen as the classifier was trained to capture 'versicolor' and 'virginica'. Library and packages from source [3] were

used to test the hypothesis. The coordinate descent in the package did not use any regularisation. The comparison was done among the primal and the dual coordinate descent SVM. The dual coordinate descent SVM showed better accuracy rates for far less training time. The values in [Table 4.1] demonstrate that it is possible that the dual coordinate descent could overfit the data for huge datasets and will require regularisation parameters in order to make it an efficient model for verifying and testing datasets.

3 Experiment

Datasets used were not sparse but had fewer dimensions. The packages and libraries used were taken from a prebuilt source [4], they are available at: [https://github.com/lenassero/linear-svm]. The datasets chosen for the tests were breast cancer(569 data points) and iris (105 data points, for versicolor and verginica binary classifier) datasets. Coordinate descent (primal) Linear SVM, and Coordinate descent (dual) Linear SVM were applied on these datasets without any regularisations as the datasets were small. Observed quantities to make inferences were- primal time taken to train, primal accuracy on the training set, primal number of iterations, dual number of iterations, accuracy by the dual on the training set, time taken to train by the dual on the training set, and additionally the time taken by the dual to find the optimal solution.

4 Results

Dual Coordinate descent irrespective of the size of the data set had better performance than coordinate descent for primal SVM. The dataset size and type influenced the training time, for finding optimum solution. The number of iterations used to converge was also dependent on the size of the dataset. The training accuracy makes it obvious that there is a strong need to use a regularization parameter that does not overfit the model to the test data.

Dataset	PI	PMA	PTime	DI	DMA	DTime	DOptTime
Breast Cancer	16	0.86	0.64	16	0.965	0.651	0.202
Iris	9	0.94	0.06	10	0.97	0.053	0.0109

Table 4.1: PI is Primal Iterations, PMA is Primal Mean Accuracy for training, PTime is Primal Training Time, DI is Iteration for Dual, DMA is Dual Mean Accuracy for training, DTime is Dual Training Time, DOptTime is Time for Dual Coordinate Descent to find the Optimal value. Breast cancer has 569 data points and for the purpose of training a binary classifier only 105 data points of the iris data set were used.

5 Conclusion and Future Work

Dual coordinate descent for dual converges faster with better accuracy. The time and the training accuracy are dependent on the size of the data. It could be concluded that the models require L1, L2 or other regularisation techniques whose accuracy, and the trained models are greatly dependent on the size of the data, the type of the data- sparse or not sparse. Different data types, linear, non-linear, sparse, not sparse, high dimensional data, low dimensional data, data with less data points, data with high data points. It would be ideal to test which of these binary classifiers work best in situations where there is severe class imbalance. BFGS with SVM could also be a great place to start a case study and pick on the intuition and the symbiosis between the data, its type, and the kind of SVM classifiers used along with its optimizers.
Algorithm for the optimizers to be used:
1. Choose optimizer
2. Pass the respective parameters, initialize learning
3. Compute loss
4. Zero gradients
5. Compute gradients
6. optimize for a step
7. Set a threshold to a desired iterative value
8. Loop 3 to 6 till the threshold/ iterative value is satisfied.

6 References

[1] Aryan Mokhtari, and Alejandro Ribeiro, "A quasi-newton method for large scale support vector machines," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[2] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen. Lin, S. Sathiya Keerthi, and S. Sundararajan, "A dual coordinate descent method for large-scale linear SVM," Proceedings of the 25th international conference on Machine learning - ICML '08, 2008.

[3] Nasser Benabderrazik, "Lenassero/linear-SVM: Linear SVM optimization from scratch.," GitHub. [Online]. Available: https://github.com/lenassero/linear-svm. [Accessed: 09-Dec-2022].

[4] Xinjung Peng, Dongjing Chen, and Lingyan Kong, "A clipping dual coordinate descent algorithm for solving support Vector Machines," Knowledge-Based Systems, vol. 71, pp. 266–278, 2014.