Assignment 2

Machine Learning

- Mayank Chauhan, MT18008

1. GridSearchCV and Support Vectors

1.1. CIFAR 10

Dataset

The training set of CIFAR 10 dataset consist of 50,000 images of size 32x32 pixels. Whereas, the test set contains 10,000 images. Total no. of classes is 10.

Preprocessing

First, the pixels are rescaled into -1 to 1 using MinMaxScaler. Then, training time of the SVM model is reduced by reducing the dimension of the dataset.

Principal Component Analysis (PCA) is used on each channel (RGB) of the image. From each channel top 50 components were choosen. The dimension was reduced to 150 from 3072 (32x32).

GridSearchCV (cv=3)

The total time taken by GridsearchCV is 135 minutes. Then using the best parameters the model is refit on the whole training data in 13.4595 minutes.

Best Parameters

Model : SVC with C=0.01 and linear kernel.

Table 1: Accuracy and Runtime

	Accuracy	Runtime (in min)
Training set	0.4397	9.45
Test set	0.4315	1.90

The new training set consist of 45294 support vectors. It took 9.94 min to train on this new training set. The new model has 45293 support vectors. The results found with the new model are as follows.

The new model uses the parameters C=0.01 and linear kernel.

Table 2: Accuracy and Runtime using new model

New Model: SVC C=0.01 and linear kernel.	Accuracy	Runtime (in seconds)
Training set	0.4398	570.01s
Test set	0.4313	113.96s

Observations:

- The lower value of C indicates the model is trying to make the margin larger by trading off the training accuracy. The no. of support vectors is very high because the re are many training points which are miscalssified. For all those points, the alpha is non zero.
- The accuracy and run time is same in both cases, becasue the support vectors of the both models are almost same.

2. Wine dataset

Dataset

The wine dataset consist of 3 classes with 13 different features. The classes are labeled as 0, 1 and 2.

2.1. Pair wise relation

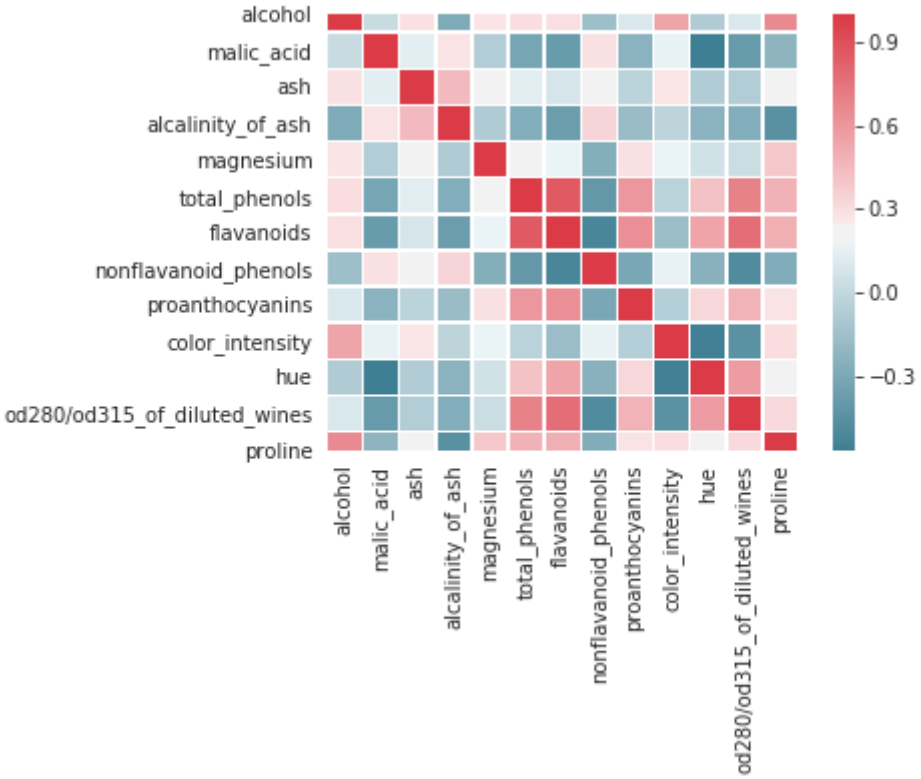


Figure 1: Correlation heatmap between the features.

Observations*:*

- Hue and malic_acid is highly negatively correlated.
- Flavanoids and total_phenols are highly positively correlated.

Pair wise plots

2.2. Evaluations of One-vs-One, One-vs-rest, Gaussian NB and Decision Tree

Dataset

The no. of training examples of individual classes are as follows.

Class 0 – 41

Class 1 – 50

Class 2 – 33

Training set size: (124, 13)

Test set size: (54, 13)

Preprocessing

Rescaling the features values into -1 to 1 using MinMaxScaler.

1-vs-1

The model used for implementing 1-vs-1 and 1-vs-rest is SVC with linear kernel. The regularization parameter was found for every binary classifier.

The weights and intercept can be accessed after the model is trained. Then, the predictions are made by calculating the distance of the test point from the hyperplane. In one ve one, we predict the class which is majority class by simply voting. The no. of folds is fixed and is equal to 3.

Table 3: Value of C for all one-vs-one SVC classifiers found using GridSeachCV

Models	Zero vs One	One vs Two	Two vs Zero
C	1	0.1	0.1

Table 4: One-vs-One Binary Classifiers

	Training accuracy	F1 (macro avg)
0 vs 1	1.0	1.0
1 vs 2	0.98	0.98
2 vs 0	1.0	1.0

1-vs-rest

The class of the test point is predicted by calculating all the three distances and predicting class of the binary classifer which has minimum distance.

Table 5: Value of C for all one-vs-rest SVC classifiers found using GridSeachCV

Models	Zero vs Rest	One vs Rest	Two vs Rest
C	2	1	8

Table 6: One-vs-Rest Binary Classifier

	Training accuracy	F1 (macro avg)
0 vs Rest	1.0	1.0
1 vs Rest	0.98	0.98
2 vs Rest	1.0	1.0

Evaluation of model

Table 7: Performance comparison on various metrics

Model	Training Accuracy	Test Accuracy	F1-Score (macro avg)	Training time
1-vs-1 SVC	0.983870	0.981481	0.9811	0.004448
1-vs-rest SVC	0.983870	0.981481	0.9828	0.006059
Gaussian Naive Bayes	0.967741	1.00	1.0	0.002684
Decision Tree with best parameters	0.983870	0.981481	0.9827	0.003486

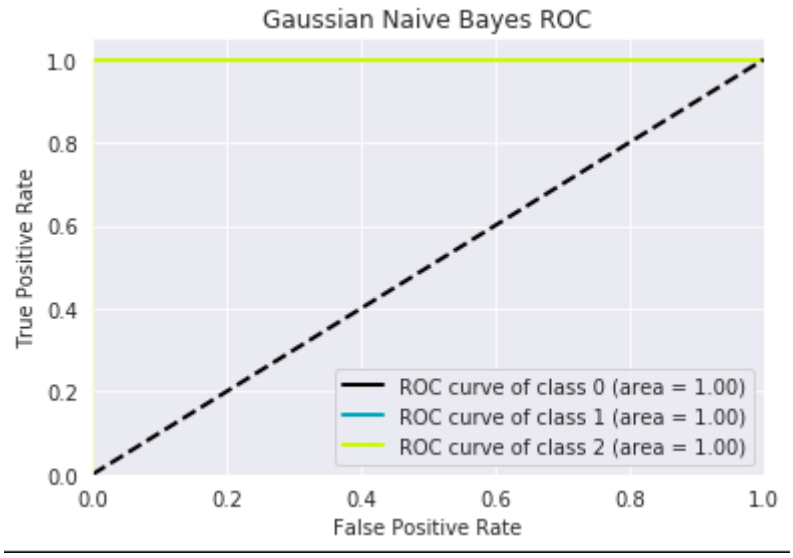
Observations:

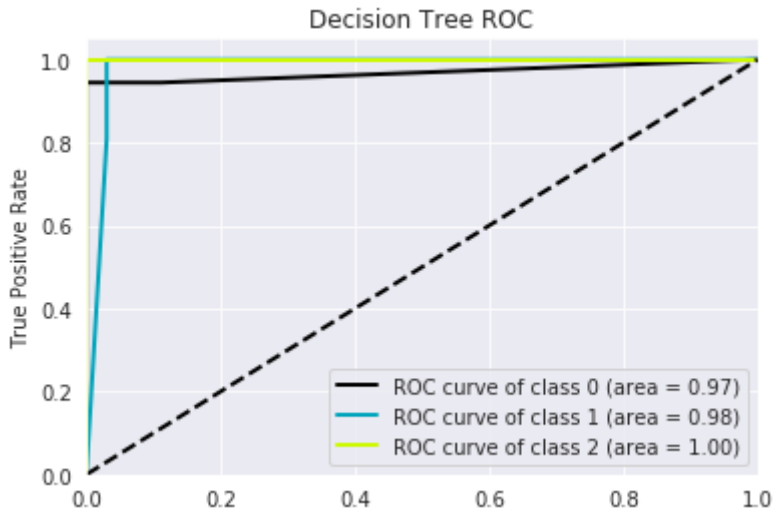
- The Decision Tree Classifier parameters like depth, min_samples_split and min_samples_leaf were tuned. The optimal values are -

Parameters	**Value**	**Default**
depth	3	-1
min_samples_split	2	2
min_samples_leaf	2	1

- The f1 score computed using the micro averaging, which means we are not favouing any particular class.
- As the training the data is not so much imbalanced. The 1-vs-1 and 1-vs-rest approach yield same results. If there were a minority class, then 1-vs-rest classifer will not able to predict well for that class.
- Computationally, the all the four classifers doesn't take too much time to train. So, we just have just to select the best model based on the model performance.
- Gaussian Naive bayes has higher more accuracy on the test set, which should not happen. It means the test is very similar to the training set.
- The best model is decision tree classifier, because the difference between training and test accuracy is not so much. And both training and test accuracies are sufficiently high. This model will result in faster predictions, as the depth of the tree is only three.

ROC





Theory Solutions

Name: MAYANK CHAUHAN

MT18008

ML: Theory A-2

Q(1) Convexity

(1) Given, $f: R^n \to R$

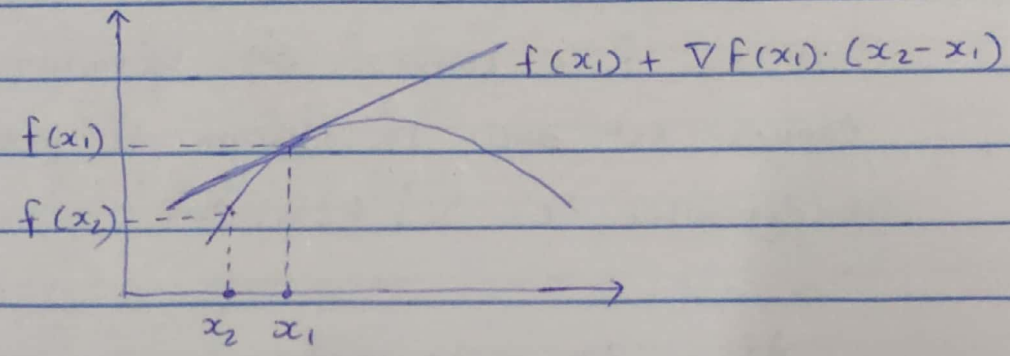
$f(x)$ is concave on a convex set S.

To prove: $\forall x_1, x_2 \in S$, $f(x)$ satisfies

$$f(x_2) \leq f(x_1) + \nabla f(x_1) \cdot (x_2 - x_1).$$

where

$$\nabla f(x_1) = \left[\frac{\partial f(x)}{\partial x_1}, \dots \frac{\partial f(x)}{\partial x_n} \right]$$



As f is concave, for any $\lambda \in [0,1]$:

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

$$f(x_1 + \lambda(x_2 - x_1)) \geq \lambda(f(x_2) - f(x_1)) + f(x_1).$$

$$f(x_1) \geq f(x_2) - \frac{(f(x_1 + \lambda(x_2 - x_1)) - f(x_1))}{\lambda}$$

$$f(x_1) \geq f(x_2) - \frac{(f(x_1 + \lambda(x_2 - x_1)) - f(x_1)) \cdot (x_2 - x_1)}{\lambda(x_2 - x_1)}.$$

put $\lambda(x_2 - x_1) \equiv \Delta x$

$$f(x_1) \geq f(x_2) - \underbrace{\frac{(f(x_1 + \Delta x) - f(x_1))}{\Delta x}} \cdot (x_2 - x_1).$$

$$\downarrow$$

$$\nabla f(x_1)$$

Therefore, $f(x_1) \geq f(x_2) - \nabla f(x_1) \cdot (x_2 - x_1)$.

$f(x_1) + \nabla f(x_1) \cdot (x_2 - x_1) \geq f(x_2)$.

$f(x_2) \leq f(x_1) + \nabla \cdot f(x_1) \cdot (x_2 - x_1)$. Au

Hence Proved.

(2)

$f(x,y,z) = x^2 + y^2 + 5z^2 - 2xz + 2axy + 4yz$.

As f is convex so its must satisfy second order condition for convexity.

That is, f is convex $\Leftrightarrow$ Domain (f) is a convex set and its Hessian is positive semidefinite. i.e $\nabla_x^2 f(x) \succeq 0$.

$$\frac{\partial f}{\partial x} = 2x - 2z + 2ay.$$

$$\frac{\partial f}{\partial y} = 2y + 2ax + 4z$$

$$\frac{\partial f}{\partial z} = 10z - 2x + 4y.$$

$$\frac{\partial^2 f}{\partial x^2} = 2 \qquad \frac{\partial^2 f}{\partial y^2} = 2 \qquad \frac{\partial^2 f}{\partial z^2} = 10.$$

$$\frac{\partial f}{\partial x \partial y} = 2a, \quad \frac{\partial f}{\partial y \partial z} = 4, \quad \frac{\partial f}{\partial x \partial z} = -2.$$

$$\nabla_x^2 f(x) = \begin{bmatrix} 2 & 2a & -2 \\ 2a & 2 & 4 \\ -2 & 4 & 10 \end{bmatrix} = A.$$

As f is convex (given) ⇒ A is positive semi definite ⇒ Eigen values are non-negative.
Product all the eigen values will be same as |A|.

$$|A| \geq 0$$

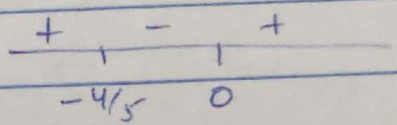
$$\Rightarrow 2[20 - 16] - 2\alpha[20\alpha + 8] - 2[8\alpha + 4] \geq 0$$

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

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

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

$$\alpha(5\alpha + 4) \leq 0$$



$$\boxed{\frac{-4}{5} \leq \alpha \leq 0}$$ Ans

Q(2) SVM

(1). Given $\phi(u) = (u, u^2)$

Take x_i and $x_i' \in R^1$.

$$K(x_i, x_i') = \phi(x_i)^T \cdot \phi(x_i')$$

$$= [x_i \; x_i^2] \begin{bmatrix} x_i' \\ (x_i')^2 \end{bmatrix}$$

$$= [x_i \cdot x_i' + x_i^2 \cdot (x_i')^2]$$

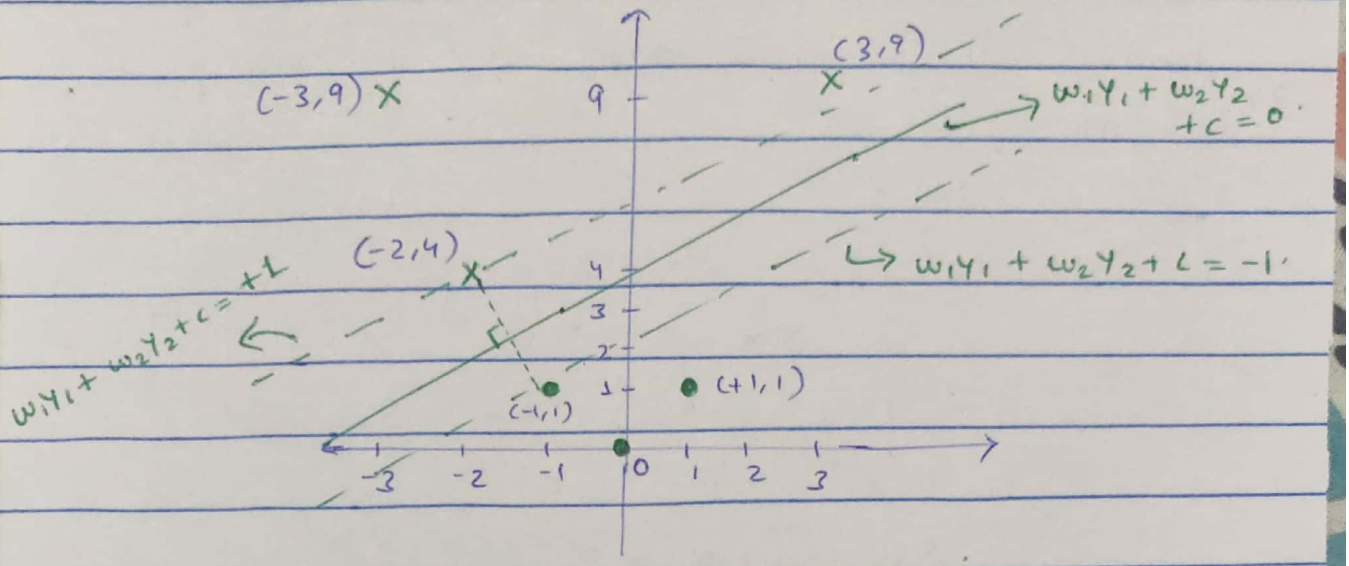
$$= (x_i \cdot x_i')[1 + x_i \cdot x_i']$$. Ans

(2)

Positive labels (+1): $x_4 = -3$, $x_5 = -2$, $x_6 = 3$.

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

Apply $\phi(4)$ on the data points —



Normal eqn line: $w_1y_1 + w_2y_2 + c = 0$

$$y_2 = -\frac{w_1}{w_2} y_1 + \left(\frac{-c}{w_2}\right)$$

slope of the normal eqn $= \frac{-1}{\text{slope of } (-2,4) \& (-1,1)}$

$$\Rightarrow -\frac{w_1}{w_2} = (-1) \times \frac{(-2+1)}{(4-1)}$$

$$\Rightarrow -w_1 = \frac{w_2}{3}. \qquad \boxed{w_2 = -3w_1} \quad \text{—(1)}$$

Also, using the two support vectors —

$$w_1(-2) + w_2(4) + c = 1 \quad \text{—(2)}$$

$$w_1(-1) + w_2(1) + c = -1 \quad \text{—(3)}$$

Subtracting (3) from (2)

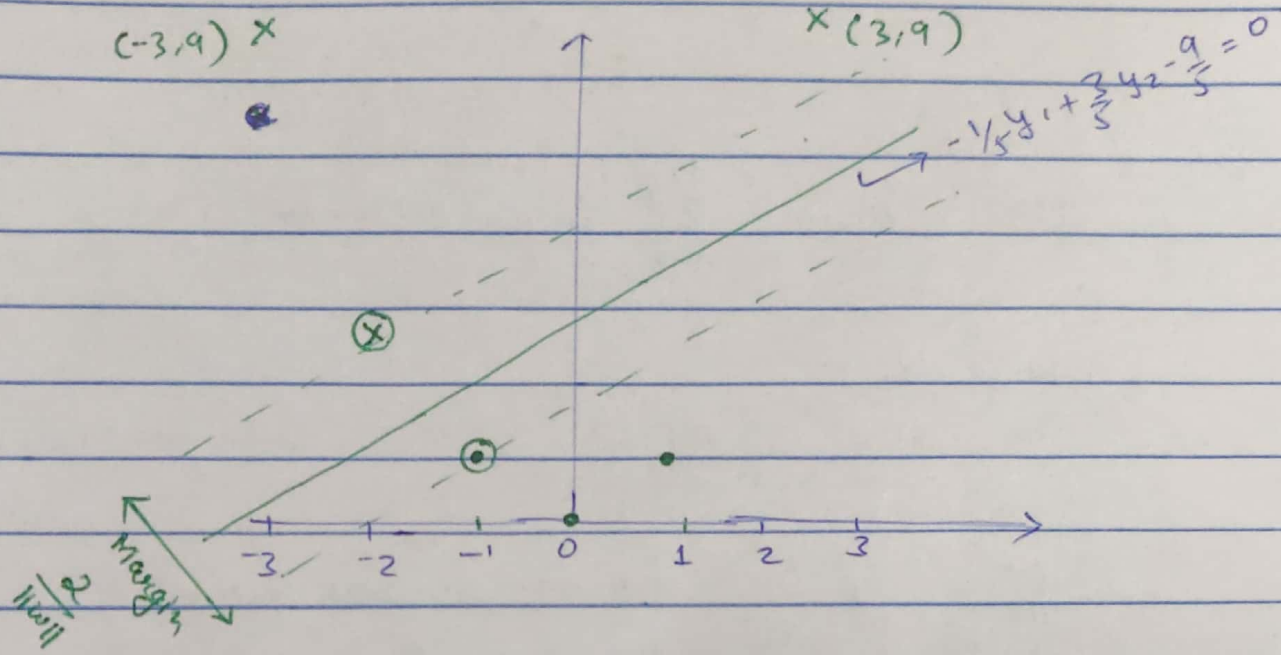
$$-w_1 + 3w_2 = 2 \qquad \text{using eq (1)}$$

$$-w_1 - 9w_1 = 2 \Rightarrow \boxed{w_1 = -\frac{1}{5}}$$

$w_2 = -3w_1 = \frac{3}{5}$ $\quad c = \frac{-9}{5}$

Normal eqⁿ line : $\frac{-1}{5}y_1 + \frac{3}{5}y_2 - \frac{9}{5} = 0$. Ans

(3)



Margin width, $\frac{2}{\|w\|} = \frac{2}{\sqrt{\frac{1}{25} + \frac{9}{25}}} = \frac{2.5}{\sqrt{10}}$

$= \boxed{\sqrt{10}}$. Ans (for previous part (2)).

(4)

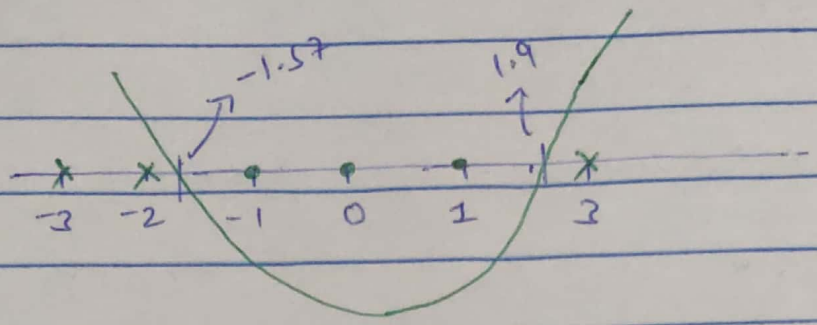
In original space,

$-\frac{y_1}{5} + \frac{3}{5}y_2 - \frac{9}{5} = 0$ $\quad$ let $x \in R^1$.

$-x + 3x^2 - 9 = 0 \Rightarrow x =$

$\Rightarrow 3x^2 - x - 9 = 0 \Rightarrow x = \frac{1 \pm \sqrt{109}}{6}$

$x = -1.57$ and $x = 1.90$.



(5)

$$y(x) = \text{sign}\left(\sum_{n=1}^{|sv|} \alpha_n y_n \cdot K(x, u_n) + b \right)$$

Computing α_i —

$$w = \sum_{i=1}^{n} \alpha_i \cdot y^{(i)} \cdot x^{(i)} \qquad i = 1 \ldots n \text{ training points}$$

$\alpha_i \Rightarrow \begin{cases} \neq 0 & \text{for non-support vectors} \\ = 0 & \text{for support vectors} \end{cases}$

$$\begin{bmatrix} -1/5 \\ 3/5 \end{bmatrix} = \alpha_1 \begin{bmatrix} -2 \\ 4 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$-2\alpha_1 + \alpha_2 = -1/5$$

$$4\alpha_1 - \alpha_2 = 3/5$$

$$\Rightarrow \boxed{\alpha_1 = \frac{1}{5}} \text{ and } \boxed{\alpha_2 = \frac{1}{5}}$$ Ans

$$y(x) = \text{sign}\left(\alpha_1 \cdot y_1 \cdot K(x, u_1) + b + \alpha_2 \cdot y_2 \cdot K(x, u_2) \right)$$

$$u_1 = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \quad u_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\Rightarrow y(x) = \text{sign}\left(\frac{1}{5}\cdot(1)\cdot[x \;\; x^2]\begin{bmatrix}-2\\4\end{bmatrix} + \frac{1}{5}\cdot(-1)\cdot[x \;\; x^2]\cdot\begin{bmatrix}-1\\1\end{bmatrix} + b \right)$$

$$\Rightarrow y(x) = \text{sign}\left(\frac{1}{5}[-2x + 4x^2 + x - x^2] + b \right).$$

for Support vector $x = [-2, 4]$

$$y\left(\begin{bmatrix}-2\\4\end{bmatrix}\right) = 1 = \frac{1}{5}[-2x + 4x^2 + x - x^2] + b$$

$y(-2) = 1 \qquad x \in R^1.$

$$1 = \frac{1}{5}[-2 + 4 + 1 - 1] + b$$

$b =$

$\Rightarrow$ Put $x = -2$.

$$y(-2) = 1 = \frac{1}{5}[+4 + 16 - 2 - 4] + b$$

$$1 = \frac{1}{5}[14] + b \Rightarrow \boxed{b = -9/5}$$

(6) Adding positive label at $x = 5$.

$\phi(x=5) = [5 \;\; 25]$.

This point lies very far away from the hyperplane.

$$\text{Distance} = \frac{-1}{5}(5) + \frac{3}{5}(25) - \frac{9}{5} = -1 + 15 - \frac{9}{5} = \boxed{\frac{61}{5}}.$$

The point doesn't affect the original support vectors, therefore after adding this point → support vectors will remain same.
Margin or Hyperplane will only change when the $|w^T \cdot x + b| < 1$ for the newly added point.