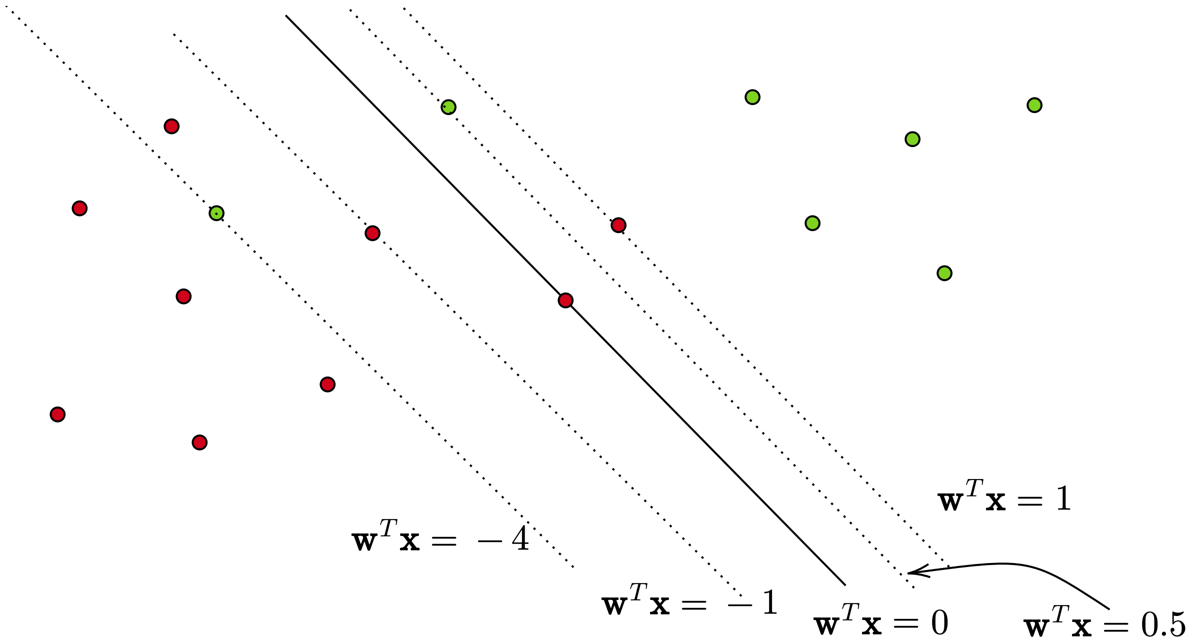
Week 11 | Solve With Instructors

Q1)


$\mathbf{w}^T\mathbf{x} = 1$
$\mathbf{w}^T\mathbf{x} = -4$
$\mathbf{w}^T\mathbf{x} = -1$
$\mathbf{w}^T\mathbf{x} = 0$
$\mathbf{w}^T\mathbf{x} = 0.5$

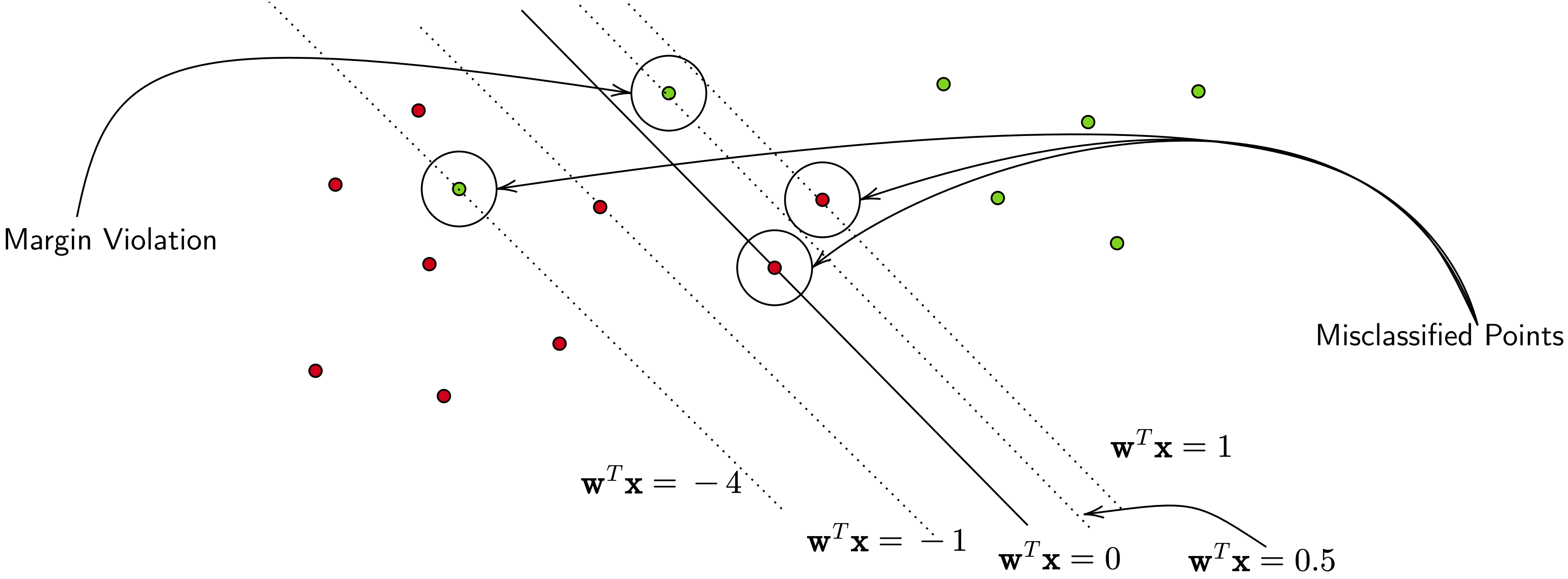
Based on the above graph, answer the following questions :

i) How many data points exhibit margin violations but are still classified on the correct side of the linear classifier?

ii) How many data points are misclassified by the linear classifier?

iii) What is the total bribe (penalty) ?

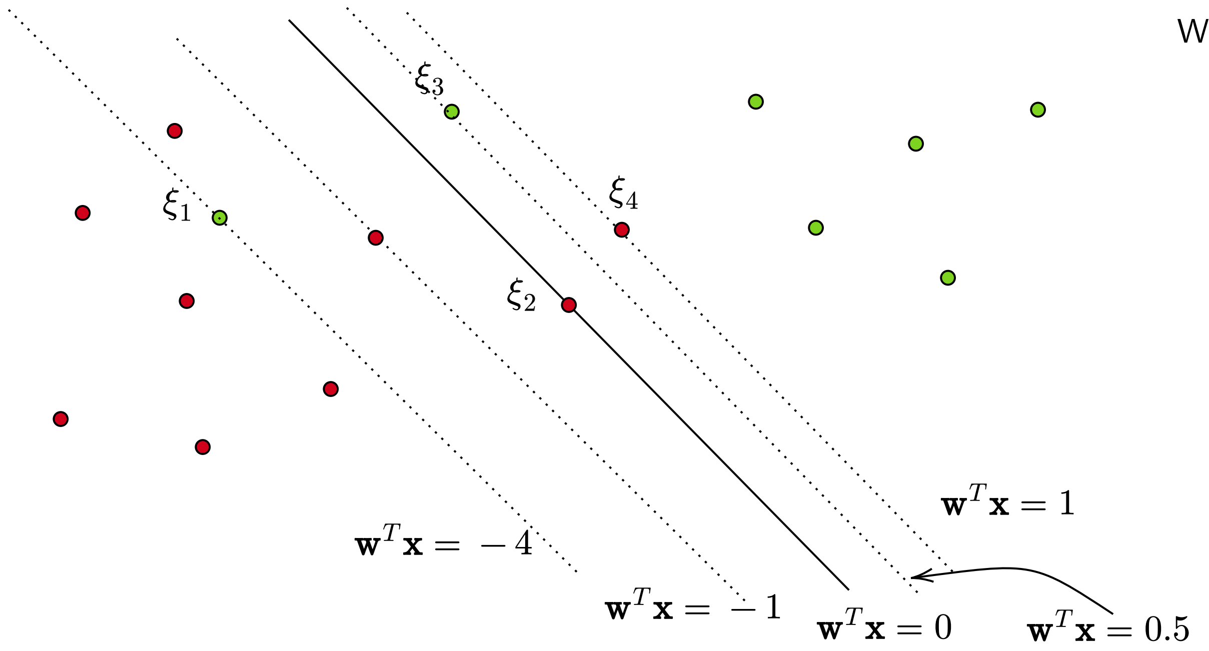
Solution:



Margin Violation
Misclassified Points
$\mathbf{w}^T\mathbf{x} = -4$
$\mathbf{w}^T\mathbf{x} = -1$
$\mathbf{w}^T\mathbf{x} = 0$
$\mathbf{w}^T\mathbf{x} = 1$
$\mathbf{w}^T\mathbf{x} = 0.5$

So there is one data point which exhibit margin violations but are still classified on the correct side of the linear classifier. While there are three data points which are misclassified by the linear classifier.

Only the points which either exhibit margin violations or misclassified will pay bribe (penalty). For others it would be zero. Let's denote the penalty paid these points as follows:



ξ_3
ξ_4
ξ_1
ξ_2
$\mathbf{w}^T\mathbf{x} = 1$
$\mathbf{w}^T\mathbf{x} = -4$
$\mathbf{w}^T\mathbf{x} = -1$
$\mathbf{w}^T\mathbf{x} = 0$
$\mathbf{w}^T\mathbf{x} = 0.5$

We know that $\xi_i = \max\Big(0, 1 - \big(\mathbf{w}^T\mathbf{x}_i\big)y_i\Big)$

$$\xi_1 = \max(0, 1 - (-4)(1)) = 5$$

$$\xi_2 = \max(0, 1 - 0(1)) = 1$$

$$\xi_3 = \max(0, 1 - 0.5(1)) = 0.5$$

$$\xi_4 = \max(0, 1 - 1(-1)) = 2$$

Hence the total bribe is $5 + 1 + 0.5 + 2 = 8.5$

Q2) Which of the following options are **incorrect** about soft margin support vector machine algorithm?

(a) Points that violate the margin and are farther away from the correct supporting hyperplane suffer greater penalty.

(b) Points that are beyond the margin and on the right side of it do not suffer any penalty.

(c) If $\|\mathbf{w}\|$ is small, then margin would also be small.

(d) If $\|\mathbf{w}\|$ is large, then margin would be small.

(e) A very small value of C encourages wider margin.

(f) Points that lie on the wrong side of the correct supporting hyperplane (margin violation) have $\alpha_i^* = C$.

Solution:

1. Points that violate the margin and are farther away from the correct supporting hyperplane suffer greater penalty.
   - Correct: The penalty for misclassification increases with the distance of the points from the correct hyperplane.

2. Points that are beyond the margin and on the right side of it do not suffer any penalty.
   - Correct: If the points are correctly classified and fall outside the margin (on the correct side), they are not penalized.

3. If $\|\mathbf{w}\|$ is small, then margin would also be small.
   - Incorrect: The small value of $\|\mathbf{w}\|$ corresponds to a wider margin.

4. If $\|\mathbf{w}\|$ is large, then margin would be small.
   - Correct: A large value of $\|\mathbf{w}\|$ corresponds to a narrower margin.

5. A very small value of C encourages wider margin.
   - Correct: A smaller C leads to a larger margin, with less emphasis on correctly classifying all points, allowing some points to be misclassified in favor of maximizing margin.

6. Points that lie on the wrong side of the correct supporting hyperplane (margin violation) have $\alpha_i^* = C$.
  - Correct: If the point lie on the wrong side of the correct supporting hyperplane (margin violation) have $\mathbf{w}^T\mathbf{x}_i < 1$.

So, for these points $\xi_i > 0$ as $\xi_i = \max\Big(0, 1 - \big(\mathbf{w}^T\mathbf{x}_i\big)y_i\Big)$.

Now from the complementary slackness condition, $\beta_i\xi_i = 0$, we will will get $\beta_i = 0$ as $\xi_i \neq 0$.

From dual constraint, $\alpha_i + \beta_i = C$, we will get $\alpha_i = C$.

Q3) Consider that the AdaBoost model is trained on the following binary classification dataset.

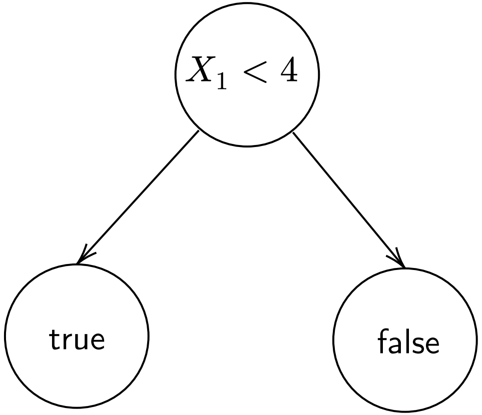
X_1	X_2	Label (y)
2	5	true
2.5	6	false
3	5	true
4	3	false
4	4	false

The dataset is split according to the first feature X_1 to create the stump. Equal sample weight are assigned to each example to create the first stump.

i) How many data points will be misclassified by the trained stump?

ii) What sample weight will be assigned to the first example to create the next stump? Don't normalize the sample weights. Write your answer correct to two decimal places.

Solution: The decision stump that we will get using the best information gain is :



The number of misclassified points is 1 which is $\begin{bmatrix} 2.5 & 6 \end{bmatrix}^T$. The error is $\frac{1}{5}$. To train next stump, we will decrease the weight of correctly classified points by $e^{-\alpha}.\, D_i$ and increase the weight of misclassified points by $e^{\alpha}.\, D_i$ where $\alpha = \frac{1}{2}\ln\left[\frac{1-e}{e}\right]$ and D_i is the previous weight of data point.

The first example is correctly classified so we will decrease its weight. The initial weight of this point is $\frac{1}{5}$ as the equal sample weight are assigned to each example to create the first stump.

$$\alpha = \frac{1}{2}\ln\left[\frac{1-0.2}{0.2}\right] = \frac{1}{2}\ln(4)$$

$$\text{Updated weight of first example} = e^{-\alpha}.\ D_i$$

$$= e^{-\frac{1}{2}\ln(4)} = \frac{1}{2}$$

So, updated weight is $\frac{1}{5} \times \frac{1}{2} = \frac{1}{10}$.

Q4) Select all the true statements about Ensemble Techniques

a) A weak learner is a model that does slightly better than a random model.

b) Ensemble techniques attempt to aggregate or combine multiple models to arrive at a decision.

c) Bagging or bootstrap aggregation is a technique that tries to reduce the variance.

d) The probability that a point appears at least once in a bag (gets picked at all) in Bagging for a very large dataset is around 36%.

e) Bagging can be done in parallel.

f) Bagging always uses decision trees as the base learners.

g) Random Forest uses all the features at each split in each tree.

h) Boosting is an ensemble method that trains models sequentially, with each new model focusing on the errors made by the previous models.

i) Mistakes in Boosting are boosted by e^{α} where $\alpha = \frac{1}{2}\ln\left[\frac{1-e}{e}\right]$.

j) In Boosting, if the classifier in round t has a higher error, it is assigned a lower weight in the final ensemble.

k) In Boosting, the final prediction is a weighted combination of all individual models, where the weights depend on the performance of each model during training.

Solution:

1. A weak learner is a model that does slightly better than a random model.
This is true. A weak learner is a model that performs marginally better than random guessing, often used as the base learner in ensemble methods like Boosting.

2. Ensemble techniques attempt to aggregate or combine multiple models to arrive at a decision.
This is true. Ensemble methods combine multiple individual models to improve overall performance, often reducing errors and increasing predictive power.

3. Bagging or bootstrap aggregation is a technique that tries to reduce the variance.
This is true. Bagging works by training multiple models on different subsets of the data (via bootstrapping) and then averaging their predictions to reduce variance and improve generalization.

4. The probability that a point appears at least once in a bag (gets picked at all) in Bagging for a very large dataset is around 36%.
This is false. In Bagging, each data point has a probability of 36% of being excluded from a bootstrap sample (i.e., not being selected).

5. Bagging can be done in parallel.

This is true. Since Bagging involves training multiple models independently, the training process for each model can be done in parallel, making it highly efficient.

6. Bagging always uses decision trees as the base learners.

This is false. Bagging can use any base learner, not just decision trees. Decision trees are a common choice, but Bagging is not limited to them.

7. Random Forest uses all the features at each split in each tree.

This is false. Random Forest uses a random subset of features at each split, rather than using all features, to improve diversity and reduce overfitting.

8. Boosting is an ensemble method that trains models sequentially, with each new model focusing on the errors made by the previous models.

This is true. Boosting methods, like AdaBoost, train models sequentially, with each new model trying to correct the errors of the previous models, thus focusing on hard-to-predict instances.

9. Mistakes in Boosting are boosted by e^{α} where $\alpha = \frac{1}{2}\ln\left[\frac{1-e}{e}\right]$.

This is true. In Boosting the weights of misclassified data points are increased by e^{α} while for correctly classified data points the weights are decreased by $e^{-\alpha}$ where $\alpha = \frac{1}{2}\ln\left[\frac{1-e}{e}\right]$.

10. In Boosting, if the classifier in round t has a higher error, it is assigned a lower weight in the final ensemble.

This is true. In Boosting, models with higher errors are given less importance (lower weights) in the final ensemble prediction.

11. In Boosting, the final prediction is a weighted combination of all individual models, where the weights depend on the performance of each model during training.

This is true. In Boosting, each model contributes to the final prediction based on its accuracy, with more accurate models being given higher weights.

$Q5$) Choose the correct option(s) assuming $w^*, \mathcal{E}^*$ to be the primal optimal solutions and α^* and β^* to be the dual optimal solutions of the soft margin SVM.

A) If $\left(w^{*^T} x_i\right) y_i \geqslant 1$, then the i^{th} data point pays a positive bribe$\left(\mathcal{E}_i^*\right)$

B) If $\left(w^{*^T} x_i\right) y_i \geqslant 1$, then the i^{th} data point does not pay any bribe$\left(\mathcal{E}_i^*\right)$

C) If $\left(w^{*^T} x_i\right) y_i \leqslant 1$, then the i^{th} data point pays a positive bribe$\left(\mathcal{E}_i^*\right)$

D) If $\left(w^{*^T} x_i\right) y_i \leqslant 1$, then the i^{th} data point does not pay any bribe$\left(\mathcal{E}_i^*\right)$

ANSWER : B

The constraints of soft margin SVM are :

$$1 - \left(w^T x_i\right) y_i - \mathcal{E}_i \leqslant 0,\ \mathcal{E}_i^*$$
$$\mathcal{E}_i \geqslant 0,\ \forall i$$

Option A, B

If $\left({w^*}^T x_i\right) y_i \geqslant 1$, then $1 - \left({w^*}^T x_i\right) y_i \leqslant 0$. The first constraint is already satisfied without bribe. Therefore, paying bribe only will increase the value of the objective function.Therefore, $\mathcal{E}_i = 0$.

Option C, D

If $\left({w^*}^T x_i\right) y_i \leqslant 1$, then $1 - \left({w^*}^T x_i\right) y_i \geqslant 0$. We would want this to be less than or equal to 0.

If $1 - \left({w^*}^T x_i\right) y_i = 0$, then $\mathcal{E}_i^* = 0$, as the first constraint is satisfied with any bribe.

If $1 - \left({w^*}^T x_i\right) y_i > 0$, then $\mathcal{E}_i^* > 0$, as the first constraint is not satisfied.

Combining these two, we get $\mathcal{E}_i^* \geqslant 0$. The data point may or may not pay bribe, depending on the actual value of $\left({w^*}^T x_i\right) y_i$.

$Q6)$ Choose the correct option(s) assuming $w^*, \mathcal{E}^*$ to be the primal optimal solutions and α^* and β^* to be the dual optimal solutions of the soft margin SVM.

A) A support vector does not pay any bribe
B) A support vector may or may not pay any bribe
C) A support vector definitely pays a positive bribe
D) If a data point pays bribe, then it is not a support vector
E) If a data point pays bribe, then may or may not be a support vector
F) If a data point pays bribe, then it is a support vector

ANSWER : B, F

The i^{th} data point is a support vector if $a_i^* > 0$

Option A, B, C

If $C \geqslant \alpha_i^* > 0$, then $C > \beta_i^* \geqslant 0$.

If $\beta_i^* = 0$, then $\mathcal{E}_i^* \geqslant 0$ and if $\beta_i^* > 0$, then $\mathcal{E}_i^* = 0$. Combining these two, we get $\mathcal{E}_i^* \geqslant 0$

Option D, E, F

If $\mathcal{E}_i^* > 0$, then $\beta_i^* = 0$. Therefore, $\alpha_i^* = C$.

7) Choose the correct option(s) assuming $w^*, \mathcal{E}^*$ to be the primal optimal solutions and α^* and β^* to be the dual optimal solutions of the soft margin SVM.

A) There could be a possibility of $\alpha_i^* > 0$ and $\mathcal{E}_i^* = 0$

B) There could be a possibility of $\alpha_i^* = C$ and $\mathcal{E}_i^* = 0$

C) Both A and B

D) None

ANSWER : A, B, C

Option A

If $C \geqslant a_i^* > 0$, then $C > \beta_i^* \geqslant 0$. If $\beta_i^* = 0$, we get $\mathcal{E}_i^* \geqslant 0$ and if $\beta_i^* > 0$, we get $\mathcal{E}_i^* = 0$.

OPTION B

If $\alpha_i^* = C$, then $\beta_i^* = 0$. Therefore, $\mathcal{E}_i^* \geqslant 0$.

PRACTICE ASSIGNMENT QUESTION 1

In a random forest model, let p be the number of randomly selected features that are used to identify the best split at any node of a tree. Which of the following is true? (d is the total number of features)

A) Increasing p reduces the correlation between any two trees in the forest.
B) Decreasing p reduces the correlation between any two trees in the forest.
C) Increasing p increases the performance of individual trees in the forest.
D) Decreasing p increases the performance of individual trees in the forest.

ANSWERS : B, C

PRACTICE ASSIGNMENT QUESTION 5

Assume that the first decision stump of the AdaBoost algorithm wrongly classifies 30 data points out of 100 data points. What will be the weights assigned to the incorrectly classified points for sampling the data points to train the second decision stump? Assume that error is defined as the proportion of incorrectly classified examples by the decision stump.

A) $\dfrac{\sqrt{7/3}}{100}$

B) $\dfrac{\sqrt{3/7}}{100}$

C) $\dfrac{\sqrt{7/3}}{30\sqrt{7/3}+70\sqrt{3/7}}$

D) $\dfrac{\sqrt{7/3}}{30\sqrt{3/7}+70\sqrt{7/3}}$

ANSWER : $\dfrac{\sqrt{7/3}}{30\sqrt{7/3}+70\sqrt{3/7}}$

equal weights in the 1^{st} round. $w_i = \dfrac{1}{1000},\ for\ all\ i$

error $= \dfrac{30}{100} = 0.3$

$$\alpha = \ln\left(\sqrt{\frac{0.7}{0.3}}\right)$$

for eg, let w_{10} be misclassified

$$w_{10} = w_{10} \times e^{\alpha} = \frac{\sqrt{7/3}}{100}$$

for eg, let w_{20} be correctly classified

$$w_{20} = w_{20} \times e^{-\alpha} = \frac{\sqrt{3/7}}{100}$$

Normalize

$$w_{10} = \frac{\frac{\sqrt{7/3}}{100}}{\frac{30\sqrt{7/3}}{100}+\frac{70\sqrt{3/7}}{100}} = \frac{\sqrt{7/3}}{30\sqrt{7/3}+70\sqrt{3/7}}$$

$$w_{20} = \frac{\frac{\sqrt{3/7}}{100}}{\frac{30\sqrt{7/3}}{100}+\frac{70\sqrt{3/7}}{100}} = \frac{\sqrt{3/7}}{30\sqrt{7/3}+70\sqrt{3/7}}$$

PRACTICE ASSIGNMENT QUESTION $6, 7$

You have been given a dataset in $1 - d$ space, which consists of 4 positive data points $1, 2, 3, 4$ and 3 negative data points $-3, -2, -1$. We want to learn a soft-margin SVM (though the dataset is linearly separable) for this dataset. Think those points on the real line.

a) If $C \to \infty,$ how many support vectors do we have?
b) If $C = 0,$ how many support vectors do we have?

ANSWERS

a) 2

b) 0

$C = 0, w = 0$

$\mathcal{E}_i^* > 0,\ for\ all\ i$

$\beta_i^* = 0$

$\alpha_i^* = C$

$\alpha_i^* = 0$