UNIT 1 : Linear Classifier and Generalizations

LECTURE 3 : Hinge loss, Margin boundaries and Regularization

Wooooooooo !!!!!!! We are back baby !!!!! Yeahhhhhh !!!!!

After the successful completion of Fundamentals of Statistics , your pal **Scooby** is back with a brand new set of handwritten lecture notes. I found that many of you liked them and appriciated the work by thanking me personally on my email. Which by the way is :

Scooby's email : **peepeepoopoo873@gmail.com**

Feel free to point out mistakes and message me anyways

For those who do not message me...
I don't care man!!! Whataever man !!!! Screw you guys !!!! Whatever!!!

(I do care, please message me. I was just pretending to be indifferent in the above line so that you can't hurt me emotionally.) Anyways.

I tried to make these notes without any mistakes and cover most of the video material.

But as Yoda has rightly said : ' *Is human, To err* '

I however **DO NOT TAKE** any responsibilty if you fail your exams following my notes. So please check the actual lecture slides and videos if you suspect something is wrong, and mail me too.

The notes **DO NOT CONTAIN** solutions to any homework problems or in-lecture exercises. Only the problems and exercises solved in the videos are included in these notes.

Here is an interesting photo I googled

CHAD KANGAROO

I ran out of made up quotes so I will write very short stories instead. These may or may not resemble true events, but will definitely have a message that will help you along the journey, that is life.

STORY TIME : A WILD LOVE STORY

Long time ago there was a lion who lived in a jungle. He was lonely. One day while hunting, he saw this beautiful deer. He approached her slowly in order to attack her. The deer figured out that she was about to be hunted, so she started running.

And so the chase began. Deer infront, lion behind her. After closing in on the distance, the lion leapt for his final attack. As he jumped the deer looked behind and the lion saw the deer's beautiful green eyes. The lion lost his balance somehow. The deer escaped.

At night when the lion, was sleeping in his bed gazing at the stars, he couldn't stop thinking about the deer. He was in love. He knew it. But he was ashamed of it. How can the king of jungle fall in love with a lowly deer, he thought to himself. But the heart wants what it wants.

So the lion started stalking the deer in the bushes for days. Finally he confessed his love. The deer was the jungle equivalent of a gold digger so they got married. The initial days of their marriage were preety good but they started having marriage troubles later.

The deer used to say the lion was lazy and wouldn't go to jungle broadway with her.
They argued about whether the babies would be raised with deer values or lion values.
For some the deer's family did not like the lion (I wonder why.) and the deer did not miss a chance to bring this fact up in every other conversation.
The deer started withholding physical intimacy to get her way. However hard the lion pretended to be civil, he was after all a beast inside.
They even went to the elephant marriage counselor but nothing worked.

The final nail in the coffin was struck one day. When at the dinner table, the dear being a vegan started arguing about how eating other animals was murder, and how th lion should go vegan, for a better life. She also made a soy dish that night and forced the lion to try it. She claimed that it was just like real meat.

The lion had had enough. The lion took out his claws and ripped apart his wifey, the deer. He had a belly full of tasty deer meat that night, and went on deer hunting the next day as usual.

THE END

MORAL : If your are a vegan, don't try to shame or convince other people to go vegan. If you do that, a lion may eat you some day.

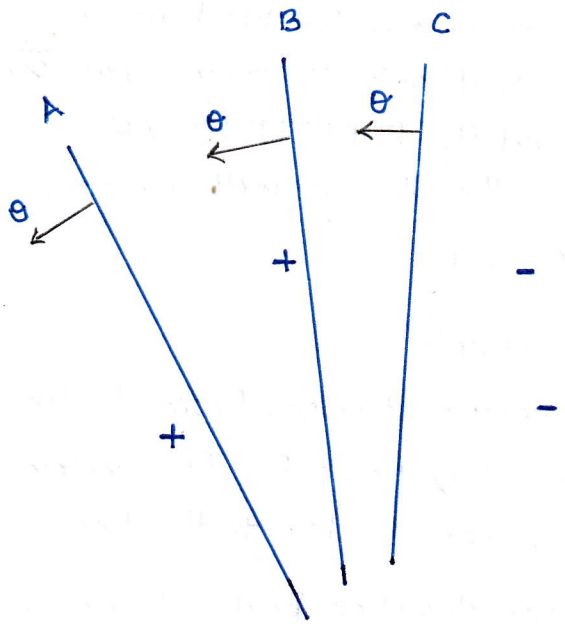
I hope these notes will help you

Let's get started

▷ Linear Classification

In the previous lectures, we discussed the idea of finding a hyperplane (line in 2d) that can classify our points. We also are equipped with an algorithm called the Perceptron, that will find a solution if the test set S_n is linearly separable.

However we never talked about how good the solution is. For example look at the following figure



Three choices are plotted for a training set. As we can clearly see that the set is linearly separable. So we don't want something like decision boundary A to be our choice because clearly we have better classifiers, like B and C.

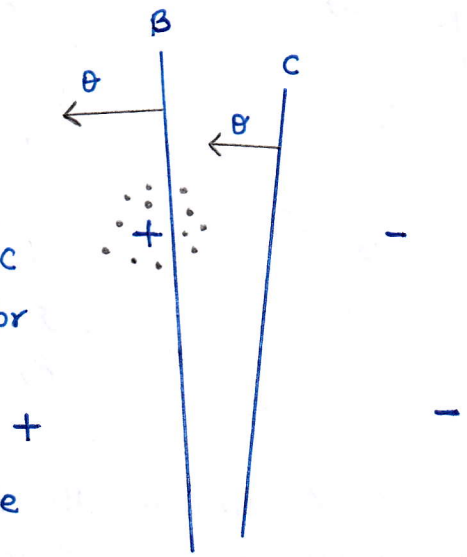
Even if we run our perceptron here, we will find a line like B or C or any other line that seperates <u>ALL</u> points correctly.

So A is out of the question, Let's look at choice B and C

Now somehow intuitively we feel that Choice C is a better choice than Choice B.

And the reason is simple. The choice B is just grazing a point labeled +1

Ie a point is very close to the decision boundary B, while all the points seem to be at a moderate distance from Choice C



We are doing all this calculation in order for us to classify new points. Suppose some new cases from the test set are such that they are located in the vicinity of this +1 label point, that is very close to the decision boundary.

These points must clearly have label +1 because, they are very close to a feature vector with +1 label. (This is not a robust claim, just a logical assumption), But we see that our choice B will misclassify many such test cases.

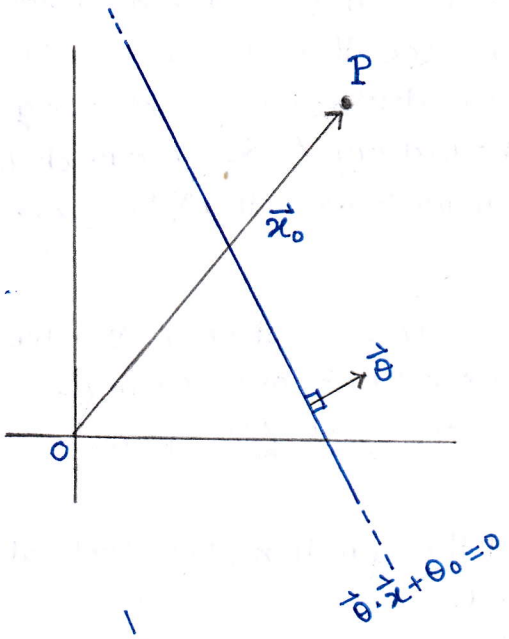
However Choice C is still far far away from these points, and therefore it will properly classify them

So we need to understand how these delicacies play out in mathematical terms, but before that we need ....

MATHEMATICAL INTERLUDE : DISTANCE BETWEEN LINE AND POINT

Throught this Lecture, we will use the concept of distance, in order to formulate our problem. Lets see how the distance between a line and a point is calculated. Although we are finding the distance between a line and a point, the procedure we will use and the result we obtained is the general form for distance bw a point in $\mathbb{R}^d$ and a hyperplane in $\mathbb{R}^d$ given by $\theta \cdot x + \theta_0 = 0$

Lets start.



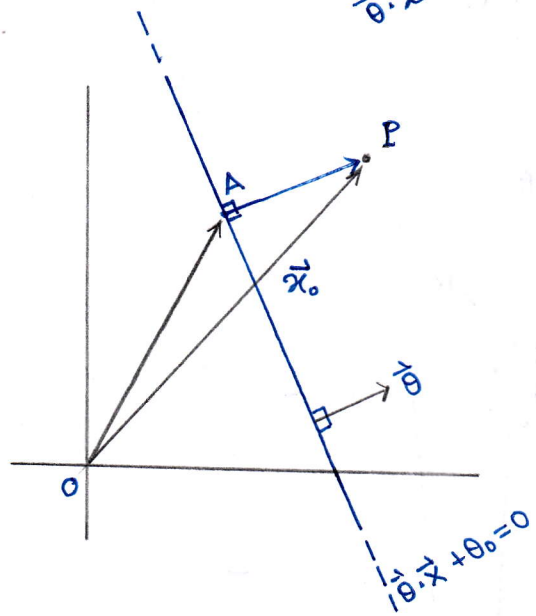
So the problem is simple. We have a point P and we denote it vectorily as $\vec{x}_0$ we wish to find its PERPENDICULAR distance from the the line with equation

$$\vec{\theta} \cdot \vec{x} + \theta_0 = 0$$

our plan is simple

> we draw a vector starting from the line at A and meeting point P. This vector $\vec{AP}$ will be perpendicular to the line

> There's only one direction perpendicular to the line. That is $\vec{\theta}$'s direction.
Therefore we can say that $\vec{AP}$ is a multiple of $\vec{\theta}$ ie $\vec{AP} = \lambda\vec{\theta}$ where $\lambda \in \mathbb{R}$



> Now by vector addition
$\vec{OA} + \vec{AP} = \vec{OP}$
$\vec{OP}$ is $\vec{x}_0$ by definition
& $\vec{AP} = \lambda\vec{\theta}$ decided above

> So we have
$\vec{OA} = \vec{x}_0 - \lambda\vec{\theta}$

> But the point A lies on the line $\vec{\theta}\cdot\vec{x} + \theta_0 = 0$ so $\vec{OA}$ vector should satisfy the equation ie : $\vec{\theta} \cdot \vec{OA} + \theta_0 = 0$

> This gives us $\vec{\theta} \cdot (\vec{x}_0 - \lambda\vec{\theta}) + \theta_0 = 0 \Rightarrow \lambda = \dfrac{\vec{\theta}\cdot\vec{x}_0 + \theta_0}{\|\vec{\theta}\|^2}$

> finally we can get the distance bw P and line by finding the length of vector $\vec{AP}$

$$\|\vec{AP}\| = \|\lambda\vec{\theta}\| = |\lambda|\|\vec{\theta}\| = \frac{|\vec{\theta}\cdot\vec{x}_0 + \theta_0|}{\|\vec{\theta}\|}$$

So finally we have

Distance between point P with vector $\vec{x}_0$ and the line $\vec{\theta} \cdot \vec{x} + \theta_0 = 0$

$$= \frac{|\vec{\theta} \cdot \vec{x}_0 + \theta_0|}{\|\vec{\theta}\|}$$

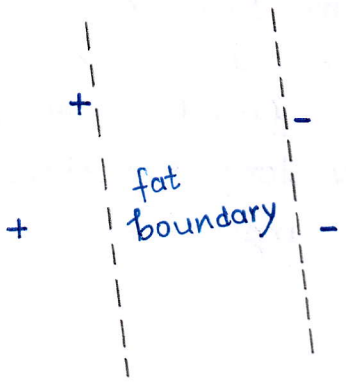
Note that during the derivation, we never mentioned ~~d~~ the dimension we are in. All equations we used were in vector form. So this is a formula in general for perpendicular distance between a point in $\mathbb{R}^d$ and hyperplane: $\theta \cdot x + \theta_0 = 0$

Now that this is out of the way we are ready to tackle our problem

▷ Learning as Optimization

Now we had the problem where we had to find a decision boundary given the training data. We will now see this problem with a different perspective. We will find $\vec{\theta}$ and θ_0 that will minimise something. This is a common practise in Maths and we will do that. But first we will set the problem up

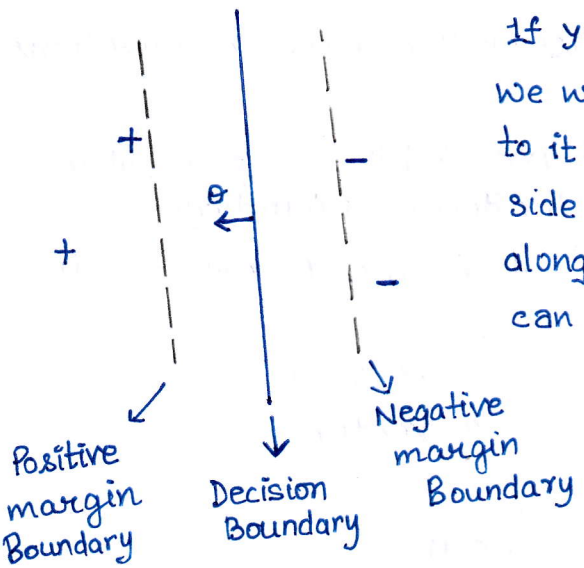
We have the training points below, and instead of fitting/finding just one decision boundary, we will fit a broad thick boundary in between the points as shown.



The right line (or hyperplane) that's close to −1 labelled points is called negative margin boundary

The left line (or hyperplane) that's close to +1 labelled points is called the positive margin boundary

The idea is simple: The actual decision boundary will lie right in between the left and the right margin boundary. Note that all these three boundaries are parallel to each other

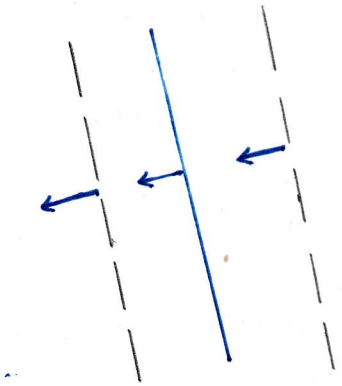


If you wish to see it another way heres the idea

We will have our decision boundary, and parallel to it will be two more boundaries, one on each side. We will somehow stretch these two boundaries along $\vec{\theta}$ direction and see how our problem's solution can be found in an optimal way.

▷ Linear Classification : Margin

Lets now try to write the equation of these margin boundaries. First we must note that these are just lines (hyperplanes in general) that are parallel to our decision boundary $\theta \cdot x + \theta_0 = 0$. So the normal vector for all boundries, decision (one) and margin (two) will be in same direction



So we can with confidence say that the equation of margin boundaries will be of the form

$$\theta \cdot x + c = 0$$

c is some real value and will differ for positive and negative margin boundaries

Now for a given θ vector and θ_0, that is for a given decision boundary we want two lines (hyperplanes generally), that are parallel to the decision boundary. Once we imposed (by construction) this parallality condition it's fixed that the equations of margins will be of form $\theta \cdot x + c = 0$

Now note that we only have one more constraint left for these margin boundaries and that is c. Once the c is fixed they are fixed, and they can't move towards or away from decision boundary along θ vector.

And now out of nowhere we claim that the equation of margins are :

positive margin classifier : $\theta \cdot x + \theta_0 = 1$

negative margin classifier : $\theta \cdot x + \theta_0 = -1$

WHAT???

How can we say that, what about the idea of moving these margins along θ vectors direction in order to find a good margin classifier. It seems that we have fixed the margins once we assign the above two equations to the two margins respectively.

But before we rush to judgement lets just try to find the perpendicular distance between any of the margin boundary and the decision boundary.

we already know the distance of point with x_0 from line $\theta \cdot x + \theta_0 = 0$ is

$$\frac{|\theta \cdot x_0 + \theta_0|}{\|\theta\|}$$

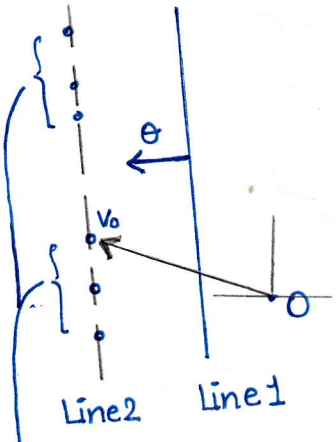
here x_0 is a vector

we can use this trick to find distance between

line 1 : $\theta \cdot x + \theta_0 = 0$ and line 2 $\theta \cdot x + \theta_0 = 1$

The ideais as following

Finding the distance bw line 1 and line 2 is equivalent to finding the perperdicular distance between ANY point on line 2, and line 1. We can take any point on line 2 because, line 1 and line 2 are parallel so distance bw any point on line 2, to that of line 1 will be distance between the two lines.



Any point can be taken

All have equal distance

So lets take a general point on line 2 with vector V_0

The distance between V_0 (Point) and line 1 will be

$$\frac{|\theta \cdot V_0 + \theta_0|}{\|\theta\|}$$

Now vector V_0 lies on Line 2 so it must satisfy

$$\theta \cdot V_0 + \theta_0 = 1$$

Substituting we get the distance bw line 1 and line 2 that is

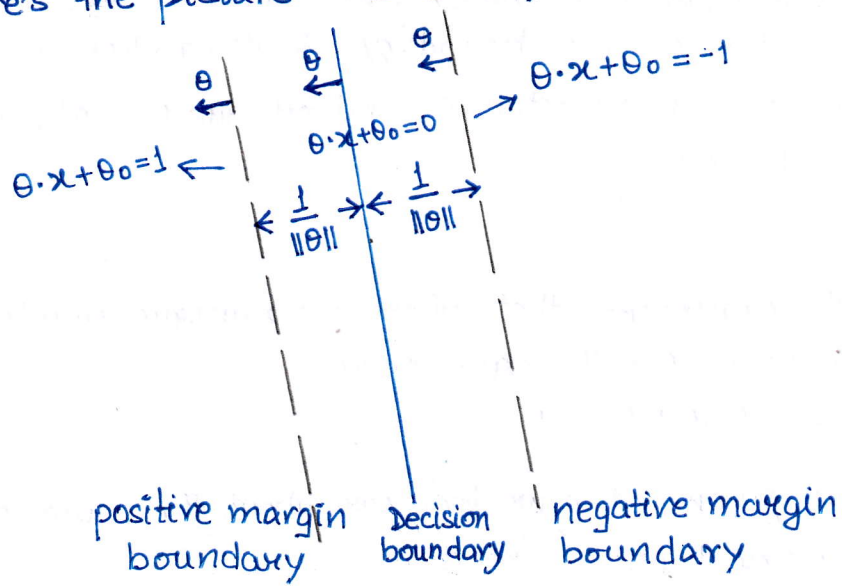
$$\frac{1}{\|\theta\|}$$

If we wanted to find the distance bw $\theta \cdot x + \theta_0 = -1$ and Line 1

It will similiarly be $\frac{|\theta \cdot V_0 + \theta_0|}{\|\theta\|} = \frac{|-1|}{\|\theta\|} = \frac{1}{\|\theta\|}$

which means both margins are equidistant from the decision boundaries & the distance is : $\frac{1}{\|\theta\|}$: And that's what we wanted. GOOD!!!

So here's the picture we finally have :

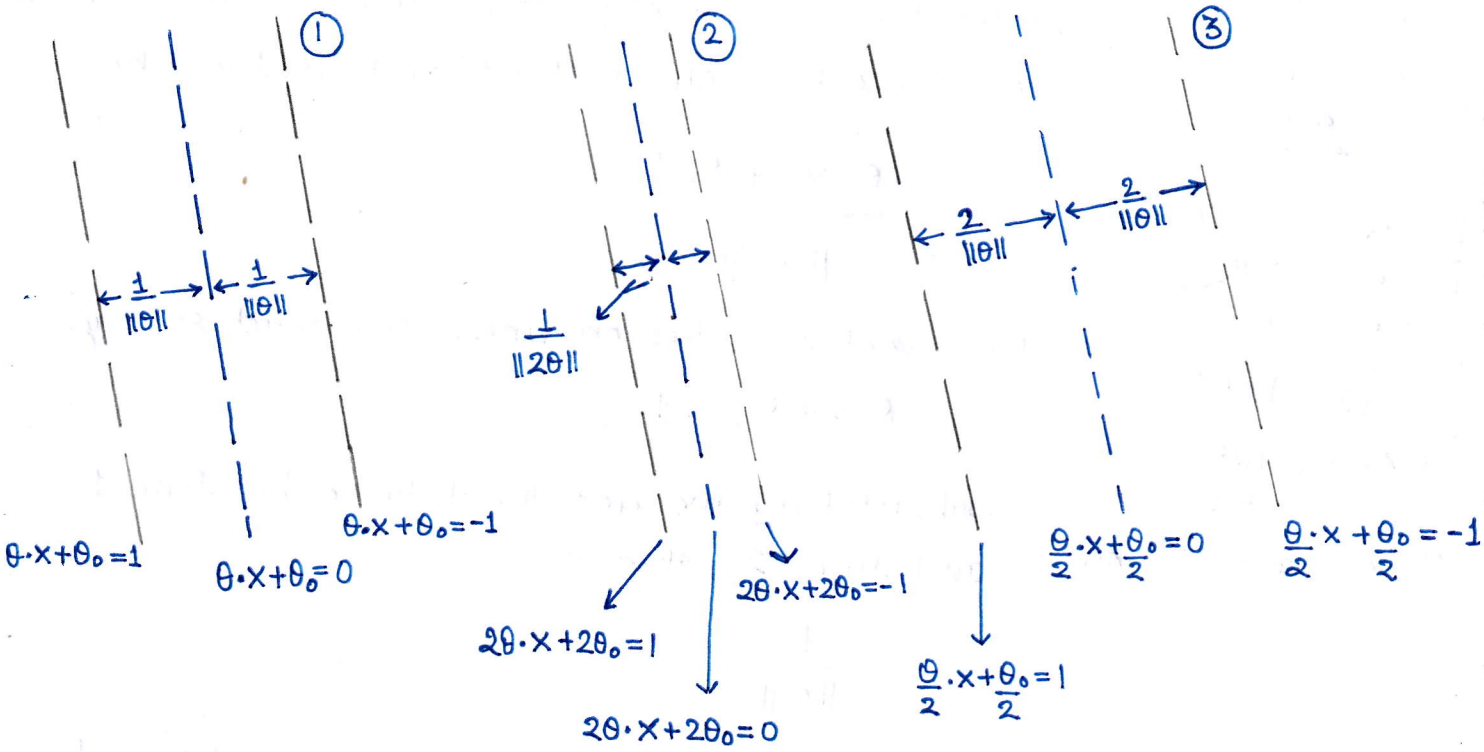


> AaaHah !!!! Here's the solution to our problem

We did nothing wrong. The choice to take the equations

$$\theta \cdot x + \theta_0 = 1 \quad \text{and} \quad \theta \cdot x + \theta_0 = -1$$

still allows us to move the margin boundaries, because we haven't fixed $\|\theta\|$ yet. This is the extra degree of freedom, we talked about in Lecture 2 earlier. For eg look at the cases below :-



In all the above three cases the equation of decision boundary is the same

$$\theta \cdot x + \theta_0 = 2\theta \cdot x + 2\theta_0 = \frac{\theta}{2} \cdot x + \frac{\theta_0}{2} = 0$$

So the direction of θ vector remains unchanged. However the distance between the margin line and decision line changes when θ is multiplied by some factor

Now we can look back about what we did earlier. We said that say, the positive margin boundary is parallel to the decision boundary. So its equation is of form

$$\theta \cdot x + c = 0$$

I can take a new constant c', and withouth loss of generality write

$$\theta \cdot x + c'(\theta_0 - 1) = 0$$

: I can divide by c'

$$\frac{\theta}{c'} \cdot x + \theta_0 - 1 = 0$$

And we would again given the argument that there is no unique equation for this plane. we may just as well write the equation as

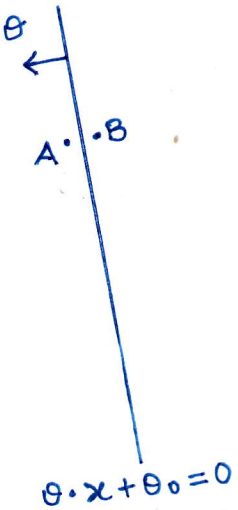
$$\theta \cdot x + \theta_0 - 1 = 0 \Rightarrow \theta \cdot x + \theta_0 = 1$$

But this argument can only be given when we don't care about the norm of the normal vector θ, which is our extra degree of freedom

▷Linear Classification : margin

Now its time to develop the idea properly. Let me provide an intuitive feel of what we are doing. Read carefully.

we had this problem of seperating points with two Labels +1 and −1. So we proposed that we can for the simplest case divide the region into two parts using a line which we called the decision boundary. The idea was that points to one side of the boundary will be labelled +1, and to the other side −1.



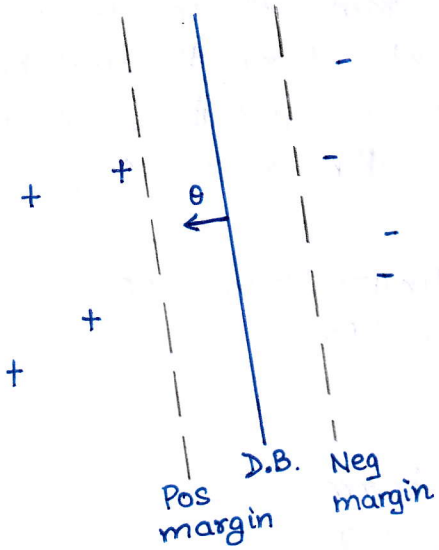
Now look at the adjoining figure. The idea in itself is not so bad, but one might say that it's absurd for points that are very close to the boundary.

So a point just to the left will be labelled +1
a point just to the right will be Labelled −1

That's not robust at all, just a small deviation of point near the boundary, and the label changes abruptly

So someone suggested that their must be some <u>margin</u> for a boundary like this, within which points can be properly labelled.

Therefore came the concept of margin boundaries. We will not only fit a single decision boundary, but two margin boundaries with this decision boundary. It would be like fitting a thick decision boundary. We will also try to fit the margin boundaries such that no or minimum points lie inside this thick boundary so that we can with confidence say that the points to the right of neg. margin boundary should be labelled −1 and the points to the left of pos. margin boundary should be labelled +1



The plan was to stretch these margins along θ's direction (negative & positive), in order to make the width (b/w neg & pos. margin) as wide as possible.

This means we want to maximise the distance between the positive/negative margin & the decision boundary

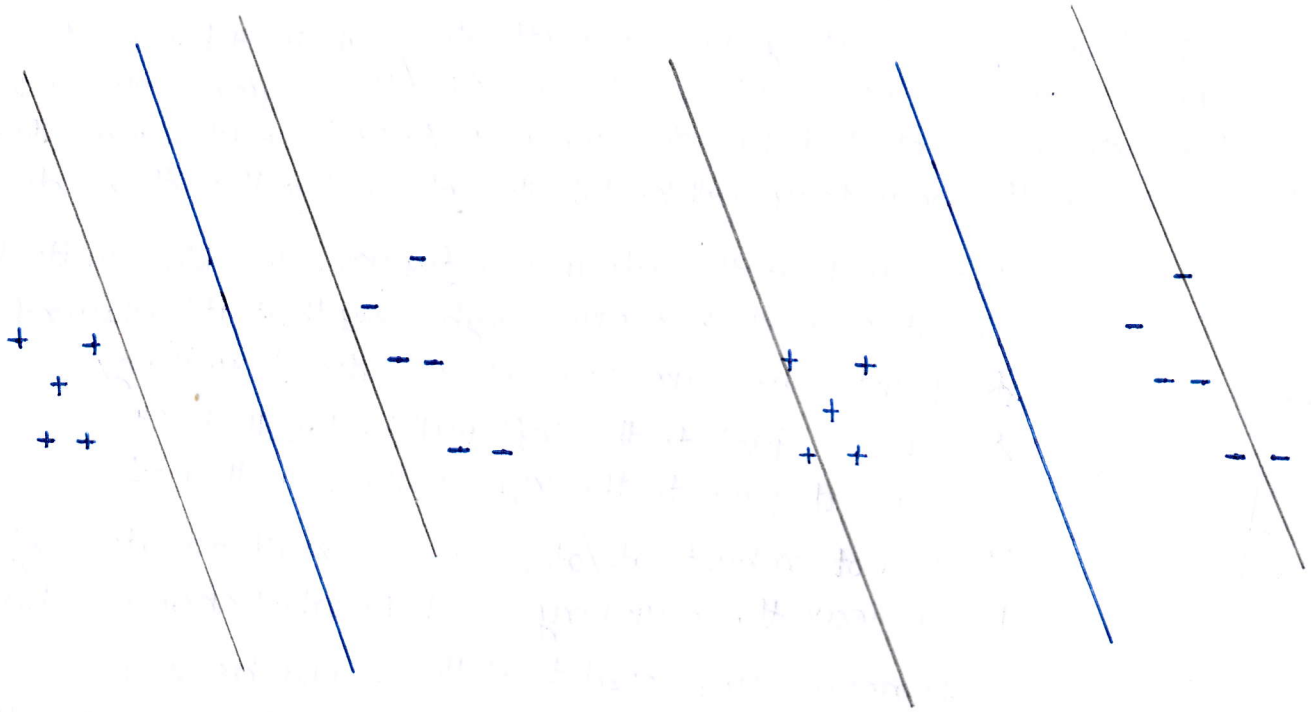
ie maximise $\frac{1}{\|\theta\|}$

or minimise $\|\theta\|$

Well if that's the thing you wan't to do here's my Brilliant plan

<u>SCOOBY'S PLAN</u> : If you want to maximise the distance b/w boundaries, why don't we just ~~it~~ one boundary to $+\infty$ and other to $-\infty$. The distance between them will be $\frac{1}{2 \cdot \infty} = 0$, the minimum possible

That's an absolutely trash idea. I can't believe a genius like me came up with it. And here's the reason why. Look below at the figures



As the distance between margin will increase. More and more points will start to come inside the fat margin. That would defeat the whole purpose of having a fat margin. We may allow some points inside but not a lot.

So this brings us to a dilemma. We want to maximise the distance between the two margins, but doing this starts taking in points inside the fat boundary

This is the issue of Regularization vs Loss

> Regularization vs Loss

Its not sufficient to tackle the problem of finding these margin boundaries, as by minimizing $\|\theta\|$. This is the regularization term that will aim for a boundary of maximum width $1/\|\theta\|$. We see a counter force that will be emerging as the margins widen. This will graphically manifest itself as more and more points falling inside the fat boundary. It will be manifested in mathematical terms, in form of a loss function.

So we will add these two functions and try to optimize / minimise that overall function in order to find optimal θ and θ_0

we will call this function the objective function

Objective function = Some terms containing Loss PART + Some terms containing regularization PART

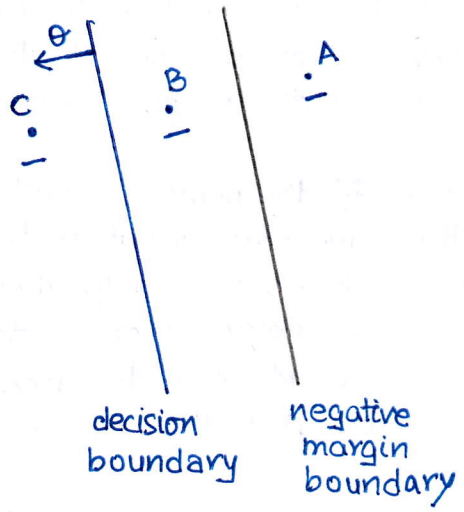
The Regularization part is a no brainer. We wanted to maximize the distance bw the boundaries, which is equivalent to maximizing $\|\theta\|$. So the term should be proportional to $\|\theta\|$ in the objective function. (or something of this form)

The loss function is what we haven't talked about till now. And I promise you, by the time you reach the end of this line, we will be talking about it.

> LOSS FUNCTION

So the idea is pretty simple. We want to penalize points. They have been naughty and it's time for them to learn who their daddy is.

To each point we will assign a LOSS. Look at it this way. Suppose we have three points with Label -1 as shown, ie points A, B, C



A: is -1 labelled point that is far away from the decision boundary. Its even to the right of negative margin boundary. There is no need for us to penalize A as its far away in the right direction. It can be labelled -1 with a lot of confidence

B: Now this point is <u>NOT</u> misclassified. It's still to the right of the decision boundary. But it's not to the right of negative margin boundary So a point between the margin and decision boundary should be penalised. Also its apparent that the more close B gets to the decision boundary, the more problem we will have classifying it correctly. So the penalization should increase as the point gets close to the decision boundary.

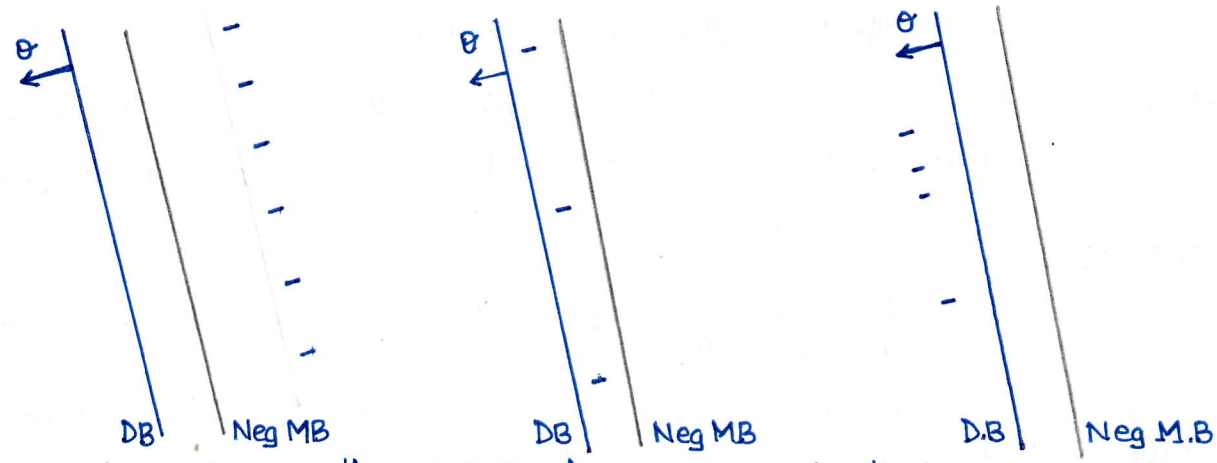
C: Now this is out of control. Not only is point C inside the margin boundary, its also inside (to the left of) the decision boundary. It is a misclassified point (by our decision boundary). One might say, why don't we adjust the three boundaries such that it's (the point) is correctly classified? well we could have done that if the data we have is Linearly seperable, but that's not always the case. So we will have to account for losses from these points too. We will have to penalize them even more strongly than the previous two points. This whole analysis can be done with +1 labelled points & positive margin boundaries too. Now.

We have been talking about penalizing points. What does punishing the points mean? As dirty as your brain is, it does not mean, what you think it means.

What we are trying to do here is simple. We for a given θ and θ_0, will go to each of the n points and see how each point contributes to an error or loss. Just like above, we will assign each point a loss, and add these up. This will them form a part of our objective function earlier. And that objective function will be minimized wrt θ and θ_0. Thus finding proper boundaries

How do we write this loss function mathematically. Lets see.

First we must note what's important and what's not



Here I have drawn three cases when, the points that we want to penalize are equidistant from the decision and margin boundary. Which is to say that they are at equal PERPENDICULAR distance from the boundary. There placement along the direction parallel to the boundaries may vary, but the perpendicular distance will be the same.

But we are interested in classification, so why do we care if the point is anywhere along the parallel direction to the boundaries. The only thing we care about is how close or far from the decision boundary is the point. ie we care only about the PERPENDICULAR distance of a point from the decision boundary, in order to decide how much to penalize it. Points with equal perpendicular distance from decision boundary will be penalized equally, no matter what their position is along the direction parallel to the line.

We already know the expression for a point's distance from a line (hyperplane)

The distance of a point with vector $x^{(i)}$ from line (hyperplane) : $\theta \cdot x + \theta_0 = 0$ is

$$\frac{|\theta \cdot x^{(i)} + \theta_0|}{\|\theta\|}$$

But this is the notion of distance, we want a thing that will tell us the classification of the point too. We in lec 2 saw that this quantity

$y^{(i)}(\theta \cdot x^{(i)} + \theta_0) > 0$ when point is correctly classified

$y^{(i)}(\theta \cdot x^{(i)} + \theta_0) < 0$ when point is misclassified

$y^{(i)}(\theta \cdot x^{(i)} + \theta_0) = 0$ when point lies on the decision boundary

We we introduce the notion of signed distance

$$\gamma_i(\theta, \theta_0) = \frac{y^{(i)}(\theta \cdot x^{(i)} + \theta_0)}{\|\theta\|}$$

Now this is a quantity that will tell us not only the perpendicular distance of the ith point from the line, but also whether the point is misclassified or not. Note that

$$|\gamma_i(\theta,\theta_0)| = \frac{|y^{(i)}(\theta\cdot x^{(i)}+\theta_0)|}{\|\theta\|} = \frac{|y^{(i)}|\,|\theta\cdot x^{(i)}+\theta_0|}{\|\theta\|}$$

Now $y^{(i)} = 1$ or -1 so $|y^{(i)}| = 1$

Therefore $|\gamma_i(\theta,\theta_0)|$ = distance (orthogonal) of the ith point from the line

But when $\gamma_i(\theta,\theta_0) > 0$, we know we are talking about a correctly classified point

$\gamma_i(\theta,\theta_0) \leq 0$ we know we are talking about a misclassified (error included at 0) point

With this being defined, we will define another quantity to help our intuition.

> Agreement

Agreement is defined as : $y^{(i)}(\theta\cdot x^{(i)}+\theta_0)$

What the HELL!!! {I can't use the F word, because little kids might be reading this lecture}

This is not a new thing, we already studied this earlier. But if we look closely enough, we find that this quantity has an alltogether new meaning, in the light of what we did in this lecture. Let me elaborate.

First lets recall that the distance bw positive or negative margin boundary to that of the decision boundary is $\frac{1}{\|\theta\|}$. We derived that earlier

Now look at the signed distance

$$\gamma_i(\theta,\theta_0) = \frac{y^{(i)}(\theta\cdot x^{(i)}+\theta_0)}{\|\theta\|} = \frac{\text{Agreement}}{\|\theta\|}$$

So Basically

Agreement = $\|\theta\|$ * Signed distance

we also know distance bw a margin boundary & decision boundary : $1/\|\theta\|$

So Agreement $= \dfrac{\text{Signed distance}}{1/\|\theta\|}$

$= \dfrac{\text{Signed distance}}{\text{dist bw M.B \& DB}}$

M.B : Margin boundary
D.B : Decision boundary

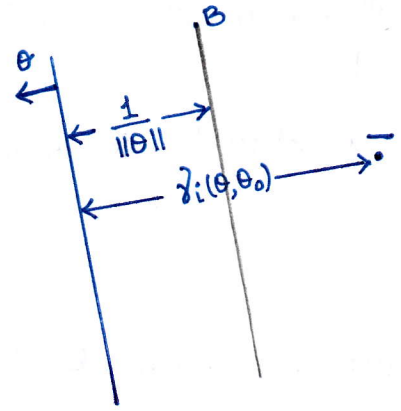
So agreement is the ratio between the two distances. Lets denote agreement by z : $y^{(i)}(\theta \cdot x^{(i)} + \theta_0)$

This means we have a way of seeing, how the ith point is placed wrt the margin. we can say that the ith point is

z times away from the decision boundary compared to the margin bound.

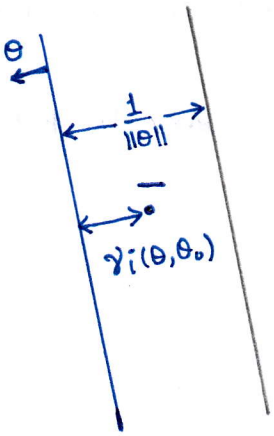
because distance of point i from the decision boundary $= z *$ distance of margin boundary from the decision boundary

Let me illustrate with some cases, Here we have a fixed θ and θ_0 for all the cases & we wanna see what agreement means.



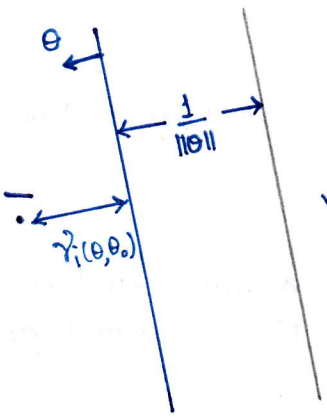
Here a -1 labelled point is to right side of decision boundary so $\gamma_i(\theta,\theta_0)$ will be positive. We can also see that it's well beyond negative margin boundary therefore $\gamma_i(\theta,\theta_0) > \frac{1}{\|\theta\|}$ for this case

Therefore the agreement will be much larger than 1 ie we strongly agree when point is to the correct side & well beyond margin boundary



Here the -1 labelled point lies in between the decision boundary and negative margin boundary. The point is still classified correctly. so $\gamma_i(\theta,\theta_0) > 0$, but we clearly see that $\gamma_i(\theta,\theta_0) < \frac{1}{\|\theta\|}$ for this case

Therefore the agreement will be smaller than 1, but larger than 0 for this case. so when the point is in between the margin (negative for -1, positive for +1), we will somewhat agree and the agreement z will be between 0 and 1



Here the -1 labelled point lies beyond the decision boundary. 1e it's missclassified. So $\gamma_i(\theta,\theta_0) < 0$. So the agreement z will also be negative. I don't know about you people, but I studied in my grammer class that a negative agreement is a disagreement. So the agreement being negative tells us that the point is on the wrong side, we disagree.
ie as point goes more and more away from the decision boundary in the wrong direction, our disagreement will become more and more strong.

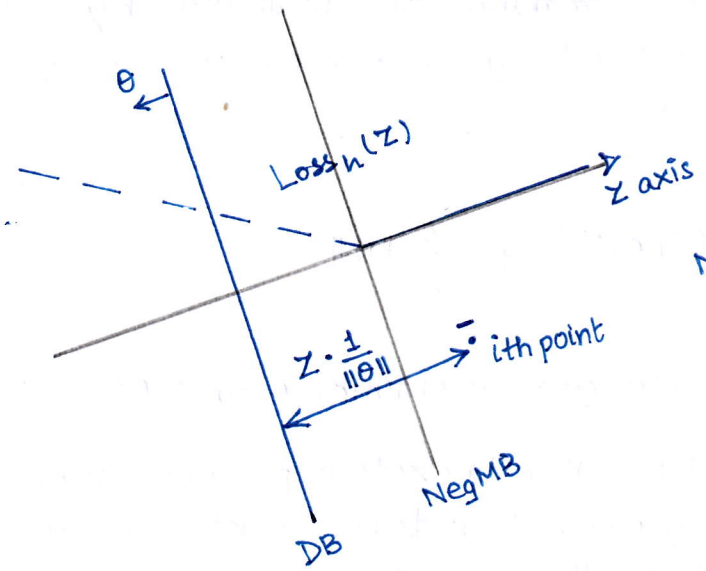
Ie agreement will go from 0 to $-\infty$ as point goes to the left.

This all can be done with a positive boundary & +1 labelled points too.

> Hinge Loss

Once we have defined the agreement z, we will penalize our points on basis of this z. We will chooze a function, that will not penalize points that are well beyond their respective boundaries (pos for +1 labelled, neg for -1 labelled) and as the points start coming closer to the decision boundary the penalizing will increase more and more. For this we choose a simple linear function

$$\text{Loss}_h(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)) \equiv \text{Loss}_h(z) = \begin{cases} 0 & \text{if } z \geq 1 \\ 1 - z & \text{if } z < 1 \end{cases}$$



No points penalized beyond margin boundary
Other points linearly penalized
Note at decision boundary. Loss is 1

Door Hinge

With the help of this function we can account for the loss each point will contribute to the optimization function. We will try to (partially) minimize this loss on average ie $\frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h(y^{(i)}(\theta \cdot x^{(i)} + \theta_0))$ & this will form one part of our objective function. The other part will be.

> Regularization term

We have already talked about this earlier, but once again we want to maximize the width of boundaries

$$\text{maximise } \frac{1}{\|\theta\|} \overset{\text{eq.}}{\Longleftrightarrow} \text{minimise } \|\theta\| \overset{\text{eq.}}{\Longleftrightarrow} \text{minimise } \|\theta\|^2$$

For some reason (explained later) we will add a $\frac{1}{2}$ factor to

$$\text{minimise } \frac{\|\theta\|^2}{2}$$ & this is the regularization term

Lets now finally make our objective function

▷ Large margin as Optimization

So we want to minimise the average loss : $\frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h(z_i)$

where $z_i = y^{(i)}(\theta \cdot x^{(i)} + \theta_0)$

And we want to minimise $\frac{\|\theta\|^2}{2}$

But both are counter to one another, making one small will by construction make other large. So instead we try to optimise a combination of the two which we call objective function

$$J(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)) + \frac{\lambda}{2}\|\theta\|^2$$

and minimising the above function will give us optimal θ and θ_0

We here introduced a term λ called the regularization parameter or hyperparameter. This weighs the regularization terms contribution to the objective function in contrast to the loss term. We will see what λ does in the next lecture.

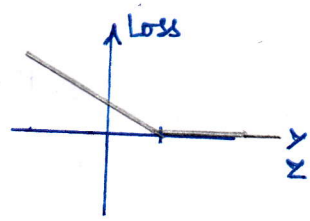
We will also see how the above objective function is minimised to find θ & θ_0 optimally.

Before we end this lecture, lets see how different points have different losses wrt the boundaries

▷ Calculating Losses: example

Let's recall the def'n of hinge loss

$$\text{Loss}_h(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)) \equiv \text{Loss}_h(z) = \begin{cases} 1 - z & \text{if } z < 1 \\ 0 & z \geq 1 \end{cases}$$

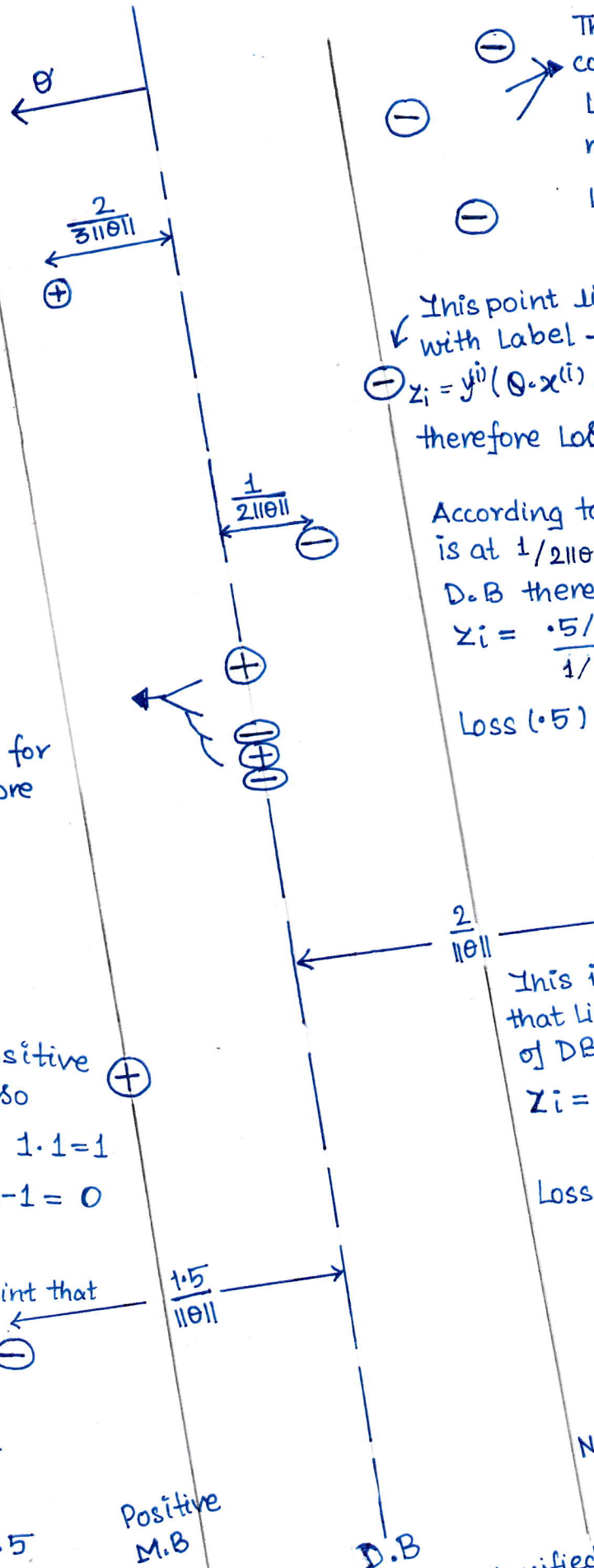


The point to note here is, as long as a point is beyond its respective margin boundary <u>and</u> correctly classified, the loss is zero. As z decreases the loss increases linearly

In the next figure we will see losses for different points placed at different positions for a given θ and θ_0

Remember that $z_i = \dfrac{\text{Distance distance of ith point from D.B}}{1/\|\theta\|}$

Let's do it



These points are correctly classified and lie beyond negative m, therefore Loss = 0 ∀ all of them

According to the graph this point is $\frac{2}{3\|\theta\|}$ away from the decision boundary but it's not misclassified

$z_i = \frac{2/3\|\theta\|}{1/\|\theta\|} = \frac{2}{3}$

Loss (2/3) = 1 − 2/3 = 1/3

This point lies on negative MB with Label −1.

$z_i = y^{(i)}(\theta \cdot x^{(i)} + \theta_0) = -1 \cdot -1 = +1$

therefore Loss(z_i) = 1 − 1 = 0

According to the graph this point is at $1/2\|\theta\|$ distance from the D.B therefore

$z_i = \frac{.5/\|\theta\|}{1/\|\theta\|} = .5$

Loss (.5) = 1 − .5 = .5

No matter the Label

$z_i = y^{(i)}(\theta \cdot x^{(i)} + \theta_0) = 0$

because $\theta \cdot x^{(i)} + \theta_0 = 0$ for all these points, therefore

Loss (0) = 1 − 0 = 1

for all points on D.B.

This is a +1 Labelled point that lies to the opposite side of DB. According to graph

$z_i = -\frac{(2/\|\theta\|)}{1/\|\theta\|} = -2$

Loss (−2) = 1 − −2 = 3

This point lies on positive M.B with Label +1 so

$z_i = (\theta \cdot x^{(i)} + \theta_0) \cdot y^{(i)} = 1 \cdot 1 = 1$

Loss (z_i) = Loss (1) = 1 − 1 = 0

This is a −1 labelled point that lies to the opposite side of the D.B. According to the graph

$z_i = -\frac{(1.5/\|\theta\|)}{1/\|\theta\|} = -1.5$

Loss (z_i) = 1 − −1.5 = 2.5

These points are classified correctly & the lie beyond positive MB. therefore $z > 1$ there fore Loss = 0 ∀ all of them

FIN.