NPTEL Week-4 Live Session

on Machine Learning and Deep Learning - Fundamentals and Applications (noc24_ee146)

Powered by:

FREE ONLINE EDUCATION
swayam
शिक्षित भारत, उन्नत भारत

A course offered by: Prof. Manas Kamal Bhuyan, IIT Guwahati

NPTEL Quiz Solution: Week-3, Practice problem on Support Vector Machine

By

Arka Roy
NPTEL PMRF TA

Prime Minister's Research Fellow
Department of Electrical Engineering, IIT Patna
Web: https://sites.google.com/view/arka-roy/home

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी • Indian Institute of Technology Guwahati •

PMRF
Prime Minister's Research Fellows
Ministry of Education
Government of India
सत्यमेव जयते

भारतीय प्रौद्योगिकी संस्थान पटना
INDIAN INSTITUTE OF TECHNOLOGY PATNA
विद्यार्थी लभते विद्याम्

The MLE for the data samples X={x1,x2,…,xi,…,xk} with the Bernoulli distribution is

$X \sim$ Bernoulli Dist.

$\rightarrow$ iid (identical and independent distribution).

(1)

$$PMF = p^{x_i}(1-p)^{(1-x_i)}$$

$$\mathcal{L}(p) = \prod_{i=1}^{k} p^{x_i}(1-p)^{(1-x_i)}$$

$$\Rightarrow \ln \mathcal{L}(p) = \sum_{i=1}^{k}\left(x_i \ln p + (1-x_i)\ln(1-p)\right)$$

$$\frac{d}{dp}\left(\ln \mathcal{L}(p)\right) = \sum_{i=1}^{k}\left(x_i \times \frac{1}{p} + (1-x_i) \times \frac{1}{1-p} \times (-1)\right)$$

$$\frac{d}{dp}\left(\ln \mathcal{L}(p)\right) = \sum_{i=1}^{k}\left(\frac{x_i}{p} - \frac{1-x_i}{1-p}\right) = 0.$$

$$\Rightarrow \frac{\sum_{i=1}^{k} x_i}{p} = \frac{\sum_{i=1}^{k}(1-x_i)}{1-p}$$

$$\Rightarrow \sum_{i=1}^{k} x_i - \cancel{p\sum_{i=1}^{k} x_i} = p\sum_{i=1}^{k} 1 - \cancel{p\sum_{i=1}^{k} x_i}$$

$$\Rightarrow \sum_{i=1}^{k} x_i = p \cdot k$$

$$\Rightarrow \boxed{p = \frac{1}{k}\sum_{i=1}^{k} x_i}$$

$$p = \frac{1}{k}\sum_{i=1}^{k} x_i$$

$\rightarrow$ MLE - The sample mean of the distribution

9) Consider single observation X that depends on a random parameter .Suppose θ has a prior distribution

Prior

$$f_\Theta(\theta) = \lambda e^{-\lambda\theta} \text{ for } \theta \geq 0, \lambda > 0$$

$$f_{\frac{x}{\theta}}(x) = \theta e^{-\theta x} |x| > 0$$

Find the MAP estimation of θ

- [x] $\frac{1}{\lambda+X}$
- [ ] $\frac{1}{\lambda-X}$
- [ ] $\frac{\lambda}{X}$
- [ ] X

Prior :- $f_\theta(\theta) = \lambda e^{-\lambda\theta}$; $\theta \geq 0$, $\lambda > 0$

Conditional density :- $f_{x|\theta}(x) = \theta e^{-\theta x}$; $|x| > 0$

$f_{\theta|x}(\theta) = \frac{f_{x|\theta}(x) \cdot f(\theta)}{}$

Total probability

Aposteriori probability $f_{\theta|x}(\theta) \propto f_{x|\theta}(x) \cdot f_\theta(\theta)$.

$\propto \theta e^{-\theta x} \cdot \lambda e^{-\lambda\theta}$

$f_{\theta|x}(\theta) \propto \lambda \theta e^{-(\lambda+x)\theta}$

$\ln f_{\theta|x}(\theta) \propto \ln(\lambda\theta e^{-(\lambda+x)\theta})$

$\ln f_{\theta|x}(\theta) \propto \ln\lambda + \ln\theta - (\lambda+x)\theta$

$\frac{d}{d\theta} \ln f_{\theta|x}(\theta)$.

$\propto \frac{1}{\theta} - (\lambda+x)$

$\frac{d}{d\theta}(\ln f_{\theta|x}(\theta)) = k\left(\frac{1}{\theta} - (\lambda+x)\right)$

$= 0$

$k\left(\frac{1}{\theta} - (\lambda+x)\right) = 0.$

$\boxed{\theta = \frac{1}{\lambda+x}}$.

6) Suppose that X is a discrete random variable with the following probability mass function: where is a parameter.
$(0 \leq \theta \leq 1)$

X	0	1	2	3
P(X)	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

Probability is multiplicative in nature.

$x_i = \{3, 0, 2, 1, 3, 2, 1, 0, 2, 1\}$

The following 10 independent observations were taken from such a distribution: $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$. What is the maximum likelihood estimate of θ ?

- 2
- 1
- 0.5
- 0

$$\mathcal{L}(\theta) = \prod_{i=1}^{10} P(X = x_i)$$

$$= P(X=3) \cdot P(X=0) \cdot P(X=2) \cdot P(X=1) \cdot P(X=3)\, P(X=2) \cdot P(X=1) \cdot P(X=0) \cdot P(X=2) \cdot P(X=1)$$

$$= [P(X=3)]^2 \, [P(X=0)]^2 \, (P(X=2))^3 \, (P(X=1))^3$$

$$\boxed{\mathcal{L}(\theta) = \left(\frac{1-\theta}{3}\right)^2 \cdot \left(\frac{2\theta}{3}\right)^2 \cdot \left(\frac{2}{3}(1-\theta)\right)^3 \cdot \left(\frac{\theta}{3}\right)^3}$$

6) Suppose that X is a discrete random variable with the following probability mass function: where is a parameter.
$(0 \le \theta \le 1)$

$P(X=1|\theta)$

X	0	1	2	3
P(X)	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

The following 10 independent observations were taken from such a distribution: $(3,0,2,1,3,2,1,0,2,1)$. What is the maximum likelihood estimate of θ?

- 2
- 1
- 0.5
- 0

$$\mathcal{L}(\theta) = \left(\frac{1-\theta}{3}\right)^2 \cdot \left(\frac{2\theta}{3}\right)^2 \cdot \left(\frac{2}{3}(1-\theta)\right)^3 \cdot \left(\frac{\theta}{3}\right)^3$$

iid

$$\mathcal{L}(\theta) = \prod_{i=1}^{10} P(X=x_i|\theta)$$

$P(x|\theta)$ class conditional density

$P(X=1|\theta)$ $\theta/3$

$$\ln \mathcal{L}(\theta) = 2\left[\ln\left(\frac{1}{3}\right) + \ln(1-\theta)\right] + 3\left[\ln\frac{2}{3} + \ln(1-\theta)\right]$$
$$+ 2\left[\ln\frac{2}{3} + \ln\theta\right] + 3\left[\ln\frac{1}{3} + \ln\theta\right]$$

$$\ln \mathcal{L}(\theta) = 5\ln\frac{1}{3} + 5\ln\frac{2}{3} + 5\ln(1-\theta) + 5\ln\theta.$$

$$\frac{d}{d\theta}\left(\ln \mathcal{L}(\theta)\right) = 0 + \frac{-5}{1-\theta} + \frac{5}{\theta} = 0$$

$$\Rightarrow \frac{5}{\theta} = \frac{5}{1-\theta}$$

$$\Rightarrow 1-\theta = \theta$$

$$\boxed{\theta = 1/2 = 0.5}$$

4) There are 18 points in an axis plane namely -

$[(0.8, 0.8)^t, (1, 1)^t, (1.2, 0.8)^t, (0.8, 1.2)^t, (1.2, 1.2)^t]$, → 1

belong to class 1;

$[(4, 3)^t, (3.8, 2.8)^t, (4.2, 2.8)^t, (3.8, 3.2)^t (4.2, 3.2)^t, (4.4, 2.8)^t, (4.4, 4.4)^t]$,

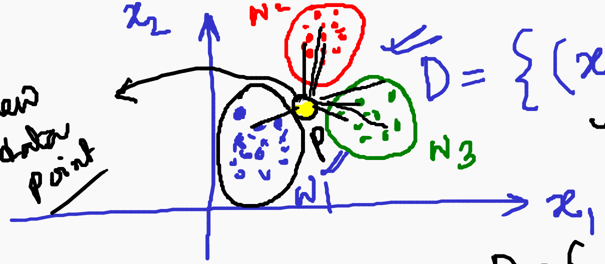
belong to class 2; → 2

$[(3.2, 0.4)^t, (3.2, 0.7)^t, (3.8, 0.5)^t, (3.5, 1)^t, (4, 1)^t, (4, 0.7)^t]$,

belong to class 3. → 3

KNN

K=5



$D = \{(x_1, x_2)_i\}_{i=1}^{18}$

$D = \{X_i, Y_i\}$

Data

Label of class

A new point

$P = (4.2, 1.8)^t$ introduces into the map. The point P belongs to which class? Use k -nearest neighbor technique with $k = 5$ to calculate the result

- Class 1
- Class 2 ✓
- Class 3
- None of the above

Euclidean distance → $d(X_1, P) =$

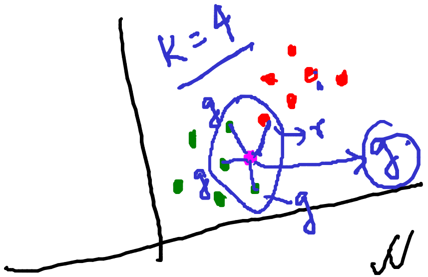
$d(X_2, P) =$

$d(X_3, P) =$

$0^2 + (-1 \times -1) \Leftarrow \begin{bmatrix} 0 \\ -1 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ -1 \end{bmatrix}$

$= 1$

a



4 nearest points from the new data point

$\text{eucdist}\left(\begin{bmatrix} 4.2 \\ 1.8 \end{bmatrix}, \begin{bmatrix} 4.2 \\ 2.8 \end{bmatrix}\right) = \begin{bmatrix} 4.2-4.2 \\ 1.8-2.8 \end{bmatrix}$

$[4.2, 1.8]^t \Leftarrow [4.2, 2.8]^t$

$d(X_8, P) = \sqrt{(4.2-4.2)^2 + (2.8-1.8)^2}$

$= 1.$

$d(x_1,P) = 3.54$

$d(x_2,P) = 3.29$

$d(x_3,P) = 3.16$

$d(x_4,P) = 3.45$

$d(x_5,P) = 3.06$

$d(x_6,P) = 1.22$

$\boxed{d(x_7,P) = 1.08} \rightarrow 2$

$\boxed{d(x_8,P) = 1} \rightarrow 2$

$d(x_9,P) = 1.46$

$d(x_{10},P) = 1.4$

$\boxed{d(x_{11},P) = 1.01} \rightarrow 2$

$d(x_{12},P) = 2.6$

$d(x_{13},P) = 1.72$

$d(x_{14},P) = 1.49$

$d(x_{15},P) = 1.36$

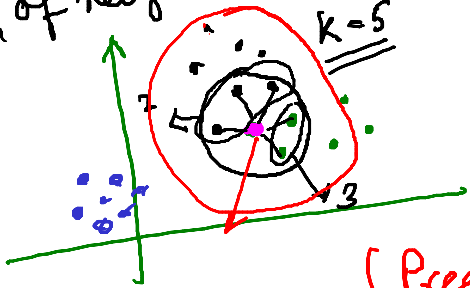
$\boxed{d(x_{16},P) = 1.06} \rightarrow 3$

$\boxed{d(x_{17},P) = 0.82} \rightarrow 3$

$d(x_{18},P) = 1.12$

K=5

No of neighbors.



$P \in$ Class 2 (Prediction)

Acc Test Sample → original lable

→ Radius : 2.6

The bandwidth parameter in the Parzen Window method determines:

The number of neighbors to consider for classification

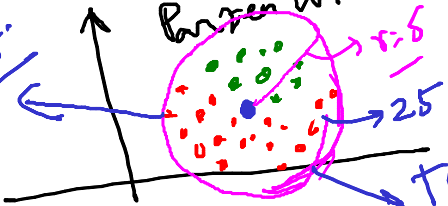
The size of the neighborhood around a test instance

The dimensionality of the feature space.

The complexity of the classifier

Size of neighbor

Parzen window



KNN

K

Total no of data points for neighbors 'bw' specific

The disadvantage of using k -NN as classifier:

- Fails while handling large dataset
- Fails while handling small dataset
- Sensitive to outliers
- Training is required

Large dataset → KNN will have computational problem

Inference Time/Test Time taken by algo. ↑↑

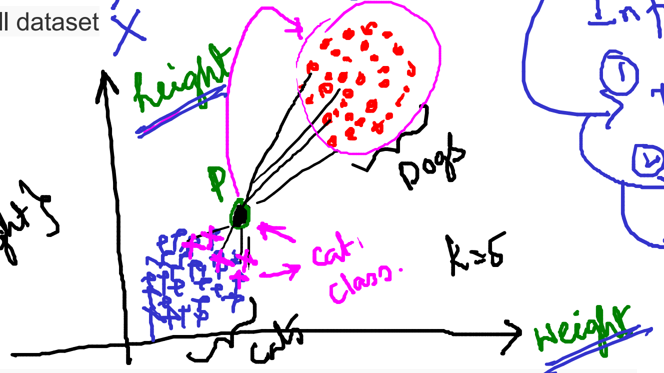
① There is no need of training.

② because this is a non parametric method.

③ And it just has to calculate the distance value from the new point to all other point

W={cat, Dogs}

x={height, weight}

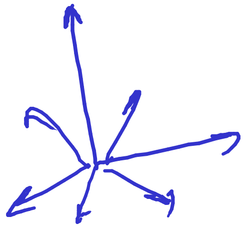


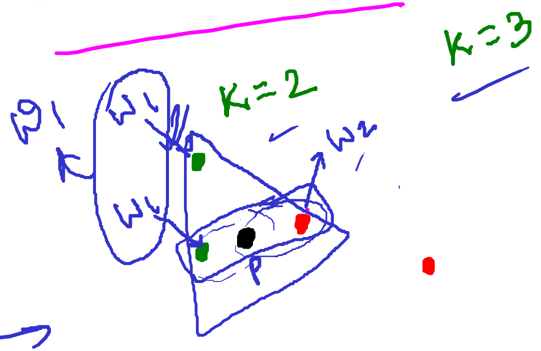
True label of (P)= Dog

Prediction is: Cat → mis classification.

Which of the following statements are true about k - nearest neighbor (KNN)-

- Odd value of "K" preferred over even values.
- Does more computation on test time rather than train time.
- Work well with high dimension. → All points become almost equally distant in higher dimension
- The optimum value of K for KNN is highly independent on the data.





If the number of data samples becomes very large.

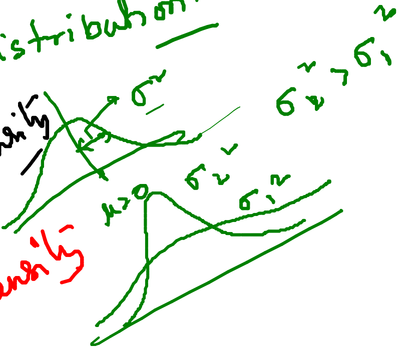
- Bayesian Estimation is worse than MLE
- Maximum Likelihood estimates are slightly bad
- Bayesian Estimation performs same as MLE
- None

$f(\theta|x) = f(x|\theta) \cdot \cancel{f(\theta)}$

aposteriori — class conditional prob. density — Prior

Data

Parameter of distribution.



MLE :- $\mathcal{L} \approx f(x|\theta)$

likelyhood estimate

for large data size

Bayesian estimate $f(\theta|x) \propto f(x|\theta)$

Similar to

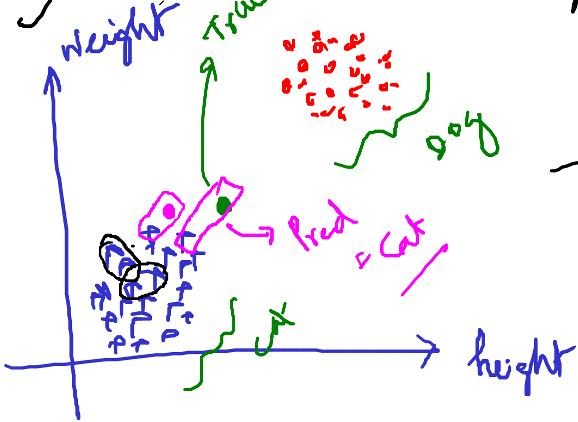
likelyhood function/density

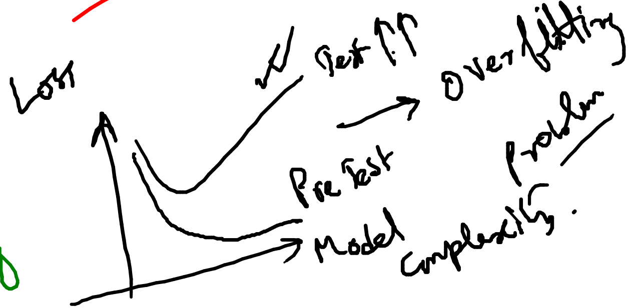
class conditional prob-density

Posterior prob.

What happens when k=1 in k -Nearest Neighbor algorithm:

- Underfitting
- Overfitting
- High testing accuracy
- All the above





Find the scalar projection of vector b = <-3, 2> onto vector a = <1,1>? SVM.

a. 0

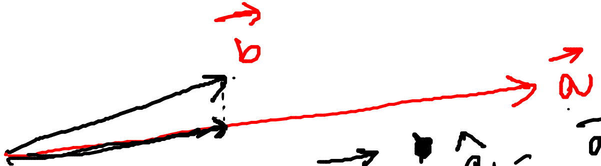
b. $\frac{1}{\sqrt{2}}$

c. $\frac{-1}{\sqrt{2}}$

d. $\frac{-1}{2}$

Dot product = Dot product of $\vec{b}$ and $\hat{a}$

$= \vec{b} \cdot \frac{\vec{a}}{\|\vec{a}\|}$



$\rightarrow \hat{a} = \frac{\vec{a}}{\|\vec{a}\|}$

$= \frac{\langle -3, 2\rangle \langle 1, 1\rangle}{\sqrt{1^2 + 1^2}}$

$= \frac{-3+2}{\sqrt{2}} = \frac{-1}{\sqrt{2}}$

Suppose there is a feature vector represented as [1, 4, 3]. What is the distance of this feature vector from the separating plane $x_1 + 2x_2 - 2x_3 + 3 = 0$. Choose the correct option.

a. 1

b. 5

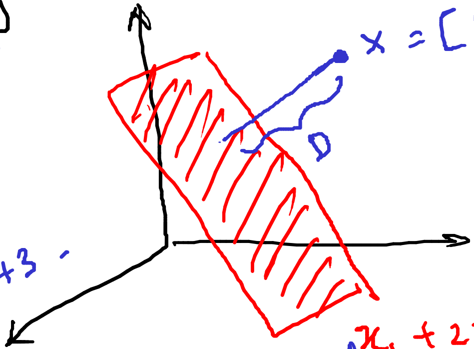
c. 3

d. 2

$X = [1, 4, 3]$

$D = \frac{W \cdot X + W_0}{\|W\|}$

$W^t X + W_0 = [1\ 2\ -2]\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + 3 = x_1 + 2x_2 - 2x_3 + 3$



$x = [1, 4, 3]$

$W^t X + W_0 = 0$ Bias

$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

$W^t = [1, 2, -2]$

$W_0 = 3$

$x_1 + 2x_2 - 2x_3 + 3 = 0$

Find the scalar projection of vector b = <-3, 2> onto vector a = <1,1>?

a. 0

b. $\frac{1}{\sqrt{2}}$

c. $\frac{-1}{\sqrt{2}}$

d. $\frac{-1}{2}$

Suppose there is a feature vector represented as [1, 4, 3]. What is the distance of this feature vector from the separating plane $x_1 + 2x_2 - 2x_3 + 3 = 0$. Choose the correct option.

a. 1

b. 5

c. 3

d. 2 ✓

$$D = \frac{W \cdot X + W_0}{\|W\|}$$

$$= \frac{1 + 8 - 6 + 3}{\sqrt{1^2 + 2^2 + 2^2}}$$

$W = [1, 2, -2]$

$X = [1 \;\; 4 \;\; 3]$

$w_0 = 3$

$$= \frac{6}{\sqrt{9}} = \frac{6}{3} = 2$$ ✓

Find the distance of the 3D point, $P = (-2, 4, 1)$ from the plane defined by $2x + 3y + 6z + 7 = 0$?

a. 3

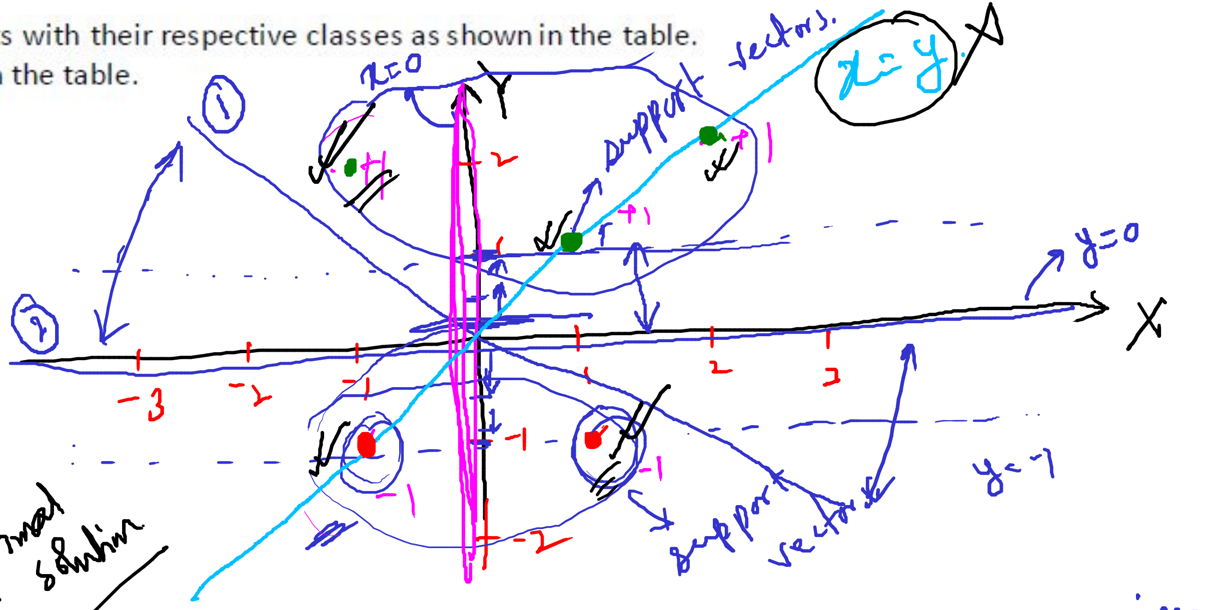
b. 4

c. 0

d. ∞ (infinity)

Suppose we have the below set of points with their respective classes as shown in the table. Answer the following question based on the table.

X	Y	Class Label
1	1	+1
-1	-1	-1
2	2	+1
-1	2	+1
1	-1	-1



optimal solution

SVM Corresponds to decision boundary which gives maximum margin.

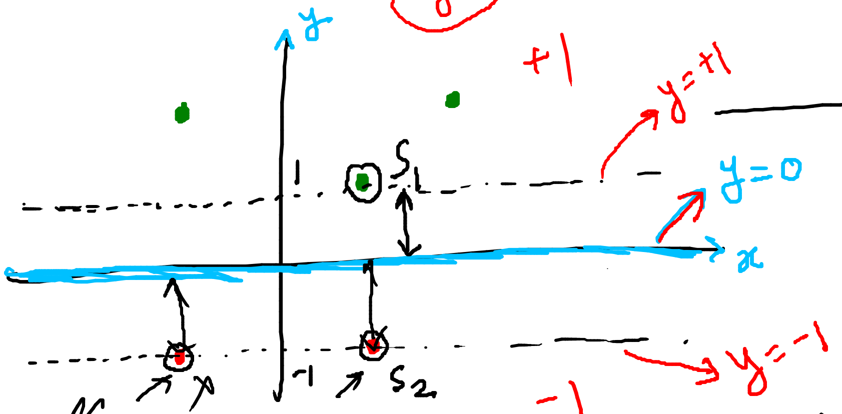
What can be a possible decision boundary of the SVM for the given points?

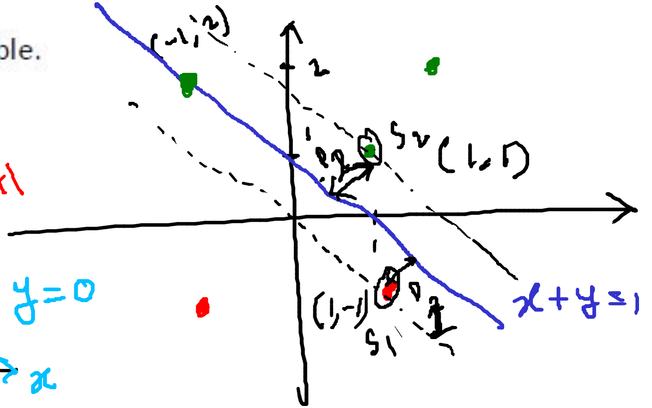
a. $y = 0$
b. $x = 0$
c. $x = y$
d. $x + y = 1$

Suppose we have the below set of points with their respective classes as shown in the table. Answer the following question based on the table.

X	Y	Class Label
1	1	+1
-1	-1	-1
2	2	+1
-1	2	+1
1	-1	-1

$y = K$





$\text{Margin}_{(y=0)} = \sum$ Distance between Decision Surface and unique set of Support vectors, residing at different location.

$= D(S_2 \text{ to Dec}) + D(S_1 \text{ to Dec})$

$= 2.$

What can be a possible decision boundary of the SVM for the given points?

a. $y = 0$ → $M = 2$.

b. $x = 0$

c. $x = y$

d. $x + y = 1$

$-1+2<1$

$\text{Margin}_{(x+y=1)} = D_1 + D_2 = \dfrac{|1-1-1|}{\sqrt{1^2+1^2}} + \dfrac{|1+1-1|}{\sqrt{1^2+1^2}} = \dfrac{1}{\sqrt{2}} + \dfrac{1}{\sqrt{2}}$

$= \dfrac{2}{\sqrt{2}} = \sqrt{2} = 1.414$

$W = [1,1];\ w_0 = -1$

$x + y - 1 < 0$

Suppose we have the below set of points with their respective classes as shown in the table. Answer the following question based on the table.

X	Y	Class Label
1	1	+1
-1	-1	-1
2	2	+1
-1	2	+1
1	-1	-1

D_1

$$D = \frac{W \cdot X + W_0}{\|W\|} = \frac{|1 - 1 - 1|}{\sqrt{1^2 + 1^2}}$$

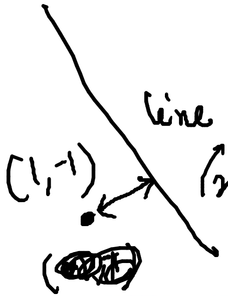
Point $(1,-1)$

$X = (1,-1)$

$w^t = (1, 1)$ $\quad x = [1, -1]$

$W_0 = -1$

$w^t X + W_0 = 0$.



$x + y - 1 = 0$

$X = \begin{bmatrix} x \\ y \end{bmatrix}$

$w^t x = [w_1 \ \ w_2] \begin{bmatrix} x \\ y \end{bmatrix}$

$= [1 \ \ 1] \begin{bmatrix} x \\ y \end{bmatrix}$

$\Rightarrow x + y - 1 = 0$

What can be a possible decision boundary of the SVM for the given points?

a. $y = 0$
b. $x = 0$
c. $x = y$
d. $x + y = 1$