### ML\_week3\_lecture3\_2-20240920

说话人1 00:00  
Here is a very important statements regarding realism elected within steps times how can decrease this in this function. Here is a simple proof. Let's signal there. We have the initial .. We use first order, previous time machine for the next point. And here we use first, here we use first order theatre thinner in spanish or see at this point as the initial point, sigma zero at sigma zero. That's why you'll see the one is equal to lcw and we take reading from l as the expand point is zero and 61, s zero. That's the first order approximation of this. At this .. This is exactly the same formula for the first order conditional complex situation. Wait for this one. We have a first order today expansion of everyone. You value at a particular apps. And here in not enough circumstances, why is it from single one? And it is perform series?

Here, according to this great reason step, sigma one minus theta zero is equal to negative alpha alpha sigma. And they all sigma zero. And we have this transpose. This is equal to elton's first, or number is 00. And since this is one equal to zero, and alpha is sufficient, small, why alpha should be sufficient small? Here, why are the sufficient schools? Because we're taking first order to the expansion, right? To make the first equation adequate enough, we need to be, which is small. We have l sigma one minus l sigma zero is equal to and alpha element squares of this winning is greater less than zero. And for sufficient smaller to make this, I could be known. So that's one. The negative readings. There is station small, early is guaranteed is this function? The mathematical to prove this, it expansion for contamination, then that is a function in really distance that the initial point matters.

Consider this figure. We have different initial points, the final convergence. It's a lot of so to prove to coordination, we run gd more on the different relation to the best need more work. It makes it fun. So learning it also matters if our eyes to be to improve more quickly with the this or the capital ever converse.

As the case is and if I was too small to use storage, so you may just patients and everybody here. Is it really was so very raise. I take about 100 fish. Unfortunately, the new mass would exist to find a perfect learning rate. So you have to test, it will come to the us this is the case in, especially in 20 euros, which is specified properly read on the same scale. We will be introduce the maybe that the computational cost for gdid situation is on over omm is a number. 20 samples rules. Is that why is that? Because we need to average the gradient over n 20 samples, right? It's prohibitive to scale up computers themselves. Then we introduce stochastic reading is an oscd for sure, which we use a combination that we need into the one.

At this situation, we rather example, a single, an example that it is reading about, basically theta. Instead, in the priest, in the previous case, instead of average, over over to an example of the environments at a little, in expectation is gd is just doing reasons, right? Expectation of ingredients or 1 group, for example, to the post estimation, we take average over the examples. The gradient is over 20,000 us, gradient reasons, aspirations. And here is some visualization of great reasons and stochastic reasons. I we have a smooth trajectory of reading this, and because the reading is average overall examples, we have various, exactly, tragically supposed to pass really decent. As you see, stochastic, decent is not guaranteed to dissent, your loss function. You may increasingly lost family and some particular decisions.

So I found it also has to be decent. It's easy to fit into memory. And computational mask is a simple example to that. In each situation, we don't need to wait to compute all agreements. The reason of all examples and adversary. The reason then we do not be must completely, but we know once we come to do not be the step has both oscillations, which means of the afternoon give some bad, no problem. And despite among the great stability, the objects are always decreasing. Each day I it takes a number of political powers. They use the amount of reservations I deal with only a single time at a time. As a compromise, we have mini batch really decent, which random sample is small subset of communication first would be, which is a subset of this. Is that b and t situation and compute the readings to our facility. Here is the it plus one relation is equal to the minus alpha. The ingredients divided at this particular subset in the match is equal to theta t minus alpha one over the calculator is the number of ten example. This is the match and the samples in the invention.

And now I feel xy divided as ft and immediate batch with reason and its random types. Bring the program extensive to the gene, different. And they're going to learn the balance of manifest with a decent as somebody introduce different relationships. This was somewhere we got this picture. How was the position? How many sets that comments function, commerce com in two conditions? One of the input formation, commerce set and they see what it should be satisfied. Methods and complexity, by definition, at first order commission, let's say, in order commission. And of the miracle, this matter, gg guarantee to decrease your loss. As long as you have a suspicious more memory, it's not scared of the process that was the average, the little over or examples you said. As gd sugar, fast computer is more degrees, have consistent costs. And in many banks, she needs a compromise, same within decent as the best within. It is a defective number used to be confirmed after those kinds of backgrounds on comments of position. But you can get to a very importance, immersion and concepts in terms of support their own peace.

Today, we've seen one examples of discriminate mean, and that's the last majestic regression, which is the maximum medical remote to another separate category.

So they're going to the second example as well, which is a purely not for this particularly dramatic approach, linear, acceptable data for law, assume the company data is linear and separable, which these are the two classes in 27 can be separate by a line or have a plane. For the case of paris data, it can be separated by the delight. The question is, which is the best separate demand. We have a lot of possible difference. Number of minds are accepted in two classes, so which make the best of it.

Now. So according to the principle of maximum margin, we define the distance between the 79 and those points as much, not here. So if this decision boundary, this is the first point also known as a margin. This is a margin. You can think of this space as an amount of great legal room for accommodating errors, investment in this paper. The idea is to separate the best of the nine is a one on mars measure margin. He puts the most distance between the first point of the decision. Here we choose for decision boundary and the corresponding marching lines, which one is the best among this form? According to the maximum market principle. So the one is a maximum investment, which is this one. And the following the follow up question time, is this the ultimate one we have a better submit line than this, yes or no? So you guys need really simple, this kind of high level question, which will make you easier to understand the alignments.

Also. Alignment. Is this the optimal certain ., according to the principle of maximum money? So this is not the optimum. We can move this up from a little bit to increase the margin until this margin touch this .. So by the percentage, there should be at least one margin point on each side of the party. Otherwise, we can move. The is it bound up a little bit to the side with no margin .? I we increase the margin, get a better. It's kind of green, according to the principle, not. Right? Therefore, the point on the margins and what is support vectors. So lots of the conditions support vectors. Support of efforts is the points nice of the market and the points. These are the ones for behind the market and we should have support papers on both sides of the market. Then the principle of maximize the margin, maximum margin.

And the first question is, we need to compute the market. So how to compute the market? So what is the market at the distance of this hyper plane after the transpose s must be with respect to a .? Is this one? Now let's prove it. Here we have this propane. And this is it's why where is this case? Why is positive? Then we can create a uniqueness that not good divided by expense of long. This is the unit that . in the center, right? I stop it. None of this to this distance, sb we move x towards the negative direction, this direction, the amount of d by doing.

So we can project this point onto this. Right? We move s to the left direction. They use this matter at the amount of dd is a business. These are marked.

So here, at least we protect the effects on this thing. And this . that's on the plane. I you must say this plan. This equation right here is four d it's w transpose s plus e minus e of the transpose. You might find the normal rather be, we look up the transpose that is equal to l known squared. L too long, square divide by l too long. This is equal to too long. And so will be they have. Now they transpose it to us. When I feel too long, we are not. Since this is why is equal to one, right? We can not the one. This is the way to compute the market. This is the market respect for one point. If the points. So you give the point less here, that's why is -, how to project this point onto the plane.

And here, this decent st we move s along this direction by the amount of d here. This becomes a path we solve for d we have left is in one in front. And this is one. The label is next one. Again, we can replace this naked one as well. We can write the two cases comparatively in this expression. Do you have these equals why I passed by another transpose, except us to they invest on this on how to compute. You need to understand the very first step in order to understand the whole support them, ready to compute the whole group, the margin of this is 1/2 of many respect that the point, right? So this is the market. So this is called the geometric. And the numerator, when I times x plus xiw plus xx that has these for the function name. Here is the next question. What is the margin of this have a plan with respect to local points? Not to the 27, not consistent, right? You define the marking respectful when you said that marking of the first point b is the minimum of or the distance, all three partners, minimal bi or I one am that's minimal xy an example why I asked it, double time is coming like that is first.

So this is the margin of this have a plan is better for 27 when you first identify the close . to the market. That's right. We have encountered this innovation problem in order to compute among the market respect. He said, so this is a margin, and we need to marx and margin, according to the principle of marx method. And lots in the market, then you get the intuition. But in order to solve as well, we encounter the maximum. Any more crazy problem, the maximum out there, the minimum in the internet. So that's rapidly explicit. Right? The maximum marking solution is found by solving this problem, maximize the margin, right? To find the optimum wb about the maximum this margin respect. And he said, this is very difficult problem. And none comments probably before. So are we not necessary? We can trump transform it into an equivalent complex relation problem by exploiting its property.

And Notice that if we make this game, w is equal to gamma w b equal gamma b the object fund is unchanged. This is what I say, scale ambiguity. If find the optimal wip if we find the optimal wip we come out by any number. So this aw and a double star and bp star is all also. If you have the same exact same object function, then we can play with the skill factor down. We can play this skill factor down, suck. Start, the margin is equal to one. We can work with a particular partner that's not this margin equal to one. This means not or function margins greater than or equal to one. The origin is not. The minimum is equal to one. The function marking for watching it now should be greater than. This opposition here becomes are similar. Wb 1 over 5 equals this photon is equal to one. And they need a minimal personal probability.

And we do not work with. This is their job. We do not work with square root. We can take the squares. The ultimate bubble will not change because its outer mass, since the spirits, by convention, we put a lot of in front. Because when we take these, the derivative, there is a tool here, which cancels this number two.

Therefore, the primal form, harder margin version of svm the primal form, harder marginal resume is this. Given a 27, we all mass, we must not believe about working from. Instead of work is no too long. You work is no too long square. This is the power comments, the function margin, for examples. If we belong to one, this is a concrete solution problem. This is not because this problem suffers from scaling ambiguity. This is broadcast also in spain. That's why we by exploring the scheme and do it. They can convert it, possibly non commerce opens from the congress and problems. So the objective b method inverse of the margin distance. I demands method to imagine margin. The quality customer be sure.

Now our points are either animal outside of the function, because which kind of function market should be is equal to one. The function one is the distance of one from the function. And some students, some students ask, why is magic number one? Why is no two, no three, or no five? For the sake of solving support vector machines, you can see any constant that should be fine. 123 equivalent, because it was found a different time. We are working with a different time, but in terms of compare different those functions, taking this function from 1 to 1, it's easier to compare or visualize the underlying those functions and as well. And compared with others function that is, as you can see. All right. That's why we say this. And prediction, given a new data point x bar, we use assignment in the function plus state map, the case of more chance to purchase. So why is marx in my so we have two aspects to something about this problem. One is the true. Make better is uncertain. Maximizing the margin allows the most uncertainty, for example, while keeping work on several possible, the data form the answer, because we made kind of a manner errors in the computer, the pressures.

So marx and marking was the most big move for the points, rather getting more than one spirits as that. Right? So we can, we do this in the boundary, slightly, a whole point of the aspect. We use the data points slightly, not very classical performance. That's why we did not imagine the market. So it's hard market enough. No, why is here? We should do cases? The first is we encounter this noise example. If you don't have this example, the decision boundary is discussed. Nice, sweetheart. If we have this example, the decision among boundary move from this national place that are so this noise example was that the decision done with a lot, which is not is very the hard market version is very sensitive. And we oriental we have some very large market to accommodate to grab in class this morning.

Now we have very much small market. This is 100 since it could. The second reason is, but this is almost being acceptable, but it's more convenience. If we work with trying to form part of our conversion, they spend a lot of final physical solution. There is a little type of thing about time satisfied with the good. There is no problem in your form. So we can use the same in your but we allow some training example to value the market. It can be inside the market. You can even be classified. The process is not, but we eat home. We need to quantity by the mark amount of valuations in china. And I do them together and penalize this violation. Right? We don't encourage, we allow it to evaluate, but we don't encourage you to evaluate. We need to get some penalty. That's why, for each training example, we define a snack barrel. If we have a little empty, not to have a total and snap levels.

So the ci is equal. Zero means sub, or I is also as the margin error is this. The examples is one of the besides, because it's outside of this market area and it's very classified related. Independent, the ci is greater than zero means the sample is inside the market. We want to make this ci as small as possible would allow it to be. We know about the most is not here has been trying from so much version, right?

In this first group, look at the conference. So originally, we want to order the function margin should be greater. Now in the black we here we give some flat surface. It can be less than one. It can be even zero or even that. Ci should be greater number zero. And this is the total amount of evolution. We need to minimize this validation. And here, you see, for these two, the c is equal to zero. For this one, the c is equal to zero. We don't have any kind of this on the boundary, c zero, boundary, c zero. And for this group, c one, c two, c three, we are within the market, or even these classified for this economics. Right? That's why we need to minimize this for those things. And we added have a plan here that's penalty for values in the market. So small value means a lot more violation part of vitamins.

Any question is? So the effect of c right? If they small, we will have more violations. You see, there are a lot of temples not within the market. And if these are, we don't encounter violation. She has a lot of violation. You see, the market is very small because we move on from 80 . 5 as it is in the market. And obviously to make the market as much as possible, but after they want. And in the case of primal form, as when we saw markets, even equipment, minimize this. So now this is established here. For proceed. There are two continents, right? To see, I should be greater than or equal to zero. And this, to see, I remove this to the left hand side, move this to the right hand side. Besides should be greater than or equal to one minus yi comes up with transpose xi must be right? Because that should be not around those values. Therefore, because the n should be greater than or equal to the maximum of these two values, because it should not allow those values is going to be step d it's going to say that it should be mother and mother.

Therefore, and we have one, we replace this could say I with its lower bound, how to understand this? Here, this is proceed. This is lower bound. The right hand side is slow bound, right? We minimize the low bond is minimization, right? We want to, instead of minimize to see, we want to minimize this low bond. We sets to see to attend the program. Right? Instead of minimize proceed refers to minimize obama, said they say too small, not what is. Therefore, we have this expression. We can move this c here to c and we kind of look also these two. We see this become one over c that's why we have an oversea, no conspiracy.

So you have to have this one, the sum of the working examples and maximum 01 s when I have the 27 last year, this is yi and this is f of xi we can define the whole term as they are similar as in the case of what is the question from it? We can treat this as some form of regret for that, right? And hinge laws, we can define the hinge laws, remark session 01, - 3. So it takes tens of years from angels, orange on this man trying to form crime from crime margin, and you have found this open margin for having step travels.

And then we establish the this organization from to be figures, this importance name from which that in the business. And this is a figure for humans. Was it right? Quality will classified, make it equally classified. We have not most linear margins. This is a linear, meaning are not supposed to balance examples. We have none, zero for samples of zero margin. We have zero zeros for samples on the correct side of the margin, 01 loss. Suppose the logistic loss and the engines are complex, relaxation of this 01 month. If they say if they is negative, this on zero, if this event is true, but it means the sample is classified. And the indicator function will give you 11 increase, in some ways, miss class. Remember, z is equal to y and fx we should have y and fx at the same time, in order to make it pretty fast, it is never be small. Decide this is the past. If the language is true, the initiative company will give you one, right? That's there.

Therefore, this is other one means. We have really missed an example. Zero needs is great enough, not everyone else. The average they want. Most of what they said is exactly the transportation area. And this is the most popular imaginary mass, because we really want to minimize the transportation in a way. But this is, in general, computational in favor. And we need for several years ago. Here we draw the figure for 01 loss, hinge loss, six, explain your loss and logistic loss, the exponential loss. It is defined as here. And they use the development of argument turn up in such a as when you get a flavor.

So here from this figure, there are several interests of relations. Why? 01 loss is complementary individual. Symbolic 01 was this communication. The gradient is zero, almost everywhere. The gradient of this 01 notes is zero, exactly zero, almost ever, except for this continues for zero, where the britain is not fine.

That's why you cannot use this as opposed function and takes a great, decent, because it's zero. I will be simply initiate the group here, if, let's say, if this is equal to one, right? If this is equal to zero, the little term is equal to one. That's why the exponential nodes to have traverse 01, and maximum 01, - 0. This is zero. The bottom is one. The engineers will traverse. They still want. Here, what's interesting is the majestic laws.

So what's the majestic laws? Majestic laws is law, one plus s omega z if we work with normal phase two, if z is zero, this is 11 + 1 is 2. Note 202 is one, right? It also should traverse this point, right? But it is below this .. The curve is below this .. Why is that? So if we draw figures like this, you should be immediately realize on the local, maybe in order to receive it. But it is true if the northeast is true, the logistic function would also traverse this .. Right? You can draw from your own version of this, not diverse, exactly this .. That's why we, in aspen, we need to set the margin to be one, to make it traverse this point, to make it easier to be compared with our cross functions. That's it. Is the same for us for same and more for logistic, the base of the not for logistic function to be equal to 27 other basis, such as natural mode. Mobile panel, any questions on this paper? Multi class comes back. And we have one for this one, numbers, two, numbers, 32, or 3.

The label will keep the parties the most votes, my favorite votes among four banner aspects. The problem is very slow where there are a lot of passports, so there are three passports with train, CC select two, which is seen on the same as a country that you have better aspects and become also the members of it, the rest of the class classification, numbers 23, first one, creature, 12 to label example. We picked the classes and largest. That's one. This is a decision boundary for three class area speaker. I think it's one versus rest. Interesting time. And here is the decision market. And if you remember for logistics program, right? So the decision number is pieceme here. Here we have for sn conversion forms.

Now, this to it, it has been summary. So classifier opinion senior give an example. It based on the sign of this in the doctrine and training remarks in the market, for training data and not in the situation the format. You need to know how to compute the margin. 100. And this 1 point, this is for 27 and how maximum about it. And in Hong Kong where this maximize automation, in addition into a complex relation problem. And we allow some attentive to violate the marketing. We just crossed by things. If you have a time to see and summary, both linear classifier, except the data using an Indian service, have a plane. There are two combinations. One is just regression. They maximize probability for management of the data. And on the line or something is majestic course. And majestic loss is the complex most option that in this course we have good. This is the comments I take the same order development of the day for this function.

And the second point where these with material and so for the commercial remarks and as a margin to have a pen on the line was from the usual. And having encouraged us to prove that the teams lost is also a complex function. A team's most function is also a complex function. And the logistical question, push the decision class to come as far as possible for all points and possible that machine, we ensure a margin of one between boundary and closed for option one.

We have already given reason why we set this quantum argument to be one instead of our concept. Right? If you understand the figure, the new cycle will single compound, even more sponsors, as long as this report is high dimension, which one has motorization here and have been always. The budget situation has no example. It is is found the decision service company beginner. So in the next level, see how to deal with non immunity, the same service, non university and office. So and we have a natural 3 . 3, which is a step by step for school or implementation or just regression. I don't have that. You can edit we use the equivalent formation where the label the label is zero. I response to the percentage loss, by the way, and how the regions working for you.

All right? And in the natural, I have made a detailed relation of a great, decent logistic equation, logistic regression error in the location system of - 1 of 1, and location system of 01, and compare different relations, which are equivalent Top. That's all for this lecture. And we are going to have, the first of all, I I think the first report is the immediate, right? Because another II had two t as are waiting for you guys. Some of. All right. So this is very fun. You can see you guys at this ..