CS771 Introduction to Machine Learning
Assignment 1

Group No. 7
Avesh Kumar Agrawal(19111020)
Kamlesh Kumar Biloniya(160317)
Sanjeev Lal(19111077)
Shashank Kumar(19111084)
Shrikant Jhajhra(19111089)

1 Lagrangian of P2

$$\underset{W\in R^d \xi\in R^n}{\arg\min}\ \frac{1}{2}\left\|W^2\right\| + C\sum_{i=1}^{n}\xi_i^2$$

$$s.t.\ y^i\langle \mathbf{W}, \mathbf{X}^i\rangle \geq 1-\xi_i\ for all\ i \in [n] \qquad (P2)$$

Let's Rewrite conditions in less than equal to format

$$1-\xi_i - y^i\langle \mathbf{W}, \mathbf{X}^i\rangle \leq 0$$

Lagrangian of P2 can be written as

$$\mathscr{L}(\mathbf{W},\xi,\alpha) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\alpha_i(1-\xi_i - y^i\langle \mathbf{W}, \mathbf{X}^i\rangle)$$

2 Dual of P2

The Lagrangian problem to solve

$$\underset{W\in R^d \xi\in R^n}{\arg\min}\ \underset{\alpha\geq 0}{\arg\max} \qquad \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\alpha_i(1-\xi_i - y^i\langle \mathbf{W}, \mathbf{X}^i\rangle) \qquad (1)$$

$$\frac{\partial \mathscr{L}(\mathbf{W},\xi,\alpha)}{\partial \mathbf{W}} = \mathbf{W} - \sum_{i=1}^{n}\alpha_i y^i \mathbf{X}^i = 0$$

$$\Rightarrow\ \mathbf{W} = \sum_{i=1}^{n}\alpha_i y^i \mathbf{X}^i$$

$$\frac{\partial \mathscr{L}(\mathbf{X},\xi,\alpha)}{\partial \xi_i} = 2C\xi_i - \alpha_i = 0$$

$$\Rightarrow\ 2C\xi_i = \alpha_i$$

$$\Rightarrow\ \xi_i = \frac{\alpha_i}{2C}$$

On simplifying (1)

$$\arg\min_{W\in R^d \xi\in R^n} \max_{\alpha\geq 0} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\alpha_i y^i\langle \mathbf{W}, \mathbf{X}^i\rangle \tag{2}$$

Putting values of $\mathbf{W}$ and ξ_i in (2)

$$\arg\max_{\alpha\geq 0} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j\langle \mathbf{X}^i, \mathbf{X}^j\rangle + C\sum_{i=1}^{n}(\frac{\alpha_i}{2C})^2 - \sum_{i=1}^{n}\alpha_i(\frac{\alpha_i}{2C})$$

$$\arg\max_{\alpha\geq 0} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j\langle \mathbf{X}^i, \mathbf{X}^j\rangle - \sum_{i=1}^{n}\frac{\alpha_i^2}{4C} \tag{$D2$}$$

3 Details of Implemented methods

3.1 Coordinate Maximization on D2

From D2 choosing only i^{th} coordinate of α_i :

$$\arg\max_{\alpha\geq 0}(\alpha_i - \frac{\alpha_i^2}{4C} - \frac{\alpha_i^2\|X\|^2}{2} - \alpha_i y^i\sum_{i\neq j}\alpha_j y^j\langle \mathbf{X}^j, \mathbf{X}^j\rangle)$$

Differentiating above equation w.r.t α_i and equating to zero we get equation of α_i as below :

$$\alpha_i^{new} = \frac{1 - y^i(\langle \mathbf{W}, \mathbf{X}^i\rangle + b) + \alpha_i\|X\|^2}{\frac{1}{2C} + \|X\|^2}$$

$$W^{new} = \alpha_i^{new} * y^i * X^i$$

$$b^{new} = \alpha_i^{new} * y^i$$

We have chosen getRandompermCoord() method for getting next random coordinates

3.2 Coordinate Ascent on D2

From D2 choosing only i^{th} coordinate of α_i :

$$h = \arg\max_{\alpha\geq 0}(\alpha_i - \frac{\alpha_i^2}{4C} - \frac{\alpha_i^2\|X^i\|^2}{2} - \alpha_i y^i\sum_{i\neq j}\alpha_j y^j\langle \mathbf{X}^j, \mathbf{X}^j\rangle)$$

Equation of Gradient :

$$\frac{dh}{d\alpha_i} = 1 - \frac{\alpha_i}{2C} - \|X\|^2\alpha_i - (y^i(\langle \mathbf{w}, \mathbf{X}^i\rangle + b) - \alpha_i\|X^i\|^2)$$

$$\frac{dh}{d\alpha_i} = 1 - \frac{\alpha_i}{2C} - y^i(\langle \mathbf{w}, \mathbf{X}^i\rangle + b)$$

Now,

$$\alpha_i^{new} = \alpha_i + \eta \frac{dh}{d\alpha_i}$$

$$W^{new} = \alpha_i^{new} * y^i * X^i$$

$$b^{new} = \alpha_i^{new} * y^i$$

We have taken initial value of η as 2.5 and updating the step length as :

$$\eta^{new} = \eta^{old} * 0.1$$

We have chosen getRandompermCoord() method for getting next random coordinates

3.3 Stochastic Gradient Descent on P1

$$f(W) = \arg\min_{W \in R^d} \frac{1}{2} \left\| W^2 \right\| + C \sum_{i=1}^{n} \left[1 - y^i \langle \mathbf{W}, \mathbf{X}^i \rangle \right]_+^2$$

Differentiating above equation w.r.t W , we get :

$$\nabla f(W) = W + 2C \sum_{i=1}^{n} g^i y^i X^i \left[1 - y^i \langle \mathbf{W}, \mathbf{X}^i \rangle \right]_+$$

$$where \quad g^i \in \nabla l_{hinge}(y^i . W^T X^i)$$

Concentrating on i^{th} coordinate of X i.e. X^i

$$\nabla f(W) = W + 2C\sqrt{n} g^i y^i X^i \left[1 - y^i \langle \mathbf{W}, \mathbf{X}^i \rangle \right]_+$$

Similarly

$$\nabla f(b) = 2C\sqrt{n} g^i y^i \left[1 - y^i \langle \mathbf{W}, \mathbf{X}^i \rangle \right]_+$$

$$W^{new} = W^{old} - \eta(\nabla f(W))$$

$$b^{new} = b^{old} - \eta(\nabla f(b))$$

We have taken initial value of η as 0.001 and updating the step length as :

$$\eta^{new} = \eta^{old} / \sqrt{t+1}$$

where, t = current iteration number

We have chosen getRandompermCoord() method for getting next random coordinate

4 Testing on the Implemented Methods

We trained our model on randomly selected 75% (15000) data points and tested on 25% (5000) data points. For all three Algorithms we ran our trained model for 1,5,10,20,30 seconds on every selected choice of random function for evaluating next random coordinate with different step length functions.

4.1 Coordinate Maximization on D2

We have tried following three methods to choose random coordinates :
- getRandpermCoord
- getRandCoord
- getCyclicCoord

From the tables we can see that for the choice of "getRandpermCoord" method our Primal objective value is conversing fast and with minimum value as compared to other methods.

Time(sec)	getRandpermCoord	getRandCoord	getCyclicCoord
1	647909	87427	302981
5	1692	2151	67365
10	1668	1668	24978
20	1662	1668	9502
30	1662	1668	33398

Table 1: Coordinate Maximization

4.2 Coordinate Ascent on D2

We have tried following three methods to choose random coordinates :
- getRandpermCoord
- getRandCoord
- getCyclicCoord

but found that "getRandpermCoord" method is giving best value of Primal objective value.

From the tables, we can also see that for the $\eta = 2.5$ and step length $\eta = \eta/10$, Primal objective value is conversing fast and with minimum value as compared to other methods.

Time(sec)	$\eta 1 = 2.5$	$\eta 2 = 2$	$\eta 3 = 3$
1	14811384	58142648	12016607
5	1357246	72378	1122961
10	32458	7739	361305
20	51134	414502	23131
30	3071	20514	3972

Table 2: Coordinate Ascent with $\eta = \frac{\eta}{\sqrt{t+1}}$

Time(sec)	$\eta 1 = 2.5$	$\eta 2 = 2$	$\eta 3 = 3$
1	24556	7129	41140
5	1690	3006	1724
10	1669	1668	1668
20	1668	1668	1668
30	1668	1668	1668

Table 3: Coordinate Ascent with $\eta = \frac{\eta}{10}$

Time(sec)	$\eta 1 = 2.5$	$\eta 2 = 2$	$\eta 3 = 3$
1	12043	4498	4839
5	2241	3789	2923
10	2382	2203	2336
20	2284	1808	1712
30	1720	1721	1678

Table 4: Coordinate Ascent with $\eta = \frac{\eta}{100}$

4.3 SGD on P1

We have tried following three methods to choose random coordinates :
- getRandpermCoord
- getRandCoord
- getCyclicCoord

but found that "getRandpermCoord" method is giving best value of Primal objective value.

From the tables, we can also see that for the $\eta = 0.001$ and step length $\eta = \frac{\eta}{\sqrt{t+1}}$, Primal objective value is conversing fast and with minimum value as compared to other methods.

Time(sec)	$\eta 1 = 0.001$	$\eta 2 = 0.01$	$\eta 3 = 0.1$
1	1411.06	1371.17	2929.24
5	1357.51	1353.728	1426.32
10	1349.35	1347.80	1369.86
20	1346.23	1345.03	1363.10
30	1345.51	1348.39	137.73

Table 5: SGD with $\eta = \frac{\eta}{\sqrt{t+1}}$

Time(sec)	$\eta 1 = 0.001$	$\eta 2 = 0.01$	$\eta 3 = 0.1$
1	1760.96	1651.67	3216105.54
5	1762.30	1538.818	3056862.77
10	1747.30	1576.13	1586.13
20	1751.76	1549.46	3929.44
30	1746.37	1560.65	3182.68

Table 6: SGD with $\eta = \frac{\eta}{t+1}$

5 Graph for the Implemented Methods

To draw the Graph, we used full data set to train the model i.e. all 20K data points and graph is showing objective value of primal objective function with respect to time. From the graph, we can see that **Coordinate Maximization Algorithm** is giving us minimum Primal objective value i.e. **5223**.

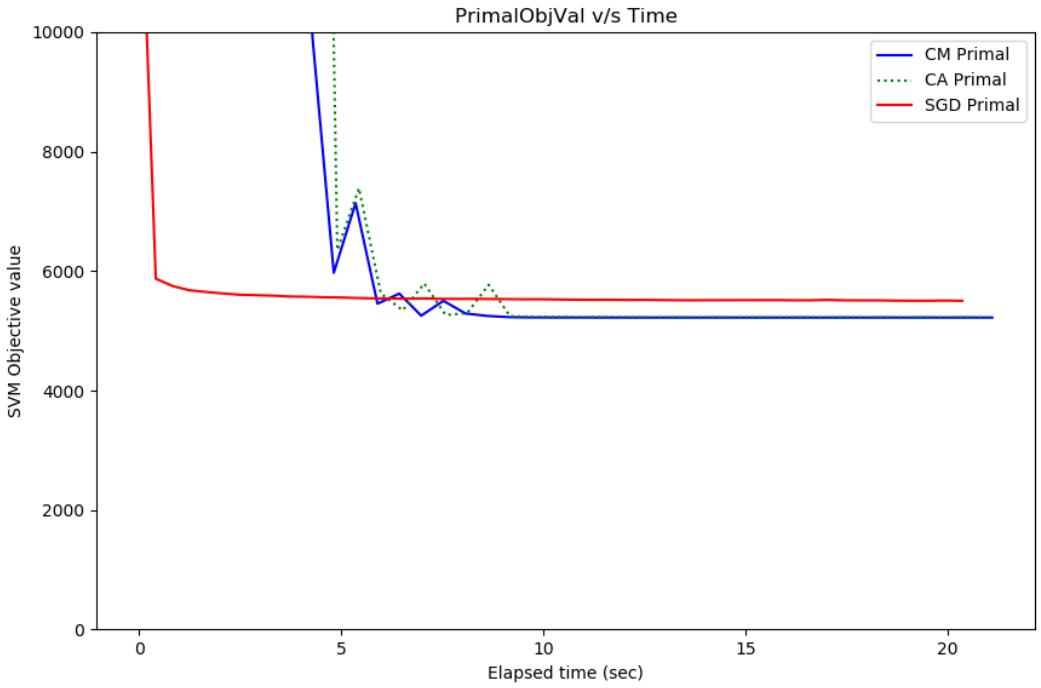

Figure 1: Convergence curves of Primal Objective Function

6 Code for Coordinate Maximization on D2

`https://www.cse.iitk.ac.in/users/kamlesh/cs771/submit.py`

7 Bonus Problem

Optimization problem

$$\underset{W\in R^d \xi\in R^n}{\arg\min} \frac{1}{2}\left\|W^2\right\| + C\sum_{i=1}^{n}\xi^2$$

subject to

$$y^i\langle \mathbf{W}, \mathbf{X}^i\rangle \geq 1-\xi_i \quad for all\, i \in [n]$$
$$\xi_i \geq 0$$

Converting to less than equal to form

$$1-\xi_i - y^i\langle \mathbf{W}, \mathbf{X}^i\rangle \leq 0 \quad for all\, i \in [n]$$
$$-\xi_i \leq 0$$

Introducing Lagrange Multipliers $\alpha_i, \beta_i (i \in [n])$ for constraints

$$\mathscr{L}(\mathbf{W}, \xi, \alpha, \beta) = \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\alpha_i(1-\xi_i - y^i\langle \mathbf{W}, \mathbf{X}^i\rangle) - \sum_{i=1}^{n}\beta_i\xi_i \tag{3}$$

Differentiating $\mathscr{L}(\mathbf{W}, \xi, \alpha, \beta)$ with respect to W

$$\frac{\partial \mathscr{L}(\mathbf{W}, \xi, \alpha)}{\partial \mathbf{W}} = \mathbf{W} - \sum_{i=1}^{n}\alpha_i y^i \mathbf{X}^i = 0$$

$$\Rightarrow \mathbf{W} = \sum_{i=1}^{n} \alpha_i y^i \mathbf{x}^i$$

Differentiating $\mathscr{L}(\mathbf{W}, \xi, \alpha, \beta)$ with respect to ξ_i

$$\frac{\partial \mathscr{L}(\mathbf{W}, \xi, \alpha)}{\partial \xi_i} = 2C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \beta_i = 0$$

$$\Rightarrow 2C\xi_i = \alpha_i + \beta_i$$

$$\Rightarrow \xi_i = \frac{\alpha_i + \beta_i}{2C}$$

Putting values of $\mathbf{W}$ and ξ_i in (3)

$$\underset{\alpha \geq 0, \beta \geq 0}{\arg\max} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j \langle \mathbf{X}^i, \mathbf{X}^j \rangle + \frac{1}{4C} \sum_{i=1}^{n} (\alpha_i + \beta_i)^2 - \sum_{i=1}^{n} \alpha_i (\frac{\alpha_i + \beta_i}{2C}) - \sum_{i=1}^{n} \beta_i (\frac{\alpha_i + \beta_i}{2C})$$

On simplifying

$$\underset{\alpha \geq 0, \beta \geq 0}{\arg\max} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j \langle \mathbf{X}^i, \mathbf{X}^j \rangle - \frac{1}{4C} \sum_{i=1}^{n} \alpha_i^2 - \sum_{i=1}^{n} \frac{\beta_i^2}{4C} - \sum_{i=1}^{n} \frac{\alpha_i \beta_i}{2C}$$

$$\underset{\alpha \geq 0, \beta \geq 0}{\arg\max} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j \langle \mathbf{X}^i, \mathbf{X}^j \rangle - \frac{1}{4C} \sum_{i=1}^{n} (\alpha_i + \beta_i)^2$$

As we know that, α and β both are greater than zero, so $(\alpha + \beta)$ is also greater than zero, then we can represent $(\alpha + \beta)$ as a single constant represented by same α

So, we get final equation as below :

$$\underset{\alpha \geq 0}{\arg\max} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^i y^j \langle \mathbf{X}^i, \mathbf{X}^j \rangle - \frac{1}{4C} \sum_{i=1}^{n} \alpha_i^2 \qquad (4)$$

Hence equation (4) is same as equation (D2)