HOMEWORK-2 (PROGRAMMING REPORT)

1) (1)After loading the dataset and scaling(normalizing) it,it is found that the number of features in the dataset is 3072,As it takes a very long time to train such huge data,I have tried for feature reduction using Principal Component Analysis PCA(0.95) for preserving 95% of data.

    The number of features is reduced to 217,So now our training dataset shape is 50,000*217 and our test dataset shape is 10,000*217.

    The parameters I tried for tuning are,

    parameters= [{'kernel':['linear','rbf'],'gamma':[10,1,1e-1,1e-2],'C':[1e-2,1e-1,1,10]}]

    Time taken for GridSearchCV is 284.02276536623634(around 4.73 hours)

    Passing this to GridSearchCV,I obtained best parameters as,

    Best scores: 0.4691

    Best params: {‘C’: 10, ‘gamma’: 0.01, ‘kernel’: ‘rbf’}

    The time taken for modeling a classifier with this best parameters and fitting on train data and test data is,

    Time taken for training is 35.34050410191218

    Now using this optimal parameters as the parameters for model,we obtain train accuracy and test accuracy as,

    The train accuracy is
    0.99334
    The test accuracy is
    0.5674
    Time taken for testing is 14.771410338083903

1)(2)

Now extracting support vectors using the model we got above,and making as a separate dataset,Using previous model training on the support vectors takes time as,

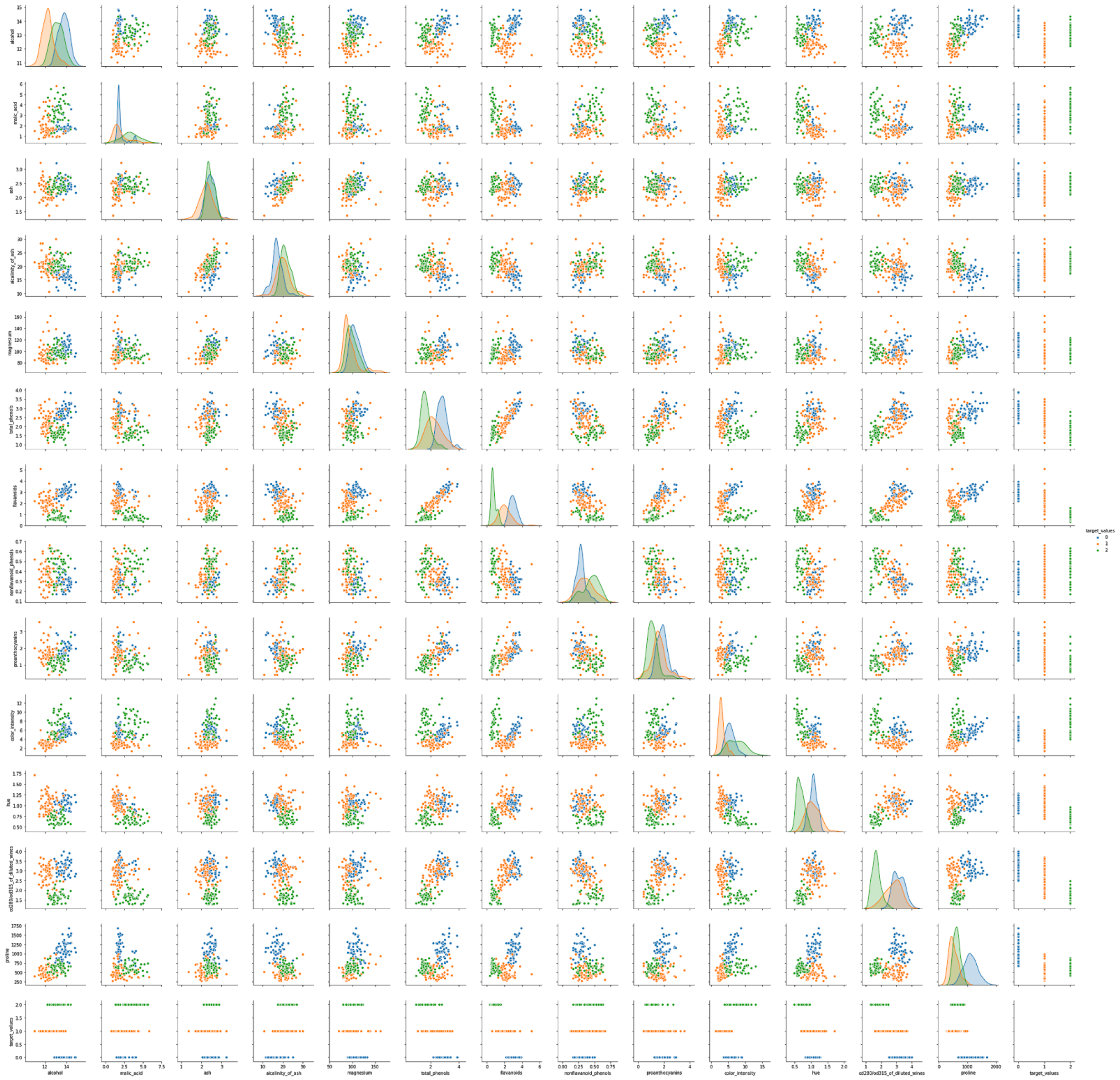
Time taken for training is 26.15493901570638
The train accuracy is
0.99334
The test accuracy is
0.5674
Time taken for testing is 14.46000481446584

1)(3)“Both for the model 1(normal one) and model 2(with support vectors) I have got the same accuracy values.

Because even when I have trained model on the total data,the hyperplane will be obtained only based on support vectors.

So,in both cases we will get the same accuracies as we will be having the same hyperplane in both cases.

2)(1)**PairPlot**



We usually use pairplot for visualizing the data and to get inferences about any two pair of features impact the data.

But here,no plot corresponding to every pair of features infer any relation between those features and the class label.So,we cannot say anything from this pairplot as the data is not separable.

2)(2)

SVM-One vs. Rest

After loading the data and scaling it using MinMaxScaler,I have used get_dummies() to separate the class labels as this is multi-class classification.

For One vs. Rest,we model ‘n’ classifiers where n is the number of different class labels to be predicted.

Here,as we have 3 class labels,we define 3 classifiers and train the data which is obtained as a stratified split from original data.

Model0-(considering class label 0 as ONE and the other labels as REST)

```
Training time of model-OVR-class-0: 1.9917 ms
The weight matrix for SVM-OVR classifier with class label 0
[[ 1.84080424  0.19032062  1.24529786 -1.84389913  0.07306259  0.39845225
   1.13496864 -0.02557918  0.04582676  0.12357082  0.07776224  1.25328193
   2.47006112]]
The intercept for SVM-OVR classifier with class label 0
[-3.50154538]
```

Model1-(considering class label 1 as ONE and the other labels as REST)

```
Training time of model-OVR-class-1: 2.9924 ms
The weight matrix for SVM-OVR classifier with class label 1
[[-2.44797226 -1.14983304 -1.67792623  1.42765172 -0.34422336 -0.31779017
   0.35565918  0.21369178  0.95454931 -2.31058156  1.45181873  0.49427444
  -2.24872224]]
The intercept for SVM-OVR classifier with class label 1
[1.75242266]
```

Model2-(considering class label 2 as ONE and the other labels as REST)

```
Training time of model-OVR-class-2: 1.9944 ms
The weight matrix for SVM-OVR classifier with class label 2
[[ 1.03297445  0.92339586  0.74199575  0.23254561  0.46757037 -0.25497548
  -1.44255656  0.15689778 -1.16059031  1.61816243 -1.38352558 -1.86471992
   0.19651787]]
The intercept for SVM-OVR classifier with class label 2
[-0.4847462]
```

Training Accuracies:-

```
The training accuracies of 3 classes:
Training accuracy for class-0  100.0
Training accuracy for class-1  99.19354838709677
Training accuracy for class-2  98.38709677419355
```

Test Accuracies:-

```
The test accuracies of 3 classes is:
Test accuracy for class-0  100.0
Test accuracy for class-1  96.29629629629629
Test accuracy for class-2  100.0
```

Total training accuracy is:-

```
The total training accuracy for SVM-OVR is:
98.38709677419355
```

Total test accuracy is:-

```
The total test accuracy for SVM-OVR is:
96.29629629629629
```

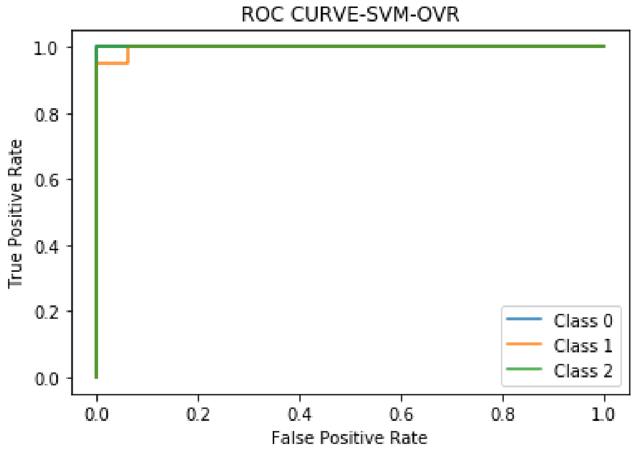
F1 Score

```
The F1 Score for SVM-OVR is:
array([0.95, 0.95, 1.  ])
```

Accuracy Score:-

```
The accuracy score for SVM-OVR is:
0.9629629629629629
```

ROC Curve:-



SVM-One vs. One

Model01-

```
The weight matrix for SVM-OVO classifier of model01
[[ 2.12141497  0.23533428  1.17783289 -1.88478602  0.05157273  0.41740306
   0.50819023  0.00416705 -0.23383709  0.96909606 -0.41221012  0.56594397
   2.39383237]]
The intercept for SVM-OVO classifier of model01
[-2.60346845]
Training time of model-OVO-classifier-01: 4.9875 ms
```

Model02-

```
The weight matrix for SVM-OVO classifier of model02
[[ 0.23387434 -0.15583137  0.1154545  -0.42468004 -0.2610917   1.02253438
   1.36649175 -0.43725651  0.77342409 -0.44749432  0.57233776  1.4399118
   0.69821596]]
The intercept for SVM-OVO classifier of model02
[-1.53907708]
Training time of model-OVO-classifier-02: 4.9846 ms
```

Model12-

```
The weight matrix for SVM-OVO classifier of model12
[[-1.05656517 -0.94741822 -0.71580855 -0.18507442 -0.48023036  0.17801473
   1.37693313 -0.16341584  1.0924949  -1.6363748   1.40167402  1.85231212
  -0.24169621]]
The intercept for SVM-OVO classifier of model12
[0.53071818]
Training time of model-OVO-classifier-12: 4.9806 ms
```

Total Accuracy:-

```
The total accuracy of SVM-OVO on test data
100.0
```

Test accuracies class wise:-

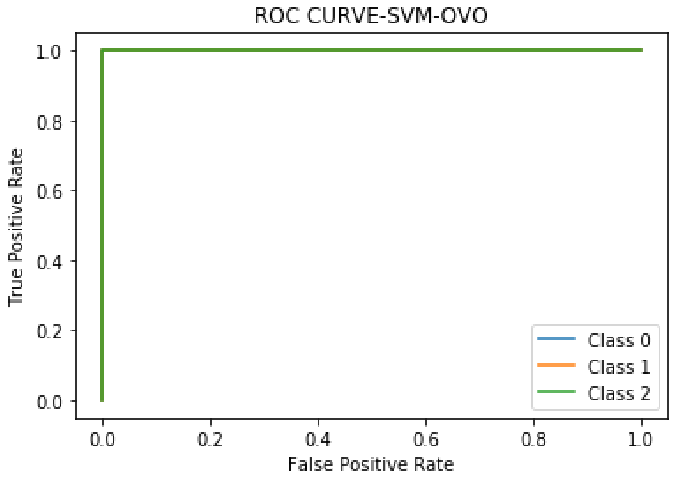
```
The test accuracies of 3 classes is:
For class 0
100.0
For class 1
100.0
```

For class 2
100.0

F1 Score:-
The F1 Score for SVM-OVO is:
array([1., 1., 1.])

Accuracy Score:-
The accuracy score for SVM-OVO is:
1.0

ROC Curve:-



2)(3)

Gaussian Naive Bayes:-

Test Accuracy:-
Total Accuracy for Gaussian Naive Bayes on Test Data is:
100.0

Class wise accuracy:-
Individual Accuracy for class-0 on Test Data is:
100.0
Individual Accuracy for class-1 on Test Data is:
100.0
Individual Accuracy for class-2 on Test Data is:
100.0

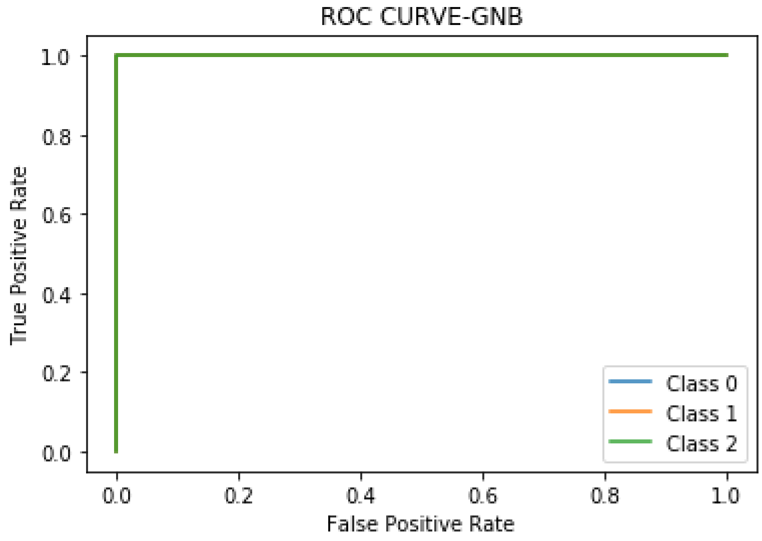
F1 Score:-

The F1 Score for Gaussian Naive Bayes is on Test Data is: array([1., 1., 1.])

Accuracy Score:-

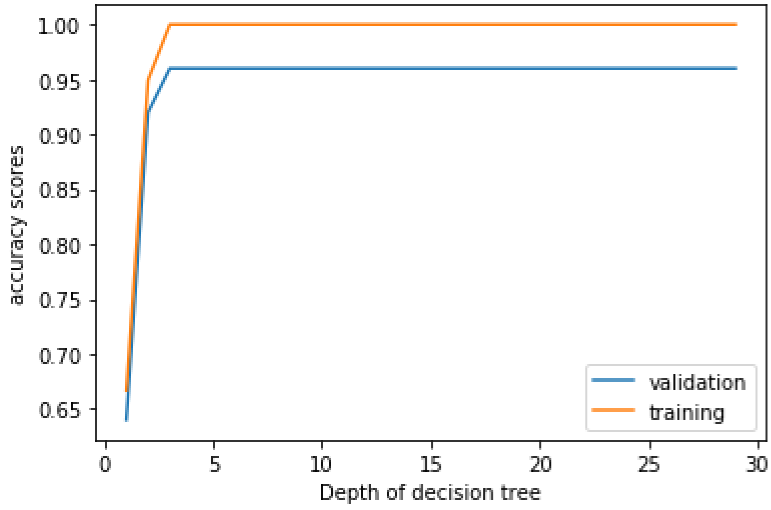
The accuracy score for Guassian Naive Bayes is on Test Data is: 1.0

ROC Curve:-



2)(4)Decision Trees:-

Parameter tuning:-



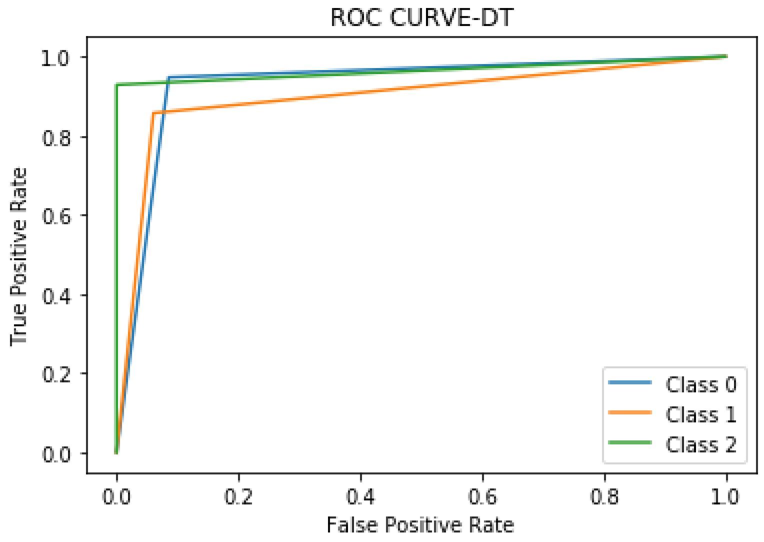
```
Run time of the optimal classifier: 1.9941 ms

Class wise accuracies:-
Individual Accuracy for class-0 on Test Data is:
96.29629629629629
Individual Accuracy for class-1 on Test Data is:
94.44444444444444
Individual Accuracy for class-2 on Test Data is:
98.14814814814815

Total accuracies are:-
The training accuracy score is:  99.19354838709677%
The test accuracy score is:  94.44444444444444%
F1 Score:-
The F1 Score is on Test Data is:
array([0.95      , 0.92682927, 0.96296296])

Accuracy Score:-
The accuracy score on Test Data is
0.9444444444444444

ROC Curve:-
```

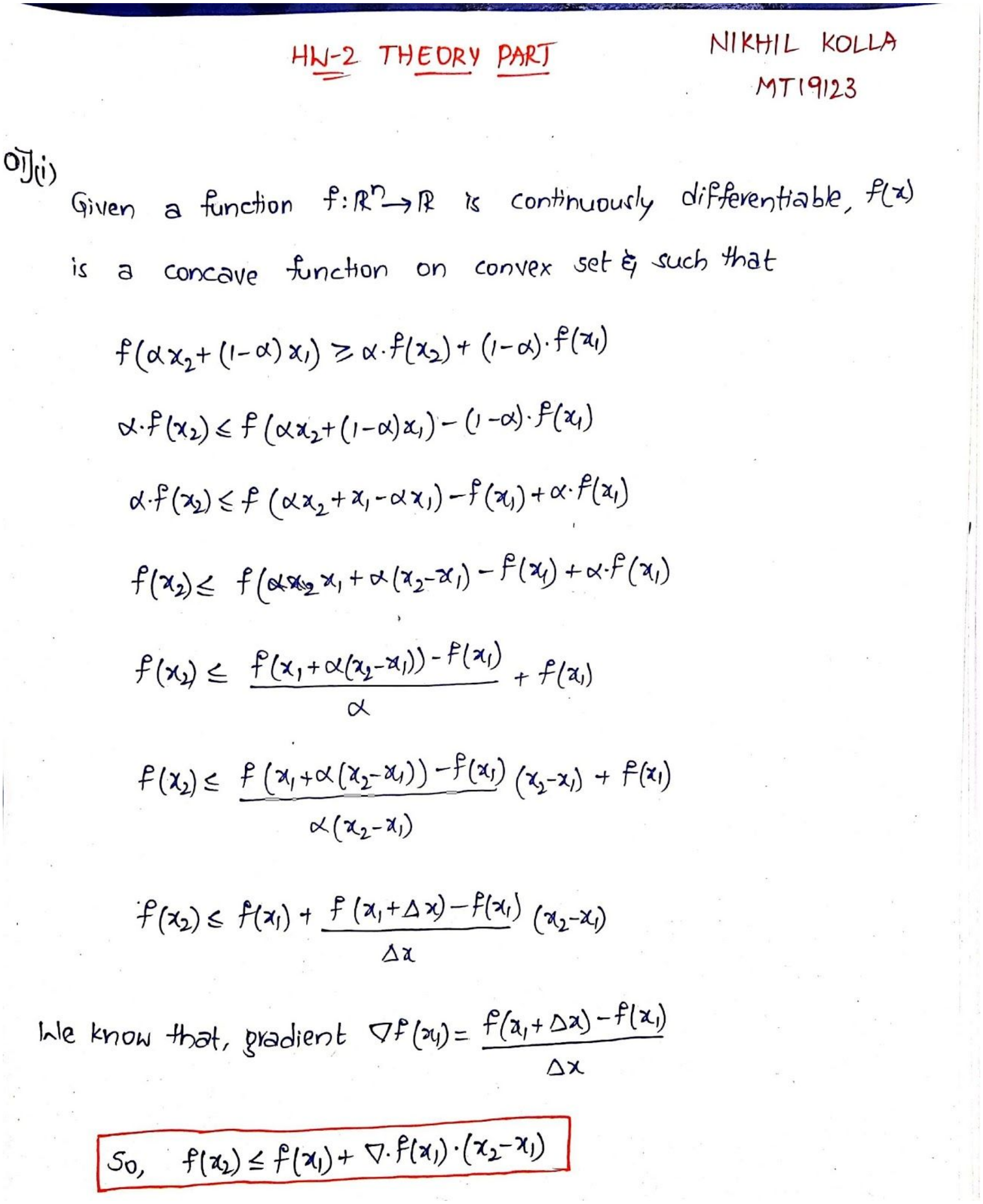


2)(5) Based on the above four models,i.e., SVM-One vs. Rest,SVM-One vs.One,Gaussian Naive Bayes,Decision trees trained on this data and based on the test accuracies obtained,

The best model is Gaussian Naive Bayes and SVM One vs. One, both with 100% accuracy on test data.

HOMEWORK-2 (THEORY PART)

1)(1)



HW-2 THEORY PART

NIKHIL KOLLA
MT19123

Q1)(i)

Given a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable, $f(x)$ is a concave function on convex set & such that

$$f(\alpha x_2 + (1-\alpha) x_1) \geq \alpha \cdot f(x_2) + (1-\alpha) \cdot f(x_1)$$

$$\alpha \cdot f(x_2) \leq f(\alpha x_2 + (1-\alpha) x_1) - (1-\alpha) \cdot f(x_1)$$

$$\alpha \cdot f(x_2) \leq f(\alpha x_2 + x_1 - \alpha x_1) - f(x_1) + \alpha \cdot f(x_1)$$

$$f(x_2) \leq f(\cancel{\alpha x_2}\, x_1 + \alpha (x_2 - x_1) - f(x_1) + \alpha \cdot f(x_1)$$

$$f(x_2) \leq \frac{f(x_1 + \alpha(x_2 - x_1)) - f(x_1)}{\alpha} + f(x_1)$$

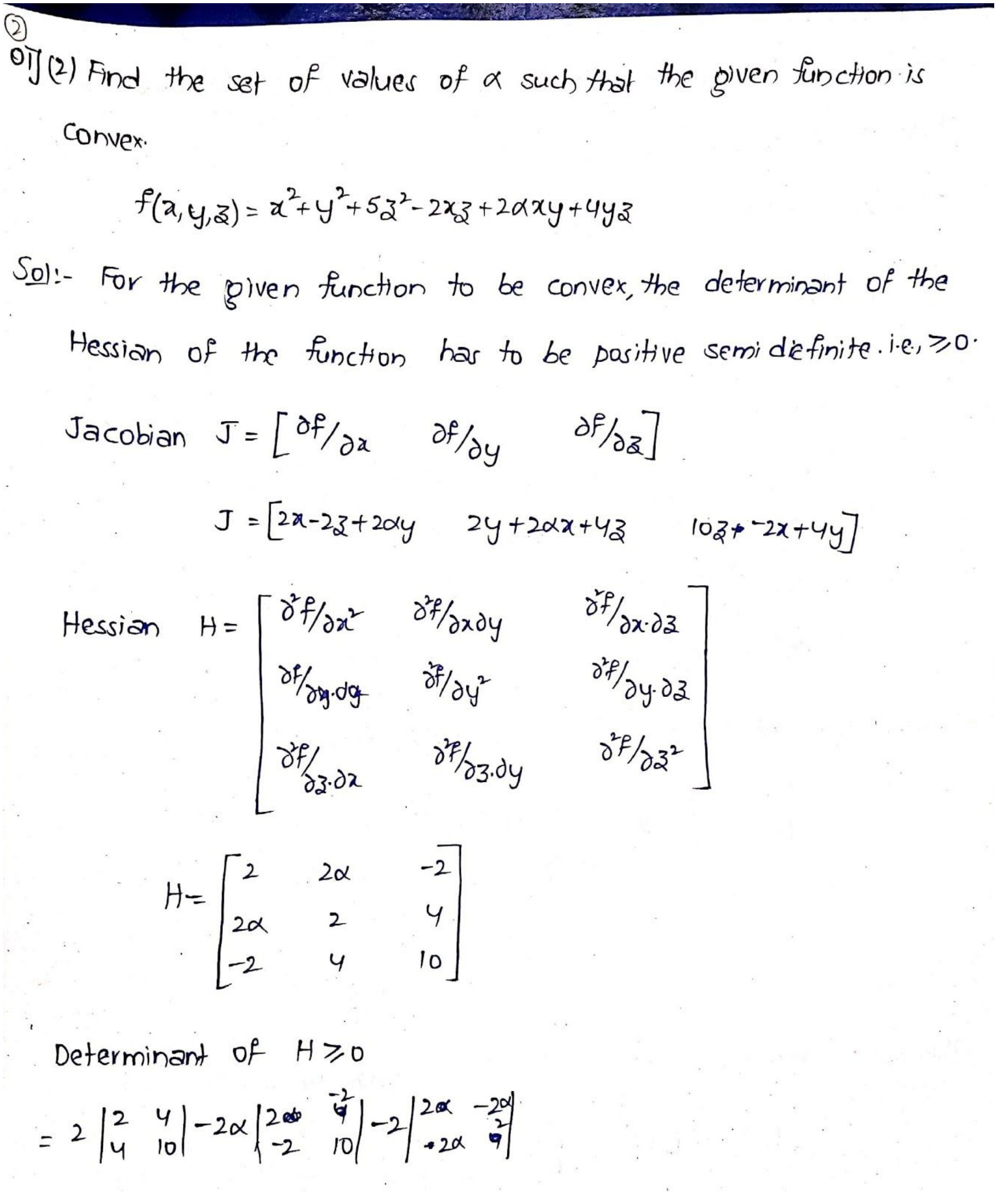
$$f(x_2) \leq \frac{f(x_1 + \alpha(x_2 - x_1)) - f(x_1)}{\alpha(x_2 - x_1)} (x_2 - x_1) + f(x_1)$$

$$f(x_2) \leq f(x_1) + \frac{f(x_1 + \Delta x) - f(x_1)}{\Delta x} (x_2 - x_1)$$

We know that, gradient $\nabla f(x_1) = \dfrac{f(x_1 + \Delta x) - f(x_1)}{\Delta x}$

$$\boxed{\text{So,} \quad f(x_2) \leq f(x_1) + \nabla \cdot f(x_1) \cdot (x_2 - x_1)}$$

1)(ii)



(2)

01](2) Find the set of values of α such that the given function is convex.

$$f(x,y,z) = x^2+y^2+5z^2-2xz+2\alpha xy+4yz$$

Sol:- For the given function to be convex, the determinant of the Hessian of the function has to be positive semi definite. i.e., ≥ 0.

Jacobian $J = \left[\partial f/\partial x \quad \partial f/\partial y \quad \partial f/\partial z\right]$

$$J = \left[2x-2z+2\alpha y \quad 2y+2\alpha x+4z \quad 10z+-2x+4y\right]$$

Hessian $H = \begin{bmatrix} \partial^2 f/\partial x^2 & \partial^2 f/\partial x \partial y & \partial^2 f/\partial x \cdot \partial z \\ \partial^2 f/\partial y \cdot \partial x & \partial^2 f/\partial y^2 & \partial^2 f/\partial y \cdot \partial z \\ \partial^2 f/\partial z \cdot \partial x & \partial^2 f/\partial z \cdot \partial y & \partial^2 f/\partial z^2 \end{bmatrix}$

$$H = \begin{bmatrix} 2 & 2\alpha & -2 \\ 2\alpha & 2 & 4 \\ -2 & 4 & 10 \end{bmatrix}$$

Determinant of $H \geq 0$

$$= 2\begin{vmatrix} 2 & 4 \\ 4 & 10 \end{vmatrix} - 2\alpha\begin{vmatrix} 2\alpha & \overset{-2}{4} \\ -2 & 10 \end{vmatrix} - 2\begin{vmatrix} 2\alpha & -2\alpha \\ 2\alpha & \overset{2}{4} \end{vmatrix}$$

⇒ Determinant has to be calculated with the minors of principal diagonal elements.

$$\Rightarrow 2(20-16) - 2\alpha(20\alpha+8) - 2(8\alpha+4) \geq 0$$

$$\Rightarrow 8 - 40\alpha^2 - 16\alpha - 16\alpha + 8 \geq 0$$

$$\Rightarrow -40\alpha^2 - 32\alpha \geq 0$$

$$\Rightarrow 40\alpha^2 + 32\alpha \leq 0$$

$$\Rightarrow 5\alpha^2 + 4\alpha \leq 0$$

Case 1:- $\alpha \leq 0$; $5\alpha+4 \geq 0$

$\alpha \leq 0$ and $\alpha \geq -4/5$

$\alpha \in \cancel{\phi}\left[-4/5, 0\right]$ — ①

Case 2:- $\alpha \geq 0$; $5\alpha+4 \leq 0$

$\alpha \geq 0$ and $\alpha \leq -4/5$

$\alpha \in \phi$

⇒ Determinant of principal minor has to be greater than or equal to 0.

$$\begin{vmatrix} 2 & 2\alpha \\ 2\alpha & 2 \end{vmatrix} \geq 0 \Rightarrow 4 - 4\alpha^2 \geq 0$$

$$1 - \alpha^2 \geq 0$$

$$\alpha^2 - 1 \leq 0 \Rightarrow \alpha^2 \leq 1$$

$$\alpha \in [-1, 1] \text{ — ②}$$

∴ The set of values of 'α' will be the intersection of ① and ②.

$$[-4/5, 0] \cap [-1, 1] \Rightarrow \boxed{\alpha \in [-4/5, 0]}$$

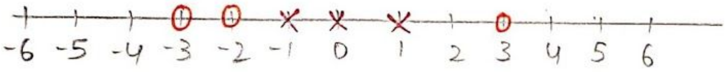
2)

Q2] Suppose you are given a dataset of six one-dimensional data points. Three of the six data points have negative labels and other three have positive labels.

- Negative labels:- $x_1 = -1$, $x_2 = 0$, $x_3 = 1$
- Positive labels:- $x_4 = -3$, $x_5 = -2$, $x_6 = 3$

Plot the dataset. You may find that the data is not linearly seperable. However, if we apply the feature map $\phi(u) = (u, u^2)$, the points in $\mathbb{R}^1$ will be transformed to new points in $\mathbb{R}^2$. Visualize the transformed data points in $\mathbb{R}^2$, now a linear seperator can seperate the points in $\mathbb{R}^2$.

Given points in $\mathbb{R}^1$.



X - Negative labels.

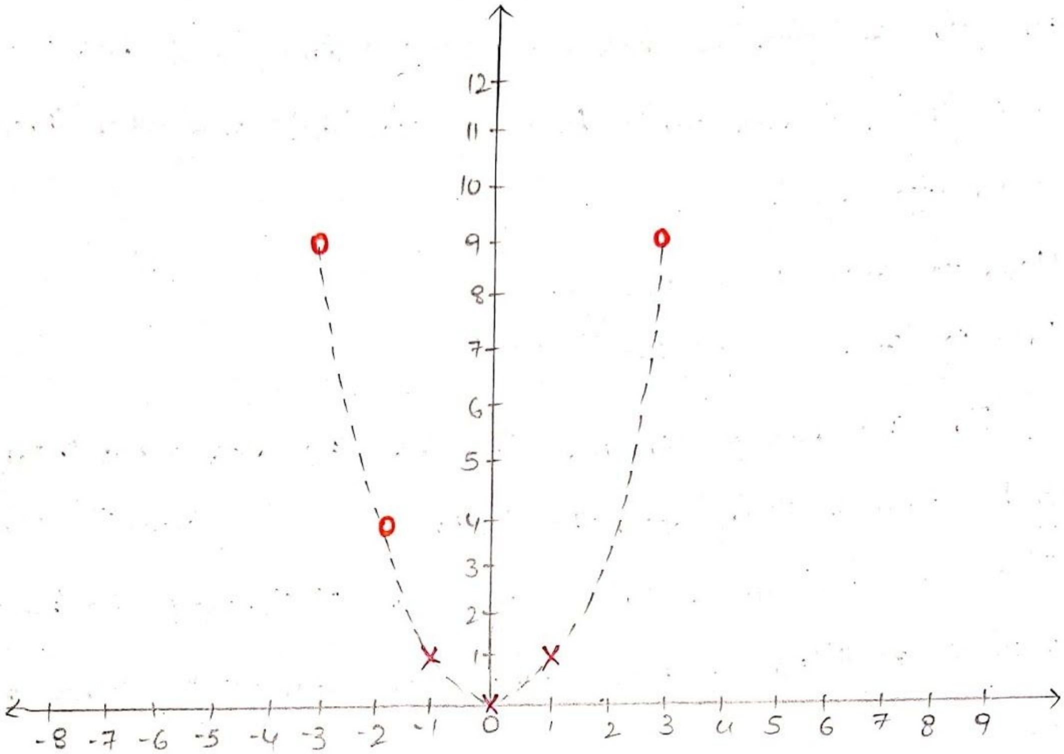
O - Positive labels.

From the above diagram, we can find that the data is not linearly seperable. So, by applying feature map, $\phi(u) = (u, u^2)$. The new data points in $\mathbb{R}^2$ are:-

Negative labels:- $x_1 = (-1, 1)$; $x_2 = (0, 0)$; $x_3 = (1, 1)$

Positive labels :- $x_4 = (-3, 9)$; $x_5 = (-2, 4)$; $x_6 = (3, 9)$

⑤



o- Positive labels ; x- Negative labels.

After transformation, it is clear that the data is linearly seperable.

(i) Give the analytic form of the kernel that corresponds to the feature map ϕ in terms of only x_1 and x_1'. Specifically define $k(x_1, x_1')$.

$$K(x_1, x_1') = \phi(x_1, x_1')$$

$$= (x_1, x_1^2) \cdot (x_1', x_1'^2)$$ as per the defⁿ of $\phi(u)$

$$= x_1 x_1' + x_1^2 x_1'^2$$

$$\boxed{K(x_1, x_1') = x_1 x_1'(1 + x_1 x_1')}$$

⑥

(ii) Based on our geometric intuition, we can clearly say that the margin maximising hyperplane (say π) has to pass through somewhere between $x_5\,(-2,4)$ and $x_1\,(-1,1)$.

But, for this π has to be the margin maximising one, it has to pass through the midpoint of $x_5\,(-2,4)$ and $x_1\,(-1,1)$.

Midpoint of x_5 and x_1 $= \left(\frac{-2-1}{2}, \frac{4+1}{2}\right)$

$$= \left(-3/2, 5/2\right)$$

So, π passes through $\boxed{\left(-3/2, 5/2\right).}$

We also have a constraint that π has to be perpendicular to the line passing through x_5 and x_1.

Slope of $\pi = -1/m$; where 'm' is the slope of line joining $x_5\,(-2,4)$ and $x_1\,(-1,1)$.

$m = \frac{4-1}{-2+1} = -3$; Slope of our hyper plane $\boxed{\pi_m = 1/3.}$

⇒ Eqn. of hyperplane passing through $(-3/2, 5/2)$ and with slope $= 1/3$.

$$(Y_2 - y_1) = m\,(Y_1 - \overset{x_1}{\cancel{y_2}})$$

$$(Y_2 - 2.5) = 1/3\,(Y_1 + 1.5) \Rightarrow 3Y_2 - 7.5 = Y_1 + 1.5$$

$$\boxed{Y_1 - 3Y_2 + 9 = 0}$$

So, the obtained values are $w_1 = 1$; $w_2 = -3$; $c = 9$

(7)

Margin distance:- Distance b/w $(-3/2, 5/2)$ to $(-2,4)$ (or) distance b/w $(-3/2, 5/2)$ to $(-1,1)$.

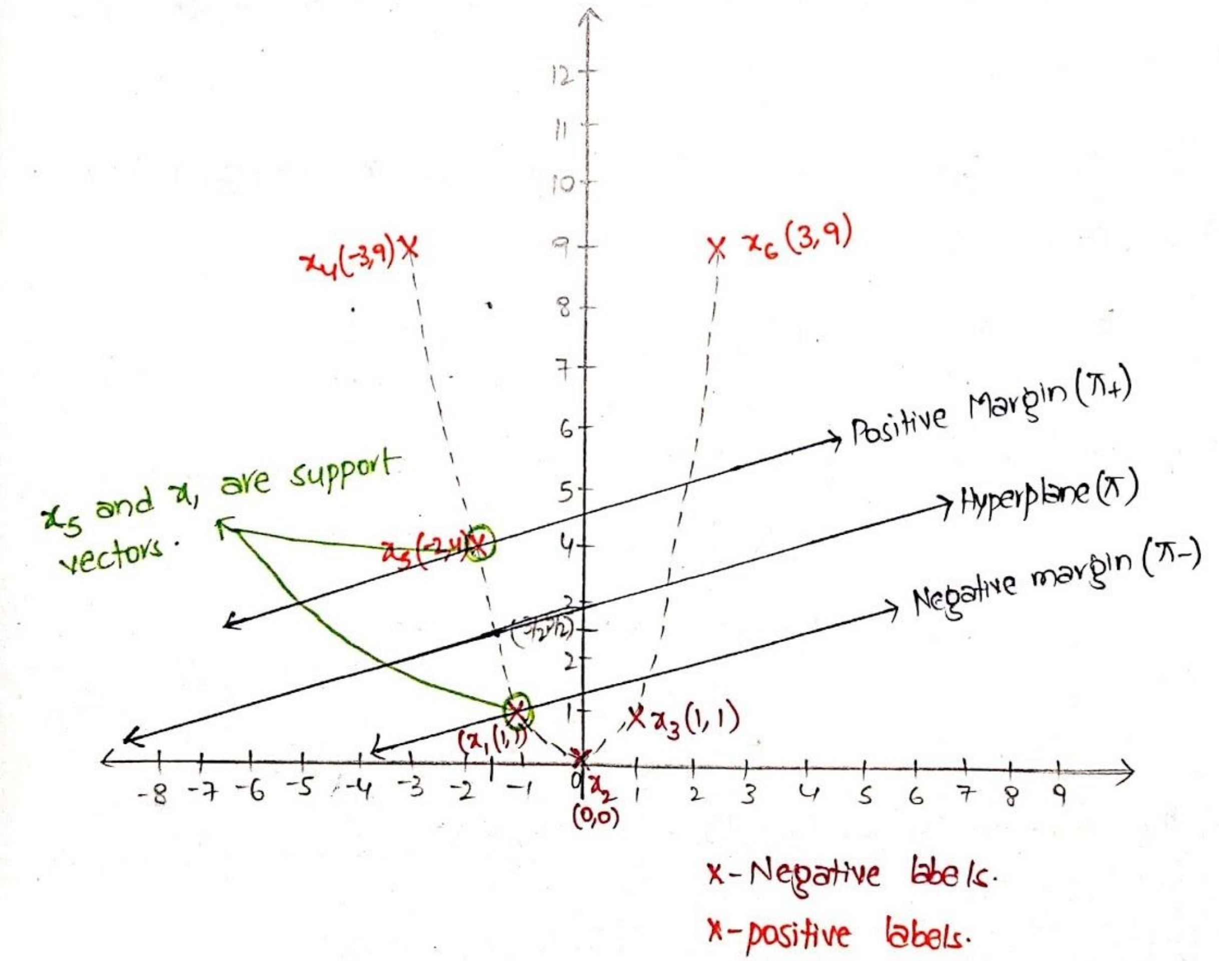
$$\text{Margin distance} = \sqrt{(y_2-y_1)^2+(x_2-x_1)^2}$$

$$= \sqrt{(4-2.5)^2+(-2+1.5)^2}$$

$$= \sqrt{2.25+0.25} = \sqrt{2.5}$$

$$\boxed{\therefore \text{Margin distance} = \frac{\sqrt{10}}{2}.}$$

03]



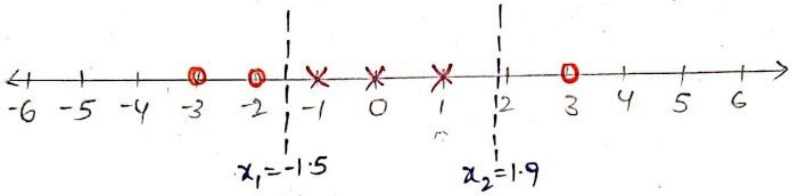
04] The points of intersection b/w the optimal hyperplane is the parabola when projected on to $\mathbb{R}^1$. feature space.

Points of intersection are obtained by solving the equation of hyperplane, $x-3y+9=0$ and eqn of parabola, $y=x^2$.

$$\Rightarrow x-3y+9=0$$

$$x-3x^2+9=0 \quad (\because y=x^2)$$

$$\boxed{-3x^2+x+9=0} \Rightarrow \boxed{x_1, x_2 = 1.9, -1.5}$$



o-positive points ; x-negative points.

05] The support vectors u_1, u_2 are $(-2,4)$ and $(-1,1)$.

Legrangian dual form:-

$$\sum_{I=1}^{|SV|} \alpha_i - \frac{1}{2} \sum_{i=1}^{|SV|} \sum_{j=1}^{|SV|} y^i y^j \alpha_i \alpha_j < u^i, u^j >$$

α- legrangian multiplier; y_i - class label of ith point.

u_n- support vectors.

$$= \alpha_1 + \alpha_2 - \frac{1}{2} [y_1 y_2 \alpha_1 \alpha_2 < u_1, u_1 > + y_1 y_2 \alpha_1 \alpha_2 < u_1, u_2 >$$
$$+ y_2 y_1 \alpha_2 \alpha_1 < u_2, u_1 > + y_2 y_2 \alpha_2 \alpha_2 < u_2, u_2 >]$$

$$= \alpha_1 + \alpha_2 - \frac{1}{2} [y_1^2 \alpha_1^2 < u_1, u_1 > + y_1 y_2 \alpha_1 \alpha_2 < u_1, u_2 > + y_2 y_1 \alpha_2 \alpha_1 < u_2, u_1 >$$
$$+ y_2^2 \alpha_2^2 < u_2, u_2 >]$$

$$\boxed{y_1 y_2 = y_2 y_1 = y_1^2 = y_2^2 = 1}$$

⑨

$<u_1 \cdot u_1> = (u_1 \cdot u_1) = (-2,4) \cdot (-2,4)$

$= 4 + 16 = 20$

$<u_2, u_2> = (u_2 \cdot u_2) = (-1,1) \cdot (-1,1)$

$= 1 + 1 = 2$

$<u_1 \cdot u_2> = (u_1 \cdot u_2) = (-2,4) \cdot (-1,1)$

$= 2 + 4 = 6$

$\Rightarrow \alpha_1 + \alpha_2 - \frac{1}{2}\left[\alpha_1^2(20) - \alpha_1\alpha_2(6) - \alpha_1\alpha_2(6) + \alpha_2^2(2)\right]$

$= \alpha_1 + \alpha_2 - 10\alpha_1^2 + 3\alpha_1\alpha_2 + 3\alpha_1\alpha_2 \mp \alpha_2^2$

$= \alpha_1 + \alpha_2 - 10\alpha_1^2 + 6\alpha_1\alpha_2 \mp \alpha_2^2$ — Ⓐ

$\Rightarrow$ We know that by partial differentiating lagrangian w.r.t b,

we get, $\boxed{\sum \alpha_i y_i = 0}$

$\alpha_1 y_1 + \alpha_2 y_2 = 0$ where $(y_1, y_2) \in (1,-1)$

$\alpha_1 - \alpha_2 = 0$

$\boxed{\alpha_1 = \alpha_2}$

Ⓐ $\Rightarrow \alpha_1 + \alpha_2 - 10\alpha_1^2 + 6\alpha_1\alpha_2 \mp \alpha_2^2$

$= 2\alpha_1 - 10\alpha_1^2 + 6\alpha_1^2 \mp \alpha_1^2$

$= 2\alpha_1 - \cancel{1}5\alpha_1^2 \Rightarrow \boxed{2\alpha - 5\alpha^2 = 0}$

For optimal ~~ws~~ α,

$\frac{\partial}{\partial\alpha}(2\alpha - 5\alpha^2) = 0 \Rightarrow 2 - 10\alpha = 0$

$\boxed{\alpha = 1/5}$

To compute the value of 'b', we consider support vector on the -ve margin.

$$-1 = \alpha_1 y_1 \cdot k(x_1, u_1) + \alpha_2 \cdot y_2 \cdot k(x_2, u_2) + b$$

$$-1 = \tfrac{1}{5}(1)(-1,1)\cdot(-2,4) + \tfrac{1}{5}(-1)(-1,1)\cdot(-1,1) + b$$

Here, u_1 and u_2 are support vectors.

$$-1 = \frac{1}{5}(6) + \left(\frac{-1}{5}\right)(2) + b$$

$$-1 = \frac{6}{5} - \frac{2}{5} + b \Rightarrow b = -1 - \frac{4}{5} \Rightarrow \boxed{b = \frac{-9}{5}}$$

(vi) No. The hyperplane do not change. by adding new positive label at $x = 5$. Because, this point will be correctly classified with the present hyperplane we considered and does not effect the hyperplane in changing the position.