Fraser, we cannot hear you. Okay. Sorry. Another thing that I want to mention is that, for the homework one, I want You guys reply some question about what is machine learning, blah blah.

And I found some of you just copy paste them from online. And the place do not do it, again later. Okay. And then, similarly, another course, I found some of my guys Do a little bit of revised also on online content.

I also detect that. So 1st, on the feedback I gave you to guys, please remember, try to avoid. In the future, if I have some kind of similar Question, for your answer, please just write what are your thoughts.

Do not Google it. Do not modify online content. Okay. Makes sense. Good. Other things, we would like, get in the first form of this. Should be fine because we totally we have a certain homework for you guys.

Right? I will only select the highest attendee. You have 3 chance to get there. So no worry about that. But at least you need to keep in mind. If you do copy it again, next time you will be up. It'll be the problem for you.

Okay. Let's just refresh our memory about last week. So, actually, last week, we do a little bit about our interview practice, and then we review some basic mass content that that we needed to use in this cost.

That's the first review about all those kind of notations. Like, so this is a small one Let me just quickly read about that.

Maybe you could start from this one. Yeah. About the norm was different, the product. And maybe the next slide. Last time, I believe we already talked about this, gravitation and the gradient parts and the Hessian metrics.

In today's class, we'll we are manually covers our rest of notations from the best way, more about probabilities Awesome. See, from the information there. Let's try to turn to the bottom left. Left. Okay.

Let's try to review what is part of it. Intuition. So you have a process. Several outcomes are possible. So when the process is repeatedly, a large number of times, Each outcome occurs with a relative frequency or so called probability.

So if a particular outcome occurs more often, we say it is it is the more probable. So the probability arise through a contest.

First of all, in actual repeated experiment, for example, You will recall the color of 1,000 cars driving by 0 of 5, 7 of them are green, And then you estimate the probability of car being green as in this equation, like, 57 divided by 1,000, so eventually, there is a small number, and this is so called the probability here.

In the idealized, conceptions of repeated process, for example, you'll consider the behavior of an unbiased The 6 side die. As expected, probability of enrolling, 5 is 1 over 6.

Right? That should be The same. Another example is that you needed to you need a model for how people's height are distribute distributed. You choose a normal distribution to represent the expected radio probabilities.

So next, try to solving machine learning problems. So you will may need to dealing with some uncertainty, qualities as well as the sarcastic problem are non deterministic parties.

In property theory, theory, it provides a mathematical framework of representing and quantifying uncertain quantities, And there are several different sources of uncertainties.

So for example, the first one can be the inherent Stochasticity in the system being modelled has, another concrete example.

So most, interpretations of quantum mechanisms describe the dynamics of, some atom atom particles as being a probabilistic.

Another reason cause the uncertainty can be the income incompleted observability. So even deterministic systems can appear stochastic when we cannot observe all the variables that the behavior of the system.

Another reason can be the incomplete modeling. So when we use our models, let a master some of the information we observed.

That is part of the information resulting in uncertainty in the model's prediction. So for example, the Discrepitization of real number of values, the dimension reduction, etcetera.

Right? Because we saw always, unnecessary information. So you would think that is unnecessary, but, to in terms of data itself, it should be useful.

That was another reason of the uncertainty of the model. Then we quickly review about the random variables. Random variable x is a variable that can take, different values.

For example, x, that is equal to drawing time. So the possible ways of x are come compromise of Step on, space, for example, also part of, outcome space that is s equal to 1 to 6 because it's like, that has 6 different, service.

So we donate the event of the state of 5 as x equal to 5 or x that, equal to 5 is this different equation.

And there's a probability is so called is donated in is this to. Also, the p file can be used to donate to make the probabilities that x times the value of y.

A probability distribution is a description of how likely a random variable is to take each of its possible states. A compass and location is common. So one p x, the probability distributional of the random for x.

Also, the notation x that is in tutor, p x can be used to donate to the random variable. X has a probability of p x. And, definitely, for the random variable, there can be is that just creates all the continuous.

The discrete random variables, have a finite number of stages. For example, the size of die, and the continuous random variables infinite number of states, for example, the highest of a person, it can be infinite numbers.

Right? Here are some, like, kind of rules of the probability. For example, the probability of an event a. Yeah. Given sample space test, the net as the PA master satisfies the following properties.

Like, that is a non negativity, so which means that probability of all event, it should be bigger than 0. And all possible outcomes is, For example, if we sum all the, probability, then it should be equal to 1.

And, yeah, another property is about Identity of the this joint event, which means that, for example, a 1 and the a 2 used to be equal to 0.

So which means that there is no, common unit between this a one and 2. And also the Probability of a one, the unit was the h two should be equal to the probability of a one then cross all probability of a 2.

Right? And next, the probability of a random variable k x must obey the exons of the probability over the possible values in the sample space s. And here's another example of our discretes of variables.

A probability distribution of discrete variables may be, described as a probability mass function. So also named the PMF. So let me you might need to remember what is PMF. That is just probability mass function as you can see here.

And our probability distribution over continuous variable may be described as an TDI, because of probability density function. For example, waiting time between the, eruption of all the, best film and the APDIA gives the broad video.

And even, if not just keep small region with the random of the that acts here to find a probability of the interval a and b where you already have of the integration between of a and b of the probability p x and with p x.

Right? Because this is how we, find this probability over kind of certain interval there. For the market variable of random variables, so we may need to consider several random variables at one time.

And then if several random process that occurred in parallel or in sequence, for example, to model that relationship is several diseases and the symptoms.

And, another example is if you process images with millions of pixels, And then each pixel is a one random variable you can see, in this way.

And that's the way we are studying the probability distribution defined over the multiple random variables, And those include the joint, the condition, and marginal distribution.

The individual random variables can be grouped together into a random vector because they represent different properties of an individual's, status to go unit.

Our market variant of random variable is a vector of multiple random variables.

So for example, I the capital x can, be represented by x 1, x2 to xm with its transport state. So now let's talk about 1st about the joint, probability probability distribution.

So the probability distribution that acts many variables at that time is annoying as a joint probability distribution to arrive, but you're able to see about all That is capital x is equal to s, small s, and, capital y is equal to small y simultaneously.

So that is Probability of x equal to smart x and the capital y would equal to small y, the less is a joint probability.

And, women Also, right, it is a p x for the initial to represent this probability. And from this Figure, you have found that as I join the probability, p one x is equal to mini event.

So the mini event is here. Right? This is a mini event. And why why is European? So you will get us a joint of probability. If tax is equal to a minimum Why is equal European? Then we cannot get it on probability of 0. 1481. Right?

So this is a we need to consider about the post case. And our distribution is called the marginal probability distribution. So for the marginal probability distribution is a problem in a probability distribution of a single variable.

So previous, but jointly, consider 2. Right? But here, we only consider about Europe and just consider about 1. It'll calculate the best time that joined them probability distribution to just push P (X,Y).

For example, we may need to use, some rule. For example, the probability of the capital x is equal to small x. So there should be a we need to sum all the probability across other kind of variable than the y.

So as you can see here, either some or or other probabilities across the y here. Right? And for continuous random variable, some measure is replaced by indication.

So here, just the indication up here. And this process is called, modernization. So as you can see, this figure So now let's say, what's the probability of the, marginal probability of the p x equal to mini van?

So That basically we need to calculate as our probability of the by x is equal to million, then try to sum All these 3 brought together.

Right? So eventually, we can get a 0. 3333 as it's minor distribution of x equal to any vector. The last, distribution we want to mention is about the conditional provision distribution.

As you can see from the NEM, so the conditional probability distribution is a probability distribution of 1 variable provides that another variable has a taken, certain value, for example, it can be donated in this query and solve probability of x is dependent on given the small y here.

And, also, it can be right in this equation.

And, also, similarly, in this figure, if we want, you can calculate the con Additional probability of y, y y is equal to Europeans and x is equal to a minivan, so we can Calculators, probability.

First of all, we need to find his joint view.

Probability that is about this one. Right? Then divide by the all other case. So eventually, we can get this this kind of probability of y is equal to, European giving x is equal to many events.

Now can I get a 0. 4 or so 3 as is contingent conditional mobility? I guess no questions. Right? You have any questions?

Okay. I think this is a nice device. The for, you know, is that is patient goal. Right? So our calculated that conditional probability for one variable, but additional probability support another variable unknown.

For example, this is a probability of x given y that is And we take ready in this secret. It's called a base rule.

And, again, it's the multiplication rule of that for the joint distribution is used. As you can see it's just I cannot move a little bit to the last side. Biasymmetry, we have another, p y of x that is equal problem.

Can you show probability of a x given y that times the probability of p y? And there's a term I'll refer as the p x here, just, we can also name it as a prior probability or the initial degree of of x.

And he has, p x y is a posterior probability. That degree of We we live after the incorporation of knowledge of y and the p y x is so called the likelihood of the y given x, and the p y here is just the evidence.

And, eventually, the is just, can be understand in this way. So later, once we Try to estimate something we were trying to reuse this likelihood.

It was kind of prior And, that will talk about the independence. So the 2 random variable x And the y are independent if it was a curious of y does not reveal any information about the curious of x.

For example, true success, true success of the rules of the die, independent. Right? So therefore, we have y, the p x of y that is just equal to p x because they are independent.

And as you guys have notation here, as notated from Independent random variables that pxy, that is just equal to the times up to px times with p y.

In other cases, the random variables are dependent. For example, getting arcane, our successive draws from our deck, the draw card is not a replacement.

And the 2 random variables, x, y, are conditionally independent to give another random variable z if and only if them p x y given z, that is equal to p x given z times of is, the probability of y given z here and also people make it a UDC clear.

Among the, continuous multivariate just means that for some concept of joint, marginal, and the con conditional probabilities apply for continuous random variables.

So the probability distributions use the indication of continuous random variables instead of us.

Summation of a dissipated random variables. Here is an example about the 3 component of motion mixture from the distribution in 3 d dimension. The expected value. So the expected value or the expectation of a function f x, Yeah.

With respect to to probability distribution p x as, average or the means of the mean, y x is drawn from p x. For example, for our discrete random variable x, it is can be calculated using pretty best way.

We And the estimation of all the px probability of x that time is a bunch of x that had a sum of it. However, usually, it is our continuous random variable facts that we are being calculated using some hidden indication here. Right?

So when the identity of the distribution is clear around the context, we can write, you know, as this kind of explanation here. So if it is clear with random variable, it's used, and we can write just write it as the expectation of.

So b is the most common matter of central tendency of a distribution for a random variable f x So I is equal to f x. That means that mu is equal to the expectation of x I. Then can be plus or rewrite at this point.

As this signal, I just missed that you get well, categorize the average of all the samples x I. Right? Other measures of central tendency, Like, for example, like a medium or a techno mode of the data.

What about the various? Varies gives the measure of how much the values of function f x, divide from Give it from the expected values as we sample values of x from p x.

And then here is the clear on how we calculate the variance here. Once the virus is low, the value of f x faster near the expected value.

And, the you already the variance is the commonly the nature they use is sigma squared here, and then the above, we create a similar to a bunch of s. F x is equal to x I minus mu here. What is the mu again? Mu is what?

Expand. Mu is Just the main line here. That is a cosmine line, this one. And we have, like, sigma square can be rewriting this equation, and this is similar to the formula of calculating the variance of sample of observation.

So as you guys might be familiar with this So I write out particular various, you know, parts of the data.

And here is how we define the the standard deviation. That's just a single here. The the best theorem as per root of the various x here. The covariance gives the merit of how much 2 random variables are linear related to each other.

And then here is the is the gradient. And, let's say, if the f x I is equal to f x I minus mu x, and the g at y I is equal to y I minus the mu y here.

And then the covariance of x y can be donated in this way and then compared to where covariance of actual samples, so you can calculate it in this way.

We needed to man notice that the covariance measures, the tenders of x y to divide from the, miss in the same or opposite direction at the same time.

For example, Is this one kind of observe some covariance between x y? No. Right? But go to Is that right?

How about this one? You probably can find a high correlation between x and y. And this should best miss a positive one. And if you like, this way should be on that relationship with an x and y. Any questions? No.

Sure. The correlation So correlation coefficient is a covariance, normalization by the standard deviation of the 2 variables that is just And the k rate device, covariance of x y is then divided by the Sigma of x times the Sigma of y.

And I also named the Pearson's correlation or or efficient at the end as, row x y here.

And there's a value This one is what? I'm running the relation, but some but you can read something that I want variable is decreased. Mhmm. So, your rate was negative. Yeah. And the y adjustment is positive.

And then it only reflects the linear dependency between variables, and it does not measure the nonlinear dependencies between the variable. So for example, in this case so So here, it's about linear dependency with noise.

From one tip place, very the higher positive than 0. 8, 0. 4, The 0, we are assuming there's no there's no kind of, correlation and, minus 4. 0. 4 and minus 0.

8, then to minus 1 was means a strong negative. And, also, here is the linear dependence without noise. It basically means that you can find a very clear relationship between each other. So what about nonlinear? That's about nonlinear.

A covariance matrix consists of Margivariant of band random variable x with states of x that is, n dimensional real values has a m by m metric, such as the covariance xij that is equal to the covariance of xi and the yi.

K. It's about it's the overall metrics. And there's a diagonal element of the covariant metrics, the various of the elements of the vector, so the the covariance of xi and, xi should be just equal to the variance of x I.

So, also, we as noted that the program metrics is the metric since the program is of x I, xj that is equal to the rest of xj and, xi here.

Was a different, probability distributions. So for example, about Barunle distribution, it is like a binary random variable x instead of 0 and 1.

So the random variable can encoders, co inflate with the common comes up 1 with a probability of b and, 0 place probability of 1 minus p.

Right? Notation should be x that is, should be should be compiled with the of p here, like your. That's what a uniform distribution, the probability of each value.

For example, I, in the one to the n is, the is equal to the p I is equal to 1 over n. So this is notation. It should be x. It should be I I could tell you it's you and here.

So what about in the figure here is equal to 5, and the probability is just 0. 2. Right? Why? Why the point is is 0. 2? Yeah. I'm gonna do this more like this system. And it's 8 with the files.

Each property is just 0. 5. Yeah. 0. 2. We also have other distributions. For example, Binomial distribution to perform a sequence of an independent experiments, each of which has a probability of P of succeeding.

For example, we are the p should be between 0 and 1. So the property of getting case successful in the end trials should be like Donate as is the and this notation should be x, the, binomial, and p here.

And we we also have poison distribution, which means that a number of event, operate independently in a fixed interval of time with a known random lambda here.

Just create a random variable x with a state of care that is is it 1 to to n as a probability of this So he has a red, the down times the average number of occurrence of event, and then you can relate as this.

And this should be most of the famous distribution that is causing distribution money.

So the most well known noise distribution, you refer to the normal distribution or the informally of the pair shape distribution, and the pair is the mean and its variance and its notation.

This is this is the full equation I've got, And then you guys should have familiar with our different and I often current distribution here.

For the multinomial distribution, it is an extension of the learning distribution for binary class into body class.

Remember, previously mentioned about learning is just a binary. Right? Like a frame or a coin, just to have two sides there.

Distribution is called the category code distribution or generalized distribution. Is probability distribution that describes the possible results of, random variables that can take in, one of k possible categories.

So a category of random variable is a discrete random variable with more than 2 possible outcomes such as the role of value.

As a, for example, in a monkey class classification in machine learning, we have set out that example x1x to your accent and the least corresponded that an example x I has a k class level. That is a y I is equal to y I 1 I 2 to 1 I k.

That k just rescale different clusters corresponding to 1 hot encoding. In one hot encoding is called, 1 hot k vector, we have one element that has the value of 1 and all other has the value of 0.

That's the name as a probability of assigned, class level to a data point as a p one to p k. So we will know that is a, p a p j should be between 0 and 1, and the sum of the p j should be equal to 1 for the different classes.

Right? From j, that is equal from 1 to k, the marginal probability of the data point to x I is the probability of x I that it's gonna be, Like, some product of all the probability here.

Right? So, similarly, we can take place of all data Examples as this kind of products of all of them together.

So lastly, we are mention up our another important part about information theory. So in the your information theory, It can study some encoding, decoding, some tranformation, some manipulation of information.

It is a branch of applied mathematics that are resolved about around the twenty point how much information is presented in different signals.

As such, information server provides a fundamental language of discussing information processing in computer systems.

For example, in machine learning applications, you know, use, cross attributes as, direct from the information theory considerations. Similar work, like, in this paragraph.

So the the actual to to find this information for the first time. So best case Originally, you might need to just study the study the message of a noisy channel such as the communication we are rendering a radio transmission.

So we actually more care about some unusual information. Yeah. More details about the, so called the self information.

So the basic intuition behind the information theory is that learning that an unlikely event has occurred is more If I'm Matthew, then learning that a lot of you have done has occurred. Hold on. This is me.

So here is a concrete example. For example, A message is saying the song rose this morning. Should be all of your ad dash, I mean, you've noticed about this one. This store a informative because that is unnecessary to be sent.

But A message is saying, that was I've told them, this morning is very informative. Right? So that, Here, for self information, we really care about some and you write information.

So based on that, it's, in duration, sharing, you find the cell information as of I event x as this equation I x equal to the log of the probability of x.

So I x is a cell of information, and the p x is a probability of human x So that you see off the information output is a piece of information received from the event x.

So for example, if we want to send the code of 0 10 or by channel. So event of 0010 is a series of code of that. And in this case, the last is And it's able to move forward.

Right? So the encoding is a bit 0 or 1 that occurs with a probability of 1 over 2. So the point is, in this case, The p the probability should be, well over 2 with the power button.

So then we we can calculate the state of information of I x. The x is 0010, and again, it's it's the equation, and here is various probability.

So eventually, we calculate this, gathers just a poppets. So that means the same information out of the event of 0. 90. No. That's a good question. Okay. So the sales information is for, to document how the information how big it is.

Mhmm. And especially care about some of the Google events. This can calculate the probability or just the big or it's not just a probability because they detect the negative log here.

So with the probability, what we are we present how important of that event. So probably What we are assigned, for example, in this message where we assign more value to this information.

Yes. But why the result is 4 weeks? So plus this event is really about the the encode. Each encode is about 1 bit. Right. So once you calculate the value, this equation, you will eventually get these buckets.

Right? So it's just the the how are you? Yes. Correct. Yeah. Good. Next is to talk about the entropy. So if I just create a random variable x, That follows a probability distribution of p with a probability mass function p x.

The expected amount of information through an entry is also called the shared entropy is defined in this equation.

So best way, it's calculated what? This is. This is 1. It wants to calculate what? This. This is what? This is what? What does this mean here mean?

Actually, this It's really it's a means about expectation of all this kind of criteria of us. Right? So based on the expectation definition that is equal to this equation, we can rewrite the entropy as this kind of equation.

It's best to me is that Mark will calculate all the event x across across all the small x p s, for example, p at probability of p x that times with log of p x.

So if x is a continuous random variable that follows a probability distribution p p here was a probated. That's the function p x for the at the attributes. So now in this case, it should So we we have identification in this case. Right?

And, for continuous random variables, the entropy is also called the differential entropy. And the for those the entropy is also very important, and, as a machine learning model, especially for the neural network, it is very important.

For example, intuitivelly, we can in term it's the self information. This one, I x as a month of surprise.

We had a we had a same a particular outcome, we are less surprised. The one See, I'm more of more frequently or open event. Similarly, we can integrate the entropy, h x here as the average amount of high entropy.

And because there is a little surprise when we we will draw some Sample from a uniform distribution sees all samples that have similar variables as you can see here.

The uniform, we have higher here.

What about KL divergence? I believe most of the guys are not familiar with this KL divergence, but it's fine. This is also called the relative entropy. It can provide a matter of how different two probability distribution are.

So for q probability distribution, px and, qy qx here over the same Random variable x, the kl diverges, the calculated use this kind of equation, and the for our discrete random variables, this formula It's equivalent about all the sum here.

One of the base 2 logarithms is used that Cal Divergence provides the amount of information in bits.

So in machine learning is a net a natural logarithm is used Based on here, and there's a amount of information is provided in that state.

So k l k l diabetes can be considered as a amount of information lost. 1 distribution queue is used to approximate the distribution of p.

For example, in the GAN, the in the red tier, the side network, The p is a distribution of the true data, and the q is the distribution of the c set data. As you want, you play again with each other.

Right? So that They want to. I recognize p and q in this case. And the for chaotic emergencies, should be a satisfied with a non negative, probability, which means that chaos would be bigger or equal to 0.

So while kl is equal to 0, if I only hear p x and q s have the same distribution. But mostly Important, probably our care diverges is that it is non symmetric, which means that this one is not only put to each other.

And because kl, which is is non negative and it matters a difference between distribution, it is often considered as a distance metric between 2 distribution.

However, kl divergence is not a true distance metric because it is not symmetric, and there's a symmetric method that these data, they are important, the equivalence your choice of whether I'm using klp over q or the klq over p.

Automatic diverges is called nonnegative, and the symmetric is called the adjacent sharing diverges is input, can be donated in this equation above the m is a average of 2 distribution. M z is equal to this just here.

And the KL divergence is very useful in some kind of special kind of machine learning. For example, In the first class, I mentioned some of my work is related to domain adaptation. Is that a case? Because we have 2 different domains.

Right? We have source domain. We have time domain. We are more more likely to reduce the chaos divergence between the 2 domains so that we can measure them together. This is a problem. Okay. How that work is?

So that is some Yeah. Is there, like, The KL Diver, is there, like, a rule of thumb, like, when reach acceptable? I'll say it again. Is there, like, a rule of thumb for the value of the KL Divergence as acceptable That's too much?

It's it's not like that way. Yeah. It's not. Shouldn't be like this one. Yeah. And then, usually, for KL Divergence, was missed that you when you when you try to use it, it definitely when you either have 2 different distributions.

Otherwise, it's not useful. Right? For example, if you have a training, let's say, in our typical machine learning, you only have act tracking sample acts here.

Right? I mean, in the domain adaptation, we will have 2 to different domains. We have the source domain. We have a fact domain. In this case, we really can merge them together, How do you reduce the payroll averages?

Okay. How's that? Answer your question? I was just asking you if there's, like, And a value that you a target value for how small you want to count the difference to be?

We are well, there's no this kind of number. Definitely, the smaller, better. Right? Yeah. Okay. So that about something you guys are familiar with. It's a cross entropy loss.

Right? For example, the cross entropy is the closer we like to the kind of k l divergence, and then it is defined as the summation of the entropy h p, and the kl that bridges klp over q here, and eventually can be rewriting in this way.

Automatically, we can write the, cross entropy in this case. So somehow, you might see this.

You create, and you might help us. Right? So you you donated this cross entropy loss there. So in machine learning, it has a SUMR classification based on set up at the point x 1x2 to xm, zed needed to be classified into k classes.

So for each of them, example x I, we have a class level y I here, which means that your level y follows that your your distribution p here.

The goal is to pass by the for example, a neural network, parameters lies by a theater here. That out, outputs are predict plus level y hat I, but it's there's a somewhat inside.

So that predict the level y hat follows the estimated distribution of q. So in this case, the cross entropy loss between the, true distribution p and the estimated distribution q is calculated using this cost and a loss here.

And, usually, we will try to minimize them Because the smaller the bag are right. Other values is maximum likelihood. So cost entropy law, is, like, Closer related to maximum, likelihood estimation is, also called MLE here.

Right? A machine learning, we want to find our model There's a package of data that, maximize the property to zed as a data is assigned as a correct class.

For example, So at max, state of p are these the given the model and data, but classification problem from previous page, we want to find the parameters data.

So that for the example here, the probability of up to the class level x y to y and x next line, Which means that we try to kind of, the correct levels, by using kind of models that that you said, for some dead examples of predict class and y I hat, we have the different, than the true class y I, But the goal is to find the data that, results in an overall maximum probability.

So overall, this is like the we're sort of the at max of p given the model of data that is proportional to this kind of mathematics, is probability.

And this is the true this the p one g p n does not depend on the parameter data, And then we can assume that we have no prior, assumption with on which set our parameter thetas that, that are better than any others.

Because that the key here is a likelihood, and therefore, the maximum likelihood Estimation, etcetera, is based on solving this kind of problem.

This is the reason why it is called the lexicon in my world. And observe data points x1 to xm is operating.

Take to the class level, data point xi is yi hat using Martin, distribution as a probability of predicting true class y I is equal to this one is a probability of x over theta that is Hamid donators, products of all y I j given y r.

Given j is becoming from 1 2 to the k. K means different k classes here. For example, we have here the problem is 3 class that is also achieved an image of a car x I, and the true level is gonna be 100.

Let's assume our product predict the level should be like this one. 0. 7, 0. 1, and 0. 2. And then the probability according to this equivalent, we should calculate Just a power to 0.

7. Right? So now we can preview this level as which class. This is which class. Yeah. But that prediction is like this way. This should be which class.

Of course. The first yeah. The first should be the. Right? Because this is the biggest number. So as soon as that is a data example, I invented the likelihood of the data. I give him the more apparent that's it.

I can't even rewrite the whole system here, And there's a logo, black hoodies, often used because it is the numerical calculations since the transmons are productive as many terms in imagine, for example, like this one.

Because here we are, change this, like, product to, like, summation. So it should be much easier to calculate. That's the real that's the reason why we use a lot of likelihood here.

So, eventually, it becomes the true summation of y I and log of y I had here. So our net view of lot likelihood allow allow us to minimize our approach that is negative blah blah.

So inventory, there should be just a equal to the But the cross entropy loss here. So last, maximizing the likelihood is the same as the minimize, cross entropy Plus, so here is a quick group of your time.

And the out to now, ideally, we already cover All the important the mass background in this forecast as the if you guys think of some part and you got a lot of familiar with that, You can go back to the slides and maybe read the other post.

Just say, make make sure you understand the majority of them, Then you should be good in this case. So this is about all the references.

In the next part, Let's really talk about the data pro preprocessing. So similar as you guys did in that kind of Homework. It that does some training, replaces some, question mark or missing data with some number.

Right? Some data quality, some manual tasks. So in the data pros processing, data cleaning, data indication, data, reduction, data Mention, discretization.

Then we have another summarize of what we're going to take. So first of all, what I about is data collection. Why reprocessing the data? The definitely, there are lots of reasons.

Right? For example, merits for the data and, quite a multi dimensional video. For example, accuracy, whether correct or wrong or accurate or not, our completeness I'm not record unavailable person that I'm just missing.

Right? Some can see the consistencies or some modified by some Not. Like and the time my, the time they updated and the ability so how trust was the datasets are corrected And the interpretability, so how easily data can be understood.

Right? There are all the different reasons that we needed to pre process the data.

So here are the major tasks that we will need to do in the, data preprocessing, for example, for in the data cleaning, we will try to fill the missing details, most noisy data identity, identify, or remove some outliers.

So what are outliers? We'll be, we got some Our website for some, to be in months, but it's, like, Most We already referred to some point that's a far away from the main.

Yeah. Right? From when you did. Yeah. And our delegation, That's an indication where we are indicates of multiple dead sets like data cubes of bios.

And that, reduction, for example, we are trying mean, use the dimensionality of the data, and the way we do some data compression, etcetera.

And the for the data transformation, and the data dispatcher, we will do some normalization and the concept that we will have some there. But first, about the data cleaning.

So that's there are several reasons that we needed to do that. Example, some data has incomplete, so lack of attribute values, lacking some certain attributes, of the interest or contain only some aggregate data.

For example, some occupations that is equal to nothing. Right? So it's missing missing data there. And noisy is that it contains some noise, some errors on outliers. For example, a salary is able to negative ten. My it's a wrong data.

Right? And a inconsistent so it contains some discrepancies in the codes On NAMS, for example, edge is equal to 42. Birthday is equal in this number. Right? For the people you have given from the edge, you have given the first day.

Right? The previous is a rating like a wide history is now someone really as of ABC. Right? It's different. And there are some other discrepancy duplicate the records. And even sometimes there are some intentional, some mistakes.

For example, like, general one as, as everyone's birthday It is wrong. Right? And there are several reasons why we needed to do our data cleaning here. For incomplete, missing data. So data is not always available.

For example, magnitude, I have no record of errors for several attributes, So there's a customer in common in some several data, missing data made due to equipment, infection, and function and the ecosystem was auto recorded data and the data are not due to the misunderstanding and the sudden data may not be considered important at a time of entering and a not register history of change of data.

Missing data may need to be inferred. How can we handle the missing data? So we were trying to engross on the germ urine. It should be very easy to understand. We're just engross them.

Right? If we have not That's how so here it's on when cost level is missing. I went through the classification in this case. So you cannot affect the percentage of missing various per attribute of various considerable.

And, another way, maybe we try to fill in the missing various manually. Maybe some kind of or invisible, but, typically, we may need a drill, like, a feeling like the mean of some missing values. Right?

So if we are in our with, like, a global consider, for example, annoying a new class As I attribute the mean like that, I said, the attribute mean of all samples belong to the same class as model by where The most probable values, so the inference is best, as the patient formula or the decision tree.

There are several ways you can do Autonomy, like, filling some missing values there. What about noise? So the in my the noisy, you just missed the miss that, Azar. Random error or variance in a matter of the variable.

So incorrect, attributes values may be due to the federal data collection instruments. So data entry problems, some data transmission problems, some technology limitation and inconsistency, namely information.

So other data problems with required data cleaning, for example, that duplicated requires incomplete data, inconsistent data, And, has more lessons that we can handle on missing data for, like, billing, like, a first to start the data and partitioning into, equal frequency based.

And then when can smoke be embarrassed, most by being medium, as most by being boundaries, etcetera.

The by clearance was most by frequent of data is on regression punches, clustering is to detect and remove our layers to combine some human inspection.

So, for example, detect some suspicious values and the check by human And the journey with some possible out of here.

In data, this purpose and detection, so So a user will try to use some metadata, like the domain branch, you just dependency distribution. So check the failure overloading.

So Check the UNIX rules, a conservative rule, and a non rule, so use commercial tools. So for the data, Sprablin uses some simple domain knowledge. For example, postcode, spell, check, detect errors and make corrections.

So in the data, auditing analysis by analysis that there's how to discover rules And, relationship to detect some violators, for example, the correlation and cluster to find some outliers, in the data migration on the integration.

So by using some tools, it will allow some transmission transformations to be So the sync file, substrate.

And subsequent, cybersecurity, was, ETF, the transformations of the loading tools allow user to, satisfies, transformations through, graphic user interface Okay.

Then integration of 2 different processes. For example, IT or interactive, different ways how to, interact with strategic process.

So next, we wanna talk about the data integration And, data integration, so it is best to me that we try to combine data from multiple sources I need to go hand the stall.

So the schema indication mean, for example, you were trying to match as a customer customer ID that equal is a decustomize, which means that, indicate the metadata from different sources.

And the entity identified problems, identify from the real entities From multiple data sources, for example, the is equal to the wind.

Is that correct? So detecting some result in some data, value complex for the same value of all the entity. Attribute values from different sources are different.

Possible reasons about different representations, different skills, that, for example, the magic combined with some practice and units. Right. Let's try to say how do we handle some redundancy data.

So redundancy data occur occur open, but in Location of multiple database. For example, in the object, identification, the same attributes or the object may have different names in another table, for example, about annual revenue there.

And, with Redundant to interviews may be able to detect it by correlation analysis and all about the conversion analysis that we just, mentioned before.

So cap careful integration of the data from multiple sources may help reduce, avoid the redundancy And the inconsistency here and the improve the money and speed and the quality.

So next, what we will talk about is, correlation analysis, normal data. So here is a chi square test. That is something you guys familiar with.

Right? So like a cost scale value, the more likely is available, I would like it. So the sale Sales that contribute to, to the most of the customer base. Those who actually companies very different from the expected accounts.

That's the you know, city is up correlating both of the purchase linked to the certain variable that is about the population. So that's the missus. Yeah. No. Yeah. Martin. Miss, courage is here. Right? So he has our class here.

And later, I have another example about it. How about this have a 5 or 10 minutes break? So later, we will try already. How tech rate is the cost there at least. Okay. Okay. Let's have a shot break here. Myself. Why no one tell me?

For all my students, I'm really online. I'm doing it myself for so long time. And so it was during the class break. Okay. Yeah. So okay. That's fine. So now now let's see the answer from Resna. That's correct. Resna. Good.

I remember your name now. So let's say the answer from This answer was that is correct as what we got from you created. So that is a passing about this about this for Why is this some city? This should be correct. Right? Yes. It's good.

Yeah. Except it's a last one. So now I just read you guys can calculate the r chi square Let's go back. Now listen to me. What is a reluctant? What I'm in a place, and there was, like, sign sign science fiction that correlated.

So that is is really missing out. Issues that, I'd like to find inspection and the pictures are correlated in the growth because in using this as you have really have a That's a number here.

Good. So let's continue. So now from this my experiment, example, you guys should know how to pick square. Right? And in the future, maybe we're interested in how can that relationship, but you can also try to take, like, the.

Just choose a relation between, design some functions there. Another way is about the correlation analysis So that's just a month you need to calculate the correlation coefficient is that we mentioned earlier.

Right? It's also named the Pearson product, the moment of television is, like, connected in this ways, and that has just a it's a meaning of a total number 2.

And, a a hat and b hat are the representative is a and b, and the Sigma just the standard division of a and b.

The Sigma just across product. If the r a b is bigger than 0, which means the a and b are positive or correlated, and And, when a and r a b is equal to 0, you use the independent you will see less than 0 with negative or later.

The several ways that you can analyze the relationship and be here.

So in this case, it's just true. We have a different relationship. For example, it is from the negative as kind of parts are showing similarity from negative one to positive one.

So really this now negative one is strong neck, Next two points and the positive one is the strong qualitative relationship between the, axis and y x axis there.

The correlation also can be viewed as a linear relationship. The correlation matters a linear relationship between objects that As I reviewed earlier, right, 2 computer correlation, we we sterilize that and then be understand.

It should be easy to calculate about and be Here. I know. Well, here is the details of a while.

You can take that for various again And the positive covariance and negative covariance and the independent dependency, the others you guys should be familiar with that. Right? So that so we can really, skip about this part.

Covariance. So this is a example that can be simplified in a, you know, computational as this covariance ID that is equal to expectation of expectation of and be the minus this ahead and be happy here.

Detailed time, I wanna let you guys this come out with example.

If you're ready, want to challenge yourself or test yourself whether you understand this coherence, you can do this example by yourself later. Then we wanna talk about the data reduction and the data processing part.

So data reduction, the is are reduced. The representation of the dataset that is much smaller in one body yet produces the same or almost the same And I think our results. So why data reduction? So why we want to do that?

So the database or data warehouse remains stored in, terabytes of data. So complex data analysis may take a very long time to be run or complete the data set. That's the reason why we want to, reduction of data. Yeah.

So, that dimension narrative reduction measures, for example, the, the PCA, and there's, feature corrections, feature subset actions, numerous key reflections or regression and the log linear models, the histogram, clustering, Then the cube aggregation, so data compression, I will first try to quickly give you a overview of those different measures.

And then in the last side, we already talk about the regression a little bit You're finally there.

So let's continue. So first of all, the dimensionality reduction, I think you guys noticed about the cross of dimensionality. So when the dynamic net increase, the data becomes increasingly sparse.

And the density and distance between points, which is Critical to clustering or offline analysis become less meaningful, the positive combination of sub service will go, exponentially, dimension energy reductions avoids the cross of dimension energy have eliminated the Irrelevant features and reduce noise, reduce time, and the space required at data mining allow easier visualization by machine learning detection techniques, Now, for example, relate to some PCA to start up.

And this is about PCA, so that means we want to find, projections that capture the largest amount of variation in data.

So in our genome, data are projected on a smaller space resulting in dimensionality reduction. We find the eigenvectors and of the covariant metrics, and there's also eigenvectors defined on the new space.

So, actually, you can see it's okay as a wrong, like, a major access of the test sets there. For the PCA, so I would like to skip this part a little bit because, And then we have another specific cost that we will review it again.

So let's try to use some skip that. So another way to reduce the dimensionality of data also So it is like, for example, you redundant attributes so duplicate much of all of the information contained in one or more other attributes.

For example, purchase price of, product and the amount of sale tax paid. You let you variant attributes, which means that Content. No information that is used for the data.

I mean, mining task. I had for example, student ID is often irrelevant to the task Outriguing a student in TGA. Right? I agree with that. Yes. You are quick. Heuristic search in the educator selection.

And, the there are 2 of, how of the possible attributes combination of the attributes, Typically, a characteristic attribute selection method, for example, best single attribute on the attribute independent assumption chooses by significant tests.

So by stepwise feature selection, those are best of in a single attribute is in the pick the first and next best, attribute of condition and effects.

And the stepwise attribute elimination, so repeated the animated, animated, worst attributes and the best combined with the attribute selection and the elimination.

So optimal branch and the bond. So use some attribute elimination and, backtracking some structure to do that. And, even for the attribute of creation, for specific way, use a feature generation.

So general and correct new attributes features that are coming captured, that important information in your data set, more effectively is that original ones.

There are 3 generated methodologies, for example, attribute the attraction to domestic recipient and maintain that into use case, seem like a data reduction problem by those kind of measures.

Attributed to cons construction combined features, Discriminated frequent embarrass in in some kind of chapter 7 there, and then this particularization there.

There are several methods that we can do that. The second, data dimensional type of clause, numerology reduction.

So we just data 1 by choosing the other net, you are some Smaller forms of data representations, parametric measures, for example, is this regression and non parametric measures, not assume models.

And your family's for, like, a Instagram, clustering, some mechanics stuff.

As the first about, first of all, our primary models is different And revised model, margin type, regression about linear regression there And, let's also skip this regression a little bit because, we will cover to the details later about Linear regression should be is easy one.

Marketing up equations, the money variable equations should be have money variables, you know, log linear models There.

And, here's about the histogram analysis. So you buy the you buy data and you track it and install the averages somehow each bucket, which means that this is like Market that we see is like you're like, each of being different beings.

Right? Each of being, for example, one thing can be represent, like, 1 to 10 like, 10 to 20 is, like, equal each of them. Right? Because that even calculates the frequencies of each of them.

That's it. The partition rules like the equal y was the means the equal bucket range. I mean, for example, this which that means that each of the whites of it should be equal.

Another way is like, equal frequency or equal this, which means, like, this cannot should be equal. This frequency should be equal for all different events.

So there's 2 different rules for Participation. Right? So now we're talking about the clustering of parties partition data into clusters based on similarity and start a class faster representations.

For example, central and, diameter only can be very effective if data is clustered. Why not? So data is Smart there. Can have, hierarchical, clustering and the b story in market dimension index structures.

There are many choices on cluster is definitions and clusters algorithms across the And lastly, so we have we study the in the depths in Dutch, chapter 10 there.

And in the assembly, assembly should be opt obtain a small number of paths to represent the plot at the end and allow Finding algorithm to run the complexity that is potentially sublinear to the size of the set.

The key principle here, try to choose our The intentives are subset.

I mean, they said simple random something may have very poor performance in the As I was through that, adaptable assembly message that is struck by assembly noted the separate may not reduce the data set.

So you're in for the sampling. The end is trying to give you a call that's at, during that if you find some parties with that.

And there are different, types of sampling, like, the simple random sampling And the sample without replacement and the sample with replacement and the strap by the sampling of the dataset The 2 are samples from each partition.

It is proportionally or approximately the 7% of that on the user conjunction with the screw data there.

Has some, quality about the ways of resolving. For example, this is our data. I can use Simple random, sample without replacement on that.

For example, case this case, and then there's some other. Yeah. And, for the sampling of clustering or straight file sampling. So this is a raw data again for clustering and basically 1 trigger classified into your small samples there.

For the data cube, aggregation, so the lower level of the data based on, Q Quboy and, the aggregated data of for an individual entity of interest.

For example, a class customer, you know, phone calling data warehouse. There are not particle levels of aggregation in data tubes Further, reduce the size of data engineering risk.

And we further reference, our provide and they also use the smart most, recommendation, which is not to solve the task. Parts regarding the aggregated information should be answered. They use some data for when and this first drop.

And there's a Australia, compression. So they are extensive, theorists, and we are doing the algorithms to control loss by Only limited manipulation is possible without explanation. For example, you know, audio or video compression.

So, typically, we're more likely, have a lossy, come compression with a progressive refinement. Sometimes, small fragments of a signal can be reconstructed or resolved or constructed the whole.

Right? So for the time sequences, as not audio of such rigorous shock, I'm very So it was time. So time measurement and, you must be a production man. Also, you can see that as a form of, compression here.

Has not a graph about data compression. This is like a original data. After compression, it's gonna move some some wanted to know some information about the process. Just know of the scale.

Just know of the scale. Just know of the scale. Just know of the scale. So, actually, yeah, the way that I said it is correct. I suppose, basically, if we have some kind of or do you get that once we use some mussels or all results?

We get some compressed data. Right? If it is above most of the messes, we can also recover the Confess data back to the original data arrived. Like, sometimes in your computer, you Zip a file.

Not a file. Eventually, the ends of it, so you get some kind of similar information. But for the last thing, this means that if you Compressed, your data, for example, your compress, your bare b b, you just model it.

So you measure who will have lost the revolution of your visual video. Right? So this is a causal Mhmm. So lastly, we were we're gonna mention about the data transformation and data dispartation theorem.

A function that maps the entire set of values of a given attribute to a new set of replacement values that is the agent's own value can be Identify with the new pairs.

There's a message, for example, smooth that this is the pre check constructions, aggregations on normalization, SMM memax, normalization, 0 score normalization, normalization by the, decked models for scaling, and, discretization is the concept, how does it, can be?

Also, stop. And he has more details about the mimics normalization And, 0s for normalization and normalization by tech remote and security by so all different ways.

And, for the first one, second one, those two messages should be some frequently used normalizing methods.

Yeah. But this parallelization is yeah. That's the 3 different time attributes for a nominal, ordinal, or numerical. For example, A number of mister Bayes from the I started, like, color, some profession.

As I'm a ordinary mister Bayes from the all the The theory of the academic rank and numerical, just real numbers should be very easy to understand, like, integer or real numbers, in the discussion or divide the value of a continual attributes in the intervals.

But interval level can be used to replace replace the act database, reduce the data size by. And surprise, we have comparison. Place the top down to merge and the dispartation can be performed because of the.

And attributes are prepared for a quasi analysis, for example, like Just kinda know your audio that had to some kind of numerical presentation. But, eventually, you cannot, like, input our text or it's through some remote.

Right? There should be definitely some numbers. And, yeah, just some measures, like, typical, like, opinions, some histogram analysis across Koreans and the synthesis, like, and that we test we've used before.

And then we have echo or whites of the partition of the histogram and the echo that's a pretty specific partition in the histogram.

And, has been a message according to the most, and he has just some kind of detail to to play the with that. For example, pin 1 should be like, this number.

This is really and you focus most by being missed or be like a lot of different 3 piece. If we try this most by being boundaries and, like, in this way. So right. There are different ways that we can. 3 of the data there.

And, yeah, just about the realization about different base, in terms of, like, this has been miss you're gonna miss the equal of the frequency. And this is the k means clustering leads to the better results eventually like that. Yes.

About some of our classification, for example, this is, like, you know, supervised The entropy determines split of points, top down, reverse split, and the correlation analysis that we already mentioned this class where This is related to the start.

The concept, the error rate generation, which just means that some concepts I should be the virus in there.

Herrickly, and, is usually associated with each dimension. Yeah. Data or warehouse, concept of hierarchy. Facilitators actually, a ruling in your data warehouse to make the data in multiple generative and reality.

Concept, hierarchy of formations recursively reduce the data by collecting or replacing our low concepts such as this numerical values for h And the high level concepts such as ROS, and the concept of can be So inspired by domain experts are and the data warehouse designer, the concepts of error can be automatically Form the well, both numerical and, nominal data for new numerical data.

So we were just using discretizing methods shown before. And the sophistication of, partition of, attributes, explicitly at the schema level by user.

For example, like, street is always smaller cities. The, country and the The, country and the specification of our of our set of various by explicitly that are growing. I use this kind of different cities in the in the state.

I know it's a specification, but only a part of a set of Attribute for example, only a street should be near a city, the others, then the automatic or generation of errors, attribute that was, by the analysis on number of just picking various, for example, for a set of attributes, a straight set of, cities, state, and the country.

So he has Just a barber.

An example of our all one is different. The case is when a country, like, how many progress, or how many states do you have, many cities, how many streets there. So let's just, quickly review about, the, you know, data processing.

It's really embarrassing that we mentioned today, like, the data qualities and data cleanings and data integration and data reduction, And, data transformation and data discretization.

I would like to say, Up to now, we already cover some very basic concepts in the Shenanigans.

So we already used to bang, around, like, 2 courses, we cover about this mass background. And also we couldn't imagine about those data preprocessing.

And those Some things that you guys should already know. Right? You shouldn't have any difficulties from now. And, from, I really from next, week, we we are ready to learn some machine learning algorithms in details.

So, today, maybe we have a little bit of time. Let's try to start with, little bit about the regression stuff. Let me try to share my. So now let's see all about the regression.

So we are very We could talk about some difficult difference between regression and the classification, then we are mentioned about the simple linear regression, Partially linear regression, some polynomial regression, some regression metrics, some classification metrics that you guys should know about that.

So now What is, what are the difference between regression and the classification? Regression returns, was Some increase on ticket value for and most vacations.

It's, that's a good situation to find our Sample to Sound Plus. If you could define classification or most classification, but task is to define all the data to Sound Plus.

Yeah. That's good. We'll answer. What about other thoughts? What are the difference between regression and the classification? They don't know what is classification?

Yeah. Okay. We catch a big surprise here. What about and as a and as a Also, so how about that given speaking regression and classification? I think for regression is that your predict Predicted target is the numerical data.

And for the classification is the, Like, your predicted target is, is not the numerical data. It it is the It is the ordinary or the or the classification data.

Is that correct? Yeah. It's kinda so I will answer this question later. So now let's first talk about this classification. The classification is ready. We want to classify things classification.

So now let's talk about what are the differences between regression and classification. For classification so, actually, The output is none. This is something you needed to man, to needed to remember.

And for regression, the output is very continuous number. There can be any continuous number. Right? But for the classification, it can be just a norm, which means that it can be 1 or Hello. Are we patient? Can can you use number?

You get the difference here? Okay. Good. For the regression, so the, typically, the easy one is just a normal linear regression, which means just So why then could you could you add x plus bit? What about the market pattern, regression?

That means that just that we have a different, regression in mice, different variables in the regression. But about binomial, binomial is missing We have different degrees of the regression. Right?

As you can see, you know, the logic is in this review, polynomial. And about the classification, some measures like, classification, It's like k k nearest neighbor or some decision tree classification, some random cross classification.

So during the classification, we're really more likely to find this boundary if you give it a class. Right? But, of course, our regression, we really want to figure out data. Right. Do you see the difference from the chat?

Any questions so far? Hello? Okay. Good. Let's continue. Our linear equation is the linear model. For example, our model is that assumes linear relation between the input variable x and the single output variable y.

So more specifically, that is y can be calculated from a linear combination of the input of variable x. So what is the a single input of variable x?

The method is to refer to a simple linear equation. Simple linear equation is we only have single input of variable x. Once there are multiple input variables that often refer to multiple linear regression.

Yeah. So now this is the so called the single linear regression problem, I see what x, a single y. The form should be y equal to b zero plus b one times with x one.

So here, x one, just so called an independent variable or IV here. The dependent variable called DV here, it should be y. So p zero is constant, and the v y is the coefficient.

So it doesn't work. What about the b zero? For sure. Yeah. In the sun. Correct? So, again, One kind of model about to predict your salary with your different experience of years. For example, here.

They couldn't. The salary is able to b 0 plus b 1 times with a variance, and then you can try to predict How much, Sarah, we are we increased during the years. Right? I think it's typically it should be just a simple My name is right.

So let them that was a good with the different years. We have a lot of that. We'll have made a goal increase above the setup. No question. Why? Yes. Okay. Good. And then here's another example.

You can just Try that. Are you interested to, calculate as a number of UHERs with the salary? So have salary again. Just, give me give me is what? I'm very good. I'm very good. I'm very independent variable.

And I guess just the case is that you can't perform the same for linear regression. For example, I didn't call the data. For example, you you were trying to import import this kind of data, and there's there is no missing data.

You can try to spread the if you're chatting on the test, Then you will feature scaling by a list of preferred here, and then you will try to correlate it to the scenario with the variance.

You try to You have predictions, then you will try to verify your predictions, then try to evaluate how good it will be on test. Let's say it's about the sustainable linear regression. What about a machine in a regression?

What about multi variable input variables. Right? So here, for x1, x2 to x, and all of those are called so called the, independent variables. Again, we have whether you have s, independent variables, r v s here.

And the for the b1, b2 to bn is so called the coefficient, and the for the b0, it's still actually been constant. And, the y is a signal that dependent on variable here.

Right? So this is about called multilinear regression. Yeah. Multiple linear regression applies at least 2 variables, which can be nonlinear, ordinal, or integral or rational level variables.

And I rule out some of the same sizes that the regression analysis require at least 2, 20 cases of variable from that analysis.

So So first, the marginal linear regression price, an issue if it's a independent, and the dependent of variables to be Linear. The the linear assumption, can best be tested at the sky cross.

The pollen tree, clouds, hopefully, we have this So I'm, I would now have the path. But anyways, here's another example that you can find you with the market Linear equation. So for example, in this case, what are the inputs?

Variables. And but it's all I was it's a problem. This year. Left before. The left before. Yes. That's correct. And the y From the the y just Okay. It's right. So now you can see that for the first three, it should be easy.

Right? Because they're just the numbers. But what about this this states? Probably one way is to try you needed to donate to this different number. For example, you have say it can be 0, and it can be 4, it can be 1.

Whatever, it can be 2. So you can, table. So then, over here, the UI is just specific. It can be represented about dumb variables. Right? So this is about multilinear regression. And, again, it's another case of a polynomial regression.

Regression equation is a polynomial regression. You could if it's a power of independent of variable is more than 1. And as I you create it below, you present as a binomial. That is y equal to a classically times with as a x over of 2.

Right? I'd spare. Right? So this is about, y case, the the polynomial regression. Once there might be a a temptation to fit a high We are a phenomena to correct at, lower errors, so this can be resolved in overfitting.

Always a process of eventually, but you see the fit and and the focus on, make Sure that the curves face the nature of the problem. So here, example about underfitting, just right on overfitting.

I think you guys should have familiar with this kind of concepts. And always, we probably want to achieve in the middle and on stage. Like, just the right. Right? And these are not on the video or not over video.

So here, it's about some metrics, about the regression problems. So for example, MAE, But, anyway, you need to get a small, like, of this different metrics. And, what about classification metrics?

So a classification task can be really replicated in many different ways to achieve the specific objects. Of course, the most important metric should be the accuracy. Right? Which is, like, donating in this way.

The number of as by divided by total number of samples that use a month one minus to this dash. Right? In another way, which means that The number of tracking the samples divided by the total number of samples.

Right? That's it. It's called accuracy. Right? I mean, there's no problem on this one. Okay. So in a cycle that can be accessed by using some building accuracy score 1, you call 1 function.

So then this is the function, and this is the y test, and this is the branches you have. And here is the prediction values. For example, we can get the video. 0. 9 quality is pretty high, but good accuracy.

Another way is to try to see about the pure metrics. Right? It is necessary to be able to differentiate between different kinds of, misclassifications. So we are considering about the binary case with conventional notation.

For for example, the 0 here, I just missed And why means the positive? Because the, relations, our relative weight is 5 different. This reason, we introduce the following definitions.

True positive is a positive sample correctly classified. The first positive means a negative sample classified as positive. And the true negative is a negative sample or correctly classified by.

Read again, try to map over this true, positive, or the true negative, and what about false positive, and false negative. Right? Because, computer metrics just, consists of all these different too.

So so last thing again. So for the first positive and the first negative, this is as the why I think about me In the many, radical prediction now why our first pass positive can be easily discovered by the protest.

Our full activity is all the neck because, the we for communications, following the consequence of this action.

For this reason, it is useful to introduce the concept of computer metrics. But typically, you can't just every it is a binary classification. Here is the computer metrics.

So here, just a poor kind of numbers there. For example, in fact, you can also try call 1 line code up from this, second, the metrics, and you call it the computer metrics. So here is how we call it, take it as a con configure metrics.

That is, y choose equal to 1 choose and the vertices of predict. Zhang will assure you how this is the If you're a metric spectrum, and we will have a 5 positive and do 4 positive.

We don't want to get combined. Anyhow, how do you take the address if that's not your PureMetrics? We can, put this Yeah. This is like because it's not I don't know.

It's something you want to be measured. I know. And guess from you? So you're getting the the sum of the diagonals or this is called? Yes. That's correct. So, Actually, it's just a 50 plus 68 then divide the sum of all that.

You take the the diagonal elements then divide the all that together as an accurate. That's it. Good. So, another use for directions is trying to measure some other metrics such as the procedure.

That means that true positive divide by the, true positive plus or false positive. And, also, you can calculate another one that says, recall recall has just equal to the true Positive divide by the true positive plus the positive.

And, this is also some useful up there. And here is also one another line. A one line code, you can call this kind of function from second learning part of the senior score, then you get a bari score.

And This ability to detect true positive examples among all the potential quality can be accessed using another mirror.

What about the ROC curve? So the ROC curve means all this receiver operating characteristics is available to to compare different class y that assign a score to their predictions.

So in general, this score can be interpreted as a probit. So as a bond bounded between 0 and 1, the plane is reflecting the light of the following figure in this case.

In the x axis means the false positive rate or specificity, and the true the y axis means the here. Right? All called the sensitivity. And, here, if you have a random guess So, yes, so have a random performance. It's just like this.

Pro lines are results of random gas. Random gas, you have 0. 5 Let's say if we go live, it's very close to this part, you will get a very good performance. Okay. Good. I think, most of you guys should know about policy code.

Right? Pleasure. This something that I already mentioned. What about the dash, line, represents, perfectly random class bias is out of the curves below this. The study called the form was the random choice, and the y is up.

1 above it performs better across the best class by has the odds that's paid as a segment to 0 0 between 0101011. And the algorithm should have been as close as a positive to this limit.

So, actually, we want to find a line that is very close to this one. Very close to this one. 01 and 11. Yeah. Get it. You got it right. To show how we create our secret with cycle learning, we are going to have a lot.

That's kind of several lines. You will know how to get the policy curve by going to try a model to detect the score for the predictions, and it can be achieved using some function or predict the method.

For example, in this case, just a few line of codes. No matter this is the first line just about how you separate the data to 0. 725 as a test of 0. 75 as the training.

And then, so in this case, we use a yellow logistic regression model LRM, and we're trying to fix the x-ray and the y-ray. So that way, we'll gather some predictions from another model so that we can compare the RC curves for them.

Also, one line of code. Okay? We call this RC curve so that you will try to process everything for you. And then it is also useful to compute those stuff, and this is also another way.

So let's say Let's see what's the curves that we just calculate about in the previous example of here. So by details of on the figure. So now is that is the figure that we really want to get again.

The x as this is just a positive positive rate, and the y axis presents for the true positive and then here this blue line again is about random guess and this red line is the results from the Not just a regression.

What does this mean? This means what? You have. Right? I want to get some answer from you.

What does this mean? Yes. That's correct. So And, also, I think for those our online students usually get it because it is very good because, it is in this area, and also the l series voice achieves 0.

97, which is pretty much, right, in many classification model. I think this actually be too much for today. I think you guys, have a very brief understanding about the regression and This, classification metrics.

And the start from, next week, we'll be ready to do some Hello? A little bit difficult understanding about mass stuff. For example, about logistic regression. I feel we already have this Can I process about logistic regression, Ryan?

Sounds a little. A little. No. No. I had about logistic regression Awesome. But it was more like the the impression. Okay. Is that the color of. Okay. Okay. Good. So then, you know, next week, We're ready.

We all go through that details about just the regression on how to calculate those different equation and step by step. And also, if we have time, we will cover you guys already know how to estimate the parameters.

I'm a put down on a and b. And do that. So I guess I see some confusion. That's mad. So how will I cover I still have to try and cover that part as, That will be a kind of another kind of view of, at least the square estimation.

Hopefully, you guys but I'm happy about that. Again, in this case, we'll have another homework 3 for this week. This case, it should be another very easy homework for you guys.

Question is about how to change our classification problem to your best problem and how we change our request problem to our classification problem. Yes. More than 300 watts there.

Well, chat. Is also No. Not. I'm not. I'm not. I'm not. I'm not. I'm on. I'm on. Because, You know, first of all, Mark, somehow you're very clear, like, why are the answers to me? There's none in that. Okay? There's 2 shots.

It's not a sort of idea. Right. So now I really want to think about your idea about how to change, for example, in classification problem, how can you change your original Problem give us they'll give me an example how you can do that.

And You guys have some kind of problem regarding this question as initially preparing me. Well, I guess, Now let's say, how do we take a classification to regression?

I'm gonna have you hear some answer from both, Just guess. Singapore. How to change your classification plan to have a recent model? Alright. Mhmm. You need to know about what the difference classification and the regression.

You guys already know. Right? So now classification is just the more like the one history for right? And father, our aggression is more like what? 1. 0. 1. 0 or so it was 0. 0. It's right there. Yes. Can be a white example.

Yeah. Similarly, you can try to say something you understand and the right line, example about how you do it. Okay. I think, yeah, that's, for today. For our online students, do you have any other questions? Well, I have one question.

It's about, well, I think the professor gave us a lot of, like, index such as the ARC or the AUC or the The the recall. And is there any standard that we can follow to use, which parameter to optimize our model.

Well, Let's for these various, usually, for, let's say, our traditional methods like those stuff, we did not use those metric to optimize on board.

We just use it to evaluate our model. You notice that, for example, in this do you see it like this logistic regression?

This model in itself is not a reliable on this RC curve. No. It's not rely on that. The RC curve here, just a metric to report How good is your linear graduating is?

Yes. But, I think my, my question is, when I try to, like, change or fine tune the the parameters in the model and which index that I Shouldn't I compare between the previous one and and the present one?

Okay. Yeah. Because you see that too many metrics.

Right? But, anyway, the first one, for example, for the classification, the accuracy, your issue would be the first one you need to report. So, Zelle, I was asking you can just because this ROC code will stop.

There's something people can't understand it. Okay? Definitely, you choose the accuracy as the first one, then you try to process on those kind of obviously curves, little stuff, people will understand how good Your model is.

Okay. So so we so we usually choose the accuracy curve. We we like to choose the accuracy. Accuracy. Right. In your table, we can see lots of accuracy directly.

So is that how is that where we're gonna see this figure. Yeah. Okay. But there's an issue that, if our dataset is, like, Unbalanced data. It it will causes the accuracy will be high, but but we can't find.

Yeah. In that case, you need to report the accuracy of each class. All the accuracy for each class. It's like Yes. For example, like, class 1 and class 2 and overall class. So in that case, people will all really understand the Yeah.

Performance for each, each class. And, also, you can try to report the accuracy, the the computer metrics. Right? Because from computer metrics, Then we'll tell you which class is good or not.

Right? If we even it is imbalanced, we can from computer metrics, we can tell. Okay. Got it. Thank you. Welcome. I guess no question. Right? So then we have finished today, and I think I'll see you guys tomorrow, man. Thanks. Awesome.