Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Aman Tiwari
Roll Number: 160094
Date: April 18, 2019

QUESTION 1

Solution 1:

$p(y_n|\mathbf{x_n},\mathbf{w}) = \frac{1}{1+exp(-y_n\mathbf{w}^T\mathbf{x})} = \mu_n$

$p(w) = \mathscr{N}(0,\lambda^{-1}\mathbf{I}) = (\frac{\lambda}{2\pi})^{\frac{D}{2}} exp(\frac{-\lambda}{2}\mathbf{w}^T\mathbf{w})$

$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}}(\log p(\mathbf{y}|\mathbf{x},\mathbf{w}) + \log p(\mathbf{w}))$

$\hat{\mathbf{w}}_{MAP} = \arg\max_{\mathbf{w}}(\sum_{n=1}^{N}\log(\frac{1}{1+exp(-y_n\mathbf{w}^T\mathbf{x})}) + (\frac{-\lambda}{2}\mathbf{w}^T\mathbf{w}))$

Taking derivative with respect to $\mathbf{w}$ and setting it to zero we get the following equation:

$\sum_{n=1}^{N}\frac{exp(-y_n\mathbf{w}^T\mathbf{x_n})}{(1+exp(-y_n\mathbf{w}^T x_n))}y_n\mathbf{x_n} - \frac{-\lambda}{2}\mathbf{w} = 0$

Let $\alpha_n = \frac{2}{\lambda}\frac{exp(-y_n\mathbf{w}^T\mathbf{x_n})}{(1+exp(-y_n\mathbf{w}^T x_n))}$

$\alpha_n = \frac{2}{\lambda}(1-\mu_n)$

Then we get the following solution:

$\hat{\mathbf{w}}_{MAP} = \sum_{n=1}^{N}\alpha_n y_n\mathbf{x_n}$

If a value of μ_n is high for a given training example then the value of $alpha_n$ is low for that training example. So in $\hat{\mathbf{w}}_{MAP}$ the examples for which μ_n is high contribute low towards the weight vector $\mathbf{w}$ and the examples for which μ_n is low contribute hight towards weight vector . This makes sense because the examples which are wrongly classified contribute more towards $\mathbf{w}$ and increase it in the direction which makes $\mathbf{w}$ to classify them correctly.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 2

Student Name: Aman Tiwari
Roll Number: 160094
Date: April 18, 2019

Solution: 2

$p(x_d|y=1) = {\mu_{d,1}}^{x_d}(1-\mu_{d,1})^{1-x_d}$

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

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

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

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

Let us consider a discriminative classifier:

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

Comparing the above model with this we get the following:

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

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

This gives us the following:

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

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

The decision boundary is linear in this case.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Aman Tiwari
Roll Number: 160094
Date: April 18, 2019

QUESTION 3

Solution 3:

The lagrangian for the given constraint optimization problem is:
$\mathcal{L}(\mathbf{w}, \alpha) = \sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x})^2 + \alpha(\mathbf{w}^T\mathbf{w} - c^2)$
This can also be written as:
$\mathcal{L}(\mathbf{w}, \alpha) = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \alpha(\mathbf{w}^T\mathbf{w} - c^2)$
$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} + 2\alpha\mathbf{w}$
Setting the derivative to zero we get:
$\mathbf{w}_* = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$
Now we can formulate the dual problem and solve it to get the optimal value of α. Let it be α_*.
This gives the solution of the lagrangian as:
$\mathbf{w}_* = (\mathbf{X}^T\mathbf{X} + \alpha_*\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$
If we formulate the least squares linear regression problem with l_2 regularization as:
$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x})^2 + \lambda(\mathbf{w}^T\mathbf{w})$
Then its optimal solution is given as:
$\mathbf{w}_*' = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$
Clearly for $\mathbf{w}_*'$ to be a solution of the lagrangian problem we can choose $\lambda = \alpha_*$.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Aman Tiwari
Roll Number: 160094
Date: April 18, 2019

QUESTION 4

Solution 4:

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

Here $\mu_{nl} = \frac{exp({\mathbf{w_l}}^T\mathbf{x_n})}{\sum_{k=1}^{K} exp({\mathbf{w_k}}^T\mathbf{x_n})}$ and $y_{nl} = 1$ if true class of example n is l and $y_{nk} = 0$ if $k \neq l$.

$\mathcal{L} = -\log(p(\mathbf{y}|\mathbf{x},\mathbf{W})) = \sum_{n=1}^{N}\sum_{l=1}^{K}\log({\mu_{nl}}^{y_{nl}})$

$\mathcal{L} = -\sum_{l=1}^{K}\sum_{n:y_n=l}\log(\mu_{nl})$

$\mathcal{L} = -\sum_{l:l\neq k}\sum_{n:y_n=l}\log(\mu_{nl}) - \sum_{n:y_n=k}\log(\mu_{nk})$

$\frac{\partial\mathcal{L}}{\partial\mathbf{w_k}} = \sum_{n:y_n\neq k}\mu_{nk}\mathbf{x_n} - \sum_{n:y_n=k}(1-\mu_{nk})\mathbf{x_n}$

The GD update equation is given as:

${\mathbf{w_k}}^{(t+1)} = {\mathbf{w_k}}^{(t)} - (\sum_{n:y_n\neq k}\mu_{nk}\mathbf{x_n} - \sum_{n:y_n=k}(1-\mu_{nk})\mathbf{x_n})$

The SGD algorithm for the problem is given as:

```
Initialize $\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_k}]$ as $[{\mathbf{w_1}}^{(0)}, {\mathbf{w_2}}^{(0)}, ..., {\mathbf{w_k}}^{(0)}]$
Pick a random $i \in \{1, 2, 3, ...., N\}$
for k in range 1 to K:
If $y_{ik} = 1$:
        ${\mathbf{w_k}}^{(t+1)} = {\mathbf{w_k}}^{(t)} + (1 - \mu_{ik})\mathbf{x_n}$
else:
        ${\mathbf{w_k}}^{(t+1)} = {\mathbf{w_k}}^{(t)} - (\mu_{ik})\mathbf{x_n}$
Do the above steps until convergence.
```

If we use the hard class assignments then the SGD algorithm is as follows:

```
Initialize $\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_k}]$ as $[{\mathbf{w_1}}^{(0)}, {\mathbf{w_2}}^{(0)}, ..., {\mathbf{w_k}}^{(0)}]$
Pick a random $i \in \{1, 2, 3, ...., N\}$
for k in range 1 to K:
If $y_{ik} = 1$ and $\mu_{ik} = 0$:
        ${\mathbf{w_k}}^{(t+1)} = {\mathbf{w_k}}^{(t)} + \mathbf{x_n}$
else if $y_{ik} = 0$ and $\mu_{ik} = 1$:
        ${\mathbf{w_k}}^{(t+1)} = {\mathbf{w_k}}^{(t)} - \mathbf{x_n}$
Do the above steps until convergence.
```

In the hard margin case once $\mu_{nk} = 1$ for $y_{nk} = 1$ the weight $\mathbf{w}_k$ is no longer updated and it remains constant. Similarly if at some stage in the SGD $\mu_{nk} = 0$ for $y_{nk} = 0$ the weight $\mathbf{w}_k$ is no longer updated.

This is not the case in soft-margin case , there the weight values are updated even if the probability of the correct class is greater is than all the incorrect classes.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 5

Student Name: Aman Tiwari
Roll Number: 160094
Date: April 18, 2019

Solution 5:

Let us define a linear function $z(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ for some vector $\mathbf{w}$ and some constant b

Consider any point $\mathbf{x}$ in the first convex hull:

$z(\mathbf{x}) = \mathbf{w}^T(\sum_{n=1}^{N} \alpha_n \mathbf{x_n}) + b$

$z(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n(\mathbf{w}^T\mathbf{x_n}) + b$

$z(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n(\mathbf{w}^T\mathbf{x_n} + b)$ using the fact that $\sum_{n=1}^{N} \alpha_n = 1$

Similarly for a point $\mathbf{y}$ in the second convex hull we can define the linear function z as:

$z(\mathbf{y}) = \sum_{m=1}^{M} \beta_m(\mathbf{w}^T\mathbf{y_m} + b)$

If the points are linearly separable: In this case there exist some $\mathbf{w}$ and b such that:

$\mathbf{w}^T\mathbf{x_n} + b > 0, n = 1, 2, ..., N$ and

$\mathbf{w}^T\mathbf{y_m} + b < 0, m = 1, 2, ..., M$

Assume that there is a point $\mathbf{x}$ in the first convex hull which is same as the point $\mathbf{y}$ in the second convex hull. Then , $z(\mathbf{x}) = z(\mathbf{y})$ for the choice of $\mathbf{w}$ and b which ensures linear separability of the two points. This means :

$\sum_{n=1}^{N} \alpha_n(\mathbf{w}^T\mathbf{x_n} + b) = \sum_{m=1}^{M} \beta_m(\mathbf{w}^T\mathbf{y_m} + b)$

But this is impossible since left hand side is negative and right hand side is positive.

If the two convex hulls intersect: Let $\mathbf{x}$ be a point in first convex hull which is equal to a point $\mathbf{y}$ in the second convex hull. This means :

$\sum_{n=1}^{N} \alpha_n(\mathbf{w}^T\mathbf{x_n} + b) = \sum_{m=1}^{M} \beta_m(\mathbf{w}^T\mathbf{y_m} + b)$

This is true for any arbitrary choice of vector $\mathbf{w}$ and constant b.

For linear separability we must have:

$\mathbf{w}^T\mathbf{x_n} + b > 0, n = 1, 2, ..., N$ and

$\mathbf{w}^T\mathbf{y_m} + b < 0, m = 1, 2, ..., M$ for some choice of $\mathbf{w}$ and b.

But if such $\mathbf{w}$ and b exists then it would violate the condition mentioned above since left side would be positive and right side would be negative.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Aman Tiwari
Roll Number: 160094
Date: April 18, 2019

QUESTION 6

Solution 6:

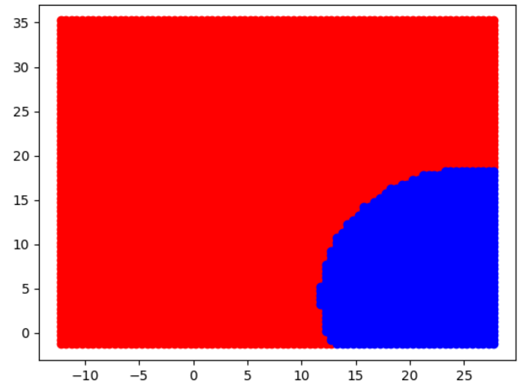
The lagrangian formed on using the modified condition is:
$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{\mathbf{w}^T w}{2} + \sum_{n=1}^{N} \alpha_n (m - y_n(\mathbf{w}^T \mathbf{x_n} + b))$
$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x_n}$
Setting the derivative to zero , we get :
$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x_n}$
$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \sum_{n=1}^{N} y_n \alpha_n$
This gives : $\sum_{n=1}^{N} \alpha_n y_n = 0$
Substituting the results we get dual problem as:
$\mathcal{L}(\boldsymbol{\alpha}) = m\boldsymbol{\alpha}^T \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha}$
Let $\beta_n = m\alpha_n$. Then we get the following:
$\mathcal{L}(\boldsymbol{\beta}) = m^2(\boldsymbol{\beta}^T \mathbf{1} - \frac{1}{2}\boldsymbol{\beta}^T \mathbf{G} \boldsymbol{\beta})$
Let $\hat{\boldsymbol{\alpha}}$ be the optimal solution for the original dual problem that is:
$\mathcal{L}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha}$
then $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}$ which means that for the given formulation the optimal solution is $\hat{\boldsymbol{\alpha}}' = m\hat{\boldsymbol{\alpha}}$
So the solution for the modified formulation is given by:
$\mathbf{w} = \sum_{n=1}^{N} m\hat{\alpha_n} y_n \mathbf{x_n}$
So the hyperplane learned by the SVM is essentially the same.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

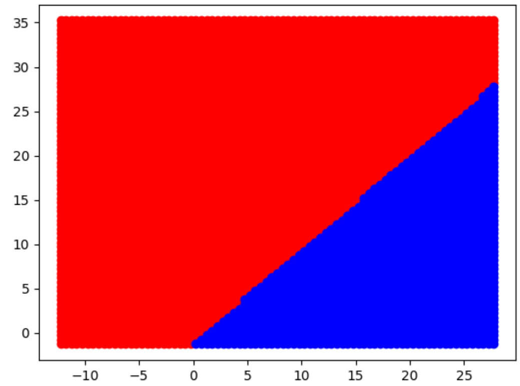
Student Name: Aman Tiwari
Roll Number: 160094
Date: April 18, 2019

QUESTION 7

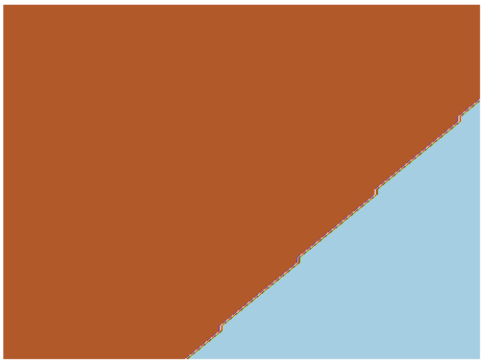
The plots obtained are as follows:
For first dataset:
Using gaussian class conditional with different covariance matrix for positive and negative examples:



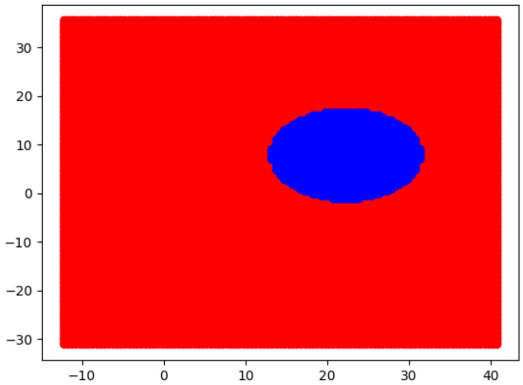
Using gaussian class conditional with same covariance matrix for positive and negative examples:



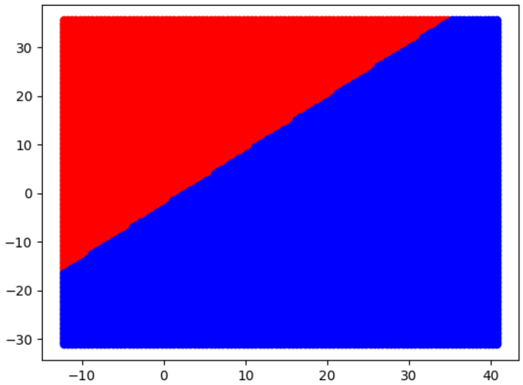
Using SVM



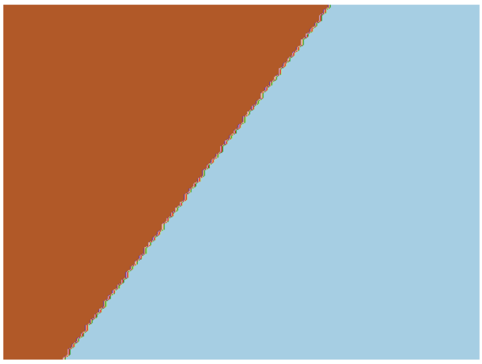
For second dataset:
Using gaussian class conditional with different covariance matrix for positive and negative examples:



Using gaussian class conditional with same covariance matrix for positive and negative examples:



Using SVM



For the first dataset all three models give good results. The results obtained are almost similar in this case. For the second dataset gaussian class conditional with different covariance matrix for positive and negative examples tends to overfit since outliers are present in data whereas the other two models generalize better.