Trung Vo – ID:109845485
CSE 512 – Homework 2

Problem 1

Dimension assumptions:

$$X:[k,n], Y:[n,1], W:[k,1]$$

From these, we infer dimensions of the rest as follow:

$$\bar{X}:[k+1,n], \bar{W}:[k+1,1], d=\bar{X}Y:[k+1,1], \bar{I}:[k+1,k+1], C=\bar{X}\bar{X}^T+\lambda\bar{I}:[k+1,k+1]$$

1.1) $J=\lambda|W|^2+\sum_{i=1}^{n}(W^Tx_i+b-y_i)^2$

$$\Rightarrow J=\lambda|\bar{W}|^2-\lambda b^2+\sum_{i=1}^{n}(W^Tx_i+b-y_i)^2$$

$$\Rightarrow J=\lambda|\bar{W}|^2-\lambda b^2+\sum_{i=1}^{n}(\bar{W}^T\bar{x}_i-y_i)^2$$

$$\Rightarrow \frac{\partial J}{\partial \bar{W}_j}=2\lambda\bar{W}_j+2\sum_{i=1}^{n}(\bar{W}^T\bar{x}_i-y_i)\bar{x}_{i,j}$$

We know $\frac{\partial J}{\partial \bar{W}}=\begin{bmatrix}\frac{\partial J}{\partial \bar{W}_1}\\ \vdots \\ \frac{\partial J}{\partial \bar{W}_{k+1}}\end{bmatrix}=2\lambda\bar{W}+2\bar{X}(\bar{W}^T\bar{X}-Y^T)^T$

Let derivative = 0, we have: $\lambda\bar{W}+\bar{X}(\bar{W}^T\bar{X}-Y^T)^T=0$

$$\Rightarrow \lambda\bar{W}+\bar{X}[(\bar{W}^T\bar{X})^T-Y]=0$$

$$\Rightarrow \lambda\bar{W}+\bar{X}\bar{X}^T\bar{W}-\bar{X}Y=0$$

$$\Rightarrow \bar{W}(\lambda+\bar{X}\bar{X}^T)=\bar{X}Y$$

$$\Rightarrow \bar{W}=(\lambda+\bar{X}\bar{X}^T)^{-1}\bar{X}Y=C^{-1}d$$

1.2) $C_{(i)}=C-\begin{bmatrix}\begin{bmatrix}x_{i,1}\\ \vdots \\ x_{i,k}\end{bmatrix} & \begin{bmatrix}0\\ \vdots \\ 0\end{bmatrix}\\ 1 & 0\end{bmatrix}\begin{bmatrix}[x_{i,1} & \dots & x_{i,k}] & 1\\ [0 & \dots & 0] & 0\end{bmatrix}$

$$\Rightarrow C_{(i)}=C-\begin{bmatrix}x_i & 0_k\\ 1 & 0\end{bmatrix}\begin{bmatrix}x_i^T & 1\\ 0_k^T & 0\end{bmatrix}=C-\bar{X}_i[1,0]\begin{bmatrix}1\\0\end{bmatrix}\bar{X}_i^T=C-\bar{X}_i\bar{X}_i^T$$

$$d_{(i)}=d-y_i\begin{bmatrix}x_{i,1}\\ \vdots \\ x_{i,k}\\ 1\end{bmatrix}=d-y_i\begin{bmatrix}x_i\\1\end{bmatrix}=d-y_i\bar{X}_i$$

1.3) $C_{(i)}=C-\bar{X}_i\bar{X}_i^T\Rightarrow C_{(i)}^{-1}=\left(C-\bar{X}_i\bar{X}_i^T\right)^{-1}$

$$\Rightarrow C_{(i)}{}^{-1} = C^{-1} + \frac{C^{-1}\bar{X}_i\bar{X}_i{}^T C^{-1}}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}$$

1.4) $\overline{\mathrm{W}}_{(\boldsymbol{i})} = C_{(i)}{}^{-1} d_{(i)} = \left(C^{-1} + \frac{C^{-1}\bar{X}_i\bar{X}_i{}^T C^{-1}}{1-\bar{X}_i{}^T C^{-1}\bar{X}_i}\right)(d - y_i\bar{X}_i)$

$$\Rightarrow \overline{\mathrm{W}}_{(\boldsymbol{i})} = \overline{\mathrm{W}} - C^{-1}y_i\bar{X}_i + \frac{C^{-1}\bar{X}_i\bar{X}_i{}^T\overline{\mathrm{W}}}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i} - \frac{C^{-1}\bar{X}_i\bar{X}_i{}^T C^{-1}y_i\bar{X}_i}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}$$

$$\Rightarrow \overline{\mathrm{W}}_{(\boldsymbol{i})} = \overline{\mathrm{W}} - C^{-1}\bar{X}_i\left(y_i - \frac{\bar{X}_i{}^T\overline{\mathrm{W}}}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i} + \frac{\bar{X}_i{}^T C^{-1}y_i\bar{X}_i}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}\right)$$

$$\Rightarrow \overline{\mathrm{W}}_{(\boldsymbol{i})} = \overline{\mathrm{W}} - C^{-1}\bar{X}_i\left(\frac{y_i\left(1 - \bar{X}_i{}^T C^{-1}\bar{X}_i\right) - \bar{X}_i{}^T\overline{\mathrm{W}} + \bar{X}_i{}^T C^{-1}y_i\bar{X}_i}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}\right)$$

$$\Rightarrow \overline{\mathrm{W}}_{(\boldsymbol{i})} = \overline{\mathrm{W}} - C^{-1}\bar{X}_i\left(\frac{y_i - y_i\bar{X}_i{}^T C^{-1}\bar{X}_i - \bar{X}_i{}^T\overline{\mathrm{W}} + \bar{X}_i{}^T C^{-1}y_i\bar{X}_i}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}\right)$$

$$\Rightarrow \overline{\mathrm{W}}_{(\boldsymbol{i})} = \overline{\mathrm{W}} + C^{-1}\bar{X}_i\left(\frac{-y_i + \bar{X}_i{}^T\overline{\mathrm{W}}}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}\right)$$

1.5) From 1.4, let $A = \frac{-y_i+\bar{X}_i{}^T\overline{\mathrm{W}}}{1-\bar{X}_i{}^T C^{-1}\bar{X}_i}$, we have:

$$\overline{\mathrm{W}}_{(\boldsymbol{i})}{}^{\boldsymbol{T}} = \overline{\mathrm{W}}^{\mathrm{T}} + \mathrm{A}\bar{X}_i{}^T(C^{-1})^T$$
$$\Rightarrow \overline{\mathrm{W}}_{(\boldsymbol{i})}{}^{\boldsymbol{T}}\bar{X}_i - y_i = \overline{\mathrm{W}}^{\mathrm{T}}\bar{X}_i + A\bar{X}_i\bar{X}_i{}^T(C^{-1})^T - y_i$$

Substitute A into this equation, it becomes:

$$\overline{\mathrm{W}}_{(\boldsymbol{i})}{}^{\boldsymbol{T}}\bar{X}_i - y_i = \frac{\overline{\mathrm{W}}^{\mathrm{T}}\bar{X}_i - \overline{\mathrm{W}}^{\mathrm{T}}\bar{X}_i\bar{X}_i{}^T C^{-1}\bar{X}_i + \bar{X}_i{}^T(C^{-1})^T\bar{X}_i\left(-y_i + \bar{X}_i{}^T\overline{W}\right) - y_i + y_i\bar{X}_i{}^T C^{-1}\bar{X}_i}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}$$

$$\Rightarrow \overline{\mathrm{W}}_{(\boldsymbol{i})}{}^{\boldsymbol{T}}\bar{X}_i - y_i = \frac{\overline{\mathrm{W}}^{\mathrm{T}}\bar{X}_i - y_i}{1 - \bar{X}_i{}^T C^{-1}\bar{X}_i}$$

1.6) For a single error of a single training example, the complexity is $O(k^3)$.
So the complexity for the whole training set is $O(nk^3)$.

Problem 2

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

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

$\Rightarrow P(Y=1|X) = \frac{1}{1+\exp\left(ln\frac{1-\alpha}{\alpha} + ln\frac{P(X|Y=0)}{P(X|Y=1)}\right)}$ (with $\alpha = P(Y=1)$) (*)

Since X_1 is a boolean variable, $P(X_1|Y=k) = \beta_{X_1,k}{}^{X_1}(1-\beta_{X_1,k})^{1-X_1}$

Since X_2 is a continuous variable (assume Gaussian distribution):

$$P(X_2|Y) = \frac{1}{\sqrt{2\pi\sigma_{X_2,k}{}^2}} e^{-\frac{(X_2-\mu_{X_2,k})^2}{2\sigma_{X_2,k}{}^2}}$$

Therefore,

$$P(X|Y=k) = \frac{\beta_{X_1,k}{}^{X_1}}{\sqrt{2\pi\sigma_{X_2,k}{}^2}}(1-\beta_{X_1,k})^{1-X_1} e^{-\frac{(X_2-\mu_{X_2,k})^2}{2\sigma_{X_2,k}{}^2}}$$

$$\Rightarrow \frac{P(X|Y=0)}{P(X|Y=1)} = \frac{\sigma_{X_2,1}\beta_{X_1,0}{}^{X_1}(1-\beta_{X_1,0})^{1-X_1} e^{-\frac{(X_2-\mu_{X_2,0})^2}{2\sigma_{X_2,0}{}^2}}}{\sigma_{X_2,0}\beta_{X_1,1}{}^{X_1}(1-\beta_{X_1,1})^{1-X_1} e^{-\frac{(X_2-\mu_{X_2,1})^2}{2\sigma_{X_2,1}{}^2}}}$$

Substitute this to (*) we can have a formula to compute P(Y|X). Parameters needed to compute P(Y|X) are those alphas, betas, mean and standard deviation in the equations above.

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

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

$\Rightarrow P(Y=1|X) = \frac{1}{1+\exp\left(ln\frac{1-\alpha}{\alpha} + ln\frac{P(X|Y=0)}{P(X|Y=1)}\right)}$ (with $\alpha = P(Y=1)$)

Since X = [X_1 ... X_d] are Boolean variables, and Bernoulli Naïve Bayes assumes that each X_i is independent for simplicity:

$$P(X|Y=k) = P(X_1 \dots X_d|Y=k) = \prod_{i=1}^{d} P(X_i|Y=k) = \prod_{i=1}^{d} \beta_{X_i,k}{}^{X_i}(1-\beta_{X_i,k})^{1-X_i}$$

So:

$$\frac{P(X|Y=0)}{P(X|Y=1)} = \prod_{i=1}^{d} \left(\frac{\beta_{X_i,0}}{\beta_{X_i,1}}\right)^{X_i} \left(\frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}}\right)^{1-X_i}$$

$$\Rightarrow ln\frac{P(X|Y=0)}{P(X|Y=1)} = \sum_{i=1}^{d} X_i \left(ln\frac{\beta_{X_i,0}}{\beta_{X_i,1}} - \frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}}\right) + \frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}}$$

$$\Rightarrow P(Y=1|X) = \frac{1}{1+\exp\left(ln\frac{1-\alpha}{\alpha} + \sum_{i=1}^{d}\left(X_i\left(ln\frac{\beta_{X_i,0}}{\beta_{X_i,1}} - \frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}}\right) + \frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}}\right)\right)}$$

$$\Rightarrow P(Y=1|X) = \frac{1}{1+\exp\left(ln\frac{1-\alpha}{\alpha} + \sum_{i=1}^{d}\frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}} + \sum_{i=1}^{d}\left(X_i\left(ln\frac{\beta_{X_i,0}}{\beta_{X_i,1}} - \frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}}\right)\right)\right)}$$

Let $\theta_{d+1} = -ln\frac{1-\alpha}{\alpha} - \sum_{i=1}^{d}\frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}}$ and $\theta_i = \frac{1-\beta_{X_i,0}}{1-\beta_{X_i,1}} - ln\frac{\beta_{X_i,0}}{\beta_{X_i,1}}$:

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

(same form)

General case:

Suppose the probability that $y = 1$ given $X = [X_1 \ldots X_d]$ is $P(y=1|X)$. Let W be an optimized set of parameters that makes an activation function $f(W^T X)$ as close as 1.

$$\Rightarrow P(y=1|X;W) = f(W^T X)$$

$$\Rightarrow P(y=0|X;W) = 1 - f(W^T X)$$

Combine these two expressions, we have:

$$P(y|X;W) = f(W^T X)^y (1-f(W^T X))^{1-y}$$

We need to find a W such that:

$$W = \underset{W}{\operatorname{argmax}} P(y|X;W)$$

$$= \underset{W}{\operatorname{argmax}} f(W^T X)^y (1-f(W^T X))^{1-y}$$

This means we need to find the estimate W for this negative log likelihood:

$$J(W) = -logP(y|X;W) = -[y\log\left(f(W^T X)\right) + (1-y)\log\left(1-f(W^T X)\right)]$$

Let $z = f(W^T X)$:

$$\frac{\partial J}{\partial W} = -\left(\frac{y}{z} - \frac{1-y}{1-z}\right)\frac{\partial z}{\partial W} = \left(\frac{z-y}{z(1-z)}\right)\frac{\partial z}{\partial W}$$

Let $s = W^T X$:

$$\frac{\partial z}{\partial W} = \frac{\partial z}{\partial s}.\frac{\partial s}{\partial W} = \frac{\partial z}{\partial s} X$$

We find z such that $\frac{\partial z}{\partial s} = z(1-z)$:

$$\Longrightarrow \frac{\partial z}{z(1-z)} = \partial s$$

Solve this equation, we have $z = \frac{1}{1+e^{-s}} = \frac{1}{1+e^{-W^T X}}$

Therefore, substitute back to $P(y|X;W)$:

$$P(y|X;W) = \left(\frac{1}{1+e^{-W^T X}}\right)^y \left(1 - \frac{1}{1+e^{-W^T X}}\right)^{1-y}$$

$$\Longrightarrow P(y=1|X;W) = \frac{1}{1+e^{-W^T X}} = \frac{1}{1+e^{-(\sum_{i=1}^{d} W_i X_i + W_{d+1})}}$$

Problem 3

3.1) 1. Dual objective: $\underset{\alpha}{\operatorname{argmax}} \sum_{j=1}^{n} \alpha_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i \alpha_i y_j \alpha_j k(x_i, x_j)$

subject to $\sum_{j=1}^{n} y_j \alpha_j = 0$ and $0 \le \alpha_j \le C$

Suppose kernel is linear form for the sake of simplicity, we have:

The dual object has the QP form as: $g(\alpha) = -\frac{1}{2}\alpha^T H \alpha + 1^T \alpha$

A kernel is the one that satisfies $\sum_{i=1}^{n}\sum_{j=1}^{n} y_i \alpha_i y_j \alpha_j k(x_i, x_j) \ge 0, \forall \alpha_n$

Let $\alpha^T H \alpha = \sum_{i=1}^{n}\sum_{j=1}^{n} y_i \alpha_i y_j \alpha_j k(x_i, x_j)$

(H is a symmetric matrix such that $H_{ij} = y_i y_j k(x_i, x_j)$)

Therefore:

$H = V^T V$ (with $V = [y_1 x_1, y_2 x_2, \dots y_3 x_3]$, assume kernel is linear for simplicity)

$f = [1,1, \dots ,1]^T$

$$A = \begin{bmatrix} -1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & -1 \\ 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix}, b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ C \\ \vdots \\ C \end{bmatrix} \text{(since } 0 \le \alpha_j \le C\text{)}$$

$lb = 0, ub = C$

$A_{eq} = Y, b_{eq} = [0, \dots, 0]$ (since $\sum_{j=1}^{n} y_j \alpha_j = 0$)

2,3. The Lagrange of the SVM is defined as:

$$L(w, b, \alpha) = \frac{1}{2}||w||_2{}^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i(w^T x_i + b))$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\Rightarrow b = y - w^T x$$

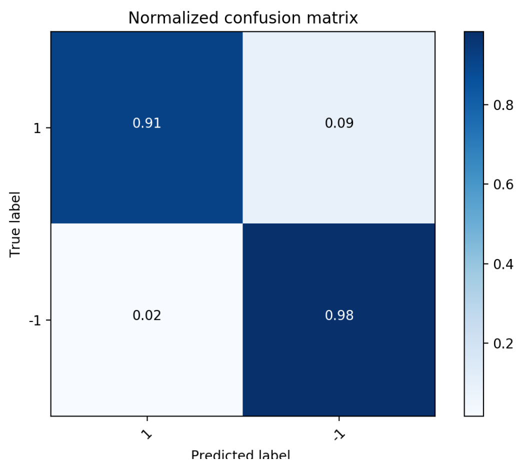
4,5.

For C = 0.1:
accuracy: 94.822888 %
train_loss/objective_value: 24.948
number of support vectors: 339

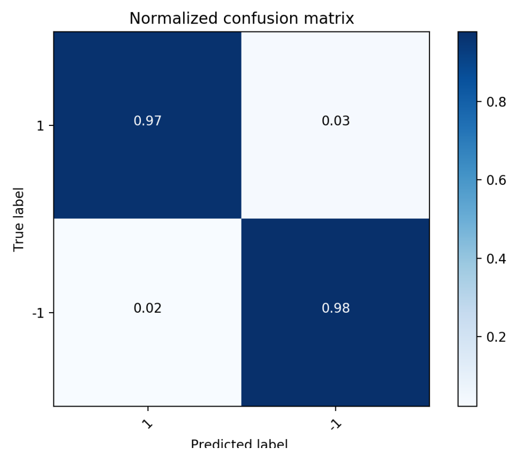
For C = 10:
accuracy: 97.275204 %
train_loss/objective_value: 112.744
number of support vectos: 124

Note: So we can obtain that with a higher C, the accuracy is higher. This is because higher C yields thinner margin, thus reduces the number of data points being mislabeled. This reduces number of support vectors as well. (assume number of SVs are points that alpha > 1e-6)

Confusion matrix:



C = 0.1



C = 10

3.2) **1.** Subgradient of L_i wrt w_{y_i}

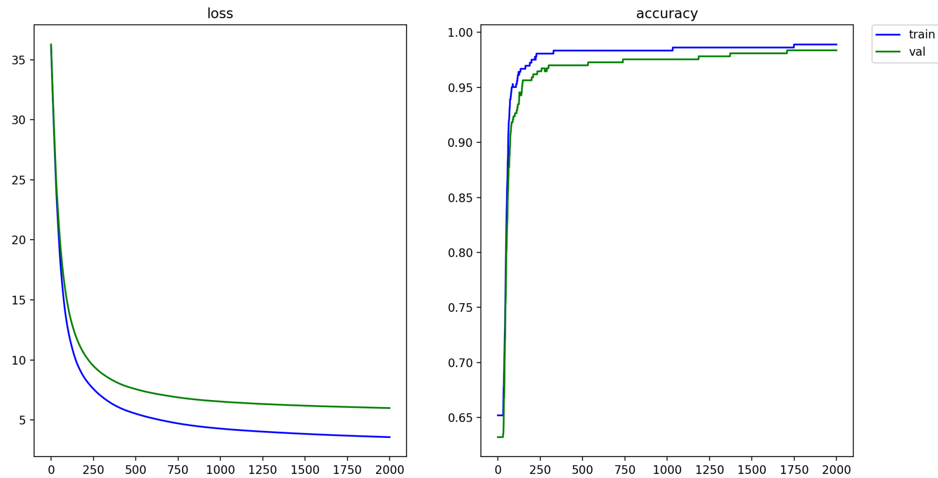
$$\frac{\partial L_i}{\partial w_{y_i}} = \frac{\partial}{\partial w_{y_i}} \max\left(0, 1 - {w_{y_i}}^T x_i + {w_{\hat{y}_i}}^T x_i\right) = \begin{cases} 0 & if\ 1 - {w_{y_i}}^T x_i + {w_{\hat{y}_i}}^T x_i < 0 \\ -x_i & if\ 1 - {w_{y_i}}^T x_i + {w_{\hat{y}_i}}^T x_i \geq 0 \end{cases}$$

2. Subgradient of L_i wrt $w_{\hat{y}_{i_i}}$

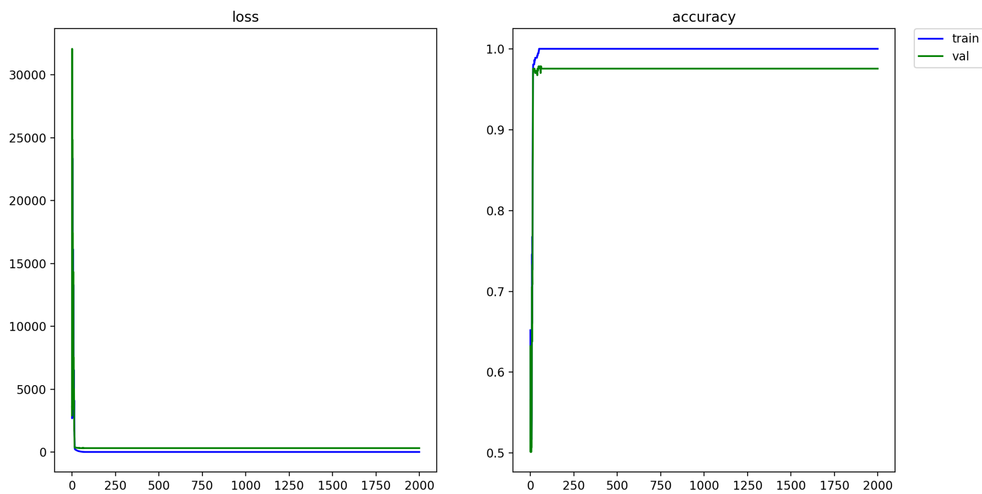
$$\frac{\partial L_i}{\partial w_{\hat{y}_i}} = \frac{\partial}{\partial w_{\hat{y}_i}} \max\left(0, 1 - {w_{y_i}}^T x_i + {w_{\hat{y}_i}}^T x_i\right) = \begin{cases} 0 & if\ 1 - {w_{y_i}}^T x_i + {w_{\hat{y}_i}}^T x_i < 0 \\ x_i & if\ 1 - {w_{y_i}}^T x_i + {w_{\hat{y}_i}}^T x_i \geq 0 \end{cases}$$

3. Combine results of those two previous questions to get subgradient of L_i wrt w_j

4,5. For C = 0.1:



For C = 10:



Numeric results:

```
epoch 1700/1999:
train_loss: 3.7191      train_acc: 98.6188 %
val_loss: 6.1008        val_acc: 98.0926 %
Best val accuracy so far: 98.0926 %
--------------------------------------------------
epoch 1800/1999:
train_loss: 3.6668      train_acc: 98.8950 %
val_loss: 6.0625        val_acc: 98.3651 %
Best val accuracy so far: 98.3651 %
--------------------------------------------------
epoch 1900/1999:
train_loss: 3.6177      train_acc: 98.8950 %
val_loss: 6.0256        val_acc: 98.3651 %
Best val accuracy so far: 98.3651 %
--------------------------------------------------
Best validation accuracy: 98.3651 %
```

C = 0.1

```
epoch 1700/1999:
train_loss: 21.9702      train_acc: 100.0000 %
val_loss: 326.7801       val_acc: 97.5477 %
Best val accuracy so far: 97.8202 %
--------------------------------------------------
epoch 1800/1999:
train_loss: 21.9696      train_acc: 100.0000 %
val_loss: 326.7713       val_acc: 97.5477 %
Best val accuracy so far: 97.8202 %
--------------------------------------------------
epoch 1900/1999:
train_loss: 21.9691      train_acc: 100.0000 %
val_loss: 326.7630       val_acc: 97.5477 %
Best val accuracy so far: 97.8202 %
--------------------------------------------------
Best validation accuracy: 97.8202 %
```

C = 10

With same value of C, loss in 3.1.4 is larger than loss we found here.

6. a) For C=0.1:
confusion matrix:

181	3
3	180

precision = (181+180)/367 = 0.9837
val prediction error = 1-0.9837 = 0.0163

For C=10:
confusion matrix:

180	4
5	178

precision = (180+178)/367 = 0.9755
val prediction error = 0.0245

b) For C=0.1:
confusion matrix:

184	2
2	174

precision = (184+174)/362 = 0.989
train prediction error = 1-0.989 = 0.011

For C=10:
confusion matrix:

186	0
0	176

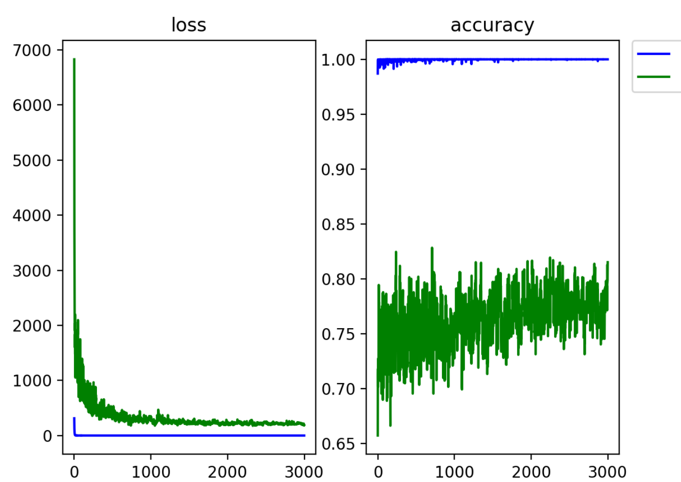
precision = (186+176)/362 = 1
train prediction error = 0

c) For C=0.1: Sum of squares of W: 58.650
For C=10: Sum of squares of W: 1425.232

7. **Test on UCF101 data for 10 classes**
- Best accuracy on validation set: 83.9151 %
- Parameters can be found in file `params_3_2.csv` file.

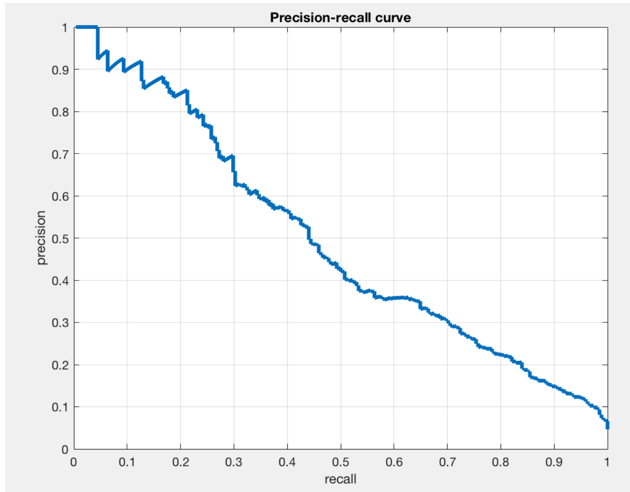
Result:

```
---------------------------------------------------
epoch 2600/2999:
train_loss: 0.0066        train_acc: 100.0000 %
val_loss: 216.5667        val_acc: 75.7075 %
Best val accuracy so far: 83.9151 %
---------------------------------------------------
epoch 2700/2999:
train_loss: 0.0084        train_acc: 100.0000 %
val_loss: 205.9336        val_acc: 79.4340 %
Best val accuracy so far: 83.9151 %
---------------------------------------------------
epoch 2800/2999:
train_loss: 0.0063        train_acc: 100.0000 %
val_loss: 206.0648        val_acc: 77.6415 %
Best val accuracy so far: 83.9151 %
---------------------------------------------------
epoch 2900/2999:
train_loss: 0.0057        train_acc: 100.0000 %
val_loss: 202.5865        val_acc: 79.0566 %
Best val accuracy so far: 83.9151 %
---------------------------------------------------
Best validation accuracy: 83.9151 %
```



Problem 4

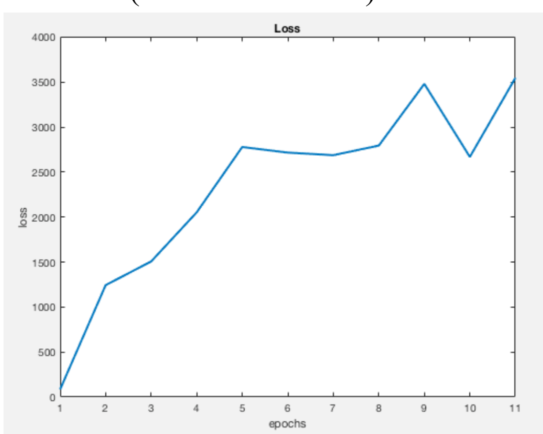
4.4) 1. Precision recall plot for validation set:

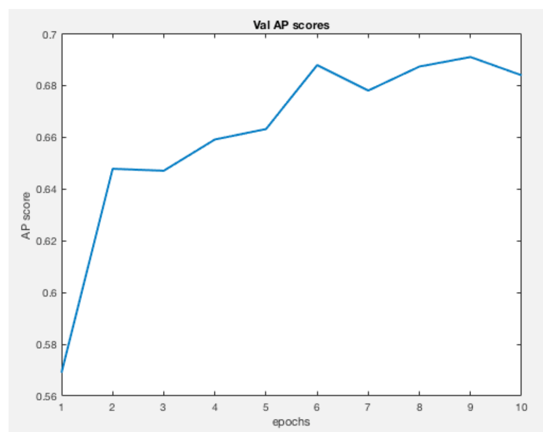


AP (average precision): ~0.4980

2,3. Hard negative mining for 10 iterations:

- Best validation AP score: 0.6981
- Loss curve (after 10 iterations):





Loss (after 10 epochs) Validation AP (10 epochs)

4.5. Result file to compute AP score for test set can be found at `109845485.mat`