Q1.1 $\underset{w,b}{\text{minimize}} \; \lambda \|w\|^2 + \sum_{i=1}^{n} (w^T x_i + b - y_i)^2$

minimize $\lambda w^T w + (Y - X^T w)^2$

$$L = \lambda w^T w + (Y - X^T w)^T (Y - X^T w)$$

$$\frac{\partial L}{\partial w} = 2\lambda w - 2X(Y - X^T w) = 0$$

$$\lambda w - XY + XX^T w = 0$$

$$\lambda w + XX^T w = XY \cancel{=0}$$

$$w(XX^T + \lambda I) = XY$$

$$\boxed{w = (XX^T + \lambda I)^{-1} XY}$$ Hence Proved

www.PrintablePaper.net

Q1.2)

$$d_{(i)} = d - x_i y_i$$

$$C_{(i)} = C - x_i x_i^T$$

(Q1:3)

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

$$C = xx^T + \lambda I^T$$

$$C_{(i)}^{-1} = (C - x_i x_i^T)^{-1}$$

$$\boxed{C_{(i)}^{-1} = C^{-1} + \frac{C^{-1}x_i x_i^T C^{-1}}{1 - x_i^T C^{-1} x_i}}$$

Q1.4

$$w = C^{-1} d$$
$$w_i = C_i^{-1} d_i$$

$$C_i^{-1} = (C - x_i x_i^T)^{-1}$$
$$d_i = d - x_i y_i$$

$$w_i = \left[C^{-1} + \frac{C^{-1} x_i x_i^T C^{-1}}{1 - x_i^T C^{-1} x_i} \right] \left[d - x_i y_i \right]$$

$$w_i = C^{-1} d - C^{-1} x_i y_i + \frac{C^{-1} x_i x_i^T C^{-1} d - C^{-1} x_i x_i^T C^{-1} x_i y_i}{1 - x_i^T C^{-1} x_i}$$

$$= w - \frac{C^{-1} x_i y_i + C^{-1} x_i y_i x_i^T C^{-1} x_i + C^{-1} x_i x_i^T C^{-1} d - C^{-1} x_i x_i^T C^{-1} x_i y_i}{1 - x_i^T C^{-1} x_i}$$

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

$$\boxed{w_i = w + \frac{C^{-1} x_i \left[-y_i + x_i^T w \right]}{1 - x_i^T C^{-1} x_i}}$$ Hence proved

Q1.5

From previous solution:

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

To prove: $\bar{w}_i^T\bar{x}_i - \bar{y}_i = \dfrac{\bar{w}^T x_i - y_i}{1 - x_i^T C^{-1} x_i}$

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

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

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

Hence proved.

Q1.6

$$w_{(i)}^T x_i - y_i = \frac{w^T x_i - y_i}{1 - x_i^T C^{-1} x_i}$$

$$C = (xx^T + \lambda I)$$

where dimensions of x are $(k+1) \times n$
y are $n \times 1$
w are $(k+1) \times (1)$
C are $(k+1) \times (k+1)$

where k is the number of features & n is the number of instances

The complexities of computing $w^T x_i = O(k+1)$
$C = n(k+1)^2$
$C^{-1} = (k+1)^3$
$x_i^T C^{-1} x_i = (k+1)^2$

Overall Complexity $= O[(k+1) + (k+1)^2 n + (k+1)^3 + (k+1)^2]$

This is for 1 instance, for n instances it will be
$$O[n(k+1) + \cancel{n^2(k+1)^2} + \cancel{n(k+1)^3} + n(k+1)^2] + O(n(k+1)^2 + \cancel{n}(k+1)^3)$$
as C & C^{-1} are calculated just once

$\therefore$ Overall Complexity $= \cancel{O}[$
$$= O[(k+1)^3 + n(k+1)^2]$$ (1)

From eq(i) in the question paper, we have

$$\bar{w} = C^{-1}d$$
$$\bar{w} = (\bar{x}\bar{x}^T + \lambda I)^{-1}\bar{x}y$$

where dimensions of x are $(k+1)(n-1)$
y are $(n-1)\times 1$
w are $(k+1)\times 1$

The complexity for computing xx^T is $(n-1)(k+1)^2 \sim n(k+1)^2$
inverse of $xx^T + \lambda I$ is $(k+1)^3$
multiplying $\bar{x}Y$ is $(k+1)(n-1) \sim n(k+1)$
multiplying inverse $\bar{x}Y$ is $(k+1)^2$

Total complexity $= O[n(k+1)^2 + (k+1)^3 + n(k+1) + (k+1)^2]$

But for n LOOCV, time complexity $= O[n^2(k+1)^2 + n(k+1)^3 + n^2(k+1) + n(k+1)^2]$

Overall Complexity $= O[n^2(k+1)^2 + n(k+1)^3]$ — (iii)

From eq (ii) & (iii), it can be seen that algorithmic complexity of computing LOOCV error using the formula in question 1.5 is less than computing it the usual way.

Q2.1 $P(Y=y|X) = \underset{y}{\text{argmax}}\ P(Y=y).\ P(X|Y=y)$

$= \text{argmax}\ P(Y=y).\ \prod P(X_i|Y=y)$

$= \alpha\ P(Y=y).\prod P(X_i|Y=y)$ — ①

since X_1 & X_2 are independent of each.

Since X_2 is a continuous variable, we have

$$P(X=X_2|Y=y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left[-\frac{(X_2-\mu_y)^2}{2\sigma_y^2}\right]$$

But for $y=0$ & $y=1$, we have

$$P(X=X_2|Y=0) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{(X_2-\mu_0)^2}{2\sigma_0^2}\right]$$

$$P(X=X_2|Y=1) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{(X_2-\mu_1)^2}{2\sigma_1^2}\right]$$

So, we have 4 parameters to estimate for $P(X=X_2|Y=y)$ those are σ_0, σ_1, μ_1, μ_0.

Now, since $X=X_1$ is a boolean variable

$$P(X=X_1|Y=y) = \theta_y^{X_1}(1-\theta_y)^{(1-X_1)}$$

But for $y=0$ & $y=1$, we have

$$P(X=X_1|Y=0) = \theta_0^{X_1}(1-\theta_0)^{(1-X_1)}$$

$$P(X=X_1|Y=1) = \theta_1^{X_1}(1-\theta_1)^{(1-X_1)}$$

So, for $X = x_1$, we have 2 parameters to estimate for $P(X = x_1 \mid Y = y)$, the parameters are θ_0 & θ_1

$\therefore$ Eq① becomes

$$P(Y=y|x) = \alpha \cdot P(Y=y) \cdot \left[\theta_y^{x_1} (1-\theta_y)^{(1-x_1)}\right] \left[\frac{1}{\sigma_y\sqrt{2\pi}} \exp\left[\frac{-(x_2-\mu_y)^2}{2\sigma_y^2}\right]\right]$$

Q2.2

1) $P(Y=1|X) = \dfrac{P(X|Y=1).P(Y=1)}{P(X|Y=1).P(Y=1) + P(X|Y=0).P(Y=0)}$

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

$$= \frac{1}{1 + \exp\left(\ln\left(\dfrac{P(X|Y=0).P(Y=0)}{P(X|Y=1).P(Y=1)}\right)\right)}$$

Let $P(Y=0) = \gamma$, we get

$$= \frac{1}{1 + \exp\left[\ln\dfrac{\gamma}{1-\gamma} + \sum_i \ln\left(\dfrac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)\right]}$$

As $x_1, x_2, \ldots, x_d$ is a vector of ~~Binomial~~ Boolean variables, they follow binomial distribution

$$P(X_i|Y=0) = \theta_{i0}^{x_i}(1-\theta_{i0})^{(1-x_i)}$$

$$P(X_i|Y=1) = \theta_{i1}^{x_i}(1-\theta_{i1})^{(1-x_i)}$$

$$\therefore\ P(Y=1|X) = \frac{1}{1 + \exp\left[\ln\dfrac{\gamma}{1-\gamma} + \sum_i \ln\dfrac{\theta_{i0}^{x_i}(1-\theta_{i0})^{(1-x_i)}}{\theta_{i1}^{x_i}(1-\theta_{i1})^{(1-x_i)}}\right]}$$

$$= \frac{1}{1+\exp\left[\ln\frac{\gamma}{1-\gamma} + \sum_i x_i \ln\frac{\theta_{i0}}{\theta_{i1}} + (1-x_i)\ln\frac{(1-\theta_{i0})}{(1-\theta_{i1})}\right]}$$

$$= \frac{1}{1+\exp\left[\ln\frac{\gamma}{1-\gamma} + \sum_i \ln\frac{(1-\theta_{i0})}{(1-\theta_{i1})} + \sum_i x_i\left[\ln\frac{\theta_{i0}}{\theta_{i1}} - \ln\frac{(1-\theta_{i0})}{(1-\theta_{i1})}\right]\right]}$$

Let $w_0 = \ln\frac{\gamma}{1-\gamma} + \sum_i \ln\frac{(1-\theta_{i0})}{(1-\theta_{i1})}$

$w_i = \ln\frac{\theta_{i0}}{\theta_{i1}} - \ln\frac{(1-\theta_{i0})}{(1-\theta_{i1})}$

$$P(Y=1|x) = \frac{1}{1+\exp(w_0 + \sum_i w_i x_i)}$$

Which is similar to Logistic Regression.

Q3.1.1

Q3.1.1

$$\max_{\alpha} \sum_{j=1}^{n} \alpha_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i\alpha_i y_j\alpha_j x_i x_j$$

$$\min_{\alpha} -\sum_{j=1}^{n} \alpha_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i\alpha_i y_j\alpha_j x_i x_j$$

$$\text{s.t.} \quad \sum_{j=1}^{n} y_j\alpha_j = 0$$

$$0 \le \alpha_j \le c \quad \forall j$$

$$\min_{x} \frac{1}{2} x^T H x + f^T x$$

$$\text{s.t.} \quad Ax \le b$$
$$Aeq.x = beq$$
$$lb \le x \le ub$$

$$x = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}_{n\times 1}$$

W.K.T. $\sum_{j=1}^{n} y_j\alpha_j = 0$

Also $Aeq.x = beq$

so.

$$Aeq = \begin{bmatrix} y_1 \ldots & y_n \end{bmatrix}_{1\times n}$$

$$beq = 0$$

$$f = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}_{n\times 1}$$

Also $0 \le \alpha_j \le c \quad \forall j$

So $lb \le x \le ub$

$A = []$

$b = []$

$$lb = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1} \qquad ub = \begin{bmatrix} c \\ \vdots \\ c \end{bmatrix}_{n\times 1}$$

$H = x_i x_j y_i y_j$

Q3.2

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

$$L(w, x_i, y_i) = \max\{w_{\hat{y}_i}^T x_i - w_{y_i}^T x_i + 1, 0\}$$

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

1) Subgradient of L_i w.r.t w_{y_i}

If $w_{\hat{y}_i}^T x_i - w_{y_i}^T x_i + 1 \leq 0$

then $\frac{\partial L_i}{\partial w_{y_i}} = \frac{w_{y_i}}{n}$

else

$$\frac{\partial L}{\partial w_{y_i}} = \frac{w_{y_i}}{n} - c x_i$$

2) Subgradient of L_i w.r.t $w_{\hat{y}_i}$

If $w_{\hat{y}_i}^T x_i - w_{y_i}^T x_i + 1 \leq 0$

then $\frac{\partial L_i}{\partial w_{\hat{y}_i}} = \frac{w_{\hat{y}_i}}{n}$

else $\frac{\partial L}{\partial w_{\hat{y}_i}} = \frac{w_{\hat{y}_i}}{n} + c x_i$

3 Subgradient of L_i w.r.t w_j for $j \neq y_i$ & $j \neq \hat{y}_i$

$$\frac{\partial L}{\partial w_j} = \frac{w_j}{n}$$

Question 3.1:

For C = 10 and epsilon = 0.1:

Accuracy: 0.978202

Objective function value: 112.1461

Confusion matrix :

179 4

4 180

Number of support vectors: 119

For C = 0.1 and epsilon = 0.0001:

Accuracy: 0.907357

Objective function value: 24.7648

Confusion matrix :

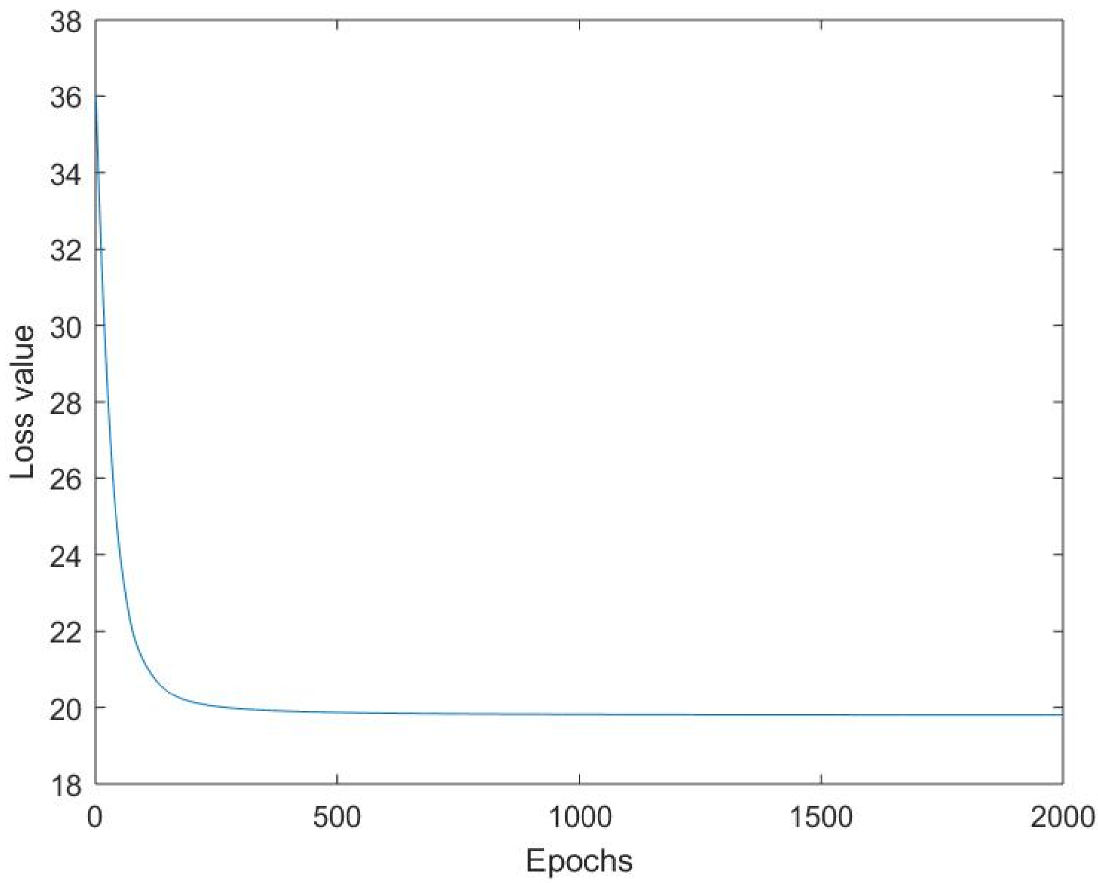
181 32

2 152

Number of support vectors: 339

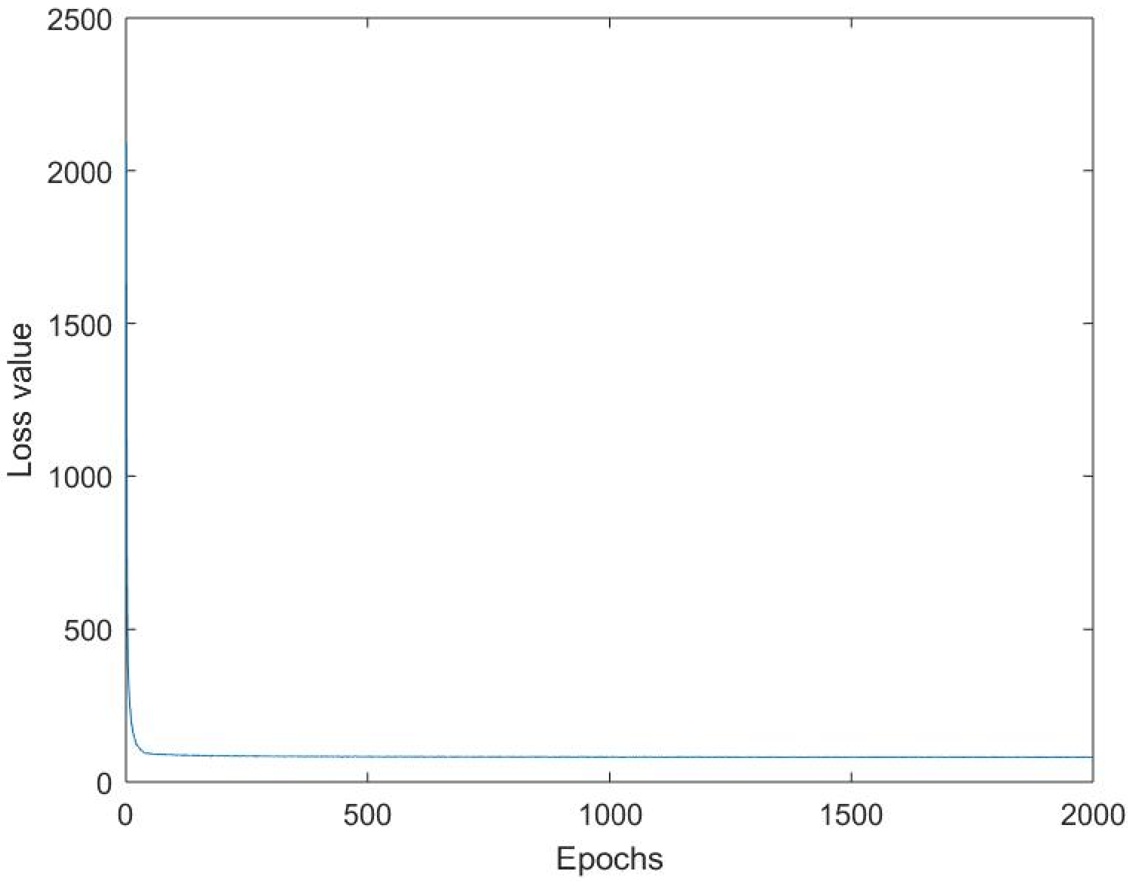
Question 3.2.5

For c = 0.1, ita_0 = 1, ita_1 = 100, iterations = 2000, the plot of loss values is



Objective function value: 19.8035

For C = 10



Objective function value: 80.7702

The objective function value calculated using stochastic gradient descent is less than the objective function value calculated using quadratic programming.

Question 3.2.6

For C = 0.1

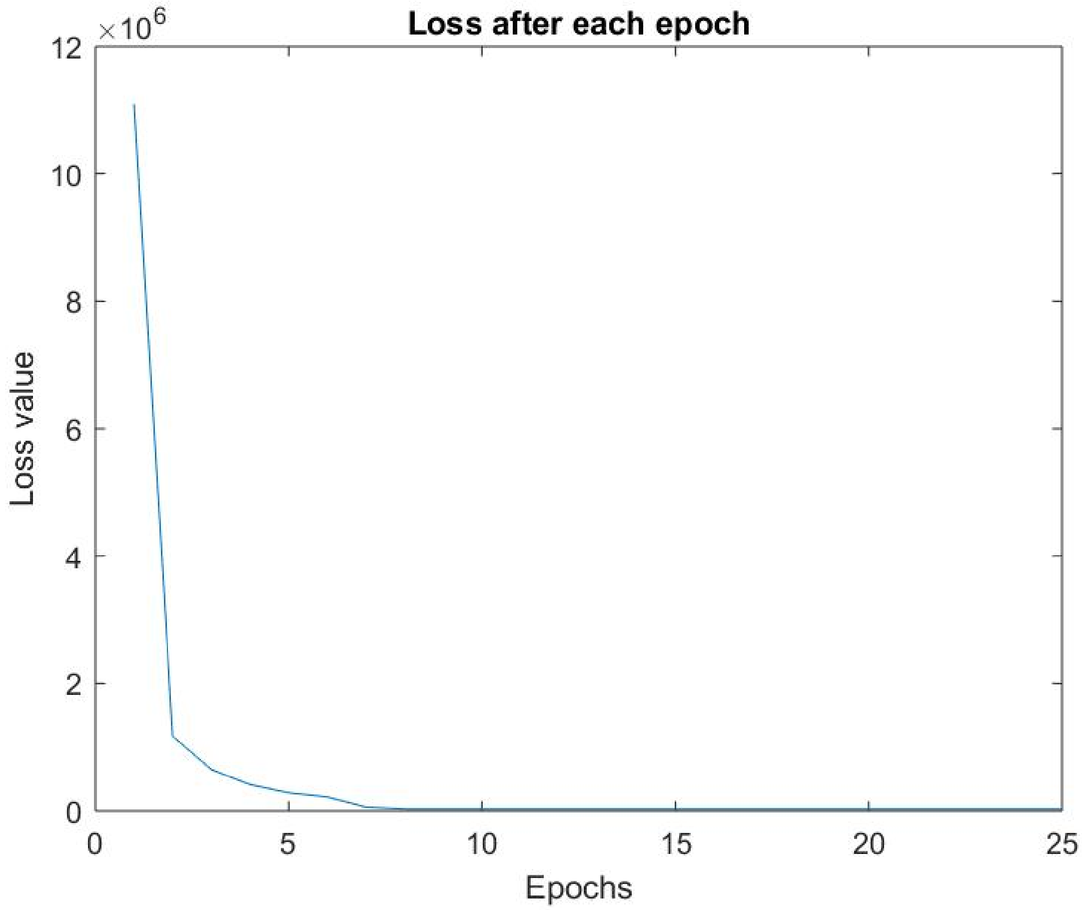
- **a) Prediction error on training data:** 0.0359
- **b) Prediction error on validation data (test error):** 0.0572
- **c) L2 norm of weights and its square:** 16.1100

For C = 10

- a) **Prediction error on training data:** 0.0055
- b) **Prediction error on validation data (test error):** 0.0354

c) **L2 norm square of w:** 121.0098

Question 3.2.7



The values on which I got the best result are: C = 8, epochs = 25, ita_1 = 1000, ita_0 = 1

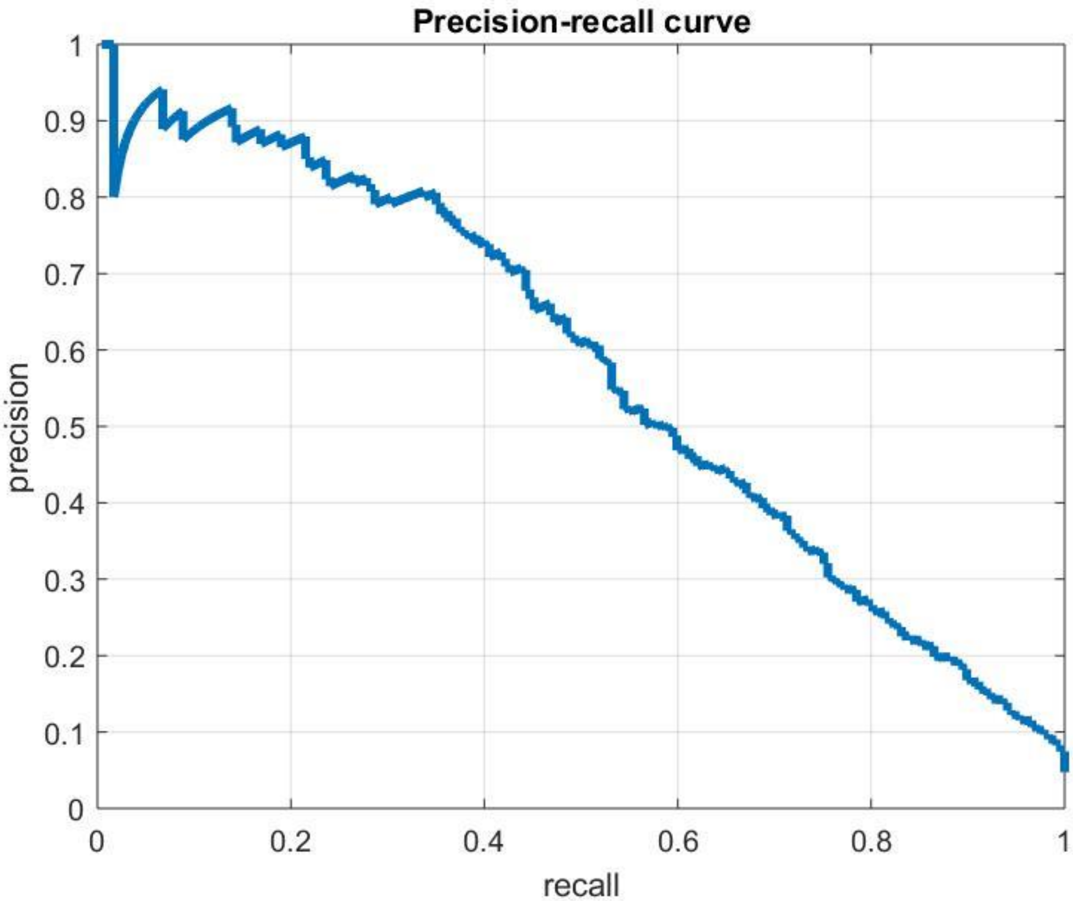
Kaggle rank: 8

Kaggle score: 0.79780

Question 4.4.1 :

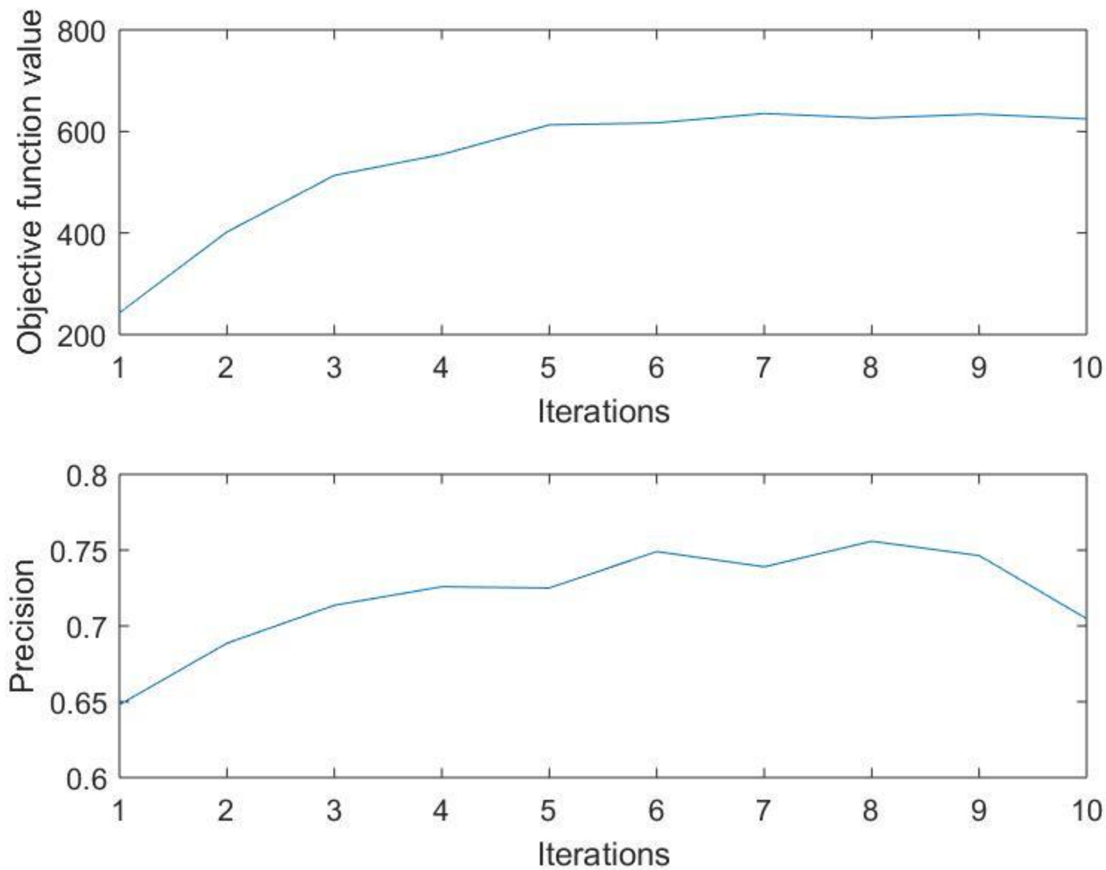
AP values : 0.578487068694247

Precision recall curve:



Question 4.4.3

Plot of Objective values and the APs



Question 4.4.4:

Value of AP: 79.63

(Best I achieved was 81.45)