### ML\_week2\_lecture2\_2-20240913

说话人1 00:01  
You see the boundary of the government ld as you can see, here. And the number three set on the test set, its a straight line. So why is a straight line? You have is high level,

说话人2 00:19  
including because

说话人1 00:20  
it's 30, they showed the parameters metrics, the program it comes come from this koreans. And since they're sure, so we can cancel a lot of reference,

说话人2 00:38  
which makes,

说话人1 00:44  
let's do this. The decision boundary concept. This is a . where they go back for the class one is equal to the normal level for the class two. And they move to be in concise. And we have cancelled a large number of terms equal because of this come on around that also the following results.

Here, let's work out this dimension. This is one by n this is n by n this is n by one. This will give you the one by n vector. One by n times n by one is just inner product, right? It just say, where is the summation of the picture matters? Not to make meaning. And those are all constant terms that are independent, not like relevant the facial benefits. So this shows up the decision found is natural in your next.

Now, let's compare nd with lde storage. And do you have some water roll? And we have already read the exact numbers for the hundreds one d and they'll give request a model and square. So the provider dependence on and makes it is slower than the during training as patient. Both model has put in probability since the progress responsible constant. They'll deal with generally more data than it is to estimate on all the open kind of secure around it. And he will give you a code drive is in boundary, and you will give yours in the decision.

Generally, classified model, it has been probably distribution features, which from this class in the future, the class is announced for several weeks.

And works with will with small amount of ability in a shoot. It makes strong assumptions on the data generation. That is just conditions and work with most of us and disadvantages. Accuracy depends on the selection of proper or reading description. So if it is real valid,

is it the ocean solutions reasonable to the normal? It's lost. So we may try some other continues

说话人2 03:38  
for the distributions.

说话人1 03:41  
Right? If the probability does not take a better way, then the absence of that. Now let's jump to not just regret, happy you. Most of you have learn or play with modest recursion. Hopefully, into the sector, we're going to have some new insights into majestic direction. So if you knew classification is generic models,

the first step is to build the 20 datas data, the vectors and corresponding class label one. And in their applications, election, classical one may be very cheap as time consuming and expensive. We learn from the last conditional situation killed at stable one. And the expense work at the hearts will be provided as possible. The training data is used to learn the last conditional and this classified.

The second group. And density estimation is a deal for the problem which density fields, how much digit we need to estimate accurately. And here is, I don't have one mirror wanting is the internet as well as when is sort programmers and spend

a week and a half to introduce as well and put it into

说话人2 05:29  
a broader

说话人1 05:30  
context or convert support and they should.

The banner map, if you say, it was an easy chance in the field of machine learning. And they suggest that when you solve your problem, try to avoid. So there's no problem as an

说话人2 05:45  
immediate study.

说话人1 05:46  
I I partially agree with his common claims, because I don't think it is a patient dependent. So ii work with vision problems, or human vision, or computer vision or machine vision. In envision problems, we have the velocity part. If you cannot solve a simple problem, you have to solve one next one. That's absorbs your simple problems and salts. When you solve a more complex one, the simple one may be sought to some reason for accuracy for number. When we saw this complex task, we may have extra information, not only eight problem solving courses of this simple problem. And moreover, so if you put the simple problem into a global context, it may becomes will formulate. And when we really want, when we really have the, you find some favorable

说话人2 07:01  
and better solutions for the same people

说话人1 07:04  
that's probably

说话人2 07:06  
around.

说话人1 07:08  
According to

说话人2 07:09  
this case.

说话人1 07:15  
Therefore, in this context, what general problem is to learn a journalism class. The simple problem is to directly learning discrimination. Last one that is resolve for the he'll write him x directly. Just impression. So about a limited technology, discriminative means not to directly discriminate. Class is a party individuals on appeal like you makes. Poy index is a one dimensional probability distribution, because y is just some class label, not much, easier, best and charity. There are models of coffee material from different classes to show x given one. X is a high dimension distribution, and linear clusters set up. We have a vector vector again in the products and dimension vector vector. And we have classes for the demand, for political classes.

Then if you want forever class and then you may see from some other textbooks, are they may use 01 notation, one for positive class, zero for every class. But they are essentially the same, the derivation of human atoms. Maybe it's not different, but they will arrive at the the premises I in our course

说话人2 08:57  
is

说话人1 08:58  
maybe one location. It will makes the comparison of the most functions that are for five different machinery albums easier. And the goal is to calculate the linear function of the vector. X is equal to double transpose x must be. That's our decision function. This is a medium product where the definition medium product, I would say the sum, it is equal to one of n of the j plus d probably is a big factor in the country with multi passage, facial value with the weight, and then add them together and pass this d as a baxter.

And the original is very if fx is greater than zero, that's positive, negative. That's negative plus. And we can reckon compact as a sine function, sine function. Sine of a is equal to one, where a is greater than 0, - 1, and this and zero. When a is except zero, when it

说话人2 10:23  
puts

说话人1 10:24  
the equal here or limit for the equal here, we may even have to save certain condition.

But last part of what matters, the only concern is three people, the fact, 3 days, the linear cluster separate the feature space into two half space.

Each corresponds to the picture, rather not to the first class condition. The class boundary is no more or a solo or a

说话人2 10:59  
particular

说话人1 10:59  
to the big lecture. And the class boundary is also for separating part. Here is one example where are released to one vector. Here. Is that right? That's two. This is one. And we started this direction and the bias to miss you. So my question is not. So can you prove, they said pretty harder plane is also pro recovery. Why it has a 90 °. And the proof is gonna be very simple. We random 2 or 2 points on the descent boundary. Let's see, basically, x one. This is base two. And since they are on the decision on the right, you must satisfy x two must be is equal to zero. Subtract one from two. We have w transpose x one minus x two is equal to zero. What is x one minus x two? It's a vector from x 2 . 2, x one. Those letter has the same direction and the same number. The inner product between we get her on this direction vector is still, which means that he has 90 °, which means that make it 30.

So some of the zimbabwe and we are done. But is it clear to prove they want to that?

说话人2 13:14  
It is a company

说话人1 13:15  
and make it? In an n dimensional figure space factors are in n dimensional equation. Double transpose x plus b finds an and n minus one dimensional linear surface. When n is equal to two, that's one in mind which has been showing the real site. When any 63 ws should be playing in generally, reported half a place is a hyper and how to save the classified primary.

We immersion we learned from any data. So class players are very objective used to learn the crime perspective. We will look at three examples. One is just a regression will be covering this lecture. The other is support the machine or message for sure the recovery majors.

Now this regression is of this approach, and this film is a geometrical. Logistic regression takes the probability approach, improper. This we need to map

说话人2 14:39  
those

说话人1 14:40  
decision value or function value. We write between zero and one. We use the same one function to do. So. Here is the model that z is equal to 1 over 1 class equals per power plan to be wearing the wheel. And this is a shape of the sigmoid power. My question is that, so what does this ship remind? So what does this ship remind you of same body? So have we encountered it? Being our probability course? What would the cultural solution? We learn the huge distribution function, cef in which follows this time.

说话人2 16:22  
Right? So

说话人1 16:25  
dao xin city, which has the similar. This means that so this city from this is what function? Maybe the cdf function of someone, right? That variable is called logistic random variable. And the variant of sigmoid

说话人2 16:53  
function.

说话人1 16:54  
Time in d time, indeed, be interpreted as the cdfon logistics around that.

I'm also have a detailed comments on this in our lecture, given a picture vectors. Probability of us is for positive is sigmoid for this function value. For that class is one minus sigmoid is something, right? And they

说话人2 17:44  
use

说话人1 17:45  
and since they should have some other one, right? That's why it's one minus single reference. One minus sigma fx is equal to sigma negative fx promote its profit. One minus sigmoid fx is equal to one minus. It's write down the extreme formula, one plus e to f to f this is equal to 1 + you need to connect to

说话人2 18:21  
afx and

说话人1 18:22  
you need to the

说话人2 18:23  
ffs

说话人1 18:26  
we multiply both. You don't remember it by eofx plus one plus e to the n and this is equal to one plus one over one plus e to the negative. This is seek more, and we are up. And therefore, we can write those two equations comparatively as poy given x is equal to sigma y times effects. When one is 0 ~ 1, as this formula one, when one is - 1 on this one.

Any questions on this? Here we directly more than the possible security will be in order to pass condition. This is the visualization of one particular classroom. Right? This is probably for dangers next time, too. We could have negative for features. We can do it for the quality. Given the data set d we learn function parameters in moe we maximize the condition of the sum of the conditional, no connect of the data. The optimal plan for double star p star is equal to our mars. Is that double b we average design works in the 26. The norm for videos of wifi even excite.

So remember here we have come on, are they? The same is gonna be. So we put w in front of the same important behind the conditional bar, meaning that we treat this rate vector as a run as a run around. Instead of just some kind of as a random variable, we may apply some prior probabilities to incorporates our priority

and b since it's behind the same economy, as just some time, if you miss it, our own campus to be estimated. Great question of this location. So private. So locations behind the same corner partners, notation behind the conditional bar and variables. Here, we know the formula for this, right? It's this formula. In this form here, I essentially have this is one over one plus x function yi fxifxi is a linear function, a proper transpose xi plus b we take a lot of it. This is one over, that's negative in front, negative more one plus as function. Why? I

说话人2 22:46  
ask a

说话人1 22:47  
real article says, I ask you about six former here.

This is average over or 25 %. 27 is this woman here. All right. So to prevent over 50 at the prior description of value, and we will assume the gaussian distribution, pow with variance c over two, and which means that pow is proportional to this formula. We congress exactly. What's w what distribution does wo here we have tow in general, follow the multivariate functions, which means inventor

说话人2 23:53  
new and

说话人1 23:55  
the grams metrics sigma. And they know the general formula for the smart grammar distribution, right? One over square roots, too high to a variable n the matrix determinant of sigma. And as a function, make one over two, w minus new vector transpose, sigma inverse, w minus mu, so that those are human instructions. Here we don't expect you, right? We don't have you.

So here, not is mu. Here. Mu is equal to zero. We assume w has zero v this structure has zero mean, which means that this is one over square roots, I to the part of n that should be terminate the sigma. And here we have that function, make 1 / 2. U is zero, w transpose, sigma inverse. Again, those probably has diamonds, silver, too. And we don't model. We assume they are independent each with better independent, which makes a sigma, the diagonal with balance of the diamonds, c c over two, but all diamonds are zero. Right? And this, if sigma exists, what's sigma inverse? What's the inverse of a diagram? We just inverse each end and some of that books, right? The sigma inverse is simply two over c two over c zero. Sometimes, if we plug this sigma into this formula, we have one over two square roots, too high for ron the matrix determinant over that matrix is essentially the product or elements on the battles. That's c over two to the power of n we have that function, 1 over 2.

And since those are bad words, right? With the same entry is just the so you can write explicitly work out this measurement application, which is equal to oversee will transpose. And this to these answers

说话人2 27:03  
in

说话人1 27:04  
we have negative one over c which is here, this has a constant. We get rid of these constants. Makes this equation for portion

说话人2 27:18  
to

说话人1 27:19  
while we're seeing at the

说话人2 27:20  
time percentage.

说话人1 27:31  
With a smaller w vector, the property will be higher, right? Those prior in harvard stuff that should be as small as possible. So that's our priority. We don't want to the s and w vector to be large. We want it to be small. That's why we have this part

说话人2 27:52  
into our

说话人1 27:54  
estimation problem. Now, we must see mice, the posterior probability of this weight better condition on one ix and parameters of id according to the chain rule of probability.

According to the base form, from this step, this step, you can write it as p of the joint probability

说话人2 28:34  
is

说话人1 28:37  
p of the joint probability, w yi given xi and plus ip divided by poyi given xi higher transpired e since the b is just faster, it may affect the location of the dissolution. But if you will, not affect the probability density value, that's why here we drop this fee in this denominator.

And so the numerator of the section in the numerator, we further adopt the chamber of this is equal to c of wen xi power of that b times po why I even xi from how do parents of that be?

Since this vector is independent of xi and it does not primary trust by d we get rid of these two terms to make it. This is w here, these terms are the same. And so from this to this, the denominator does not depend on all partners. We can get rid of the denominator. And the product. Local product is some of those that's our marks till w one over two, the average more. In numbers, those are the local methods. Those are prior, those are the local methods. I know we look up this follows.

P we lose. Our p of w follows as function of that one over c double transpose w the log of it is simply one over c of the transpose started. And this is negative one over m the sum as one over m nor one plus as function. And the one I they transpose xi plus b and since we have those negative sign in front, we can get rid of this negative sign, take the pass and turn our marks into the rmb program.

Arrive at this moment. We get over this next time. We make our marks to ask me. And the first term is a recognization term for the panelists. Countries in double from getting too large. And c is a regularization, have a plan for how to see the loss, not providers, and smaller cities is power to underwriters. And the second term is a bit of a term. He wants to make the parameters double b will fit the data. And we can define here, this is foxi so we define ci as a product of yi and if xi would have the following occupation, if they are extreme, is both, which is that they should be mostly positive for both 11.

For positive fx the prediction is a big against positive. For that to the fx the predictions, I think that is an analyst on tuesday. Then it means that the sample exercise about the gravity. And if you have people, if they are is negative, which means about this, and this has opposite side, it means that somebody is possible incorrectly. If there is zero, the sample is on class ￡5.

And then we can write this as long one plus next function, negative cr this is known as log one plus x function. Negative z is known as a majestic loss function. This is the shape of a majestic function. I I is negative. It's incredible past month. For the I is positive, is correct aspect.

说话人2 34:13  
E

说话人1 34:15  
when ai is smaller and smaller, we have another laws for badly discounted examples. For ai is positive, but small. We have non zero losses. It it's not be e equal to zero. We have not zero for gravity can ask for samples that are near boundary. And it's clear from this formula. This should always be greater than zero. If we have a lot to should be always greater than zero, which means that we are always kind of high the laws.

For more examples, non zero loss, but we see is not for the loss. The loss should be approach zero. That's why we have to do this curve like this. If we define the whole term as the most function email with respect to double and b and there is no close for this. Most function, no closed on solution is no simple 19 equations to solve for dot mb we may have to use it the target table, nation program to find the ultimate solution.

For example, we use great, decent step downhill in each direction where l is a positive function. W is a manufacturer. It has an early rate, how far to step in each step. And if we have already learned those majestic regression function of the arrows data, with this rate rate, the values on this bias term, how we interpret it. We have not prevalence to make fx for the two. So not prevalence is associated with what class. On selecting the recognization, I have a planner how this c has a peak effect on the decision boundary, and the accuracy how to estimate how to select the bankruptcy.

When c is 10,000 accuracy, 7 . 921 100 . nine two ten 0 . 961 hundred and six, and so on. Here we may select ccoc because it has a highest efforts. So homo city to see, so one typical approach to select happy finders on small scale, 27 is to use cross validation. So we run many experiments on the 27 to see which campus work on different versions of data, different versions of the 20 data. We first partition the data and photo, printing and regulation. We try a range of sea levels on each photo as reputation and keep the value of the world best for all foods. Here let's see we have already set the seat 100, and we speak the 27 temples.

On the first round, we use the first book as validation, the remaining nine as training within the model. It is accuracy on this first book, alice 90 3 %. And on the second round, we use the third second book as a vacation set the remaining time. He said, we retrain the model to get the accuracy on the same food as 90 % on we repeatedly persist and pass using each food as a validation and many as a training. We average those ten by patient accuracy as a performance indicator of this particular seat. And we can try it for different range of c to compare this average accuracy. I'm going to write these accuracy to say it to the basic.

Now, let's summarize this cross validation procedure into this argument, we first select a range of c values. Suppose we can select a range of

h

说话人2 39:30  
thank you.

说话人1 39:31  
Hc values. We repeat key months to split the training data into training and vacation. We learn class time for each value of c we record the accuracy of the vacation for each c and we select the value that has the highest average activity over a month. We will change the classroom in orders. They don't under the same city. That's a whole position.

My question is that coherence is the whole procedure. How many class fires have three trends? How many number of class fires have?

说话人2 40:12  
Have you trained

说话人1 40:13  
during the whole procedure? Think about it. So for each of the key ones, we need to try h different c values. And we in the whole world, right? After that, we can identify the best c and the step four, after let's see, we between the customer with all data and selectively. That's key. That's h plus one. H is the

说话人2 41:00  
number of

说话人1 41:01  
possible severe value in time.

说话人2 41:06  
B

说话人1 41:08  
but when we deal with, the large students said, we don't have those competition resources when we need 6, 18 and training, recitation and passive, and set half a parameter on the fixed right data set.

And maybe return the model, two or three parts for 2 or 3 times instead of t times c possible. Here shows this example. When we have a team's average state accuracy for a range of sea violence, we simply keep in us the highest one as the best city. So far we have there are only a cluster for two classes. This is known as a better class for the negative. For more than two classes, we can speak the problem to several other aspect classification problem. And the key people will do. So. It's one person's rest for each class between a class variable, that class versus another classes. If you have three between three values and aspects, one versus 232 versus 132, versus 12, prediction. We calculated the over cdt for each family classified.

And so then the task has appropriate. And in a piece of three, for the first class where you have this probability, why I is equal to one given degrees. And for the second battery class where we have this, first of all, the yi is two given c and we have for the third one of pyi equals three given c to repeat the one with another possibility and how to the corresponding. So that's if you move an example on three class, aristotle, we have add, received an expert errors classes.

These are the individual versus by the aspect. Here we have the first class as opposed to remaining as a negative. The second class, as opposed to the recruiter remaining selective,

说话人2 43:53  
the third, as

说话人1 43:54  
opposed to remain selective classes. Those are the same problems. And based on the prediction rule, we can combine those three. You say, the 3 % boundary into peaceful it is. Another way we'll get a multi class price to in terms of suggestion regressions, define the multi class of it. We have one big vector for each class c here reforming the barriers term for each class, for the things. And they can absorb the best in into the best venture as well. And as an expert feature domination,

说话人2 44:46  
which have

说话人1 44:48  
only take down by the one. And we define probabilities. We solve the so called salt mass function, is analogous to see more functions for better than this work.

And they see more from Is so smart company to find follows. Poy is equal to ckmx is equal to x function. How do you subsidy transpose six? If I have over those that's on my virus, the class with a larger response, WC that was, as you have the highest opinion, and will be you can pass. And here is a final class based on the individual breakfast and the first as a positive, remaining as a negative and so on. And individual ways, we'd better work together to partition in space. You see, from those decent boundary are pretty are similar, as, as we say before, you mean, one versus the rest possible. Regarding geometry, no, just a regression is explicitly designed to have an inner peace in boundary. The final case, because we rely on the linear decision, neither decision function in the model class, multi class case, the decision model is simply peaceful, similar, because you can consider it a are the marks of set our parts of the set of many in your functions, which will give you a piece lesson.

So regarding speed, learning, logistic regression, in class, in travel to the American condition. Right?

说话人2 47:18  
And

说话人1 47:20  
the very basic one is pretty decent. And in

说话人2 47:23  
the

说话人1 47:24  
next nature, or maybe

in week four, we are going to tap deeply into great decisions. We are going to prove mathematically. Why would it be used to go to where the function provide? Is not if you learning it is very small, it kind of tea to lower stopping this week, isn't it? And we got a storage. The model requires on the model. Many partners, the same order company, but how much is not go is open squares, right? Remember, you need to estimate the full share from various metrics which consist of

说话人2 48:14  
and two parts

说话人1 48:14  
and parts. One by two factors are these on portal n squared. And probability the importance of feature exchange can be understood in terms of it's not being learned with, let's do a or just regression summary classifier. We define a inner function as our inner decision function that must be.

And do you guys know how to absorb the speed into this red vector? Right here, our w is number one, number two, wn right? We can augmented in two, and then + 1, and then

说话人2 49:09  
in the

说话人1 49:11  
w zero, w ten, w one, where w zero is equal to b to the last term for the x for the x oriental is x one up to xn we can similarly augmented

说话人2 49:35  
as

说话人1 49:36  
x zero after xn where x zero is equal to one, constantly equal to 11. Here, the product between x zero and half of zero will give you the past 20. Not how you also the past, you will get your matrix to buy, but I

说话人2 50:00  
this

说话人1 50:03  
non facial remedy. That's not easy. Given a facial vector, the probability of the class is seek more than FX lecture is one minus single model that this is also equal to sigma of lecture f this is the formula of the sigmoid function, and it can be interpret as the cdf human distribution function, or logistic and variable.

During our derivation, we have come up with. We have identified this majestic, hostile. And training is a marks mathematical training data we use of recognization to prevent or taking the ads, propulsion on the big character to prevent large are doing values.

And since my working is small school later, we use cross regulation, pick the organization of a crime. When they not, they said is March, we may use a fixed speaking of training of training, reservation, and testing. And the TV research can be 6 to 2 or 712, or 811. And if they, this is not enough, we may have 90 . 50 . 5, because we want to train on as many examples as possible.

Those are the key formations to speak again. Regarding classification, given a new sample, xr would keep the class with the highest possible community. If it's battery that if the ratio is within a one quality, thank you.

说话人2 52:15  
Otherwise,

说话人1 52:17  
we can directly use this linear. Is it a function in this positive? For the past, making the past about equals assign this one? As the same as week one, we have a lecture . 3, which is the opinion. I have some notebook, a Free, a certain content. I have some of the book we introduce you with where she learning Python. So we have introduce or, say, New York with the passive English,

have some packages that are relevant to, of course, and the average state set in examples so that you can reduce some of the figures. Assess, right? If you are not from numerous Python, the highly progress to work through it, take some time to work through it. Right? So that's all, for this lecture, I must say it has to be.