Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 1

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 30, 2018

According to the given problem:

$$p(\mathbf{w}) = C * exp(\frac{-\lambda}{2}\mathbf{w}^T\mathbf{w}) \tag{1}$$

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} \frac{1}{1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)} \tag{2}$$

Therefore, MAP estimate will be:

$$\hat{\mathbf{w}}_{MAP} = argmax_{\mathbf{w}} log(p(\mathbf{y}|\mathbf{X}, \mathbf{w})) + log(p(\mathbf{w})) \tag{3}$$

$$\Rightarrow \hat{\mathbf{w}}_{MAP} = argmax_{\mathbf{w}} \sum_{n=1}^{N} -log(1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)) + \frac{-\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{4}$$

$$\Rightarrow \hat{\mathbf{w}}_{MAP} = argmin_{\mathbf{w}} \sum_{n=1}^{N} log(1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{5}$$

Therefore, to minimise the negative log likelihood the partial derivatives w.r.t. to **w** yields.

$$\lambda\mathbf{w} + \sum_{n=1}^{N} \frac{-y_n\mathbf{x}_n exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)} = 0 \tag{6}$$

$$\Rightarrow \hat{\mathbf{w}}_{MAP} = \frac{1}{\lambda}\sum_{n=1}^{N} \frac{exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)} y_n\mathbf{x}_n \tag{7}$$

$$\therefore \alpha_n = \frac{1}{\lambda}(\frac{exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)}) \tag{8}$$

From the expression of α_n, we can see that it specifies an scaled version of non-class probabilities for xn. This make sense as we can see from the probability expression for the right-class as the α_n increases the probability for the right-class will increase, which means if probability of wrong-class is high it will give us a new estimate of **w** such that probability of right class will increase.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 30, 2018

QUESTION 2

As per the problem:

$$p(y=1)=\pi \tag{9}$$

$$p(\mathbf{x}|y=1)=\prod_{d=1}^{D}\mu_{d,1}^{xd}(1-\mu_{d,1})^{1-xd} \tag{10}$$

$$p(\mathbf{x}|y=0)=\prod_{d=0}^{D}\mu_{d,0}^{xd}(1-\mu_{d,0})^{1-xd} \tag{11}$$

$$\therefore p(y=1|\mathbf{x})=\frac{\pi\prod_{d=1}^{D}\mu_{d,1}^{xd}(1-\mu_{d,1})^{1-xd}}{\pi\prod_{d=1}^{D}\mu_{d,1}^{xd}(1-\mu_{d,1})^{1-xd}+(1-\pi)\prod_{d=0}^{D}\mu_{d,0}^{xd}(1-\mu_{d,0})^{1-xd}} \tag{12}$$

$$=\frac{1}{1+\frac{1-\pi}{\pi}\prod_{d=1}^{D}[\frac{\mu_{d,0}}{\mu_{d,1}}]^{\mathbf{x}_d}[\frac{1-\mu_{d,0}}{1-\mu_{d,1}}]^{1-\mathbf{x}_d}} \tag{13}$$

$$=\frac{1}{1+f(\mathbf{x})} \tag{14}$$

where $f(\mathbf{x})=\frac{1-\pi}{\pi}\prod_{d=1}^{D}[\frac{\mu_{d,0}}{\mu_{d,1}}]^{\mathbf{x}_d}[\frac{1-\mu_{d,0}}{1-\mu_{d,1}}]^{1-\mathbf{x}_d}$
It can be easily shown that:

$$p(y=0|\mathbf{x})=\frac{1}{1+f(\mathbf{x})^{-1}} \tag{15}$$

Therefore, this makes a discriminative model with its distribution as $Bernoulli[g(f(\mathbf{x}))]$ where $g(f(\mathbf{x}))=\frac{1}{1+f(\mathbf{x})}$ and $f(\mathbf{x})=\frac{1-\pi}{\pi}\prod_{d=1}^{D}[\frac{\mu_{d,0}}{\mu_{d,1}}]^{\mathbf{x}_d}[\frac{1-\mu_{d,0}}{1-\mu_{d,1}}]^{1-\mathbf{x}_d}$

For the decision boundary we can equate $p(y=1|\mathbf{x})=p(y=0|\mathbf{x})$. Which gives:

$$\pi\prod_{d=1}^{D}\mu_{d,1}^{xd}(1-\mu_{d,1})^{1-xd}=(1-\pi)\prod_{d=0}^{D}\mu_{d,0}^{xd}(1-\mu_{d,0})^{1-xd} \tag{16}$$

Or equivalently we can write, $f(\mathbf{x})=1$ as the decision boundary! Therefore, here we get an exponential decision boundary!

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 30, 2018

QUESTION 3

Let us first solve the Lagrangian for the given condition. We get:

$$\hat{\mathbf{w}} = argmax_{\lambda}[argmin_{w}[(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda(\mathbf{w}^T\mathbf{w} - C)]] \tag{17}$$

Solving the above dual problem for argmin case we get:

$$-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda\mathbf{w} = 0 \tag{18}$$

$$\Rightarrow \mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{19}$$

If we recall the solution for the l_2 regularized least square linear regression model. There also we got the similar solution for with λ being the hyperparameter for the regularization term! Therefore solving the above dual problem completly and finding the value for λ will gives us the equivalent l_2 regularized solution! If we use the same λ as the hyperparameter.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 30, 2018

QUESTION 4

As per the problem $p(\mathbf{y}|\mathbf{x},\mathbf{W})$ can be written as:

$$p(\mathbf{y}|\mathbf{x},\mathbf{W}) = \prod_{n=1}^{N}\prod_{k=1}^{K}[\frac{exp(\mathbf{w}_k^T\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T\mathbf{x}_n)}]^{1\{y_n=k\}} \tag{20}$$

where, $1\{y_n = k\}$ is equal to 1 when $y_n = k$ and 0 when $y_n \neq k$. Therefore, taking the loglikelihood give us:

$$LL(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K} 1\{y_n = k\}log(\frac{exp(\mathbf{w}_k^T\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T\mathbf{x}_n)}) \tag{21}$$

$$= \sum_{n=1}^{N} 1\{y_n = k\}log(\frac{exp(\mathbf{w}_k^T\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T\mathbf{x}_n)}) + \sum_{n=1}^{N} 1\{y_n \neq k\}log(1 - \frac{exp(\mathbf{w}_k^T\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T\mathbf{x}_n)}) \tag{22}$$

$$\Rightarrow \frac{LL(\mathbf{w})}{\partial\mathbf{w}} = \sum_{n=1}^{N} 1\{y_n = k\}\mathbf{x}_n(1 - \frac{exp(\mathbf{w}_k^T\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T\mathbf{x}_n)}) + \sum_{n=1}^{N} 1\{y_n \neq k\}(-\mathbf{x}_n\frac{exp(\mathbf{w}_k^T\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T\mathbf{x}_n)}) \tag{23}$$

$$= \sum_{n=1}^{N} \mathbf{x}_n(1\{y_n = k\} - \frac{exp(\mathbf{w}_k^T\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T\mathbf{x}_n)}) \tag{24}$$

Therefore GD update rule will be:

$$\mathbf{w}_k^{t+1} = \mathbf{w}_k^t + \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n(1\{y_n = k\} - \frac{exp(\mathbf{w}_k^{t,T}\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^{t,T}\mathbf{x}_n)}) \tag{25}$$

SGD update rule will be:

$$\mathbf{w}_k^{t+1} = \mathbf{w}_k^t + \mathbf{x}_n(1\{y_n = k\} - \frac{exp(\mathbf{w}_k^{t,T}\mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^{t,T}\mathbf{x}_n)}) \tag{26}$$

$$= \mathbf{w}_k^t + \mathbf{x}_n(1\{y_n = k\} - \mu_{nk}) \tag{27}$$

SGD Algorithm:

```
for k in range(K):
  while not converges:
    i = rand(1,N);
    p = exp(dot(w[k].T, x[i])/sum(exp(dot(W,x[i])));
    if y[i] == k:
      w[k] = w[k] + x[i]*(1 - p)
    else:
      w[k] = w[k] - x[i]*(p)
```

For the hard class case. μ_{nk} of sgd will change to our new μ_{nk}. Which can be written as: $\prod_{j=1}^{K} 1\{\mathbf{w}_k^T\mathbf{x} > \mathbf{w}_j^T\mathbf{x}\}$ Therefore, our new SGD update rule will be:

$$\mathbf{w}_k^{t+1} = \mathbf{w}_k^t + \mathbf{x}_n(1\{y_n = k\} - \prod_{j=1}^{K} 1\{\mathbf{w}_k^T\mathbf{x} > \mathbf{w}_j^T\mathbf{x}\}) \tag{28}$$

SGD Algorithm:

```
while not converges:
  i = rand(1,N);
  if y[i] == argmax(dot(W, x[i])):
    w[y[i]] = w[y[i]] + x[i]
  else:
    w[y[i]] = w[y[i]] - x[i]
```

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 30, 2018

QUESTION 5

We say two sets $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ are linear separable if we can find $\mathbf{w}$ and $\mathbf{b}$ such that: $\mathbf{w}^T\mathbf{x}_n + \mathbf{b} > 0$ and $\mathbf{w}^T\mathbf{y}_n + \mathbf{b} < 0$ for every $\mathbf{y}_n$ and $\mathbf{x}_n$. Now lets prove the given statement!

Case 1: $ConvexHullIntersects \Rightarrow NotSeperable$
If the two convex hull intersects then we can find a common point $\mathbf{z}$ such that:

$$\mathbf{w}^T\mathbf{z} + \mathbf{b} = \mathbf{w}^T \sum_{n=1} \alpha_n \mathbf{x}_n + \mathbf{b} > 0 \tag{29}$$

Also:

$$\mathbf{w}^T\mathbf{z} + \mathbf{b} = \mathbf{w}^T \sum_{n=1} \alpha_n \mathbf{y}_n + \mathbf{b} < 0 \tag{30}$$

But they both contradicts each other, therefore this means $\mathbf{x}_n$ and $\mathbf{y}_n$ are not seperable!

Case 2: $Seperable \Rightarrow ConvexHullNotIntersects$
Since they are seperable if a point $\mathbf{z}$ lies inside the convex hull of $\mathbf{x}_n$ then it will satisfy:

$$\mathbf{w}^T\mathbf{z} + \mathbf{b} = \mathbf{w}^T \sum_{n=1} \alpha_n \mathbf{x}_n + \mathbf{b} > 0 \tag{31}$$

And if a point $\mathbf{z}^{'}$ lies inside the convex hull of $\mathbf{y}_n$ then it will satisfy:

$$\mathbf{w}^T\mathbf{z}^{'} + \mathbf{b} = \mathbf{w}^T \sum_{n=1} \alpha_n \mathbf{y}_n + \mathbf{b} < 0 \tag{32}$$

Therefore, $\mathbf{z}^{'}$ will never be equal to $\mathbf{z}$. Which means the convex hull will not intersects!

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 30, 2018

QUESTION 6

Writing the lagrangian objective function for this case:

$$\mathbf{L}(\mathbf{w}, \mathbf{b}, \alpha_n) = \frac{||\mathbf{w}||^2}{2} + \sum_{n=1}^{N} \alpha_n (m - y_n(\mathbf{w}^T \mathbf{x}_n + \mathbf{b})) \tag{33}$$

Now, substituting $\mathbf{w} = m\mathbf{w}_s$ and $\mathbf{b} = m\mathbf{b}_s$ in the above objective our new objective can be written as:

$$\mathbf{L}(\mathbf{w}, \mathbf{b}, \alpha_n) = \frac{||\mathbf{w}_s||^2}{2} + \sum_{n=1}^{N} \alpha_n (m - y_n(m\mathbf{w}_s^T \mathbf{x}_n + m\mathbf{b}_s)) \tag{34}$$

$$= \frac{||\mathbf{w}_s||^2}{2} + \sum_{n=1}^{N} m\alpha_n (1 - y_n(\mathbf{w}_s^T \mathbf{x}_n + \mathbf{b}_s)) \tag{35}$$

$$= \frac{||\mathbf{w}_s||^2}{2} + \sum_{n=1}^{N} \alpha_{s,n} (1 - y_n(\mathbf{w}_s^T \mathbf{x}_n + \mathbf{b}_s)) \tag{36}$$

where $\alpha_{s,n} = m\alpha_n$.
This come out to be the same objective as that for the SVM case. Therefore, the parameter learned in this newer case will be $\mathbf{w} = m\mathbf{w}_s$ $\mathbf{b} = m\mathbf{b}_s$. Therefore the newer seperating hyperplane will be:

$$\mathbf{w}^T \mathbf{x} + \mathbf{b} = 0 \tag{37}$$

$$\Rightarrow m(\mathbf{w}_s^T \mathbf{x} + \mathbf{b}_s) = 0 \tag{38}$$

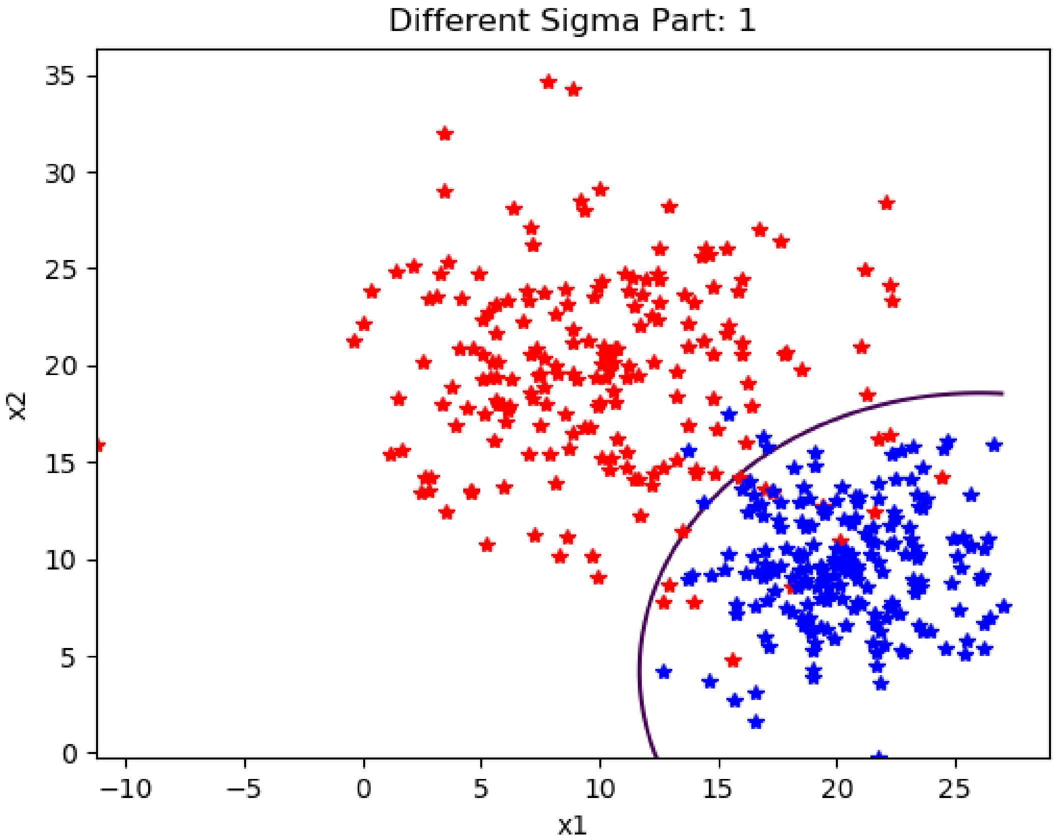
which is the same hyperplane as that for the SVM case!

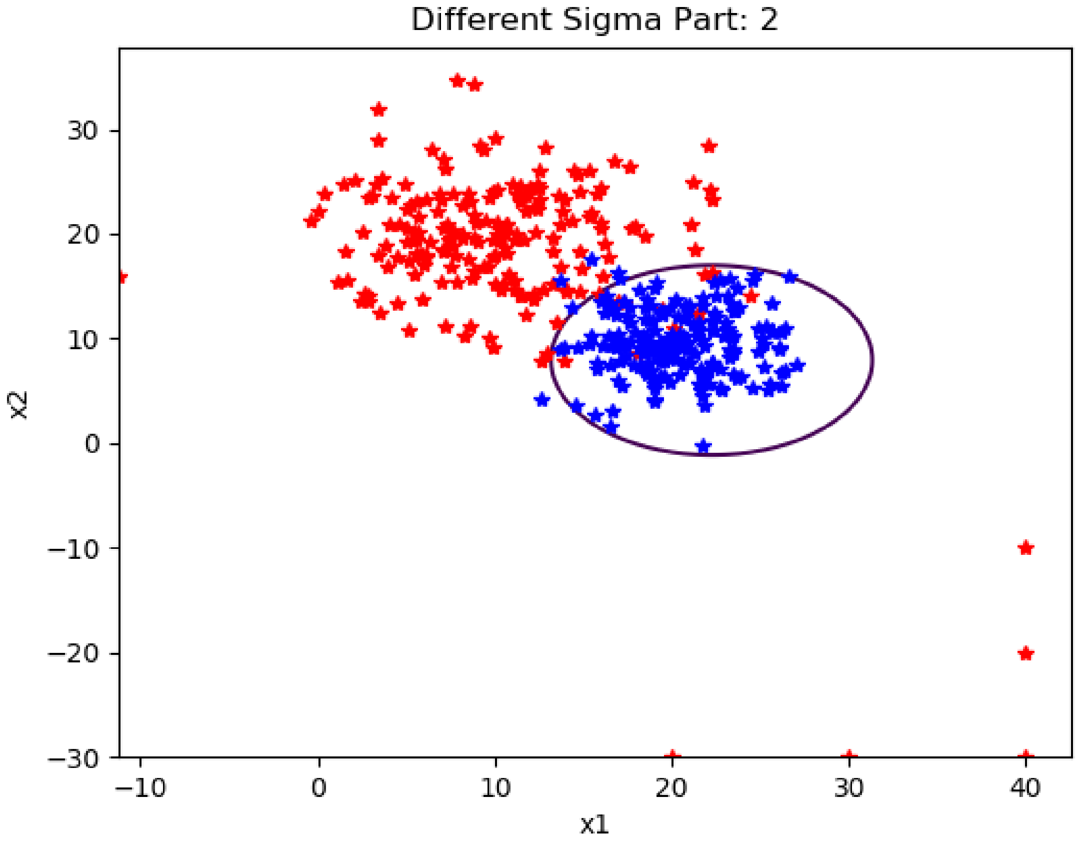
Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 2

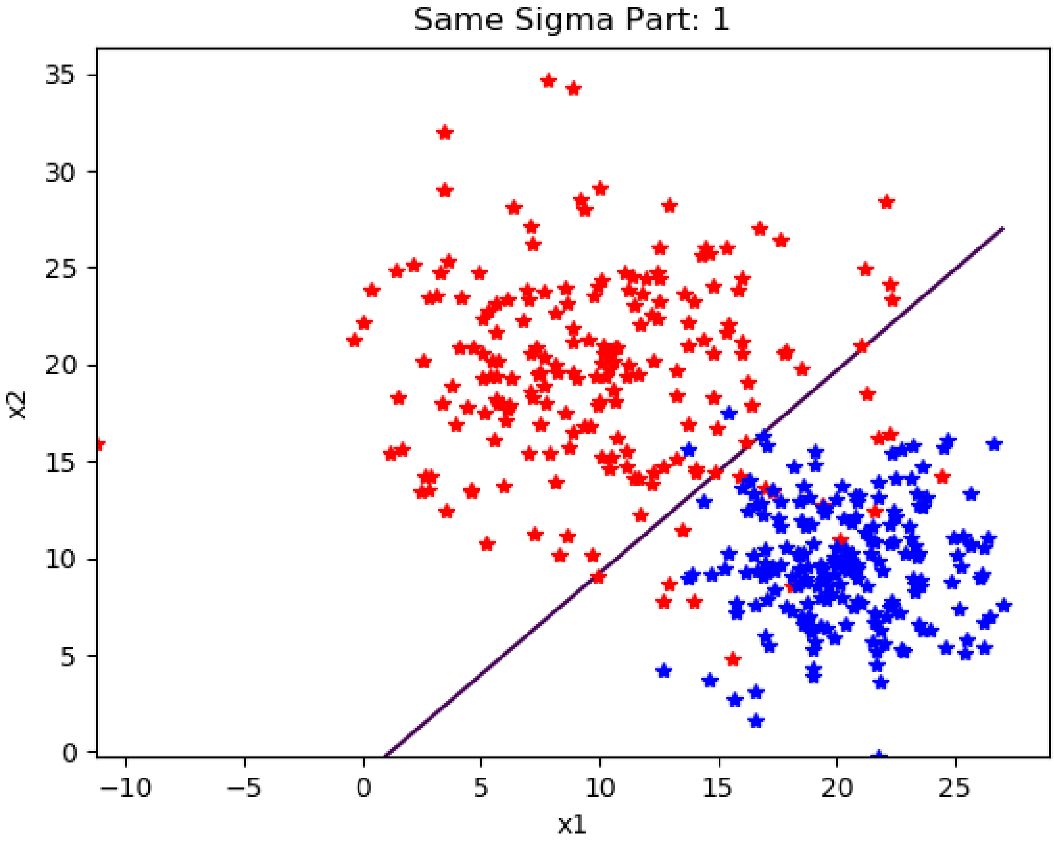
Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 30, 2018

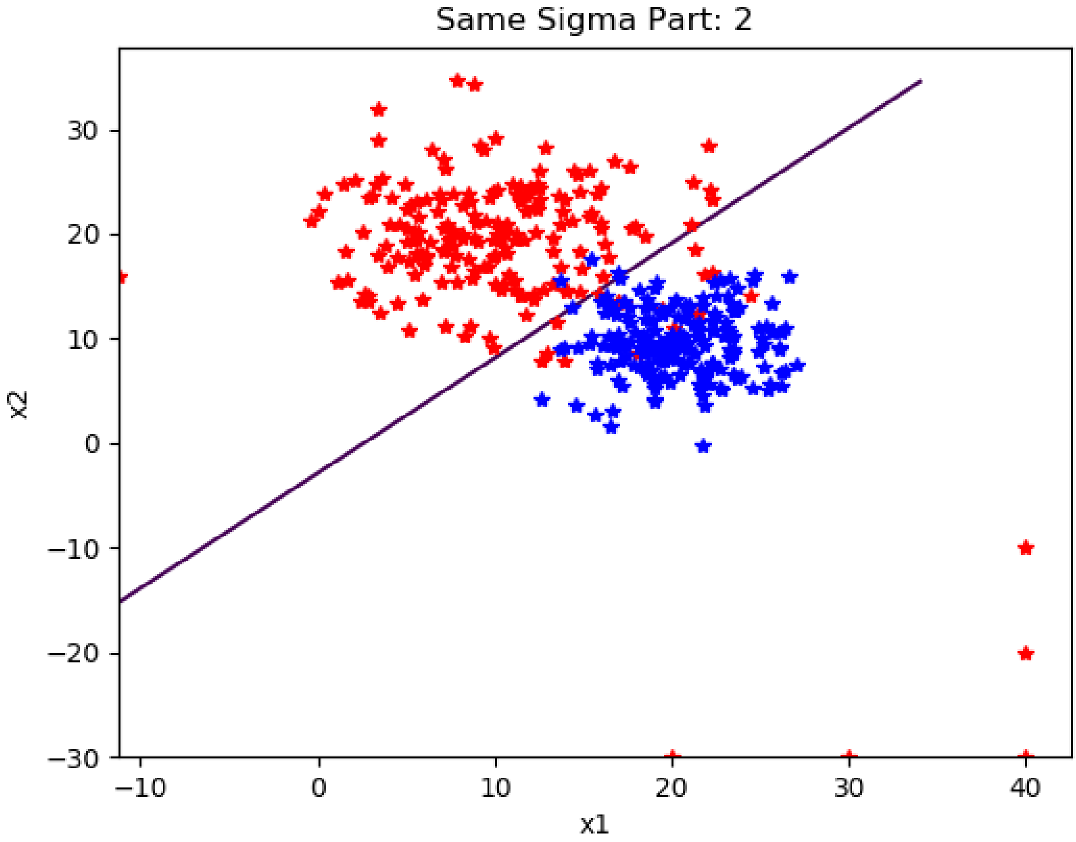
QUESTION 7

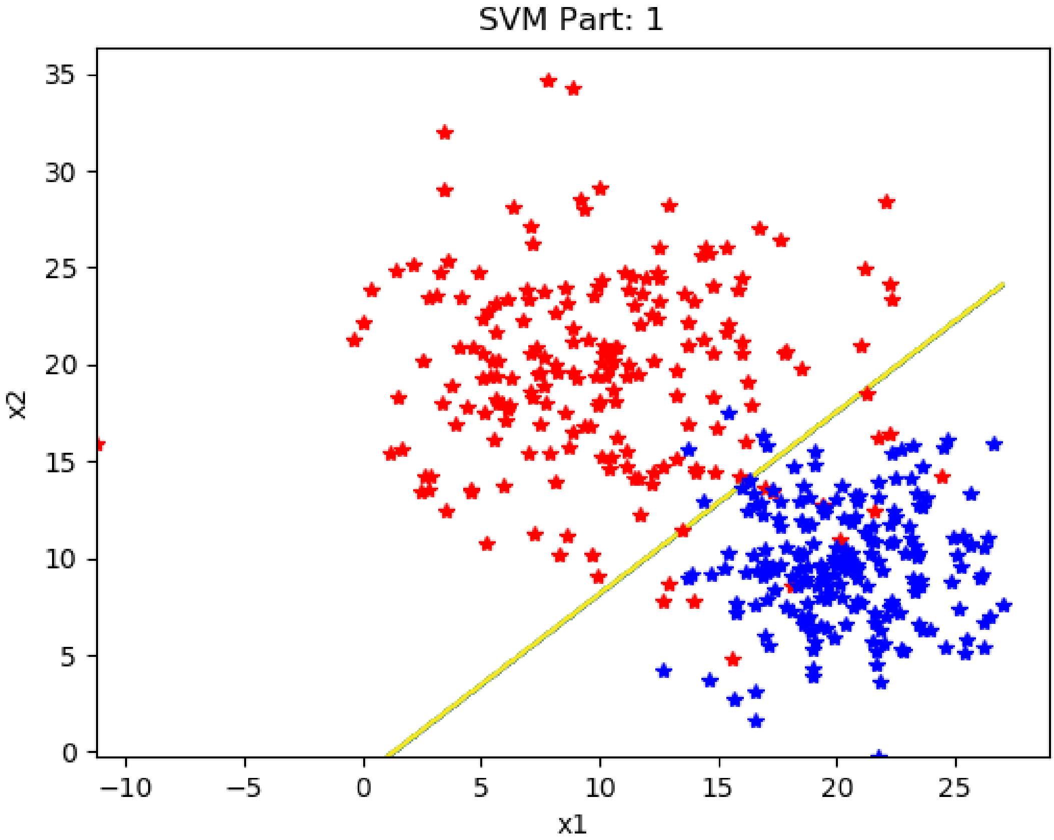
For the second part when binclassv2.txt was used which contains some outliers points for the red points.
In generative case, the decision boundary comes more close to the blue clusters, possible reason could be that since sigma variance for red will increase due to the outliers so the boundary will come closer to the blue one.
But for SVM case there was a little shift toward red clusters, possibly due to increase in the no of red crossovers the new decision had tried to shift a little to decrease other crossovers! which we can see from the image that there is more no of points on the decision boundary for part2

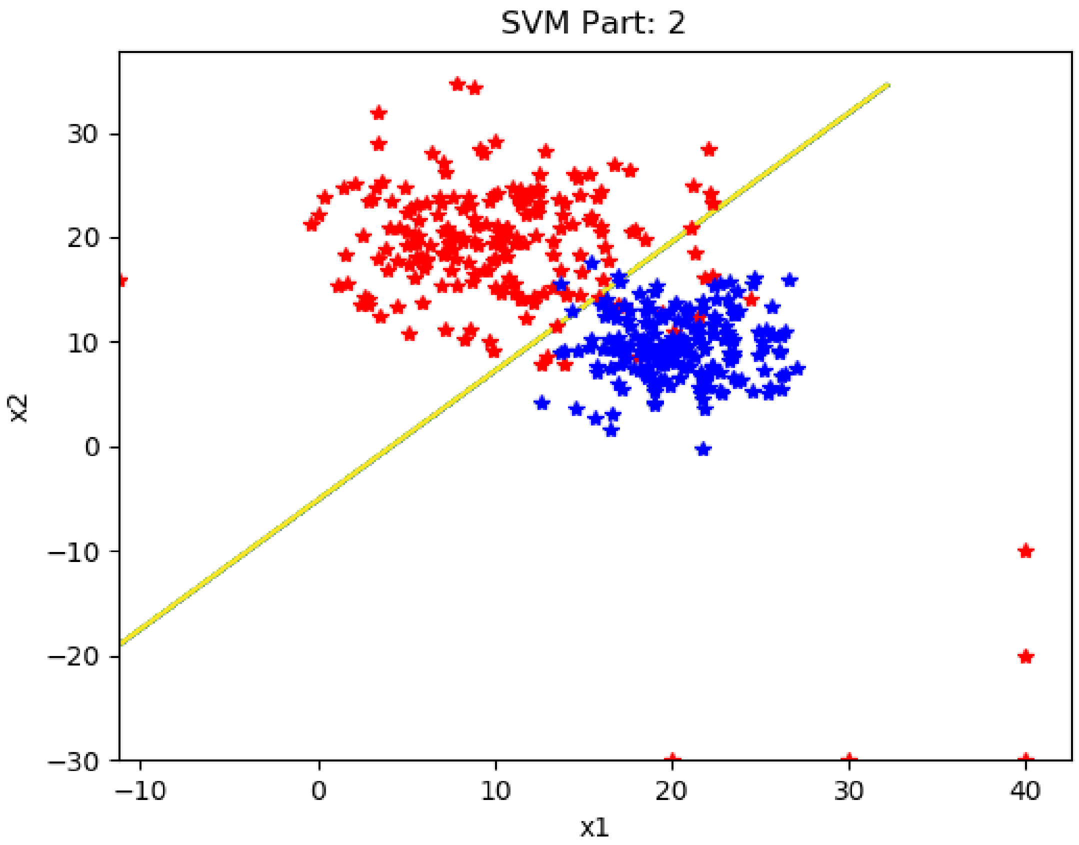



Different Sigma Part: 2
x2
30
20
10
0
−10
−20
−30
−10
0
10
20
30
40
x1


Same Sigma Part: 1
35
30
25
20
15
10
5
0
−10
−5
0
5
10
15
20
25
x1
x2


Same Sigma Part: 2
x2
30
20
10
0
−10
−20
−30
−10
0
10
20
30
40
x1


SVM Part: 1
x2
35
30
25
20
15
10
5
0
−10
−5
0
5
10
15
20
25
x1


SVM Part: 2
x2
x1
30
20
10
0
−10
−20
−30
−10
0
10
20
30
40