### ML\_week1\_lecture2\_20240906

说话人1 00:00  
Basically, just by the transpose, we have already 100 transpose. In what? Matrix inverse for this matrix trace, and what's matrix run? What's matrix determinant to stop that? We have a lot of metric concept, but i've already told you that. So the most fundamental concept

说话人2 00:25  
in

说话人1 00:26  
article is this. So you need to go check or recap in what's linear independence, how it is related to matrix rank, matrix inverse, a matrix concept, such as ip value, ip vectors, and also vanity stuff, definition, probability distribution. So probability distribution, if someone asks you, what's the probability distribution? The answer should be simple and neat. A probability of a probability distribution is a function to

说话人2 01:17  
that's all

说话人1 01:18  
the probability emerging is a function.

And this function for function, we need to figure out what's happening domain. What are the output? Right? So that the input domain of this function is what's so called a sample space. The sample space is input domain. The output domain is do a number. That's a function. This function should satisfy. The following conditions is non negativity. P of equal omega should be greater than equal to zero. For all omega. In its input domain, we use sample space. The second is normalization, p of elements in the whole type of space. If omega is should be one as normalization, the third is activity. If omega prime represents to subset of the sample space, the subset of the sample space in probability series for the events, events is a subset

说话人2 02:24  
space

说话人1 02:25  
for a big decision.

So if two events are destroyed, me, some intervene, their family don't overlap the last instrument.

Therefore, the probability of the union is equal to the sum of the respect appropriate. Those are the three conditions to constrain those functions. These three conditions are known as the three axis for me. What's on the matter? Don't answer. The random variable is the right. The answer is also be simple. At least. A random variable is a function, or a random variable is defined by a function. That's it. All right. A random variable is defined by a function f subscript x plus maps each element dot omega of the sample space. Therefore, this function has input domain as the same as probability function, the sample space

说话人2 03:38  
as the input domain

说话人1 03:40  
to the value x in a set of credit script x called the range of grammar. You can define your own range. It's map the sample space to the range of this landmark.

So by doing so, for each little s is in the range of this land rival, this is known as an event. We can interpret this as an event. Event means a subset as a subset of the sample space. This event response to this event, little omega, omega is in the sample space, where if x is or omega is equal to omega, therefore, we are able to write something like this.

Remember, in our definition of probability, probability is defined over the sample space. It's not defined over random errors. Probabilities defined over the sample space. Through this function were able to transform the traditional probability over the sum of space to the definition of events are defined by the random. Right? That's the important and subtlety. Subtle aspect. But actually, we also space on probability through for land arrival. If we already know that is defined by online function within the x directly instead of wearing online sample space, that the only condition must go. What this one probability for this one probability mass function you have, this will allow it to zero. They should sum up one for continuous probability density function. But again, you should do it, allowing you to do it. One, and the interval should be one.

Then what's the difference between disciplined mass function and continuous

说话人2 06:03  
density function?

说话人1 06:05  
The symbolic they are all represented by

what's the difference? For discipline, probability, mass function? It should be greater than zero. This number one for mass function. But for density function, pf big x is equilibrium, can be greater than one, and can even be positive beginning. Lots density function. It is this density function times is a small error, not to give you the probability mass of approximation. Those

说话人2 06:56  
quantity should be based on it.

说话人1 07:04  
And you should be familiar with the following fundamental concept from province area. Modernization, conditioning, based role expectation. The classic distribution is absolutely much normal. Dose you remember thousands solution? Multiple solutions? Right? In lecture, 1 . 2 are going to talk about two fundamental class cards. One is ten, and one is based on optimal class card. So classification here is

说话人2 07:54  
the

说话人1 07:55  
classification possible test, given where you spend or spend or have. Here is the example of spam email. It is an example time, and you cannot directly fade a email as input to a motion model. We need to have a facial extraction forces to come to represent this email as a vector of numbers of the official vector and slow this pitch vector for mercenary average, lots of input from

说话人2 08:29  
mercenary average.

说话人1 08:31  
In modern term, this is known as organization.

If you work, if you happen to know, or workplace is not enough more. This is one organization. And this is a binary example. This is like an example for that, even is given the capital essence, which critical capital as far as. So this image of iris flowers of high dimensions. And from this image, we expect two features, paper lens and separate. And we use this two dimensional vector vector represent

说话人2 09:10  
of dge

说话人1 09:11  
this is a scanning board of the x axis is evidence. The last axis is the simple ways. And give. Each different color includes a different types of iris flowers, as you see.

When we use these two features, they have we also forbidding in these two classes, right? May it's good for us to

说话人2 09:38  
have a mercenary

说话人1 09:39  
model as for those who have to average dollars. But if you come up with your own

说话人2 09:45  
two reasons,

说话人1 09:47  
and you draw this kind of part, there is a significant overlap between the two classes. So this indicate that. So those two figures

说话人2 09:56  
are bad

说话人1 09:58  
for classification. You may come up with better features that is better stability in the two classes. Then I go to five some classic models, get better

说话人2 10:10  
classic team members.

说话人1 10:14  
Now, let's first define the classification task, given which vector x is in rn not describe or represent the object belongs to one of the abc classes from the center telescope. Why? So here in our course, we are always going to use these types of courtesy letters. They set and predict the class the object belongs to. In the previous example, x classification, x is headlines. Several ways is two dimensional vector where n is equal to two for the class set, not only two classes, right? But the color gd com where the cardinality of this set, we put single word capacity around this set. It means we come to the number of elements they set, is known as a cardinality of the set. And the cardinality is a set is equal to c but c classes, in this case, is ee equal to two. And equivalently, we can treat versa color as + 1, which become as two

说话人2 11:40  
and put

说话人1 11:42  
one for part two in the last second. And the definition of class patterns given a better set, of example, pairs.

So remember, curtis swift d represent a set of 27, which consists of mpm examples. In our course. We are always gonna using dm to indicate the number of

说话人2 12:13  
examples in

说话人1 12:14  
the train set, where we use superscript I with a bracket to index the ic example,

说话人2 12:24  
sen,

说话人1 12:25  
where xi is an dimensional feature vector. Yi is class table. The goal is to learn a function from rn to the class set.

说话人2 12:41  
But I

说话人1 12:41  
release class table for any expression. That's definition, ask that there. So based on this definition, here's my question. What does this symbol represent? Xij any answer?

说话人2 13:17  
Actually,

说话人1 13:18  
it is the test issue of the as company. Right? Xi is an dimensional vector. G the g is entry of the n dimensional, which means that is the g is feature in the

说话人2 13:36  
past month.

说话人1 13:44  
We are going to introduce important indicator function. The indicator of a is equal to one. If event a occurs or event, a is true, or otherwise, it returns you.

Here, we are going to give you some intuitive examples on the indicator function, which will be sensible use in our the indicator function of three is greater than five. What's the result of this zero? Because this is false indicator of 470, give you zero. What's the indicator function of three? 111, because this stands for larger than or equal to. It's all relations with us is equal to one. Today is wednesday. This is 01, this friday.

You get the experience so that you get a function based on the indicator function, I want to define the classification areas. So given the data set, de consists of m examples and the classification function l the classification error rate f on d is defined as follows.

One over m the sum of I from one to m then you get a function when I is more equal to the exact. Just think about this expression. What's fxi what is the fxi it is as predictable, right? Why is a guangzhou state? Here the event is the predicted label is not equal to the guangzhou stable if this is true, it means that the classification model inaccurately incorrectly predict the essay, right? If this event is true, then this will be equal to one, which means that we add one. We added by one. We can't stay in credit class for examples. We added by one. And the sum means that we can't the number of in correctly predicted examples and we normalized by the end observations. M is a number, for example, knots, classification average. And similarly, by replacing inequality with equality will give us vacation accuracy. If predict label is equal to the true label, this is the event. If this event is true, we add this return one, we added by one. Here, this summation gives you the number of correctly predicted examples, no more or less by them. Expectation accuracy.

Any questions on this side? Kenya savers, or can, for sure, the can class card is one of the most basically essential class card image. It stores the train the whole training state and classify each new instance using the majority of over its key years neighbors. I'm here to represent the key years neighbors. We use labor index set. We don't store the neighbors, we store the index, the labor stuff, labor index set. The index should be one to up to m you have a little, many samples. Therefore, the labor index set is a subset, and the use of k requires choosing the distance function. The number of papers came, and here is the examples.

So in these examples, we have for each data, we use two features to represent it. The triangle is class b the star is in class a this is the new good example to class. The same first step is to calculate the distance between these new examples and all examples

说话人2 19:07  
in the 27.

说话人1 19:09  
If this training set has any examples, we need to compute m distances, and then we sort the distances, and then identify the k nearest neighbors be based on these

说话人2 19:25  
reasons. Right?

说话人1 19:26  
In this case, if we set k is equal to three and we identify this as the newspapers. And based on this newspapers, we perform major vote. In this case, we have two examples of that plus b one plus a therefore, this example is predict to predict s plus b by the principle of majority. The question of this example.

So to read come mathematically and compact in the united states very extreme, because two more massive papers, the one is maximum paper, the maximum available function, and x divide over some income domain return state, maximum value of this function, where max respectable fx with respect to x is the set consist of fs subject to that, this x is in the input domain. This is a symbol for, and this is simple for all, for all. Why? As another example is also, indeed, fy should be this sign or equal to fx which will makes fx not just by it. Unless it consists of only one element. We also have works known as the are not so great. The other amounts of little function is done over.

Some input domain gives The set of points for which fx which is a month in a while. We have outmarks of fx with fx which returns you a set of x instead of fx these are input argument. Hence, the name arc marx. Arc stands for argument. Input arguments. Subject to x is in the input domain. And for all y as another example, in the domain. If why is this lots of difference between marx and our marx? Here, i'm gonna give you some intuitive examples. What's mass of negative x minus two squares plus five?

What's the mountain lion?

255. It's five, right? What's are the marks? The r mass, in fact, two, it is two, not the twins. The maximum value of five up marks return the input argument that achieves the maximum values.

And why we we write something like this here, the set of points for which fx which is the maximum. Here, i'm gonna give you a visual illustration here. I i'm gonna define my function visually. This is x this is the effects. And

说话人2 23:40  
this is

说话人1 23:40  
225. This is three. What's maps of this function? Three, what are the maps of this function? 225. The mark marks of this function is, in fact, the interval not the single point, the interval 2 to 5 gonna any point in this interval will attain the maximum value. That's why we have a set of points, which fx which is maximum. Instead of a single point, then can be read company compact as follows.

This can classifier takes the future vector as input and predicts the class class label. That's why we have arguments here.

As mass gives input arguments, which is c it will produce either one to up to fix it.

The input describe the output is the rest is pick out this parts for all I is in the labor index set. We cant how many numbers, how many examples belongs to class one, how many numbers become? And is class two? How many numbers? How many examples? Class c class three, and so on? We take the maximum one. I return its argument. Last class right here, i'm gonna give you a visual example. This is a this is class b this is example, 123456. This is a due, in instance, to classify. And we said they need k is equal to three.

Therefore, my question is, what's the first group ks in this case? It's 234, right? It's an index of the nearest neighbors, the index of these three examples, 234. Now, here we check. Here class a is class one, class b is class two. We check whether if the sum of I is in prescript kxiyi is equal to one, this will give you what's the result of this function two? It will give you two, because when I is true, the class hit, the class is one. This is true. You give me 13. This is true, give me one. For this is 40 ~ give 0. The sum will give you two. What's the sum of I from the neighbouring that says iyi is equal to two. This will give you one, right? Due to the majority vote, the maximum value is two. The input argument that attain the maximum value is class one. And we returned in document and we returned last one.

So that's the simple, neat expression to summarize all aspects

说话人2 28:08  
of

说话人1 28:08  
this name, with the help of art marks or future, with the help of indicator function. In general, ken can work with any piece of function, such as non negativity and identity in discernibles. Generally, the more structure, the distance type, symmetry time employee is more structured. You can explore in developments if a distance function that satisfy this. And these are called a metric in mathematics or

说话人2 28:56  
a metric.

说话人1 29:01  
Here is a well known example of distance measure, coding, corpus distance, or lp not induced distance measure. Given two data points, we call this team with primer p is a proper metric. Problematic means he decides as a previous four conditions as follows, dpxx prime is lower, is computed by the alp norm, norm of the x minus x five, and is computed by the absolute value, absolute value of xj minus xj prime.

Take to the power of p and we sum over and entries, and they take this summation to the power one minus p that's the calculation of

说话人2 30:04  
normal introduce measure.

说话人1 30:05  
And don't for forget this absolute function here. A special case include depleting distance from pc to do my half distance with pc for one.

Then check here business when p is equal to infinity, when p is equal to infinity, this is equal to mass. This is equal to mass jxj minus xj prime. We just take to, we not just at the salute value of the largest decisions.

说话人2 30:45  
When t is 20,

说话人1 30:55  
given any distant function, brute force can work by computing or distance and solve the distance to form the neighbors. And do they make a ritual of

说话人2 31:06  
this situation?

说话人1 31:07  
And here is from which examples when the

说话人2 31:11  
number of the

说话人1 31:12  
papers is 137, 15, as you see, as the number of papers increases, the decision boundary becomes smoother and smoother. We may encounter some misclassification cases here. And here is my question. When k is equal to one, what's the training area for

说话人2 31:36  
kn

说话人1 31:38  
one, k is equal to one. What is the training? We are not talking about the validation year, tested about talking about 20 year. So when k is equal to what's the 20 year, the training area is always gonna be zero. Because in the training set, the nearest neighbor for one point is itself. Right? So it's already done. Give you

说话人2 32:13  
addiction.

说话人1 32:16  
Instead of giving all the killers equal weight in the majority of a decent major countries or assault distance where the distance is defined as follows.

Here, this x function means the natural number to the power of the apt are not dependent. Those are similar patients. E to the power of negative power. Yeah. When the small is closer, wi is not a closer points, contribute more. Those are points. And instead of good for seriously research, data structure like pd tree, can be constructed to save computational computation to some computations. In this lecture are not going to have given katy trees. You should be able to learn it interests in courses and data structure or advanced program. But I will give you the high level ideas when we have already compute the distance between a and b and they are very large, we have already computed the distance between b and c they are very small.

There is no need to compute the distance between a and c it should be much if c cannot be the us paper so big. But we kd tree allows you to explore this kind of relationships

说话人2 33:51  
to

说话人1 33:51  
save calculations.

说话人2 33:53  
Right?

说话人1 33:57  
So the advantage of can those training periods involved, it's easy learning. We store the whole data set. There is no need to make in the model. We have new 20 data as we go and convert to the directly in the service as they are posting it ready for. The disadvantage is startled for with 90 %. Performing newspaper requires a lot of memory, takes a lot of time, and more importantly does not work with dimensions, because it becomes difficult to computer distance. And moreover, everything is far from far from everything else in the high

说话人2 34:38  
conditions. This

说话人1 34:39  
is known as the person that in

说话人2 34:41  
china

说话人1 34:41  
because we have so dimensions.

So each definition, if each time give you a little bit distance with those small distance across all dimensions, it's not gonna be a huge on, right? That's why everything is possible. The third is based on optimum classifier. So it's one type of classifier, probabilistic, as far as 1 type of transfer model, the data model, how data is jerry to use for free distributions? I it's generative models. When we model the power

说话人2 35:24  
generating process,

说话人1 35:26  
we first define the general models. The world has object patterns of various classes, the observed mayors, features or observations from the optical patterns, such feature inspection process. I will represent the objects page vector, and each class of objects has a particular distribution, which is that the distribution of features is conditioning on the class label for different class that have different distributions, its class conditional.

For class label, we store all possible classes in the class set.

For example, in iris, yeah, I said we have two classes. In the real world, the frequencies are as well as given by the practical vdpy in our score is if poy is equal to one, is 0 . 4, poy is equal. There are . 6. So those numbers should be summed up to one more generally. So if cbc is in the y and p or y is equal to c should sum up one to obey

second axiom of probability. And this means that in the world, I was from 40 % + 1 and 60 % + 2. We can learn it from this. Right? So focus on the numerator. If c is equal to one, the numerator simply counts the number of examples that denounced + 1. We divided by them, which will give you the ratio. You if you replace this is two, then you readers simply counts a number of examples that comes to question

说话人2 37:25  
normalized by n

说话人1 37:31  
for observation model, we observe feature vector x which means that the feature vector, for each example, is given by and the value of the picture and the class, you should, which gives you this class conditionally class conditional solution.

P of x given y here, this is a high dimensional for b distribution, because x in general, it is a

说话人2 38:00  
damage in space,

说话人1 38:01  
is higher dimensional

说话人2 38:02  
for business school.

And

说话人1 38:08  
we have again learned from our data. So here we only work with n is equal to one feature. In the iris floors, we draw his grant of feature pedal lengths for each class. If you do something like this, the problem is that it looks a a little bit noisy. And we can focus on y is equal to one, how to understand this feature? This one here. This means that there is only one example that is pedal and three.

There is five examples start with peddlings

说话人2 38:58  
for

说话人1 38:59  
at least how you understand this is

说话人2 39:02  
this one. And

说话人1 39:03  
we can normalize by the number of normalized by the number of each class. It is imperial distribution. And the promise, is it how to solve this problem? We can assume for 50 model, for this class conditions, we estimate the practice in the provision model. And in this case, we will show each class is modeled as a separate motion distribution, given by this formula. Here, pay attention to this location. This is a conditional bar. This is a

说话人2 39:49  
semicolon.

说话人1 39:50  
If some variable is behind a conditional bar, we treat this as a random. If some parameter, if some locations behind the same economy treated as it released 5 years to be an estimate instead of primary, if we write something like this, x given y comma uc sigma c squared between this, or as random. If there are random variables, we may have some variable v for mu, c variable v for sigma c squared.

But here, we treat these simple aspects. And we know the function distribution. 5 hours through the two passengers were. And if we call this formula, and each class has these only advanced factors. In the areas. For our case, we have two class. Each class, the class condition is modeled as possible distribution. And for each possible distribution, we have two factors. Therefore, there is a total of four parameters it is made in this class conditions. This is a typical figure for gaussian distribution will feel like in shape, with view, is control is locations, and sigma control is spread. We can use maximum that estimation to estimate factors. The customization of receptive factors is to maximize the network of the sample for rc and here, we gather all examples up in our class c into the d subscript c class, where here yi is always gonna be equal to sub, equal to class c here, dc and that's all examples of the last two only denounced only two classes equals it, therefore, valuable to press the falling of position problem of mass problems. But as the new system, as worse, we are not respect. The new system are serious.

I from one to mc the node, negative p of xi given one I Semicolon, new system as is worse. This is modeled as a Gaussian distribution in the previous night, right? We can put the formula here and how to maximize it. We take the gradient of this formula, respect from uz and sigma six words and sits and to zero, and result from UC and sigma squares, which will give you these two formulas, empirical mean, and empirical variance, as expect. All right? Here, when we use this formula as a function of x when I am, we call it class conditional p or probability of the speaker picture lecture came in each class. But when we view the same formula as a function of this problem, we instead call its negative function of the data. And we

说话人2 43:44  
can

说话人1 43:46  
use another location for this last location

to explicitly present up with you this as a function of time that's to use for the same particular expression.

One from data point, this class conditions, one from private point, is negative function. Ask, if you would have to know the details how to solve this problem with this, I will upload a electrode, which contains a detail by detail, step, by step proof, not for unique approaching discretion, but a multi round approach

说话人2 44:29  
is

说话人1 44:30  
more general is we have to check out.

Here is a learned gossip class conditions. It's much smaller than the bigger background. Right? Now, here comes the basis decision rule of the Alpha. Force bathing is a rule makes the optimal decisions on probably more probability for uncertain if you meaning as a whole problem of making predicting error. So to prove the optimality of vision, optimum classifier. So there is 1 page of very elegant and straightforward and also and derive it and upload it, including into the network

说话人2 45:22  
and up.

说话人1 45:23  
In fact, we are having time to get to check out the proof of manage of the failing task, optimal task. Given operations, it reflects give the class c with the largest posterior b that's at the Mark cpoy equals cpos here, if this is not like this, we predict as class one, if this is smaller than this, it has class two. Right? Due to, I better than the senior more according to the not just poster for you.

And the problem is we don't have we don't have pure recommends. We only have class class mode of pure y or private solution or class conditional POS given. That is where the Beijing is in poor comes in, right? And the possibility can be accurate using a very cool. P of y gamma x is equal to p of the joint probability divided by p over x the joint probability can divide according to the chain rule of probability as p of x given y times p of y given p of x and the p of x can be computed by the the rule of total propagation. P of x is equal to the sum of c is in the class at p of xy this is equal to c is in are set pox given y times poy then over for expressions are expressed in terms of past conditions. And

说话人2 47:25  
that would be more than

说话人1 47:27  
by what compute

possible probability of the denominator is a probability of x which can be computed by the law of b and denominator makes poy and Xi of

说话人2 47:42  
one.

说话人1 47:46  
Here is the conference example of how to compute the denominator in class. Said the company is only doing that. This is the class conditional. This is posterior probability. What is the descendant boundary? The decision boundary of the 2 % probability are equal. This is since we only have two classes, it should be equal to 0 . 5, because those two should sum up one. We only have two classes. They got equal, should be 0 . 5.

And in this case, the decision boundary is something around 4 . 8 for cattle is less than for complaint we predict as us one marker line.

At class two, this is a decision part. And bearing what we made it, those are the denominator is independent of the classical one, right? The only it's only a function of possible. It's only a function of which return. Therefore, we compare the bridge. The number is, in fact, the joint lab Jones for me to pure that same one, we have ber basis and war. You can join that. If this is nothing like this, plus one other, otherwise plus two. And pdr based on posterior probability, pdr based control, that will give exactly the same decision model. We can also apply them on according to increase the function of growth comparison. Right? Using no collection, if this is harder than this + 1 on west class two, this is numerically more stable when the lab plus more, because for joints that we need to multiply numbers rather less than one.

So you the numbers to be multiplied are large. How do you like to be number on? But not makes a the product into

说话人2 50:18  
a summation

说话人1 50:20  
that makes more emergency. That's why immersion learning. And you variable position or you variable in the Afghan, but always prefer to work with not that instead of back to itself. And here is a joint lab group. Not that good, could give you %. Is it not as expected? So here is the basic conspire summary. Training. We first played training data for each class. For each class, the estimate of class conditional pox given yzc we select a formal distribution in the previous history, use using consumer culture institution. We may assume other continuous profit distributions and estimate their parameters using Mark mle marks and map like this nature. We asked me to pass five using and we have four for classification, given a new example of starving, happiness is negative, and we picked the class is not just a stereo for me.

Maybe equivalently, your strong connections

说话人2 51:36  
or

说话人1 51:37  
the sum of the black we do have a only time we do have an x one three, which is true to look, which will give you a gentle introduction of. I pass on those Google passenger more. So for those are followed from units have some I have encouraged to check it out. We need a little bit increase in the next 1 . 3,

说话人2 52:09  
which will be stop it. Right?

说话人1 52:14  
We don't have to follow this week. And for the major followers, we will make a big announcement,

说话人2 52:19  
which

说话人1 52:21  
has within the

说话人2 52:22  
switching,

说话人1 52:24  
with so teaching. That's all for this lecture, hope to see all of you next weekend.