CSE 512 - Machine Learning HW2

FAIZAAN CHARANIA - 111463646

Q1

The objective function for Ridge Regression is

$$\min_{\omega, b} \lambda\|\omega\|^2 + \sum_{i=1}^{n} (\omega^T x_i + b - y_i)^2 \quad \text{— (i)}$$

1.1 Converting the above eq^n in vector form, we get

$$\min_{\bar{\omega}} \lambda\|\bar{\omega}\|^2 + \|\bar{X}^T\bar{\omega} - y\|^2 \quad \text{— (ii)}$$

To solve for $\bar{\omega}$, we take derivative of (ii) w.r.t $\bar{\omega}$ & equate to 0.

$$2\lambda\bar{\omega} + 2\bar{X}(\bar{X}^T\bar{\omega} - y) = 0$$

$$\lambda\bar{\omega} + \bar{X}(\bar{X}^T\bar{\omega} - y) = 0$$

$$\bar{X}y = (\bar{X}\bar{X}^T - \lambda\bar{I})\bar{\omega}$$

$$\bar{\omega} = (\bar{X}\bar{X}^T - \lambda I)^{-1}\bar{X}y$$

given $d = \bar{X}y$ & $C = (\bar{X}\bar{X}^T - \lambda\bar{I})$

$\cancel{\bar{\omega} = (\bar{X}\bar{X}^T - \lambda I}$

$$\therefore \boxed{\bar{\omega} = C^{-1} d}$$

Hence, proved.

1.2 Let $C_i, d_i, \bar{\omega}_i$ correspond to matrices when we remove training data $\bar{x}_i$

For C_i, each term will be missing a factor of $\bar{x}_i$ which is $\bar{x}_i \bar{x}_i^T$

$$\therefore C_i = C - \bar{x}_i \bar{x}_i^T \quad \text{—(iii)}$$

Similarly, for d_i we get

$$d_i = d - \bar{x}_i y_i \quad \text{—(iv)}$$

1.3

The Sherman-Morrison formula is

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u} \quad \text{—(v)}$$

We know that

$$C_i^{-1} = (C - \bar{x}_i \bar{x}_i^T)^{-1} \quad \text{(from (iii))}$$

From (v), we get,

$$(C - x_i x_i^T)^{-1} = C^{-1} - \frac{C^{-1}(-\bar{x}_i \bar{x}_i^T)C^{-1}}{1 + \bar{x}_i^T C^{-1}(-\bar{x}_i)}$$

$$\therefore \quad C_i^{-1} = C^{-1} + \frac{C^{-1} \bar{x}_i \bar{x}_i^T C^{-1}}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} \quad \text{— (vi)}$$

Q1.4

We know that, $w = C^{-1} d$

$$\therefore w_i = C_i^{-1} d_i$$

$\therefore$ from (vi) & (iv)

$$\bar{w}_i = C_i^{-1} d_i = \left[C^{-1} + \frac{C^{-1} \bar{x}_i \bar{x}_i^T C^{-1}}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} \right] (d - x_i y_i)$$

$$\bar{w}_i = \left[C^{-1} d - C^{-1} x_i y_i + \frac{C^{-1} \bar{x}_i \bar{x}_i^T C^{-1} d}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} - \frac{C^{-1} \bar{x}_i \bar{x}_i^T C^{-1} \bar{x}_i y_i}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} \right]$$

$\therefore \quad \bar{\omega}^* = C^{-1}d$

$$\bar{\omega}_i = \bar{\omega} + \frac{C^{-1}\bar{x}_i\bar{x}_i^T\bar{\omega}}{1-\bar{x}^T C^{-1}\bar{x}} - C^{-1}\bar{x}_i y_i - \frac{C^{-1}\bar{x}_i\bar{x}_i^T C^{-1} x_i y_i}{1-\bar{x}_i^T C^{-1}\bar{x}_i}$$

$$= \bar{\omega} + \frac{C^{-1}\bar{x}_i\bar{x}_i^T\bar{\omega}}{1-\bar{x}_i^T C^{-1}\bar{x}} - \frac{C^{-1}\bar{x}_i y_i - C^{-1}\bar{x}_i \cancel{y_i \bar{x}_i^T C^{-1}\bar{x}_i y_i + C^{-1}\bar{x}_i\bar{x}_i^T C^{-1}}\bar{x}_i y_i}{1-\bar{x}_i^T C^{-1}\bar{x}_i}$$

$$= \bar{\omega} + C^{-1}\bar{x}_i\left[\frac{-y_i + \bar{x}_i^T\bar{\omega}}{1-\bar{x}_i^T C^{-1}\bar{x}_i}\right]$$

$$\therefore \bar{\omega}_i = \bar{\omega} + C^{-1}\bar{x}_i\left[\frac{-y_i + \bar{x}_i^T\omega}{1-\bar{x}^T C^{-1}\bar{x}}\right] \quad \text{— (vii)}$$

Hence, proved

1.5 Leave one out error can be written as

$$\bar{x}_i^T \bar{w}_i - y_i \qquad (\because \bar{w}^T \bar{x}_i = \bar{x}_i^T w)$$

from (iii) we know that

$$\bar{w}_i = \bar{w} + \frac{(C^{-1}\bar{x}_i)(-y_i + \bar{x}_i^T \bar{w})}{1 - \bar{x}_i^T C^{-1} \bar{x}_i}$$

$$\therefore \bar{x}_i^T \bar{w}_i - y_i = \bar{x}_i^T \left[\bar{w} + \frac{(C^{-1}\bar{x}_i)(-y_i + \bar{x}_i^T \bar{w})}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} \right] - y_i$$

$$= \bar{w}^T \bar{x}_i + \frac{(\bar{x}_i^T C^{-1} \bar{x}_i)(-y_i + \bar{x}_i^T \bar{w})}{1 - \bar{x}_i^T C^{-1} \bar{x}_i} - y_i$$

$$= \frac{\bar{w}^T \bar{x}_i - \bar{w}^T \cancel{\bar{x}_i \bar{x}_i^T} C^{-1} \bar{x}_i + \bar{x}_i C^{-1} \cancel{\bar{x}_i^T} (-y) + \bar{x}_i C^{-1} \cancel{\bar{x}_i} \bar{x}_i^T \bar{w} - y_i + y_i \cancel{\bar{x}_i^T C^{-1} \bar{x}_i}}{1 - \bar{x}_i^T C^{-1} \bar{x}_i}$$

$$= \frac{\bar{w}^T \bar{x}_i - y_i}{1 - \bar{x}^T C^{-1} \bar{x}_i}$$

$$\therefore \bar{x}^T \bar{w}_i - y_i = \frac{\bar{w}^T \bar{x}_i - y_i}{1 - \bar{x}^T C^{-1} \bar{x}_i}$$

1.6

In the previous question, we get

$$\bar{\omega}_i^T x - y_i = \frac{\bar{\omega}^T x_i - y_i}{1 - \bar{z}_i^T C^{-1} \bar{x}}$$

In this we need to calculate "C^{-1}" once, with the complexity k^3 & we do the error calculation 'n' times with matrix multiplication complexity k^2

$\therefore$ Complexity $= nk^2 + k^3$

In the usual way of computing ~~we will have to~~ per iteration complexity $= k^3$ (for inverse of c)

$\therefore$ Total complexity $= nk^3$ (for n iterations)

Q2

2.1 ~~Na~~ N.B with Boolean & continuous variables

$X \to Y$

$X = (X_1, X_2)$

$X_1 \sim \text{Bernoulli}(\theta_{x_i})$

$X_2 \sim \mathcal{N}(\mu_i, \sigma_i)$ $\quad \forall i \quad i = \{0,1\}$

X_1 & X_2 have 3 parameters for each class (2 classes)

$\therefore$ num-params = $3 \times 2 = 6$ (Params for X)

Total Params = ~~Σ~~ Params for X + Param for (Y)

$= 6 + 1 = 7$

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(x)}$$

$$= \frac{P(X_1|Y) \cdot P(X_2|Y) \cdot P(Y)}{P(X_1|Y=0) \cdot P(X_2|Y=0) \cdot P(Y=0) + P(X_1|Y=1) \cdot P(X_2|Y=1) \cdot P(Y=1)}$$

For $Y=0$, we get,

$$P(Y=0|x) = \frac{\theta_{x_1 0} \times \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \times (1-\theta_y)}{\left[\theta_{x_1 0} \times \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \times (1-\theta_y)\right] + \left[\theta_{x_1 1} \times \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \times \theta_y\right]}$$

Q2

2.1 ~~Na~~ N.B with Boolean & continuous variables

$X \to Y$

$X = (X_1, X_2)$

$X_1 \sim \text{Bernoulli}(\theta_{x_i^{\circ}})$

$X_2 \sim \mathcal{N}(\mu_i, \sigma_i)$ $\forall_i \; i = \{0,1\}$

X_1 & X_2 have 3 parameters for each class (2 classes)

$\therefore$ num-paras $= 3 \times 2 = 6$ (Params for X)

Total Params = ~~S~~ Params for X + Param for (Y)

$= 6 + 1 = 7$

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(x)}$$

$$= \frac{P(X_1|Y) \cdot P(X_2|Y) \cdot P(Y)}{P(X_1|Y=0) \cdot P(X_2|Y=0) \cdot P(Y=0) + P(X_1|Y=1) \cdot P(X_2|Y=1) \cdot P(Y=1)}$$

For $Y=0$, we get,

$$P(Y=0|x) = \frac{\theta_{x_1 0} \times \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \times (1-\theta_y)}{\left[\theta_{x_1 0} \times \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} \times (1-\theta_y)\right] + \left[\theta_{x_{11}} \times \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \times \theta_y\right]}$$

Similary, we can find $P(Y=1|X)$

$$P(Y=1|X) = \frac{\theta_{x11} \times \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \times \theta_y}{\left[\theta_{x11} \times \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \times \theta_y\right] + \left[\theta_{x10} \times \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}} \times (1-\theta_y)\right]}$$

2.2

$X = (x_1 \dots x_d)$

$X_i = \text{Bernoulli}(\gamma_{ij}) \quad \forall i = \{1 \dots d\} \quad \forall j = \{0,1\}$

$Y = \text{Bernoulli}(\gamma_y)$

$$P(Y=1|X) = \frac{P(X|Y=1) \cdot P(Y=1)}{P(X|Y=1) \cdot P(Y=1) + P(X|Y=0) \cdot P(Y=0)}$$

$$= \frac{1}{1 + \exp\left(\log\left[\frac{P(Y=0) \cdot P(X|Y=0)}{P(Y=1) \cdot P(X|Y=1)}\right]\right)}$$

Consider, $P(X_i|Y=k) = \cancel{\theta_{ik}^{x_i}(1-\theta_{ik})^{(1-x_i)}}$

$$= \gamma_{ik}^{x_i}(1-\gamma_{ik})^{(1-x_i)}$$

$$P(X_i \mid Y=k) = \frac{1}{1 + \exp\left[\log\left(\frac{1-\gamma_y}{\gamma_y}\right) + \sum \log\left(\frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)}\right)\right]}$$

$$= \frac{1}{1 + \exp\left[\log\left(\frac{1-\gamma_y}{\gamma_y}\right) + \sum_{i=1}^{d} \log\left(\frac{\gamma_{i0}^{x_i} \cdot (1-\gamma_{i0})^{(1-x_i)}}{\gamma_{i1}^{x_i} \, (1-\gamma_{i1})^{(1-x_i)}}\right)\right]}$$

$$= \frac{1}{1 + \exp\left(\underbrace{\log\left(\frac{1-\gamma_y}{\gamma_y}\right) + \sum_{i=1}^{d} \log\left(\frac{1-\gamma_{i0}}{1-\gamma_{i1}}\right)}_{-\theta_{d+1}} + \sum_{i=1}^{d} x_i \underbrace{\left[\log \frac{\gamma_{i0}}{\gamma_{i1}} - \log\left(\frac{1-\gamma_{i0}}{1-\gamma_{i1}}\right)\right]}_{-\theta_i}\right.}$$

$$\therefore P(Y=1 \mid X) = \frac{1}{1 + \exp\left(-\left(\sum_{i=1}^{d} \theta_i x_i + \theta_{d+1}\right)\right)}$$

$$\theta_{d+1} = -\left[\log\left(\frac{1-\gamma_y}{\gamma_y}\right) + \sum_{i=1}^{d} \log\left(\frac{1-\gamma_{i0}}{1-\gamma_{i1}}\right)\right] = \log\left(\frac{\gamma_y}{1-\gamma_y}\right) + \sum_{i=1}^{d} \log\left(\frac{1-\gamma_{i1}}{1-\gamma_{i0}}\right)$$

$$\theta_i = -\left[\log\left(\frac{\gamma_0}{\gamma_1}\right) - \log\left(\frac{1-\gamma_{i0}}{1-\gamma_{i1}}\right)\right] = \log\left(\frac{\gamma_1}{\gamma_0}\right) - \log\left(\frac{1-\gamma_{i1}}{1-\gamma_{i0}}\right)$$

Q 3

3.1 Kernel SVM using quadprog

1. Based on the function definition of quadprog in MATLAB, we get

$H = \text{diag}(Y) \times \text{linear\_kernel}(X) \cdot \text{diag}(Y)$

$f = -1 \times \text{ones}(1, n)$

$$\left.\begin{array}{l} A = [\ ] \\ b = [\ ] \end{array}\right\} \because \text{there are no inequality constraints}$$

$beq = 0$

$Aeq = Y^T$

$ub = C \times \text{ones}(n, 1)$

$lb = \text{zeros}(n, 1)$

where n = no. of samples
Y = Training label vector $(n \times 1)$
X = Training data matrix $(d \times n)$

$\text{linear\_kernel}(X) = X^T X$

5) $C = 10$

Validation Accuracy : 97.27

Objective value of SVM : 112.1461

No. of support vectors : 123

Confusion matrix :

$$\begin{array}{c|cc} & -1 & 1 \\ \hline -1 & 178 & 6 \\ 1 & 4 & 179 \end{array}$$

Q3.2

$$L_i = \frac{1}{2n} \sum_{j=1}^{k} \|w_j\|^2 + C\, L(W, x_i, y_i)$$

1)

$$\frac{\partial L_i}{\partial w_{y_i}} = \frac{1}{\cancel{2}n} \times \cancel{2}\, w_{y_i} + C \frac{\partial}{\partial w_{y_i}} L(W, x_i, y_i) \quad \text{—(viii)}$$

$$\because L(W, x_i, y_i) = \max(w_{\hat{y}_i}^T x - w_{y_i} + 1, 0) \quad \text{—(ix)}$$

where $\hat{y}_i = \underset{j \neq y_i}{\operatorname{argmax}} (w_j^T x_i)$

Since (ix) is not differentiable,

we get

$$\cancel{\frac{\partial}{\partial w_{y_i}} L(W, x_i, y_i)}$$

$$\therefore \frac{\partial L_i}{\partial w_{\hat{y}_i}} = \frac{1}{\cancel{2}n} \times \cancel{2} w_{\hat{y}_i} + C \times \frac{\partial}{\partial w_{\hat{y}_i}} L(W, x_i, y_i)$$

$$= \begin{cases} w_{\hat{y}_i}/n + C x_i & , w_{\hat{y}_i}^T x_i - w_{y_i}^T x_i + 1 > 0 \\ w_{\hat{y}_i}/n & , \text{otherwise} \end{cases}$$

3.

$$L_i = \frac{1}{2n} \sum_{j=1}^{k} ||w_j||^2 + C \times L(W, x_i, y_i)$$

consider $j \neq y_i$ & $j \neq \hat{y}_i$

$$\frac{\partial L_i}{\partial w_j} = \frac{1}{2n} \times 2 w_j + C \times \frac{\partial}{\partial w_j} L(W, x_i, y_i)$$

$\because L(W, x_i, y_i)$ is independent of w_j (from (ix))

$$\frac{\partial L(W, x_i, y_i)}{\partial w_j} = 0$$

$$\therefore \frac{\partial L_i}{\partial w_j} = \frac{w_j}{n}$$

4. code submitted

5. Objective value

	3.1	3.2
C = 0.1	24.76	19.802
C = 10	112.14	80.62

Plots attached. (~~Compared to the value in 3.1.4~~
SGD ~~performs marginally better in t~~) X

6.

6.

Considering $C = 10$ (Assuming prediction error is loss or ~~&~~ accuracy)

~~a)~~

~~Accur~~	Hinge Loss	Accuracy
a) Validation	56.14	97.27
b) Train	2.01	100

c) $\sum_{j=1}^{k} \|w_j\|^2 = \cancel{60.64}\ 16.10$

7) Validation accuracy = 80.33%

Test accuracy = 73.2%

Parameters

epochs = 30

~~eta~~ $\eta_0 = 0.05$

$\eta_1 = 100$

$C = 0.07$

batch size = 20

(Plot Attached)

Q4

4.4.1 AP = 0.5785 (validation data)

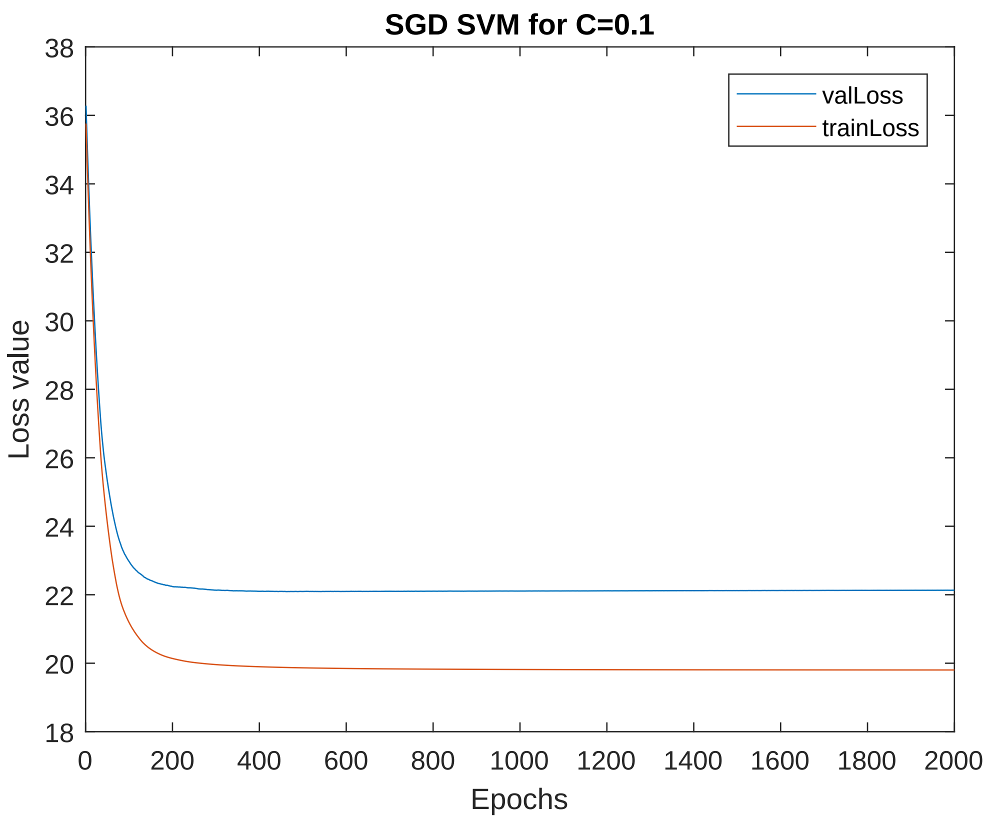
Precision-recall plot attached.

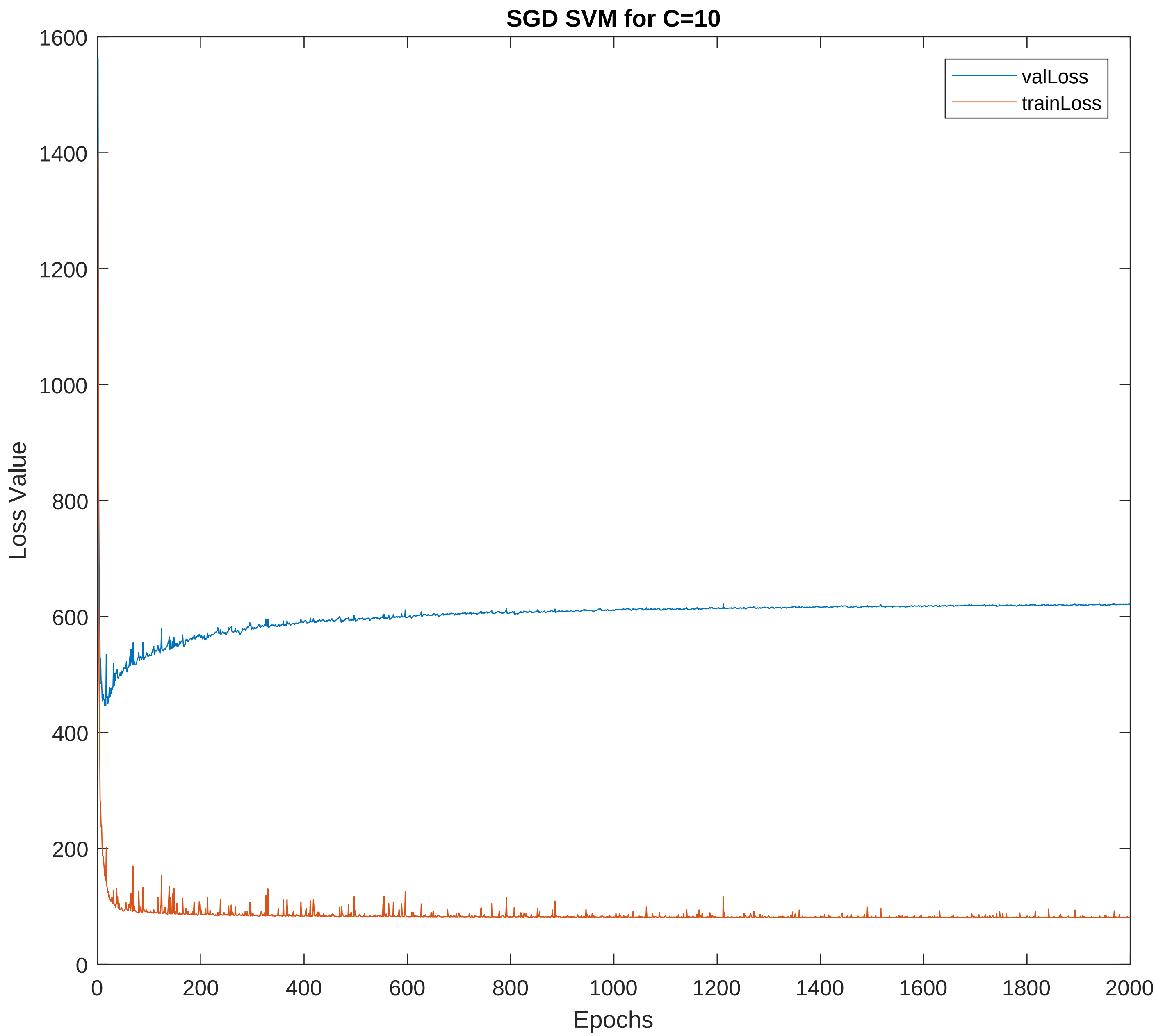
4.4.2 code submitted

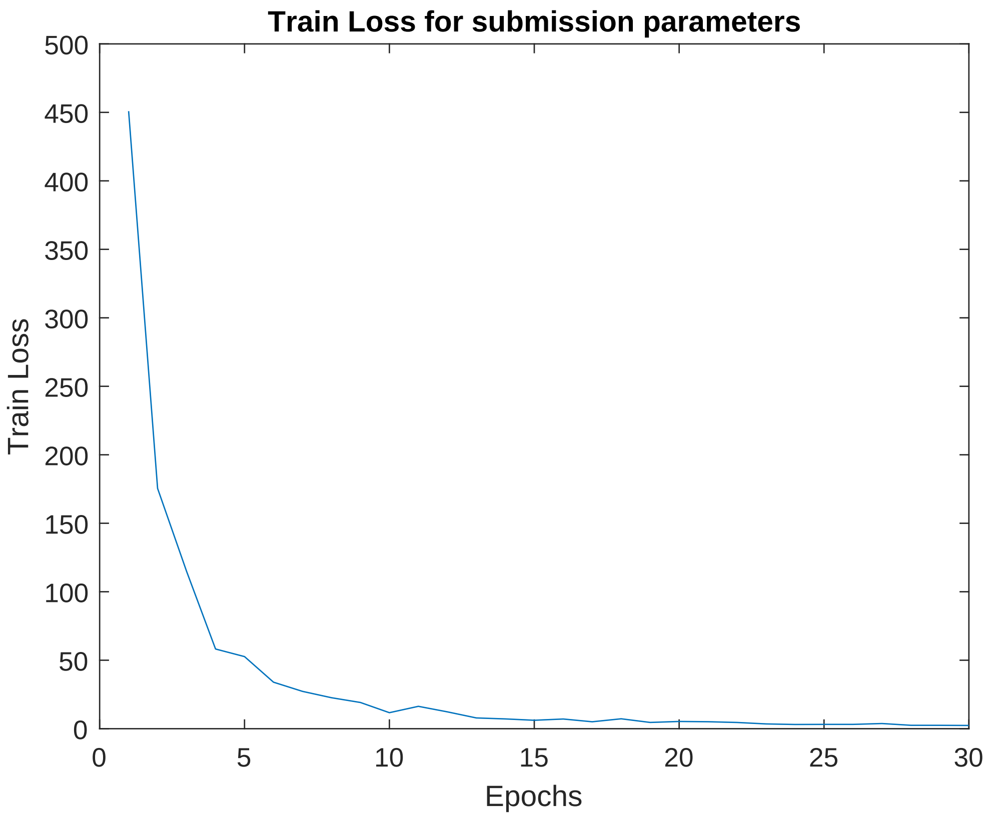
4.4.3 Plots attached

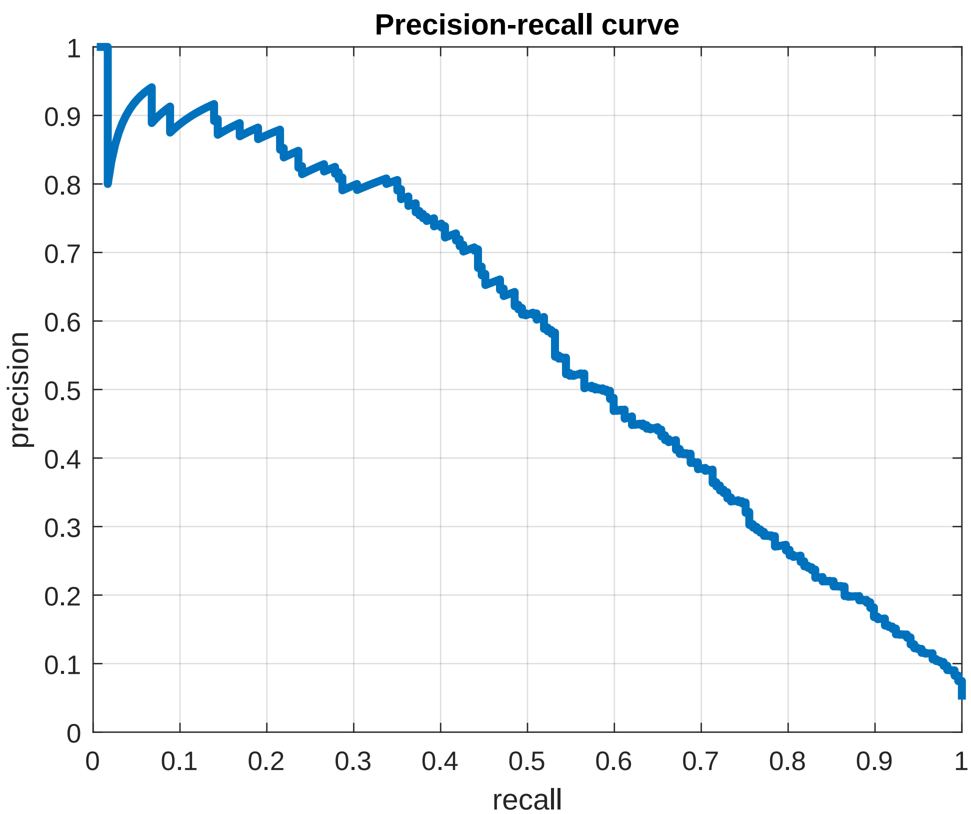
4.4.4 AP% = 84.90%

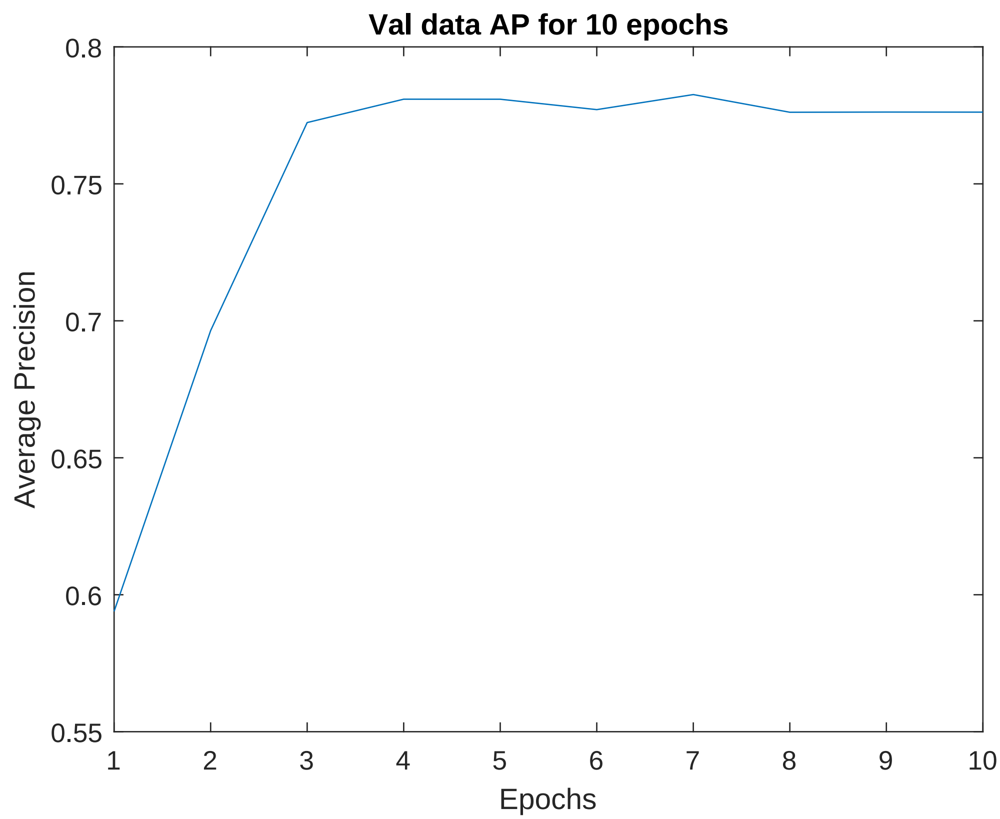
SBU ID → 111463646

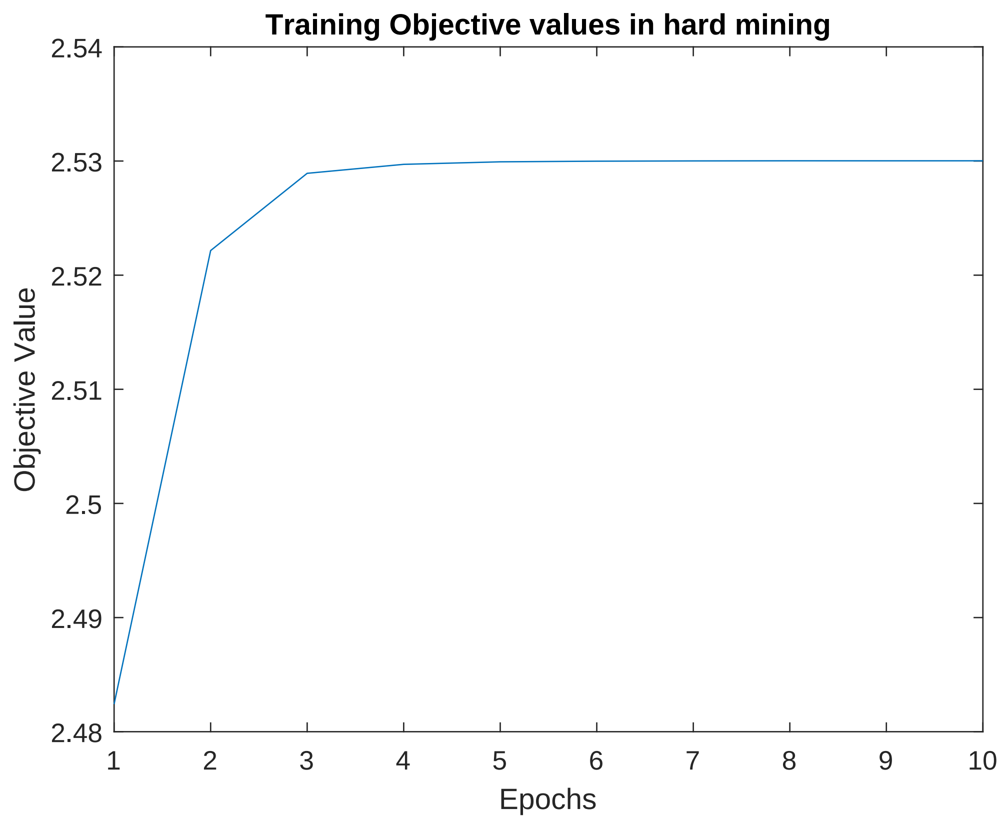

SGD SVM for C=0.1
valLoss
trainLoss
Loss value
Epochs
38
36
34
32
30
28
26
24
22
20
18
0
200
400
600
800
1000
1200
1400
1600
1800
2000


SGD SVM for C=10
valLoss
trainLoss
Loss Value
Epochs
0
200
400
600
800
1000
1200
1400
1600
1800
2000


Train Loss for submission parameters
Train Loss
Epochs
500
450
400
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30


Precision-recall curve
precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Val data AP for 10 epochs
Average Precision
0.8
0.75
0.7
0.65
0.6
0.55
1
2
3
4
5
6
7
8
9
10
Epochs


Training Objective values in hard mining
Objective Value
Epochs
2.54
2.53
2.52
2.51
2.5
2.49
2.48
1
2
3
4
5
6
7
8
9
10