### ML\_week2\_lecture2\_1-20240913

说话人1 00:00  
Two of them are training around and the third is discriminate I is that what we talk about to generally to ask that that is my face in your history and minus. So those two masters are both focusing on very important last. So we can now we have already up to the 1 page to of the optimality of the phasing of having got to take a look at it and development statement during the sick, the especially are going to compare what are the approximation made by each? What are the differences for this person who, but maybe a review, the definition of the classification task, given a facial vector, need for both x so remember, not have to worry or know it's gonna use people code letters denotes a common letter, not describe or represents the object that belongs to one of the seat pass from the seat.

First of the wire, that means this class a consist of 12 up to bc maintenance, which is not the calculated of this set. That means the number of elements, except is equal to the pieces. The revolution is natural stages that goes pretty, which is us the object answered. We also have has a definition to ask our learning. They remember that set of examples. B where begins another training levels and we use brand super sweet to index, as example, very excited worker as one possible. And those is asked as far from the domain islands about the domain are set, not have to predict a stable.

And the definition of class, if you ever read this on the computer function, given a data set, given a training set, d and the function have the classification areas of this master. Data set is, you knows, by air as if we view this as a function, this function takes a training aspect, as you said, and produce a number. It is the one over m the sum from I is one and m the indicator function, when I small t to fx sign, what is the fx sign? That's a prediction.

For the as example, why I want to support, as example, if this event, it's true, it will be part. We incorrectly, if it has found. Therefore, we at the number of this class, for example, by one, because this the account will be 21, the whole sum simply counts the number of these customs is. He said, if you went back in as areas, and the baseball team asked very loosely, each function fbos which is equal to the r r marks between the e document of the maximum value. If it turns into c pass it, we could say that, but choose the marks in the skill of the po minus people see you in the future.

According to the base rule, we have this is equal to why is equal? We will see common x divided by poby for the new major, we can further expand according to the chief of the uiispo if you are x given y is equal to equal cy come to your right? We will see that.

And from this to this, we get rid of the you don't make your pos pos is easy on the constable. It's just a constant. And if we draw the constant, they is still the same. So the input, the maximum, the maximum value, for sure, with the change, we they feel that, but the equal part of it contained much matters. It's gonna stay the same, which means that this is the equation is no approximation. That's an equation. Then we can have a strictly on the function. Our marks will stay and here. So the pdr pays the same voice for the story will be the basis for joint medical receipt, for the sample normal language. Or we did the same problem. And the base of ultimate classified deep is the best over basic classes among all possible class stars. And we have a full statements, mathematics towards. And however, it's more useful to the difficult as in this past commission.

First of all, this last commission is your high dimension for the distribution, right? That's used in the international space, but this is an information for pdpdc and from the data perspective. But a lot of movie is adequate to sufficient power growth and dimensional space. In order for us to make average probability against the model. And from the model part, we may need, they have much more with the large number of planners on the .. And those finals are very difficult as are not talking about. As come doctor more needs in order to have the capacity. Complex, high dimension for the business. So for both from their perspective, from moral perspective, the high dimensional probability function. It's very difficult to estimate, in fact, and that's why the nine base and the minimum analysis make approximation to this high energy.

So let's start giving too many ways. So how to deal with multiple features? I we were mentioned about only working with one big round we use. It usually have an option is usually more. So what about normal provisions? So in the non base of assumptions, when I base master close meets the baby important concert using a simple form, this ask for this promotion of all of the future dimensions are statistically independent statement of radicals, right? And this is also known as conventional independence condition on this class label. Those two features are in high statistical independence. They need the drone probability can be fat dressed in people profile of marking. If the two events are dependent each other, the draw probability should be equal to the product of the marginal group. Lots of definition. Here we are in added and the rivals and the condition of our condition. So in the law be in the law locally, we simply accumulate evidence from each way. It allow us to model each dimensional, the future vector, or each dimension of the operation of this vector, is a simple unit management solution.

On the means. Instead of estimating a dimensional probability density function, we down to estimating a total the end in development, which is much easier to do that.

And then, therefore, the general form for this classification function is they have a kind of various senses, substrate sub subscript x are the marks. You can see it as it. Po lines, you will see. And originally, for the base optimal conspire, the last condition as to scny and due to make us conditional independence by nonsense if that person or not or marginal, the request on this time, we'll consider the two dimension. Our state is showing the opinion direction. You have two features. One is the difference. Another 6 weeks and different are in the 50 classes and the most selective of available for me, this, you said, now that the test is tested, we are only going to test our model practice on the 27th and leave the test alone for testing.

Learn how she value basic model. We treat each dimension as an independent notion and certain notion and the space employee. This is for future value that can take real numbers. So it's beyond the world is aa real manage manner. That's that's why we use. Sorption is looking quick if you are in the school in my paragraph, which means that this one are only taken 012 letters. We can probably saw it in the movie, right? When it is a movie. And if those feature writer power, so if the test feature can only take our medium categories, 12 rc we cannot use caution toward either when we start over with my community to power in the first sites. Here we have view to be conditional how to create those joins class conditioning for this patient, the first time the fatal doctrine. And for this facial editor, for this picture, we take adoption and we multiply them together to create this edc I got more two d month, two d caution solution more is gonna have the c ds.

This is peaceful as well for each class. And for each picture dimension, more way, how do you do that about the same here? Y is equal to two. We create a notion solution, a bit of solution, and we multiply those one or two solution, and it forms are solutions. Right? Here is my question. So you see the principal access of these leaves is in parallel with coordinate access. The principal accent of this is we see power to hear this. If you have a business or an access, this principle access is how to receive basic on the access. Why is that any answers? This is because we treat each dimension, and we don't model the future relationship among those specials. That's why if you are always on the expect those paper lines, it is in my books. And we reviewed the poster for the poy these people seeking x we can be computed, using, available and computing.

So those dash, that curve is a decision boundary, and where they describe is equal to 0 . 5. They only have two classes. She should sum up one. And this curve is poy is equal to given x is equal to this number 0 . 9. For this purpose, why is your two given as equal to 01? In terms of the labor sets? Points, those curves have the same or be equal to or not. This on training set, this is on asset. And as expected, we may expect more misclassified examples on the testing compared to the chinese.

And as you take a closer look at this term, it's more or less following the variety shape. This is a moment. Now let's prove line in sex provider curve. What's the geometry? Is it normal? Now? You base rule is, this is class model. This is class conditional for looking at this cycle is assumption is that present for us. Here at one time, one of the number answers. And we take off to make the products into some laws. We look the classifications here, right? That's a new environment talking. School knowledge that they explicit how we drive this to this. Here we have p of xj given y is equal to c is equal to one over square roots to I sigma jc squared. It has function that is x minus mu gc squared divided by two sigma gc squared. Here, those substrate is to indicate that for each task are going to estimate this.

And in cyber invasion of those families, and given each class for each feature dimension of going estimate, is very and here, we use this commission. Mark indicates the classes commission. And this is just a subscript subscript. You can those are equivalent terms. You can write as j comma six words. You can even write g sigma, gc without comma squared. Those are the same. Just to make it. This sigma is depend on the us. People should depend on the data dimension. That's it.

All right. And lots of what's this term is equal to one over square root is 1/2, two high sigma g's. Six words to the power of 1/2. Screws is the power of 1/2, right? One over some value to the power of 1/2 is equal to two pi sigma tc squared that 1/2. Right? If that's the first term, right? If we take laws, so the first term becomes this. This big one up is in front of the law, and these are all who pass it on. Gs is worse. That's why there is a man making my heart in front. So the neck is because the square root is in the colonial. The 1/2 is because of this route and for the second term, not as castles and that universe each other, which is negatives in exchange minus mu to give us all gc squared divided by two sigma gc squared. That's the second term.

Any question? This time? So then these are just simple art, world and relations. No, please. The decision boundary consists of a set of points, xyx two with those supposed to normal output of class one is equal to the local level class two. There's more levels should be equal to each other. And we something about this to replace this, replace this in the previous site. And then we can combine common terms. It's not hard to say about the decision boundary. It is a project function of x one, x two with the form. This right here, let's say this.

To a simple example, those two Are independent of x it goes to the constant c those two are independent of x those are close to the constant c and what are the provided term? B comes from here, right? The xg squared. And the inner terms comes from this for that. You start here as well. That's why the final decision found has a branch. You have interest. You have figure out that is that expression for a one, b two, b one, b two, and c from this instruction combine. So we normally need to, again, my new base model for the wooden vectors, which means that the new dimension, the state economy becomes two others, zero, and one zero for absence, one more presence. Therefore, p of txc is equal to one, given y is equal to c is equal to 1 to 5. T absence is equal to one minus away. I'm here we only have one time estimates, because those 24 b is to come on.

Now, here is my question. How to write comparatively, using one equation to combine these equations, same value. How to use one equation. Combine this, what is coxg given? Why is when x is equal to one? It becomes this when x is equal to zero, comes one minus. But ii think this kind of question is very important to encourage you guys to see to treat problems in a political same problem in u five, the line is line is wifi jc to the power of theta function, xj is equal to one times one matters wifi, gdc the power of indicator function. It's gdc zero.

If xj is equal to e is indeed equal to one, this event is true. You will return one. This event is force. You will return zero. Someone is the power of zero is equal to one. Some power, some values of power one is equal to itself. And that's why I use it. If xj is equal to zero, this indicate function. We will return zero. This indicate function will be one, and we have one minus like gc which is six. Is that clear? We have mle estimating for those farmers. It just a number counting game where mtc is the number. For examples, c in class c campaigns, this feature exchange within campaigns, feature activity is extreme, is equal to one. And anything is simply the number of many times us. And right to compute, they join transformation according to nbis product of marginal transformation, because they are assumed to be conditional given as single. And you can take lot of it. The paradox of the probability becomes its side is more probability.

And for and in the examples and normal for people features presence, given a city as and there is a subtle issue here, some features may not be present, in addition examples for any parts, which means that in the previous formula, for estimating those where igc the you greater mgc is equal to this whole time. With 16 equals one plus c which, according to the formula we estimate by 1960, this means are the examples of us and we will administration. This can be a problem since we see, when we look at seeing the example with our patient, due to the meanings of these common examples in readings, expand those training examples when you may see a training examples in class c with this figure feedback to present.

So in general, for being a statistics and emotionally, it is very dangerous to estimates the probability of some event to be 70, you may as well. You wage set the provision of the interval, very small, positive number that tend to the power of - 8. I never said in which you may expect some peculiar where behavior is more profiles. You can get profiles. We have a smooth moe we add a smooth plan for alpha, alpha. You can view as a virtual count without some virtual numbers in the 27.

Therefore, the estimation former become this ngc that's are for the gas picture. We add alpha where our virtual examples where x is equal to one, and we have a total n facial dimensions, right? For each of facial dimension, we add rr where xj is equal to one. That's why the memory is alpha. The number is and is for the number of features. If alpha is equal to one, this is or not pass smoking, and the whole it is also known as advocacy. Add some virtual parts for each rate, for each of your dimension. If whether I think more smooth and asking helps to prevent.

So we're facing with china. What if next she has a finite set of category? Instead of only two here, the cardinality of this set, the number of elements that is still equivalent, speak for us. We can do. Then we can use a lot more civilized categories. We can start commission where poxg is equal to x even one is equal to c equal one by xgc we have a limit behavior, mathematical vision, because we have three letters in substitute, but it's necessary because c is for class neighbor j is for this picture I mentioned. And x is the possible values picture. They should sum up one. For example, if this polymorphic of the set is equal to four, how many numbers do we estimate? It's 4 - 1 is equal to 3, right? Because we have four or wifi, those four should sum up one, which means that the last parameters can be represented by one matters. Yes. Wifi. That's 4 - 103.

And again, how to we have how to write this competitive using one point. What is p of xj given y is equal to c not. This is just a generalized owner of what you have to previous with the written number. Right? This is equal to the two products. Or equal c is in the range of this number. And we have wifi xjc to the power of indicator company, xj is equally importance. And here we're going to use a company examples to demonstrate this.

Let's say xc consists of 12334 survivors are this figure can take. And here this becomes wifi one, g cxj is equal to one, wifi two, gc in hit functioning exchange equal to two, 1 . 3 gdc nuclear functions, cd equals three. If xg is needed one, this is one, this is zero. This is zero, and this becomes 1 × 1, gc x is equal to one by the one is, if x is indeed the two, this is zero, this is one. This is zero. We have no 1 by 2. Gc remains 22. So those are the unified terms. Understanding individual or babies. Any questions on this?

Again, ma how to estimate wifi xgc simple number of company program. It's nsgcufnc where ngc is a number of training examples in class c with feature xg is equal to one. Right here, we have the whole data set. So the three camera represents coding coming back to be, we have a subset with class c this is subset with class c and for each class c this is where xj is equal to one. This is where sg is equal to two. And so on. We simply come from. For example, if x is equal to a simple count, the number of examples in this subset, i'm divided by the number, of example, you have seen as singapore number of time program.

Any general question regarding my best? Right? You jump to. They need a discriminant analysis, linear discriminant analysis is a classification technique due to feature that dates back to 1930s. Fisher, a was a very famous sentence station family. It can be interpreted as a different of approximation, the basing open customer for real validation, which means that we can compare linear equivalent analysis with adoption in my career.

And instead of products of independent motions in my face, so they'll be able to shoot this joint class conditional follows a multivariate gao xi distribution. Ways for each class, you assume it estimates a different mean, better. And for across all classes, it estimates a common share of the variance matrix. Remember, this is in rn the mean matter is also in partner. And the covariance matrix is n by n symmetric matters. This is a formula e for multivariant function distribution. P of x given y is equal to c is equal to one over too high to the borrow nn is a bigger dimension and features the sigma. And this is known as the matrix determinant, mathematics. The matrix determine it, takes the matrix as an increment returns to the number. Do you think about? It has some volume scaling properties. You can review as a volume scaling factor or in the transform describe that is a particular manner. Here is the first term. Constant term. That's one over two x minus uc transpose sigma universe. It's not uc let's work out the dimension.

So next, then by one, you see, in the transpose 1 by 1, sigma n by n sigma inverse, this next one is matrix inverse. Let consumers n by n this is n by one. The sum of these three, the applications, 396 will give you 1 × 1 matrix, just a number, right? A this is just a number probability that is better.

Then the class classification function is as follows where we deserve this class conditional basis, more common problems with. Here, to contrast, are you fair and linear distribution analysis through the lens and multi element solution?

So we can, for 9 years, we have also, right, it's joined us commission as the most relevant motion dispersions for each class. In a single way, we will estimate a separate mean electron. For each class, it will ask me of the runners matrix. That's why we put the subscript c here. What is asked? However, this covariance matrix has a very special structure report. Can you figure out the elements of this covariance matrix for this money in money? The young person dialogue. It is a dialogue matrix, which means that on dialogues, we have sigma. On that, we have sigma one, c squared, sigma two, c squared. After the sigma nc spurs, all diamonds, all zero.

Through the lens of multi rather approaching distribution. Here, you see the difference between my face and go and it's human analysis. Right? Linear is criminalized as estimate the full programs measure, but share this full parameters networks over classes. But for my intense for each class, it will estimate separate if you write this matrix. But those large matrices are devil. There's some model with relationships among issues. Lda assume models, the relationship between features, but assume the dark features. The relations are identical all classes. That's why they share the grammar methods all classes. If any questions on the comparison between md and ld so the pencil multiple and culture dissolution.

Now let's work out. The total number of parameters like md should estimate here. For class model, the prior poyj is equal to c we're going to estimate cdcc is the number of classes - 1, because those dc privacy sum up to one, which makes the prime number at one. This is for prime, for each class, for class conditions, for each class are going to estimate vector last time. And for each class, we are going to estimate the variance methods, which is family last time. This is the total number of planets are going to estimate, not get small, right? Rather be balance. But now, for lda c minus one plus four prime and the same prime, for each class, it will estimate separate mean matter, not c times begin, plus, it will as no shared covariance matrix.

How many numbers do they estimate your commanders matrix? Remember, the covariance matrix is symmetric, which means that sigma ij is equal to sigma ji the adjacent trees with js then how many numbers they are remaining in the careers method? Same value. It's then have n plus one divided by two. They just do this figure. Since it's symmetric for the first row, we can only estimate this one element for the second row. We estimate two elements. The third rule estimates is entertainment. For the pencil, we estimate the rule.

That's one + 2 + 3 + 2 + 1, the sample, 1 + 2 + 1. Let emerge assignment has asked me as a question. There are two, the partner on this side on the counting of the number of banners to estimate there is money, whereas I okay, you see, this is on the order of is on the order of linear event. This is quadratic. And in practice, n is when we measure the number of classes, the number of features is way enough for the number of customers. And by the number of passers, you can take the intuition, not lac much more. It is more difficult question because it has more class. Right? Maybe more difficult question.

As with 90 minutes will be a promise can be learned with Emily which reduced to the sample estimates. You have already learned this formula, the class label, and those are the class means. And these are the covariance matrix. Of xg minus new writing Xi knew that why is the label corresponding to this example excite? We can work on this dimension. This is n times one. This is one constant with this code, will give you sometimes the matrix, the sum open times the matrix. If you want to know the step by step the ratios to arrive at those foreigners, in this, I will up know I will update on that lecture notes to give you a step by step mlmle relations of the us as the so DS.

So you can clearly see how we provide that those expressions. But it is option, but I haven't published to do this kind of exercise here with the two d class conditions. But as I mentioned before, into the Gaussian is always gonna have a beliefs here. But this time this edition is not aligned with the code axis, it's oriented somehow. But these two ships, so for each class, they have the identical ships, identical size, identity orientation. Why is that? Because they share the same careers, but they have different locations, because the s for each part of that lesson, a separate means the new member controls the location of this use.

And the parameters not just compose, the shape and orientation is easy use. Right? You see the beliefs oriented in this way, which means that this paper lens and a simple, which is two features are positive, create with a larger performance. We're having acting to associate with a larger several weeks with half of a so if we have another Netherlands, but with a small, separate, they called intensity is gonna be very small. That's why eld with essential model with paradise, facial relationships, that of the terms is violence. All right. Let's take a 10 minute break from the world is in time.