Machine Learning Assignment 1
Linear Models & Kernel Methods

Submission deadline: October 26, 2024

1 Problem 1. Ridge Regression (10 points)

In a regression task, we have vectors $\mathbf{x} \in \mathbb{R}^D$, target values $y \in \mathbb{R}$ associated with them, and some model $f(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$ to predict the target values for arbitrary vectors in $\mathbb{R}^D$.

Suppose we have a training dataset $\{\mathbf{\Phi}, \mathbf{t}\}$, where $\mathbf{\Phi} \in \mathbb{R}^{N \times D}$ is the design matrix in which each row is a feature vector $\phi(\mathbf{x})$ of a training point $\mathbf{x}$, and $\mathbf{t} \in \mathbb{R}^{N \times 1}$ is the vector with target values for the training points. N is the number of points in the training dataset, and D is the dimensionality of the feature space. Suppose that each entry in the last column of $\mathbf{\Phi}$ is equal to 1. Your task is to derive the closed form solution for the optimal parameters of a ridge regression model.

- State the equation of a ridge regression model and identify the model parameters
- State the equation for the loss function (mean squared error) with an ℓ_2 regularization weighted by λ
- State which condition should be met in order to find the model parameters
- Find the ideal model parameters under the proposed loss function.

Note: You can use $\|\cdot\|$ as the Euclidean norm of a vector; $\phi(\mathbf{x}_n)$ and t_n are the n-th rows of $\mathbf{\Phi}$ and $\mathbf{t}$ respectively.

2 Problem 2. Feature Engineering (10 points) and Basic Concepts (10 points)

Suppose you have the following set S of 2D points, $S_n = (x_n^{(1)}, x_n^{(2)})$.

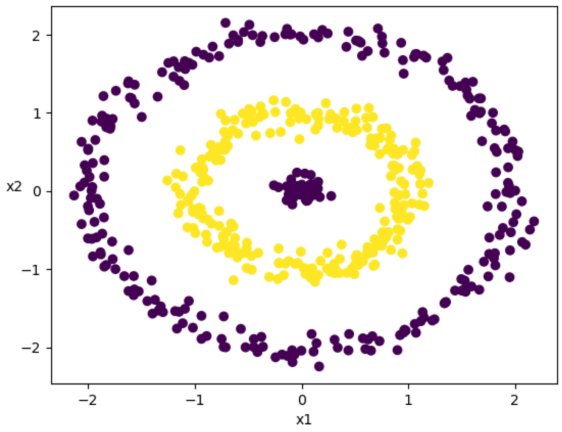


Figure 1: Color denotes the class attribution of a point: blue points belong to the class C_1, yellow points belong to the class C_2.

- Explain in detail 2 classification algorithms that could solve the problem. Discuss advantages and disadvantages of each.
- Propose new features for points in S based on $x^{(1)}$ and $x^{(2)}$. In this new feature space, classes C_1 and C_2 should be linearly separable. Come up with 2 different solutions meeting the stated criteria.

Problem 3. Kernel Functions (10 points)

Consider the following function $f : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$:

$$f(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{x})(\mathbf{x}^T\mathbf{y})(\mathbf{y}^T\mathbf{y})$$

Prove that f is a valid kernel or prove the opposite.

The only rules allowed to use without proof are the following:

- Kernel functions are *linear* and positive.
- Kernel functions can be expressed as an inner product
- A kernel function of 2 inputs can be expressed as another kernel of a transformation of those inputs (into a potentially different space).

Begin by formalizing those rules and apply them to prove or disprove the statement.

Problem 4. SVM (10 points)

Consider the following training data:

Class	x_1	x_2
+	1	1
+	2	2
+	0	2
−	1	-1
−	-1	0
−	0	0

1. Plot the six training points. Are the classes $\{+, -\}$ linearly separable?
2. Construct the weight vector of the maximum margin hyperplane by inspection and identify the support vectors.
3. If you remove one of the support vectors, does the size of the optimal margin decrease, stay the same, or increase?
4. Is your answer to (3) also true for any dataset? Provide a counterexample or give a short proof.

PROBLEM 1

$x \in \mathbb{R}^D, y \in \mathbb{R} \quad f(x): \mathbb{R}^D \to \mathbb{R}$

TRAINING DATASET $\{\mathbf{T}, t\}: \mathbf{T} \in \mathbb{R}^{N \times D}$ DESIGN MATRIX, $t \in \mathbb{R}^n$

$\mathbf{T}_D(x_i) = 1 \; \forall i = 1, \dots, N$

(a)

LET $w = \begin{vmatrix} w_0 \\ \vdots \\ w_D \end{vmatrix} \in \mathbb{R}^D$ BE THE VECTOR OF EACH GRADE'S WEIGHT

$y(x, w) = w^T \phi(x) \qquad \phi(x_n) = \begin{vmatrix} 0 \\ x \\ \vdots \\ x^{D-1} \\ 1 \end{vmatrix} \quad \forall n = 1, \dots, N$

$\Rightarrow \; y(x, w) = \sum_{i=0}^{D} w_i x^i = \sum_{i=0}^{D-1} w_i x^i + w_D$

NOTICE THAT w_D BEHAVES AS THE FUNCTION'S BIAS, THAT WE ASSUMED NOT TO BE CONSIDERED

(b)

$$\tilde{E}(w) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2} (y(x_n, w) - t_n)^2 + \frac{\lambda}{2} \|w\|^2 =$$

$$= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2} (w^T \phi(x_n) - t_n)^T (w^T \phi(x_n) - t_n) + \frac{\lambda}{2} w^T \cdot w, \text{ WHERE } \lambda > 0$$

(c)

IN ORDER TO FIND THE MODEL PARAMETER, WE NEED AN EQUATION WRITTEN AS "SOMETHING" $\cdot w =$ "SOLUTION", WHERE "SOMETHING" IS EXPECTED TO BE A SQUARE, INVERTIBLE MATRIX

$$\tilde{E}(w) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(w^T\phi(x_n)-t_n)^T(w^T\phi(x_n)-t_n)+\frac{\lambda}{2}w^T\cdot w$$

$$= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(w^T\phi(x_n)-t_n)^2+\frac{\lambda}{2}w^T\cdot w$$

$$\frac{d\tilde{E}}{dw} = \left(\frac{\partial E}{\partial w_0},\dots,\frac{\partial E}{\partial w_D}\right) > 0$$

$$I = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{vmatrix} \in \mathbb{R}^{D\times D}$$

IN GENERAL, $\frac{\partial E}{\partial w_i} = \frac{1}{N}\sum_{n=1}^{N}(w_i x_n^i - t_n)\cdot x^i + \frac{\lambda}{2}(w_i \otimes I + \Phi \otimes w_i)$

$$= \frac{1}{N}\sum_{n=1}^{N}(w_i x_n^i - t_n)\cdot x^i + \lambda(w_i \otimes I)$$

$$\frac{d\tilde{E}}{dw} = \left(\frac{1}{N}\sum_{n=1}^{N}(w_0 x_n^0 - t_n)\cdot x^0 + \lambda(w_0 \otimes I),\dots,\right.$$

$$\dots, \frac{1}{N}\sum_{n=1}^{N}(w_{D-1} x_n^{D-1} - t_n)\cdot x_n^{D-1} + \lambda(w_{D-1}\otimes I),$$

$$\left., \frac{1}{N}\sum_{n=1}^{N}(w_D \overset{\sim 1}{x_n^D} - t_n)\cdot \overset{\sim 1}{x_n^D} + \lambda(w_0 \otimes I)\right) =$$

$$T = \begin{vmatrix} t_1 \\ \vdots \\ t_n \end{vmatrix}$$

$$= \frac{1}{N}(w^T\cdot\Phi^T - T^T)\Phi + \lambda w^T I > 0$$

$$w^T\Phi^T\Phi + \lambda w^T I > T^T\Phi$$

$$\Phi^T\Phi\, w + \lambda I w > \Phi^T T$$

$$(\Phi^T\Phi + \lambda I)w > \Phi^T T$$

$\det I = 1 \neq 0$ ✓ $\quad \det(\Phi^T\Phi) = \det(\Phi)^2 \neq 0$ ✓

⇒ DESIGN MATRIX MUST BE INVERTIBLE

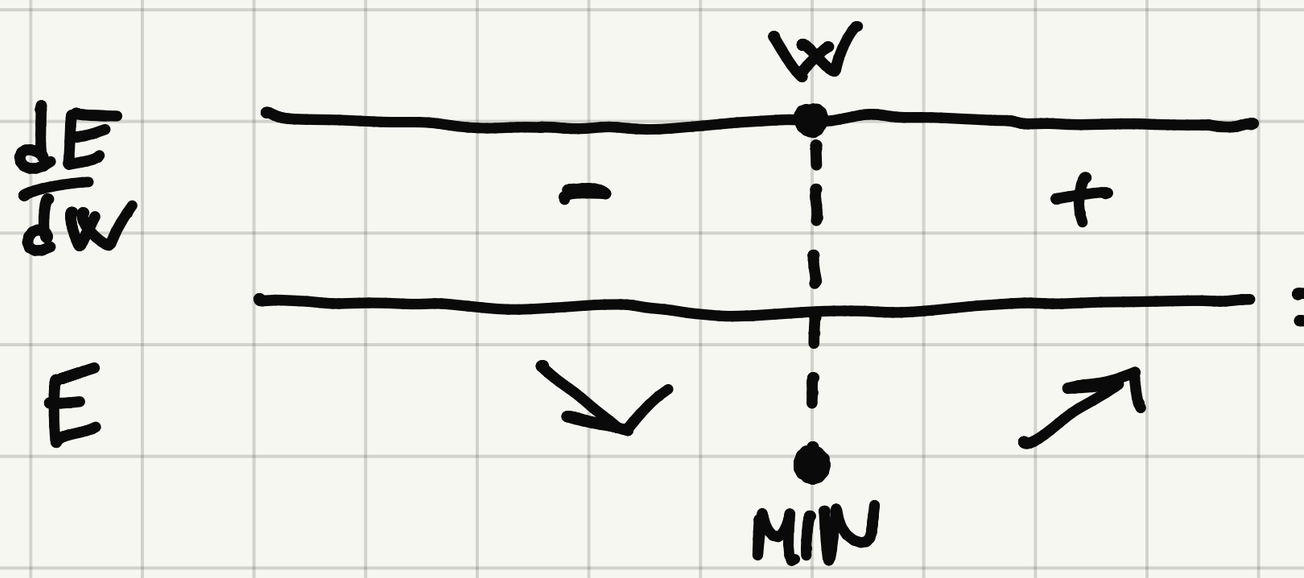
THIS CONDITION CAN BE EXTENDED FOR ANY W, NOT ONLY FOR THE OPTIMAL

IT IS ALSO IMMEDIATE TO VERIFY THAT $\Phi^T\Phi \in \mathbb{R}^{D\times D}$, SINCE

$\Phi^T \in \mathbb{R}^{D\times N}$, $\Phi^T \in \mathbb{R}^{N\times D}$ $\Big\}$ $\Phi^T \cdot \Phi \in \mathbb{R}^{D\times N\cdot N\times D} \in \mathbb{R}^{D\times D}$, HENCE $\Phi^T\Phi + \lambda I$, WHERE $I \in \mathbb{R}^{D\times D}$ IS THE IDENTITY MATRIX, IS PERMISSIBLE

(d)

$$(\Phi^T\Phi + \lambda I)W > \Phi^T T \Rightarrow W > (\Phi^T\Phi + \lambda I)^{-1}\Phi^T T$$



$\Rightarrow W_{MIN} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T T$

PROBLEM 2

(a)

1ST) K-NEAREST NEIGHBOUR (K-NN)

ASSIGN A POINT'S CLASS BY COMPUTING THE DISTANCES FROM THE OTHERS. THE CLASS TO WHICH THE NEAREST POINT BELONG WILL BE ASSIGNED TO THE POINT OF INTEREST

ADVANTAGES:

- EASY TO IMPLEMENT, INTUITIVE
- NON-PARAMETRIC

ISSUES:

- REQUIRES TO COMPUTE ALL THE DISTANCES FOR ALL THE CASES
- ANOMALOUS POINTS CAN "DISTURB" THE CALCULUS, ESPECIALLY FOR LOW VALUES => SENSITIVE TO OUTLIERS
- CURSE OF DIMENSIONALITY => IN HIGH-DIMENSIONAL SPACES, THE "NEAREST" CONCEPT BECOMES ALMOST IRRELEVANT BECAUSE OF A "POOR COVERAGE" OF THE SPACE BY THE SAMPLES

2ND) SUPPORT VECTOR MACHINES (SVM)

BUILD AN HYPERPLANE THAT OPTIMIZES THE CLASSES DISTINCTION, BY MAXIMIZING THE MARGIN BETWEEN THE DATA BELONGING TO THE TWO CLASSES.

ADVANTAGES:

- IT CAN EASILY BE USED WITH NON-LINEAR KERNELS WHEN SOLVING NON-LINEAR PROBLEMS, AS THE GIVEN ONE
- LESS SENSITIVITY TO OUTLIERS, SINCE IT WORKS WITH THE NEAREST POINTS TO THE HYPERPLANE (= SUPPORT VECTORS) ONLY
- THANKS TO KERNEL METHODS, IT IS WELL SUITED FOR HIGH-DIMENSIONAL SPACES

ISSUES:

- WORKING WITH SEVERAL DATA CAN STILL LIMIT THE PERFORMANCES
- KERNEL'S CHOICE IS NOT EASY AND TO UNDERESTIMATE; AN INCORRECT CHOICE MIGHT HIGHLY REDUCE THE PERFORMANCES

(b)

APPROACH 1: POLYNOMIAL FEATURE

FOR FEATURES WITH POTENTIALLY INFINITE DIMENSIONS, FOR MAKING THE CLASSES LINEARLY SEPARABLE WE CAN INTRODUCE THE RADIAL BASIS FUNCTION

RBF $\rightarrow$ $K(x^{(1)}, x^{(2)}) = e^{\left(-\frac{\|x^{(1)} - x^{(2)}\|}{2\sigma^2}\right)^2}$ WITH FIXED σ

TO MANAGE KERNEL'S WIDTH

TAYLOR'S EXPANTIONS CAN ALLOW TO REWRITE THE EXPONENTIAL AS

A POWER SERIES YELDING TO THE FOLLOWING FEATURE SPACES:

$$\phi(x^{(1)}) = \left[e^{-\frac{1}{2}\|x^{(1)}\|^2} \cdot \frac{x_1^{n_1} \cdot x_2^{n_2}}{\sqrt{(n_1!\, n_2!)}}\right]$$

$$\phi(x^{(2)}) = \left[e^{-\frac{1}{2}\|x^{(2)}\|^2} \cdot \frac{x_1^{n_1} \cdot x_2^{n_2}}{\sqrt{(n_1!\, n_2!)}}\right]$$

APPROACH 2: INTRODUCTION OF SLACK VARIABLES

SVM GENERALLY WORKS UNDER THE FOLLOWING CONDITIONS:

$$\begin{cases} \min(w^T w) \\ t_n(w^T \phi_n + w_0) \geq 1 \end{cases}$$

IN CASE OF OVERLAPPINGS, WE CAN INTRODUCE SLACK VARIABLES ε_n

$$\varepsilon_n = \begin{cases} 0 & \text{SAMPLE CORRECTLY CLASSIFIED} \\ & (t_n(w^T\phi + w_0) > 1) \\ |t_n - y(\phi_n)| & \text{OTHERWISE} \end{cases}$$

THEY CAN ALLOW SOME POINTS TO VIOLATE THE MARGIN WITHOUT ACTUALLY COMPROMISING THE CLASSIFICATION'S EFFICIENCY. CONDITIONS WILL BE REWRITTEN AS FOLLOW

$$\begin{cases} \min_{w}\left(\frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\varepsilon_n\right) \\ t_n(\vec{w}^T\phi_n + w_0) \geq 1-\varepsilon_n,\ \varepsilon_n \geq 0 \end{cases}$$

WHERE C DENOTES THE TRADEOFF BETWEEN MARGIN WIDTH AND CLASSIFICATION ERROR

PROBLEM 3

- KERNEL FUNCTIONS ARE LINEAR AND NON NEGATIVE

$$\forall K \in \mathcal{K}(x,y),\ z^T K z \geq 0 \quad \forall z \in \mathbb{R}^D$$

- KERNEL FUNCTIONS CAN BE EXPRESSED AS INNER PRODUCT

$$k(x,y) = \langle x,y \rangle \quad x = \begin{vmatrix} x_1 \\ \vdots \\ x_D \end{vmatrix} \quad y = \begin{vmatrix} y_1 \\ \vdots \\ y_D \end{vmatrix} \quad \langle x,y \rangle = x_1 y_1 + \ldots + x_D y_D$$

- A KERNEL FUNCTION OF 2 INPUTS CAN BE EXPRESSED AS ANOTHER KERNEL OF A TRANSFORMATION OF THOSE INPUTS (INTO A POTENTIALLY DIFFERENT SPACE)

GIVEN $\phi: \mathbb{R}^D \to \mathcal{F}$, WITH $\mathcal{F}$ A POTENTIALLY DIFFERENT SPACE

$$k(x,y) = k(\phi(x), \phi(y))$$

PROOF THAT $k(x,y) = (x^T x)(x^T y)(y^T y)$ IS A VALID KERNEL

START WITH THE INNER PRODUCT $k(x,y) = \langle \phi(x), \phi(y) \rangle$. FOR OUR $k(x,y)$

- $x^T x = \langle x,x \rangle$
- $x^T y = \langle x,y \rangle$
- $y^T y = \langle y,y \rangle$

DEFINE $\phi(x) = \begin{vmatrix} x_1 \\ \vdots \\ x_D \\ \|x\|^2 \end{vmatrix}^T \implies k(\phi(x), \phi(y)) = \|x\|^2 (x^T y) \|y\|^2$
$= k(x,y)$ ✓

NOW, WE CAN EXPLOIT THIS TO PROVE THAT $z^T K z \geq 0$

$$z^T K z = \sum_i \sum_j z_i \langle \phi(x_i), \phi(x_j) \rangle z_j =$$

$$= \left\langle \sum_i z_i \phi(x_i), \sum_j z_j \phi(x_j) \right\rangle = \left\| \sum_i z_i \phi(x_i) \right\|^2 \geq 0 \checkmark$$

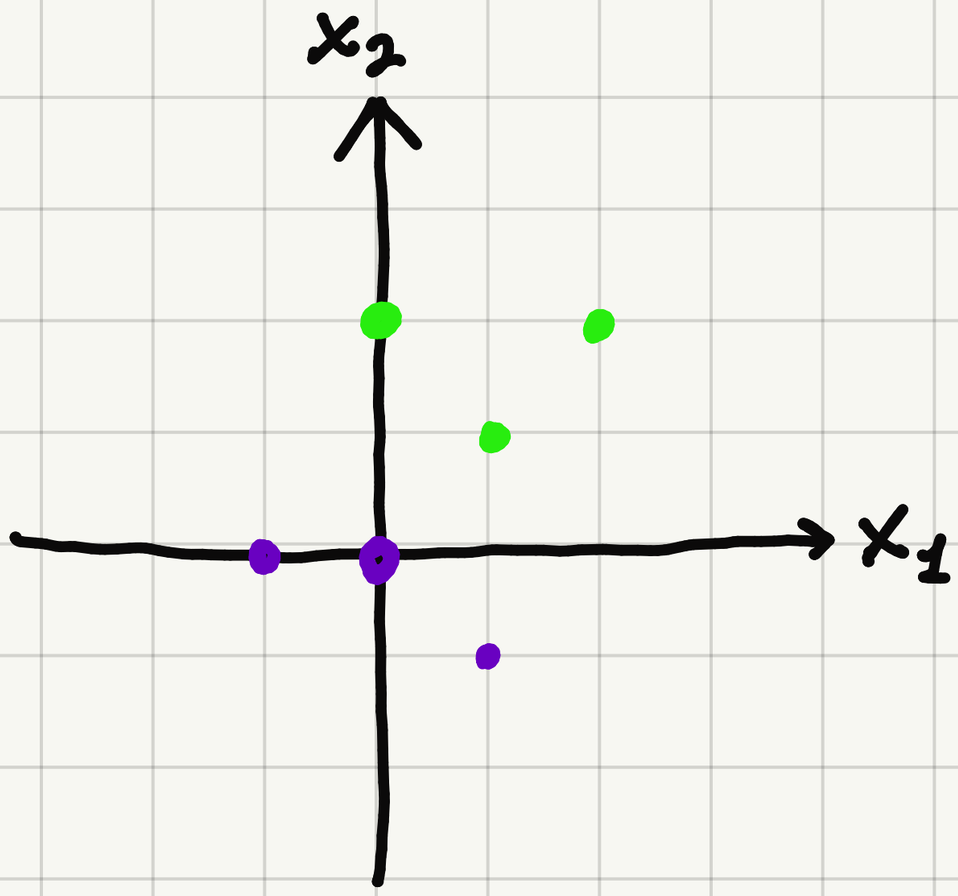
$\Rightarrow$ $k(x, y)$ IT IS LINEAR AND NON NEGATIVE SINCE IT CAN BE REWRITTEN AS AN INNER PRODUCT AS SHOWN BEFORE

MOREOVER, $k(x, y)$ CAN BE EXPRESSED AS A KERNEL FUNCTION IN A TRANSFORMED SPACE

THE THREE RULES ARE SATISFIED

$\Rightarrow$ $k(x, y) = (x^T x)(x^T y)(y^T y)$ IS A VALID KERNEL

PROBLEM 4

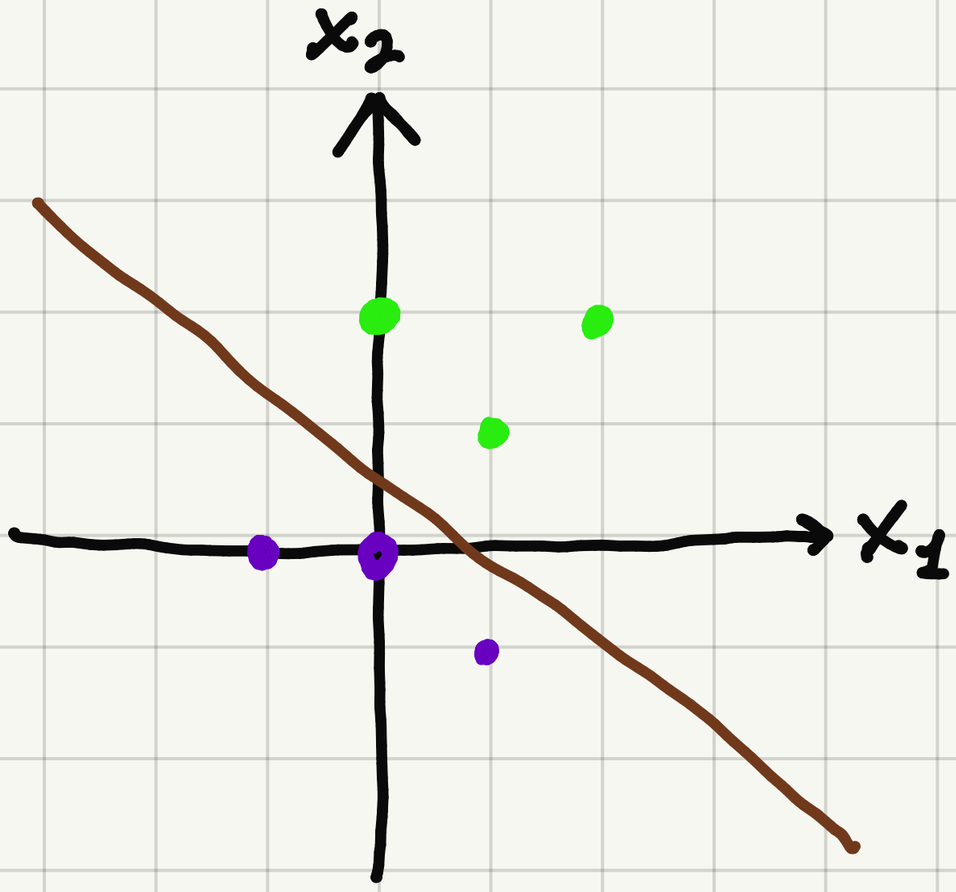


CLASS +

CLASS −

a) THE CLASSES SEEM TO BE LINEARLY SEPARABLE

b)



GIVEN THE WEIGHT VECTOR $w=\begin{vmatrix} w_1 \\ w_2 \end{vmatrix}$

AND THE BIAS b, THE HYPERPLANE

WILL BE OF THE FORM $w^T x+b=0$

$\rightarrow$ $w_1x_1+w_2x_2+b=0$

AND, GIVEN A GENERIC POINT x_i,

$$\begin{cases} w^T x_i + b \geq 1 & x_i \in \text{CLASS} + \\ w^T x_i + b \leq -1 & x_i \in \text{CLASS} - \end{cases}$$

WE CAN OBSERVE THAT $x_2 > 0 \ \forall x \in$ CLASS + AND

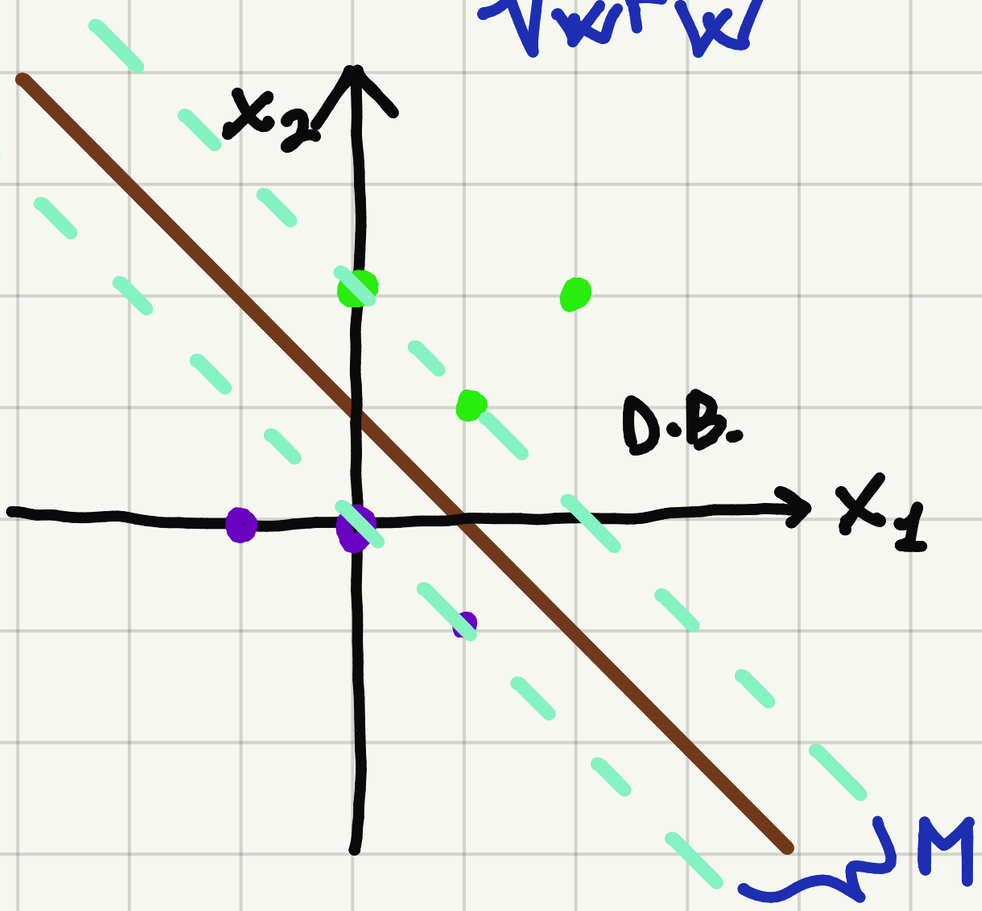
$x_2 \leq 0 \ \forall x \in$ CLASS − . SO, AS SVM, WE CAN CHOOSE

(1,1),(0−2) FOR CLASS + AND (0,0),(1,−1) FOR CLASS −

WE CAN SEE THAT $x_1+x_2-1=0$ IS ACTUALLY AN IDEAL SOLUTION

$$\Rightarrow w = \begin{vmatrix} 1 \\ 1 \end{vmatrix} \quad b = -1$$

MARGIN $M = \frac{2}{\sqrt{w^T w}} = \sqrt{2}$



$$d_{(1,1)_{D.B.}} = \frac{|1+1-1|}{\sqrt{1^2+1^2}} = \frac{\sqrt{2}}{2}$$

$$d_{(0,0)_{D.B.}} = \frac{|0+0-1|}{\sqrt{1^2+1^2}} = \frac{\sqrt{2}}{2}$$

$$d_{(-1,0)_{D.B.}} = \frac{|-1+0-1|}{\sqrt{1^2+1^2}} = \sqrt{2}$$

$$d_{(2,2)_{D.B.}} = \frac{|2+2-1|}{\sqrt{1^2+1^2}} = \frac{3\sqrt{2}}{2}$$

$$d_{(0,2)_{D.B.}} = \frac{|0+2-1|}{\sqrt{1^2+1^2}} = \frac{\sqrt{2}}{2}$$

$$d_{(1,-1)_{D.B.}} = \frac{|1-1-1|}{\sqrt{1^2+1^2}} = \frac{\sqrt{2}}{2}$$

$$M = (1,1) - (0,0) = \sqrt{(1-0)^2 + (1-0)^2} = \sqrt{2}\ \checkmark$$

c) SINCE WE HAVE FOUR SVM IN THIS CONTEXT, REMOVING ONE DO NOT AFFECT THE OPTIMAL MARGIN

d) THIS IS NOT TRUE IN GENERAL. IN FACT, IF WE GET JUST ONE SVM, REMOVING IT WILL FORCE TO RECOMPUTE ALL THE DISTANCES AND FIND THE NEW NEAREST POINT TO THE HYPERPLANE