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

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 30, 2018

QUESTION 1

We are given the logistic regression model $P(y_n|\mathbf{x}_n, \mathbf{w}) = \frac{1}{1+exp(-y_n\mathbf{w}^T\mathbf{x}_n)}$ with zero mean Gaussian prior $p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$. Also, it is mentioned that $y_n \in \{-1, +1\}$.

Now,

$$P(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} P(y_n|\mathbf{x}_n, \mathbf{w})$$

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

And,

$$P(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$$

$$P(\mathbf{w}) = \frac{\lambda^{1/2}}{(2\pi)^{D/2}} exp(-\frac{\lambda}{2}\mathbf{w}^T\mathbf{w})$$

MAP Estimation for w :

$$\hat{\mathbf{w}}_{map} = \underset{w}{\text{argmax}} log(P(\mathbf{w}|\mathbf{y}, \mathbf{X}))$$

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{X}, \mathbf{w})P(\mathbf{w})}{P(\mathbf{y}|\mathbf{X})}$$

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto P(\mathbf{y}|\mathbf{X}, \mathbf{w})P(\mathbf{w})$$

$$log(P(\mathbf{w}|\mathbf{y}, \mathbf{X})) \propto log(P(\mathbf{y}|\mathbf{X}, \mathbf{w})) + log(P(\mathbf{w}))$$

$$log(P(\mathbf{w}|\mathbf{y}, \mathbf{X})) \propto \sum_{n=1}^{N} log(\frac{1}{1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)}) - \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

$$\hat{\mathbf{w}}_{map} = \underset{w}{\text{argmax}} \sum_{n=1}^{N} log(\frac{1}{1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)}) - \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

$$\hat{\mathbf{w}}_{map} = \underset{w}{\text{argmax}} \sum_{n=1}^{N} -log(1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)) - \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

$$\hat{\mathbf{w}}_{map} = \underset{w}{\text{argmin}} \sum_{n=1}^{N} log(1 + exp(-y_n\mathbf{w}^T\mathbf{x}_n)) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

Taking the derivative w.r.t $\mathbf{w}$ and equating it to zero

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

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

Comparing with $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$ gives

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

$$\alpha_n = \frac{1}{\lambda}(1 - P(y_n | \mathbf{x}_n, \mathbf{w}))$$

Let $\mu_n = P(y_n | \mathbf{x}_n, \mathbf{w})$. Therefore α_n has the form

$$\alpha_n = \frac{1}{\lambda}(1 - \mu_n)$$

Interpretation of α_n:

If we talk about the perceptron model, then α_n denotes the total number of mispredictions for a training example $(\mathbf{x}_n, y_n)$. In this question also, suppose for a training example $\mathbf{x}_n$, y_n is 1 but the above algorithm mispredicts it, then μ_n will be close to 0 and therefore, $1 - \mu_n$ will be close to 1. So, in the above algorithm more weightage will be given to the mispredicted training examples while updating the value of $\mathbf{w}$. Therefore, this interpretation of α_n makes sense.

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

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 30, 2018

QUESTION 2

Given,
Class Marginal Distribution as $P(y=1)=\pi$
Class Conditional Distribution as:

$$P(\mathbf{x}|y=1)=\prod_{d=1}^{D}P(x_d|y=1)=\prod_{d=1}^{D}\mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}$$

$$P(\mathbf{x}|y=0)=\prod_{d=1}^{D}P(x_d|y=0)=\prod_{d=1}^{D}\mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d}$$

Derivation of the expression for $P(y=1|\mathbf{x})$

$$P(y=1|\mathbf{x})=\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x})}$$

$$P(y=1|\mathbf{x})=\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=1)P(y=1)+P(\mathbf{x}|y=0)P(y=0)}$$

$$P(y=1|\mathbf{x})=\frac{\pi\prod_{d=1}^{D}\mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}}{\pi\prod_{d=1}^{D}\mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}+(1-\pi)\prod_{d=1}^{D}\mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d}}$$

Decision Boundary Computation

$$P(y=1|\mathbf{x})=P(y=0|\mathbf{x})$$

$$\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x})}=\frac{P(\mathbf{x}|y=0)P(y=0)}{P(\mathbf{x})}$$

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

Taking log both sides

$$log(\pi)+\sum_{d=1}^{D}x_d log(\mu_{d,1}+(1-x_d)log(1-\mu_{d,1})=log(1-\pi)+\sum_{d=1}^{D}x_d log(\mu_{d,0}+(1-x_d)log(1-\mu_{d,0})$$

implies

$$\sum_{d=1}^{D}x_d(log(\mu_{d,1})-log(1-\mu_{d,1}))=\sum_{d=1}^{D}x_d(log(\mu_{d,0})-log(1-\mu_{d,0}))+k_1$$

for some constant k_1.
Therefore, the decision boundary is linear.

Expressions for the parameters

For the probabilistic discriminative model,

$$P(y=1|\mathbf{x})=\frac{exp(\mathbf{w}^T\mathbf{x}+b)}{1+exp(\mathbf{w}^T\mathbf{x}+b)}$$

Comparing this with the following equation

$$P(y=1|\mathbf{x}) = \frac{\pi \prod_{d=1}^{D} \mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}}{\pi \prod_{d=1}^{D} \mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d} + (1-\pi)\prod_{d=1}^{D} \mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d}}$$

which can also be written as

$$P(y=1|\mathbf{x}) = \frac{\frac{\pi}{1-\pi} \prod_{d=1}^{D} \frac{\mu_{d,1}}{\mu_{d,0}}^{x_d}(\frac{1-\mu_{d,1}}{1-\mu_{d,0}})^{1-x_d}}{1 + \frac{\pi}{1-\pi} \prod_{d=1}^{D} \frac{\mu_{d,1}}{\mu_{d,0}}^{x_d}(\frac{1-\mu_{d,1}}{1-\mu_{d,0}})^{1-x_d}}$$

Therefore,

$$exp(\mathbf{w}^T\mathbf{x} + b) = \frac{\pi}{1-\pi} \prod_{d=1}^{D} \frac{\mu_{d,1}}{\mu_{d,0}}^{x_d}(\frac{1-\mu_{d,1}}{1-\mu_{d,0}})^{1-x_d}$$

Taking log both sides

$$log\Big(\frac{\pi}{1-\pi}\Big) + \sum_{d=1}^{D} x_d log\Big(\frac{\mu_{d,1}}{\mu_{d,0}}\Big) + (1-x_d) log\Big(\frac{1-\mu_{d,1}}{1-\mu_{d,0}}\Big) = \sum_{d=1}^{D} w_d x_d + b$$

Therefore,

$$w_d = log\Big(\frac{\mu_{d,1}}{\mu_{d,0}}\Big) - log\Big(\frac{1-\mu_{d,1}}{1-\mu_{d,0}}\Big)$$

$$b = log\Big(\frac{\pi}{1-\pi}\Big) + \sum_{d=1}^{D} log\Big(\frac{1-\mu_{d,1}}{1-\mu_{d,0}}\Big)$$

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

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 30, 2018

QUESTION 3

Given,

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2$$

s.t. $\|w\| \leq c$ where $c > 0$

Now,

$$\|w\| \leq c$$

is equivalent to

$$\|w\|^2 - c^2 \leq 0$$

Therefore, above problem is equivalent to solving the following optimization problem using lagrangian method

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \underset{\alpha \geq 0}{\operatorname{argmax}} \left(\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \alpha(\|w\|^2 - c^2) \right)$$

Now, as both the terms in this expression are convex. So, the solutions to both the primal as well as the dual problem will be the same. As a result, the optimal value of α will be independent of $\mathbf{w}$. Let the optimal value of α be $\hat{\alpha}$

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left(\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \hat{\alpha}(\|w\|^2 - c^2) \right)$$

Again solving this problem is equivalent to solving the following optimization problem

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \hat{\alpha} \|w\|^2$$

which is same as logistic regression with l_2 regularization with hyperparameter $\frac{\lambda}{2}$ replaced by $\hat{\alpha}$

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

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 30, 2018

QUESTION 4

We are given that $P(y_n = k|\mathbf{x}_n, \mathbf{W}) = \mu_{nk} = \frac{exp(\mathbf{w}_k^T \mathbf{x}_n)}{\sum_{l=1}^{K} exp(\mathbf{w}_l^T \mathbf{x}_n)}$.

Derivation of MLE Solution :

$$P(y|\mathbf{X}, \mathbf{W}) = \prod_{n=1}^{N} P(y_n|\mathbf{x}_n, \mathbf{W}) = \prod_{n=1}^{N} \prod_{l=1}^{K} \mu_{nl}^{y_{nl}}$$

where $y_{nl} = 1$ iff $y_n = l$ and 0 otherwise.
Taking log both sides

$$log(P(y|\mathbf{X}, \mathbf{W})) = \sum_{n=1}^{N} \sum_{l=1}^{K} y_{nl} log(\mu_{nl})$$

$$log(P(y|\mathbf{X}, \mathbf{W})) = \sum_{n=1}^{N} \sum_{l=1}^{K} y_{nl} log\left(\frac{exp(\mathbf{w}_l^T \mathbf{x}_n)}{\sum_{m=1}^{K} exp(\mathbf{w}_m^T \mathbf{x}_n)}\right)$$

$$log(P(y|\mathbf{X}, \mathbf{W})) = \sum_{n=1}^{N} \sum_{l=1}^{K} y_{nl} (\mathbf{w}_l^T \mathbf{x}_n - log\Big(\sum_{m=1}^{K} exp(\mathbf{w}_m^T \mathbf{x}_n)\Big)$$

Derivate the above equation w.r.t $\mathbf{w}_k$

$$\frac{\partial log(P(y|\mathbf{X}, \mathbf{W}))}{\partial \mathbf{w}_k} = \sum_{n=1}^{N} y_{nk} \mathbf{x}_n - \sum_{n=1}^{N} \sum_{l=1}^{K} \frac{exp(\mathbf{w}_k^T \mathbf{x}_n)}{\sum_{m=1}^{K} exp(\mathbf{w}_m^T \mathbf{x}_n)} y_{nl} \mathbf{x}_n$$

which is equivalent to

$$\frac{\partial log(P(y|\mathbf{X}, \mathbf{W}))}{\partial \mathbf{w}_k} = \sum_{n=1}^{N} y_{nk} \mathbf{x}_n - \sum_{n=1}^{N} \frac{exp(\mathbf{w}_k^T \mathbf{x}_n)}{\sum_{m=1}^{K} exp(\mathbf{w}_m^T \mathbf{x}_n)} \mathbf{x}_n$$

$$\frac{\partial log(P(y|\mathbf{X}, \mathbf{W}))}{\partial \mathbf{w}_k} = \sum_{n=1}^{N} \mathbf{x}_n \left(y_{nk} - \frac{exp(\mathbf{w}_k^T \mathbf{x}_n)}{\sum_{m=1}^{K} exp(\mathbf{w}_m^T \mathbf{x}_n)}\right)$$

$$\frac{\partial log(P(y|\mathbf{X}, \mathbf{W}))}{\partial \mathbf{w}_k} = \sum_{n=1}^{N} \mathbf{x}_n (y_{nk} - \mu_{nk})$$

as $\sum_{l=1}^{K} y_{nl} = 1$ Therefore, the gradient descent step can be written as :

$$\mathbf{w}_k^{(t+1)} = \mathbf{w}_k^{(t)} + 1 * (\sum_{n=1}^{N} \mathbf{x}_n (y_{nk} - \mu_{nk}))$$

For SGD, pick a training example $\{\mathbf{x}_n, y_n\}$ randomly from $n = 1, 2, ..., N$ and approximate the gradient step as follows :

$$\mathbf{w}_k^{(t+1)} = \mathbf{w}_k^{(t)} + 1 * \mathbf{x}_n (y_{nk} - \mu_{nk})$$

For the special case, replace μ_{nk} by the following term :

$$\frac{exp(\mathbf{w}_{y_n}^T \mathbf{x}_n)}{\sum_{m=1}^{K} exp(\mathbf{w}_m^T \mathbf{x}_n)}$$

and then the gradient descent step can be written as :

$$\mathbf{w}_k^{(t+1)} = \mathbf{w}_k^{(t)} + 1 * \left(\sum_{n=1}^{N} \mathbf{x}_n \Big(y_{nk} - \frac{exp(\mathbf{w}_{y_n}^T \mathbf{x}_n)}{\sum_{m=1}^{K} exp(\mathbf{w}_m^T \mathbf{x}_n)} \Big) \right)$$

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

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 30, 2018

QUESTION 5

We need to show that the set of $\mathbf{x}'s$ and the set of $\mathbf{y}'s$ are linearly separable if and only if the convex hulls do not intersect.
To show this we have to show these two conditions :
1. If $\mathbf{x}_n$ and $\mathbf{y}_m$ are linearly separable then their convex hulls do not intersect.
2. If $\mathbf{x}_n$ and $\mathbf{y}_m$ have non intersecting convex hulls then they are linearly separable.

One Way :
As the $\mathbf{x}_n$ and $\mathbf{y}_m$ are linearly separable, therefore, there exists $\mathbf{w}$ and b s.t. for all $\mathbf{x}'_n s$ and $\mathbf{y}'_m s$

$$\mathbf{w}^T\mathbf{x}_n + b > 1$$

and,

$$\mathbf{w}^T\mathbf{y}_m + b < 1$$

Now, assume that the two convex hulls intersect then there exists γ s.t. $\gamma \in \mathbf{x}$ and $\gamma \in \mathbf{y}$ where

$$\gamma = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n = \sum_{m=1}^{M} \beta_m \mathbf{y}_m$$

Also, $\sum_{n=1}^{N} \alpha_n = 1$ and $\sum_{m=1}^{M} \beta_m = 1$
Multiply the above equation by $\mathbf{w}^T$ and add b to both the sides, we get

$$\sum_{n=1}^{N} \alpha_n \mathbf{w}^T\mathbf{x}_n + b = \sum_{m=1}^{M} \beta_m \mathbf{w}^T\mathbf{y}_m + b$$

which is equivalent to

$$\sum_{n=1}^{N} \alpha_n \mathbf{w}^T\mathbf{x}_n + \sum_{n=1}^{N} \alpha_n b = \sum_{m=1}^{M} \beta_m \mathbf{w}^T\mathbf{y}_m + \sum_{m=1}^{M} \beta_m b$$

Therefore,

$$\sum_{n=1}^{N} \alpha_n (\mathbf{w}^T\mathbf{x}_n + b) = \sum_{m=1}^{M} \beta_m (\mathbf{w}^T\mathbf{y}_m + b)$$

Now, the L.H.S is always > 1 and R.H.S is always < 1. Therefore, there is a contradiction. So, the convex hulls cannot intersect.

Other Way Around:
Now suppose the convex hulls don't intersect. Then, there exists pair of points $\mathbf{x}^*$ and $\mathbf{y}^*$ s.t. $\mathbf{x}^* \in \mathbf{x}$ and $\mathbf{y}^* \in \mathbf{y}$ where $d(\mathbf{x}_n, \mathbf{y}_m)$ is minimum for $\mathbf{x}_n = \mathbf{x}^*$ and $\mathbf{y}_m = \mathbf{y}^*$.
Consider the perpendicular bisector of line joining $\mathbf{x}^*$ and $\mathbf{y}^*$, it divides the region into 2 parts one containing $\mathbf{x}^*$ and other $\mathbf{y}^*$. Now, if there is any point belonging to $\mathbf{x}$ s.t. it lies in the region containing $\mathbf{y}^*$ or vice-versa i.e. any point belonging to $\mathbf{y}$ s.t. it lies in the region containing $\mathbf{x}^*$, then the condition that $d(\mathbf{x}_n, \mathbf{y}_m)$ is minimum for $\mathbf{x}_n = \mathbf{x}^*$ and $\mathbf{y}_m = \mathbf{y}^*$ is violated, so it's not possible. Therefore, we have found a plane separating the set of points belonging to two convex hulls $\mathbf{x}$ and $\mathbf{y}$, so we can say that they are linearly separable.

Therefore, both the conditions 1 and 2 hold true.

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

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 30, 2018

QUESTION 6

Let the hyperplane learned for the condition $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$ be $\hat{\mathbf{w}}^T\mathbf{x}_n + \hat{b} = 0$
And now the condition is modified to $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq m$ where $m \neq 0$
which is equivalent to

$$y_n(\frac{\mathbf{w}^T}{m}\mathbf{x}_n + \frac{b}{m}) \geq 1$$

Let $\mathbf{w}^{'} = \frac{\mathbf{w}}{m}$ and $b^{'} = \frac{b}{m}$
So, the above equation becomes

$$y_n(\mathbf{w}^{'T}\mathbf{x}_n + b^{'}) \geq 1$$

which is the same as the first equation, so the solution will be $\mathbf{w}^{'} = \hat{\mathbf{w}}$ and $b^{'} = \hat{b}$
or, $\mathbf{w} = m\hat{\mathbf{w}}$ and $b = m\hat{b}$
So, the learned hyperplane will be

$$(m\hat{\mathbf{w}})^T\mathbf{x}_n + m\hat{b} = 0$$

which is equivalent to

$$\hat{\mathbf{w}}^T\mathbf{x}_n + \hat{b} = 0 \qquad \text{as } m \neq 0$$

So, the learned hyperplane still remains the same.

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 30, 2018

QUESTION 7

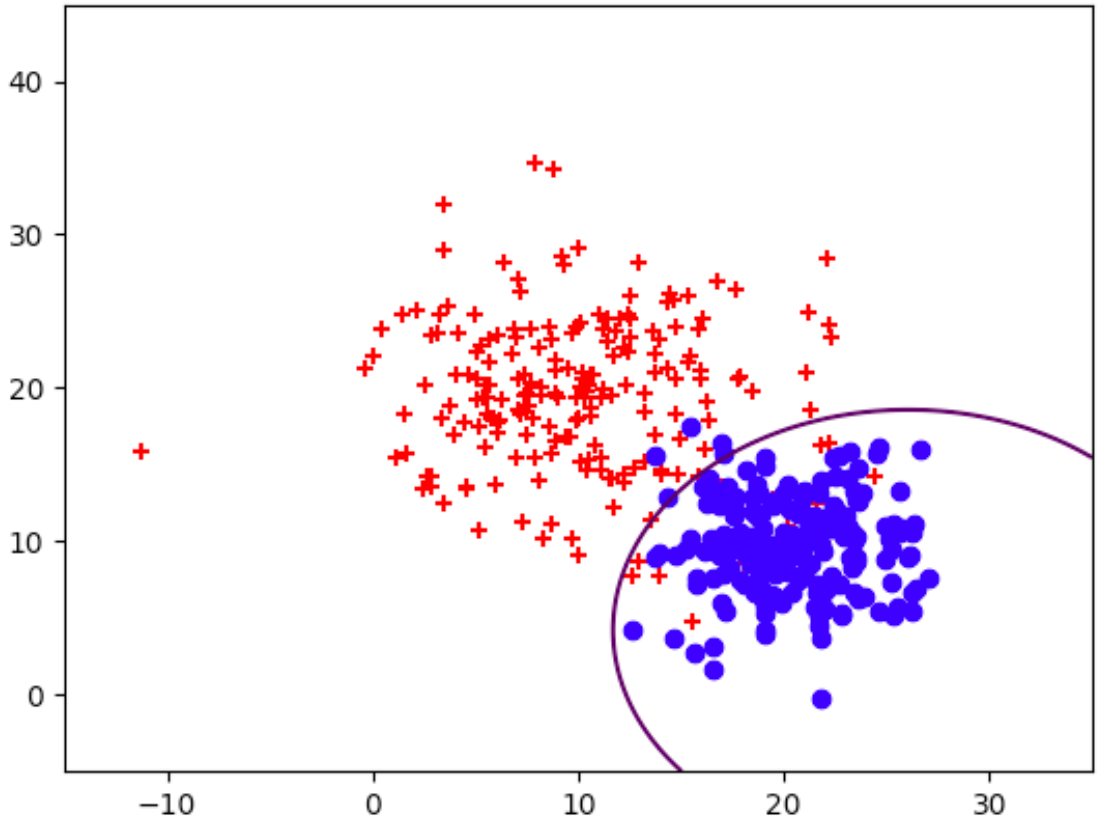

Figure 1: Learned Decision Boundary for binclass.txt

Analysis :

For the first one i.e. binclass.txt, the SVM model performs better in comparison to generative classification with Gaussian class conditional. On the other hand, for second one i.e. binclass2.txt, generative classification with Gaussian class conditional performs better than SVM model. This is primarily because of presence of outliers in the second one which are absent in the first one.

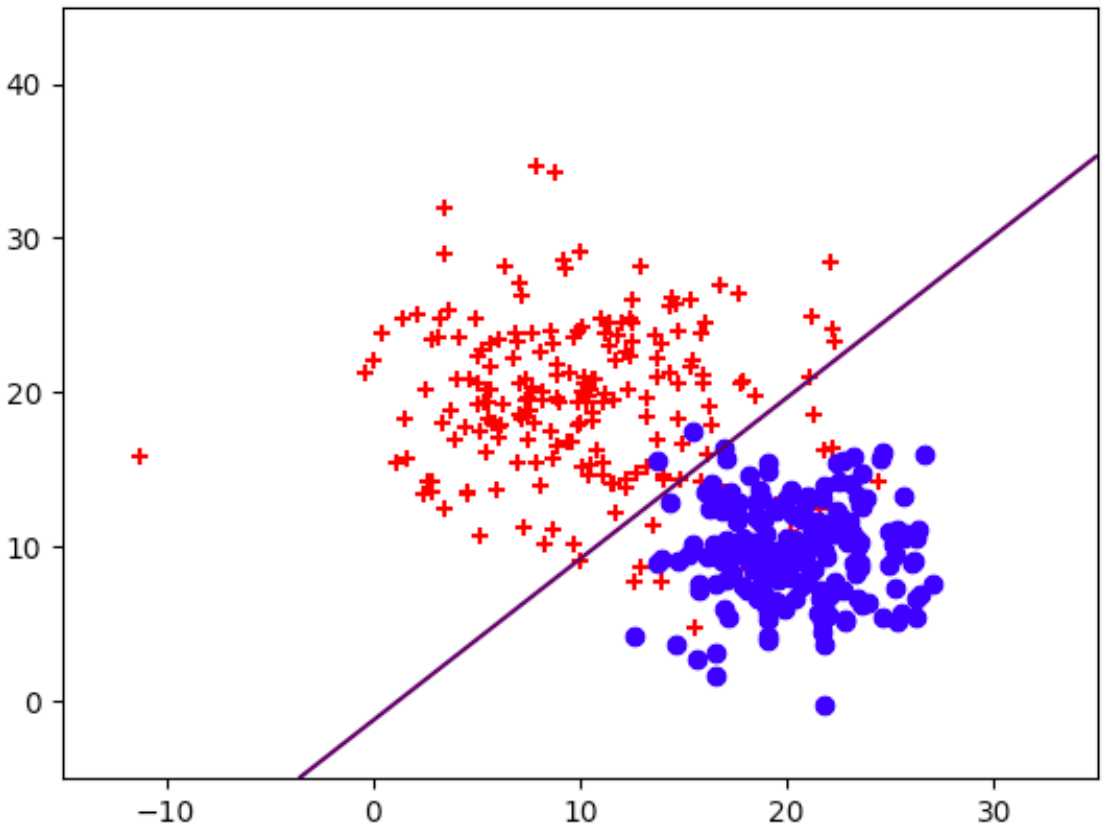

Figure 2: Learned Decision Boundary with $\sigma = \sigma_+ = \sigma_-$ for binclass.txt

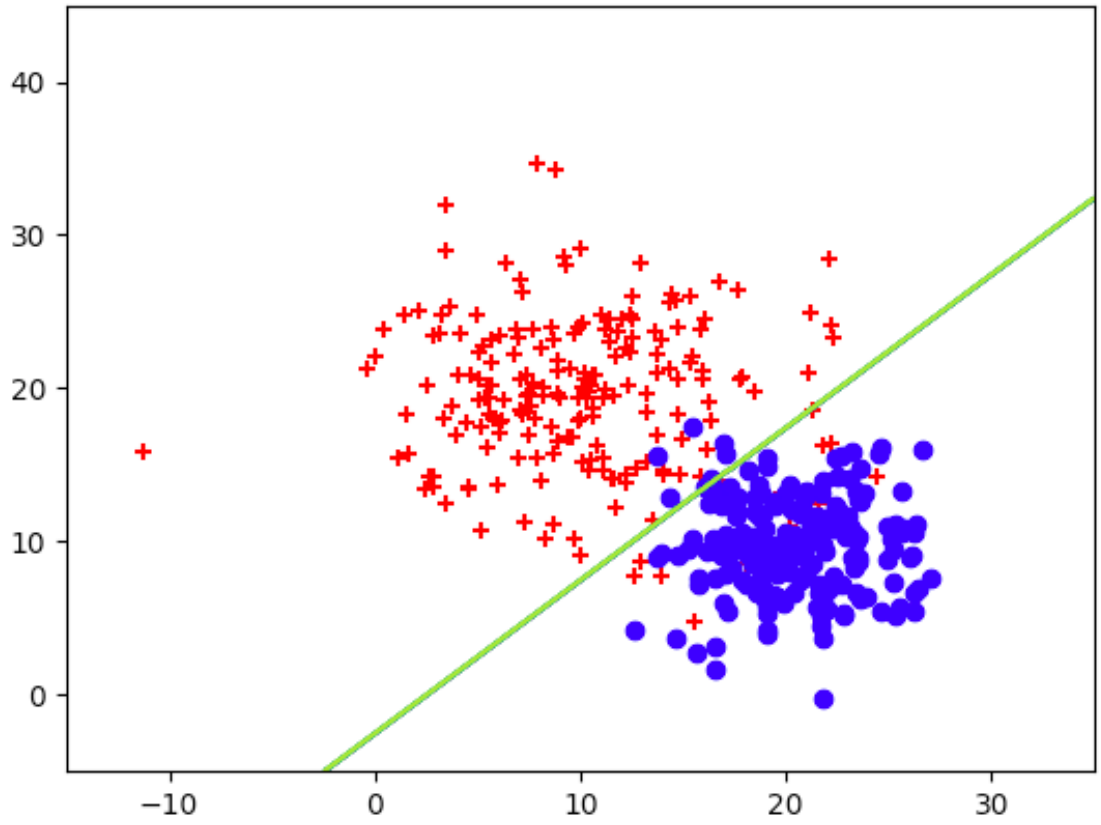

Figure 3: Using SVM Classifier for binclass.txt

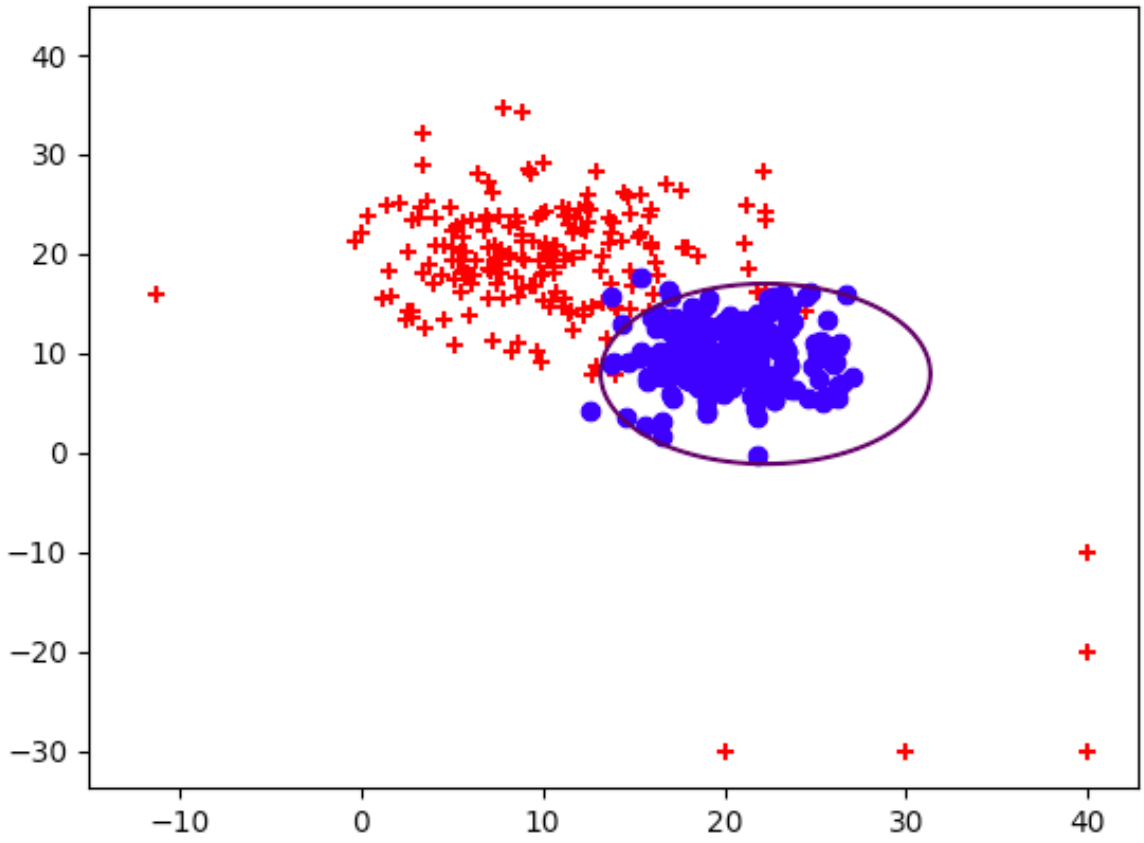

Figure 4: Learned Decision Boundary for binclass2.txt

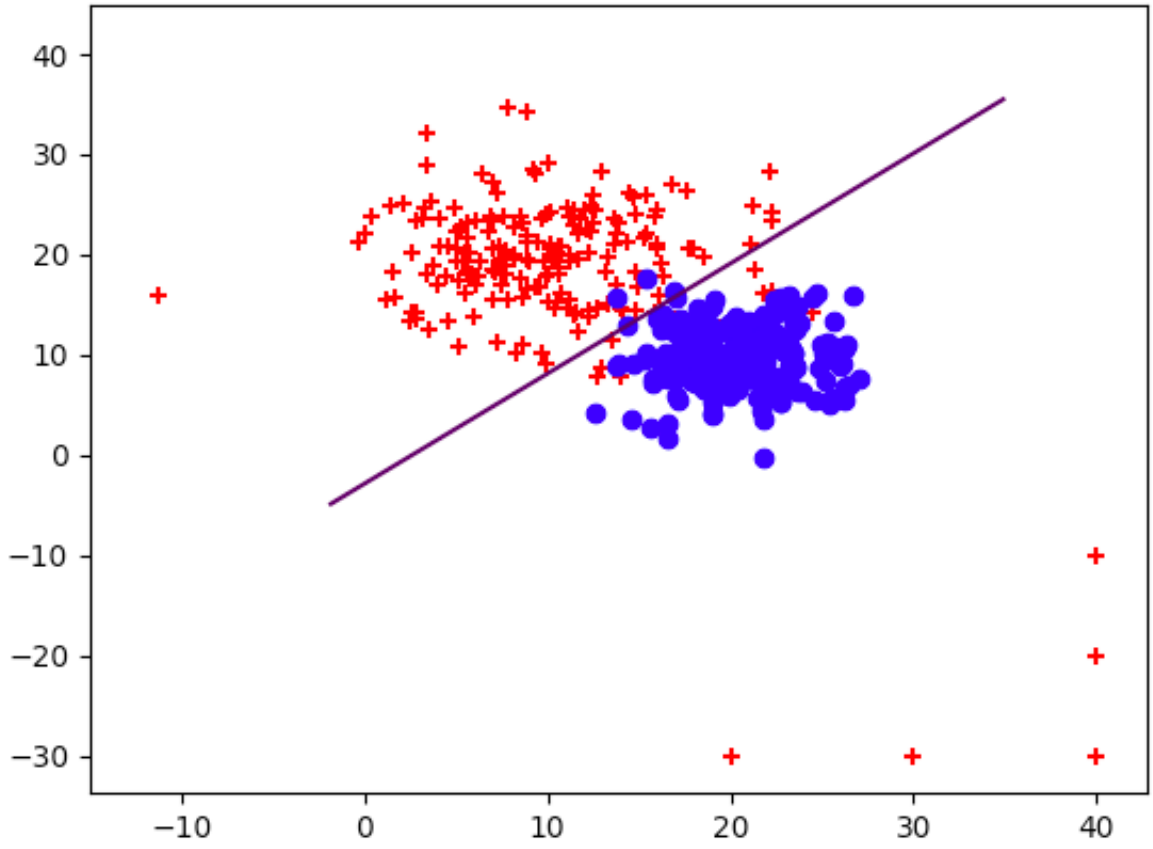

Figure 5: Learned Decision Boundary with $\sigma = \sigma_+ = \sigma_-$ for binclass2.txt

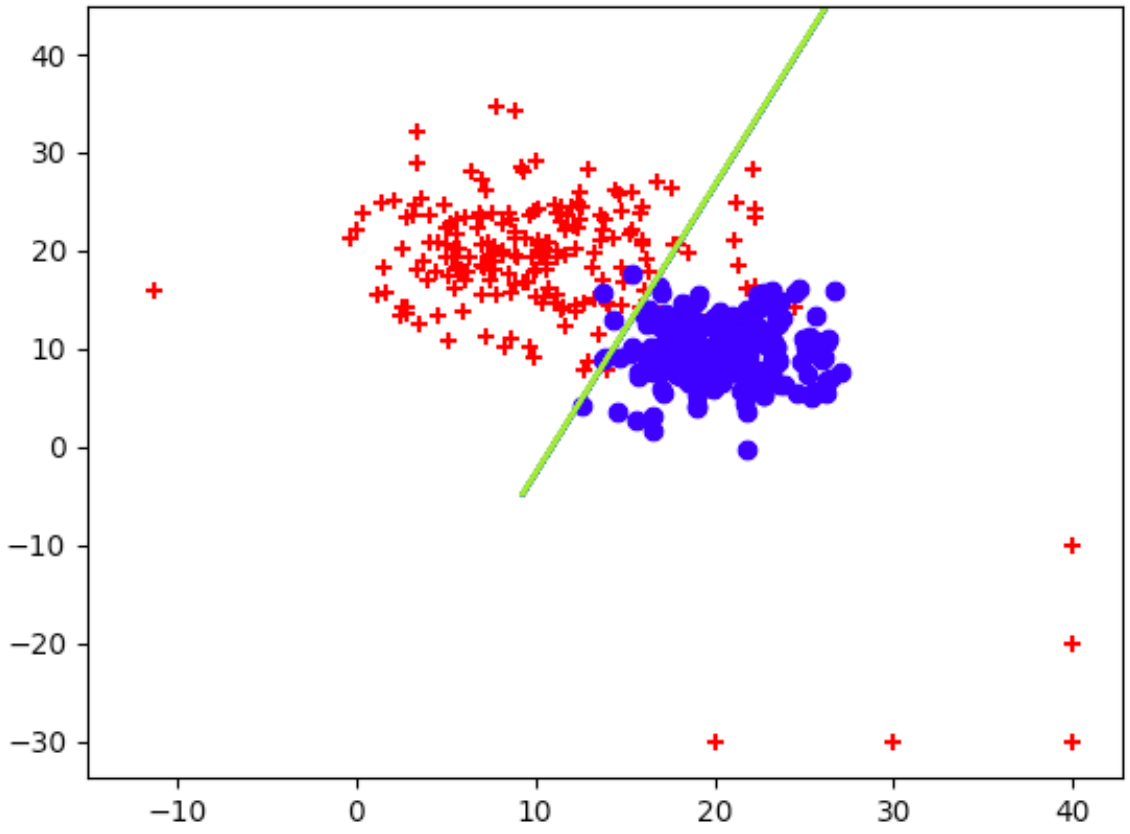

Figure 6: Using SVM Classifier for binclass2.txt