MachineLearning-Lecture04

Instructor (Andrew Ng):Okay, good morning. Just a few administrative announcements before we jump into today's technical material. So let's see, by later today, I'll post on the course website a handout with the sort of guidelines and suggestions for choosing and proposing class projects.

So project proposals – so for the term project for this class due on Friday, the 19th of this month at noon – that's about two weeks, two and a half weeks from now. If you haven't yet formed teams or started thinking about project ideas, please do so.

And later today, you'll find on the course website a handout with the guidelines and some of the details on how to send me your proposals and so on.

If you're not sure whether an idea you have for a project may be a appropriate, or you're sort of just fishing around for ideas or looking for ideas of projects to do, please, be strongly encouraged to come to my office hours on Friday mornings, or go to any of the TA's office hours to tell us about your project ideas, and we can help brainstorm with you.

I also have a list of project ideas that I sort of collected from my colleagues and from various senior PhD students working with me or with other professors. And so if you want to hear about some of those ideas in topics like on natural [inaudible], computer vision, neuroscience, robotics, control. So [inaudible] ideas and a variety of topics at these, so if you're having trouble coming up with your own project idea, come to my office hours or to TA's office hours to ask us for suggestions, to brainstorm ideas with us.

Also, in the previous class I mentioned that we'll invite you to become [inaudible] with 229, which I think is a fun and educational thing to do. So later today, I'll also email everyone registered in this class with some of the logistical details about applying to be [inaudible]. So if you'd like to apply to be [inaudible], and I definitely encourage you to sort of consider doing so, please respond to that email, which you'll get later today.

And finally, problem set one will also be posted online shortly, and will be due in two weeks time, so you can also get that online.

Oh, and if you would like to be [inaudible], please try to submit problem set one on time and not use late days for problem set one because usually select [inaudible] is based on problem set one solutions. Questions for any of that?

Okay, so welcome back. And what I want to do today is talk about new test methods [inaudible] for fitting models like logistic regression, and then we'll talk about exponential family distributions and generalized linear models. It's a very nice class of ideas that will tie together, the logistic regression and the ordinary V squares models that we'll see. So hopefully I'll get to that today.

So throughout the previous lecture and this lecture, we're starting to use increasingly large amounts of material on probability. So if you'd like to see a refresher on sort of the foundations of probability – if you're not sure if you quite had your prerequisites for this class in terms of a background in probability and statistics, then the discussion section taught this week by the TA's will go over so they can review a probability.

At the same discussion sections also for the TA's, we'll also briefly go over sort of [inaudible] octave notation, which you need to use for your problem sets. And so if you any of you want to see a review of the probability and statistics pre-reqs, or if you want to [inaudible] octave, please come to this – the next discussion section.

All right. So just to recap briefly, towards the end of the last lecture I talked about the logistic regression model where we had – which was an algorithm for [inaudible]. We had that [inaudible] of [inaudible] – if an X – if Y equals one, give an X [inaudible] by theta under this model, all right. If this was one over one [inaudible] theta, transpose X. And then you can write down the log like we heard – like given the training sets, which was that.

And by taking the riveters of this, you can derive sort of a gradient ascent interval for finding the maximum likelihood estimate of the parameter stated for this logistic regression model.

And so last time I wrote down the learning rule for [inaudible], but the [inaudible] has to be gradient ascent where you look at just one training example at a time, would be like this, okay. So last time I wrote down [inaudible] gradient ascent. This is still [inaudible] gradient ascent.

So if you want to favor a logistic regression model, meaning find the value of theta that maximizes this log likelihood, gradient ascent or [inaudible] gradient ascent or [inaudible] gradient ascent is a perfectly fine algorithm to use.

But what I want to do is talk about a different algorithm for fitting models like logistic regression. And this would be an algorithm that will, I guess, often run much faster than gradient ascent.

And this algorithm is called Newton's Method. And when we describe Newton's Method – let me ask you – I'm actually going to ask you to consider a different problem first, which is – let's say you have a function F of theta, and let's say you want to find the value of theta so that F of theta is equal to zero.

Let's start the [inaudible], and then we'll sort of slowly change this until it becomes an algorithm for fitting mass and likelihood models, like [inaudible] reduction.

So – let's see. I guess that works. Okay, so let's say that's my function F. This is my horizontal axis of [inaudible] of theta, and so they're really trying to find this value for theta, and which F of theta is equal to zero. This is a horizontal axis.

So here's the [inaudible]. I'm going to initialize theta as some value. We'll call theta superscript zero. And then here's what Newton's Method does. We're going to evaluate the function F at a value of theta, and then we'll compute it over to the [inaudible], and we'll use the linear consummation to the function F of that value of theta. So in particular, I'm going to take the tangents to my function – hope that makes sense – starting the function [inaudible] work out nicely.

I'm going to take the tangent to my function at that point there to zero, and I'm going to sort of extend this tangent down until it intercepts the horizontal axis. I want to see what value this is. And I'm going to call this theta one, okay. And then so that's one iteration of Newton's Method.

And what I'll do then is the same thing with the dec point. Take the tangent down here, and that's two iterations of the algorithm. And then just sort of keep going, that's theta three and so on, okay.

So let's just go ahead and write down what this algorithm actually does. To go from theta zero to theta one, let me call that length – let me just call that capital delta.

So capital – so if you remember the definition of a derivative [inaudible], derivative of F evaluated at theta zero. In other words, the gradient of this first line, by the definition of gradient is going to be equal to this vertical length, divided by this horizontal length. A gradient of this – so the slope of this function is defined as the ratio between this vertical height and this width of triangle.

So that's just equal to F of theta zero, divided by delta, which implies that delta is equal to F of theta zero, divided by a prime of theta zero, okay.

And so theta one is therefore theta zero minus delta, minus capital delta, which is therefore just F theta zero over F prime of theta zero, all right.

And more generally, one iteration of Newton's Method precedes this, theta T plus one equals theta T minus F of theta T divided by F prime of theta T. So that's one iteration of Newton's Method.

Now, this is an algorithm for finding a value of theta for which F of theta equals zero. And so we apply the same idea to maximizing the log likelihood, right. So we have a function L of theta, and we ant to maximize this function.

Well, how do you maximize the function? You set the derivative to zero. So we want theta [inaudible]. Our prime of theta is equal to zero, so to maximize this function we want to find the place where the derivative of the function is equal to zero, and so we just apply the same idea. So we get theta one equals theta T minus L prime of theta T over L double prime of T, L double prime of theta T, okay.

Because to maximize this function, we just let F be equal to L prime. Let F be the [inaudible] of L, and then we want to find the value of theta for which the derivative of L is zero, and therefore must be a local optimum. Does this make sense? Any questions about this?

Student:[Inaudible]

Instructor (Andrew Ng):The answer to that is fairly complicated. There are conditions on F that would guarantee that this will work. They are fairly complicated, and this is more complex than I want to go into now. In practice, this works very well for logistic regression, and for sort of generalizing any models I'll talk about later.

Student:[Inaudible]

Instructor (Andrew Ng):Yeah, it usually doesn't matter. When I implement this, I usually just initialize theta zero to zero to just initialize the parameters to the – back to all zeros, and usually this works fine. It's usually not a huge deal how you initialize theta.

Student:[Inaudible] or is it just different conversions?

Instructor (Andrew Ng):Let me say some things about that that'll sort of answer it. All of these algorithms tend not to – converges problems, and all of these algorithms will generally converge, unless you choose too large a linear rate for gradient ascent or something. But the speeds of conversions of these algorithms are very different.

So it turns out that Newton's Method is an algorithm that enjoys extremely fast conversions. The technical term is that it enjoys a property called [inaudible] conversions. Don't know [inaudible] what that means, but just stated informally, it means that [inaudible] every iteration of Newton's Method will double the number of significant digits that your solution is accurate to. Just lots of constant factors.

Suppose that on a certain iteration your solution is within 0.01 at the optimum, so you have 0.01 error. Then after one iteration, your error will be on the order of 0.001, and after another iteration, your error will be on the order of 0.0001. So this is called [inaudible] conversions because you essentially get to square the error on every iteration of Newton's Method.

[Inaudible] result that holds only when your [inaudible] cause the optimum anyway, so this is the theoretical result that says it's true, but because of constant factors and so on, may paint a slightly rosier picture than might be accurate.

But the fact is, when you implement – when I implement Newton's Method for logistic regression, usually converges like a dozen iterations or so for most reasonable size problems of tens of hundreds of features.

So one thing I should talk about, which is what I wrote down over there was actually Newton's Method for the case of theta being a single-row number. The generalization to Newton's Method for when theta is a vector rather than when theta is just a row number is the following, which is that theta T plus one is theta T plus – and then we have the second derivative divided by the first – the first derivative divided by the second derivative.

And the appropriate generalization is this, where this is the usual gradient of your objective, and each [inaudible] is a matrix called a Hessian, which is just a matrix of second derivative where HIJ equals – okay.

So just to sort of – the first derivative divided by the second derivative, now you have a vector of first derivatives times sort of the inverse of the matrix of second derivatives. So this is sort of just the same thing [inaudible] of multiple dimensions.

So for logistic regression, again, use the – for a reasonable number of features and training examples – when I run this algorithm, usually you see a conversion anywhere from sort of [inaudible] to like a dozen or so other [inaudible].

To compare to gradient ascent, it's [inaudible] to gradient ascent, this usually means far fewer iterations to converge. Compared to gradient ascent, let's say [inaudible] gradient ascent, the disadvantage of Newton's Method is that on every iteration you need to invert the Hessian.

So the Hessian will be an N-by-N matrix, or an N plus one by N plus one-dimensional matrix if N is the number of features. And so if you have a large number of features in your learning problem, if you have tens of thousands of features, then inverting H could be a slightly computationally expensive step. But for smaller, more reasonable numbers of features, this is usually a very [inaudible]. Question?

Student:[Inaudible]

Instructor (Andrew Ng):Let's see. I think you're right. That should probably be a minus. Do you have [inaudible]? Yeah, thanks. Yeah, X to a minus.

Thank you. [Inaudible] problem also. I wrote down this algorithm to find the maximum likely estimate of the parameters for logistic regression. I wrote this down for maximizing a function. So I'll leave you to think about this yourself.

If I wanted to use Newton's Method to minimize the function, how does the algorithm change? All right. So I'll leave you to think about that. So in other words, it's not the maximizations. How does the algorithm change if you want to use it for minimization? Actually, the answer is that it doesn't change. I'll leave you to work that out yourself why, okay.

All right. Let's talk about generalized linear models. Let me just say, just to give a recap of both of the algorithms we've talked about so far. We've talked about two different algorithms for modeling PFY given X and parameterized by theta.

And one of them – R was a real number and we are sealing that. And we sort of – the [inaudible] has a Gaussian distribution, then we got [inaudible] of linear regression.

In the other case, we saw that if – was a classification problem where Y took on a value of either zero or one. In that case, well, what's the most natural distribution of zeros and ones is the [inaudible]. The [inaudible] distribution models random variables with two values, and in that case we got logistic regression.

So along the way, some of the questions that came up were – so logistic regression, where on earth did I get the [inaudible] function from? And then so there are the choices you can use for, sort of, just where did this function come from?

And there are other functions I could've plugged in, but the [inaudible] function turns out to be a natural default choice that lead us to logistic regression. And what I want to do now is take both of these algorithms and show that there are special cases that have [inaudible] the course of algorithms called generalized linear models, and there will be pauses for – it will be as [inaudible] the course of algorithms that think that the [inaudible] function will fall out very naturally as well.

So, let's see – just looking for a longer piece of chalk. I should warn you, the ideas in generalized linear models are somewhat complex, so what I'm going to do today is try to sort of point you – point out the key ideas and give you a gist of the entire story. And then some of the details in the map and the derivations I'll leave you to work through by yourselves in the intellection [inaudible], which posts online.

So [inaudible] these two distributions, the [inaudible] and the Gaussian. So suppose we have data that is zero-one valued, and we and we want to model it with [inaudible] variable parameterized by phi. So the [inaudible] distribution has the probability of Y equals one, which just equals the phi, right. So the parameter phi in the [inaudible] specifies the probability of Y being one.

Now, as you vary the parameter theta, you get – you sort of get different [inaudible] distributions. As you vary the value of theta you get different probability distributions on Y that have different probabilities of being equal to one. And so I want you to think of this as not one fixed distribution, but as a set where there are a cause of distributions that you get as you vary theta.

And in the same way, if you consider Gaussian distribution, as you vary [inaudible] you would get different Gaussian distributions. So think of this again as a cost, or as a set to distributions.

And what I want to do now is show that both of these are special cases of the cause of distribution that's called the exponential family distribution. And in particular, we'll say that the cost of distributions, like the [inaudible] distributions that you get as you vary theta, we'll say the cost of distributions is in the exponential family if it can be written in the following form. P of Y parameterized by theta is equal to B of Y [inaudible], okay.

Let me just get some of these terms, names, and then – let me – I'll say a bit more about what this means. So [inaudible] is called the natural parameter of the distribution, and T of Y is called the sufficient statistic. Usually, for many of the examples we'll see, including the [inaudible] and the Gaussian, T of Y is just equal to Y. So for most of this lecture you can mentally replace T of Y to be equal to Y, although this won't be true for the very fine example we do today, but mentally, you think of T of Y as equal to Y.

And so for a given choice of these functions, A, B and T, all right – so we're gonna sort of fix the forms of the functions A, B and T. Then this formula defines, again, a set of distributions. It defines the cause of distributions that is now parameterized by [inaudible].

So again, let's write down specific formulas for A, B and T, true specific choices of A, B and T. Then as I vary [inaudible] I get different distributions. And I'm going to show that the [inaudible] – I'm going to show that the [inaudible] and the Gaussians are special cases of exponential family distributions. And by that I mean that I can choose specific functions, A, B and T, so that this becomes the formula of the distributions of either a [inaudible] or a Gaussian.

And then again, as I vary [inaudible], I'll get [inaudible], distributions with different means, or as I vary [inaudible], I'll get Gaussian distributions with different means for my fixed values of A, B and T.

And for those of you that know what a sufficient statistic and statistics is, T of Y actually is a sufficient statistic in the formal sense of sufficient statistic for a probability distribution. They may have seen it in a statistics class. If you don't know what a sufficient statistic is, don't worry about. We sort of don't need that property today.

Okay. So – oh, one last comment. Often, T of Y is equal to Y, and in many of these cases, [inaudible] is also just a raw number. So in many cases, the parameter of this distribution is just a raw number, and [inaudible] transposed T of Y is just a product of raw numbers. So again, that would be true for our first two examples, but now for the last example I'll do today.

So now we'll show that the [inaudible] and the Gaussian are examples of exponential family distributions. We'll start with the [inaudible]. So the [inaudible] distribution with [inaudible] – I guess I wrote this down already. PFY equals one [inaudible] by phi, [inaudible] equal to phi. So the parameter of phi specifies the probability that Y equals one.

And so my goal now is to choose T, A and B, or is to choose A, B and T so that my formula for the exponential family becomes identical to my formula for the distribution of a [inaudible].

So probability of Y parameterized by phi is equal to that, all right. And you already saw sort of a similar exponential notation where we talked about logistic regression. The probability of Y being one is phi, the probability of Y being zero is one minus phi, so we can write this compactly as phi to the Y times one minus phi to the one minus Y.

So I'm gonna take the exponent of the log of this, an exponentiation in taking log [inaudible] cancel each other out [inaudible]. And this is equal to E to the Y. And so [inaudible] is to be T of Y, and this will be minus A of [inaudible]. And then B of Y is just one, so B of Y doesn't matter.

Just take a second to look through this and make sure it makes sense. I'll clean another board while you do that.

So now let's write down a few more things. Just copying from the previous board, we had that [inaudible] zero four equal to log phi over one minus phi.

[Inaudible] so if I want to do the [inaudible] take this formula, and if you invert it, if you solve for phi – excuse me, if you solve for theta as a function of phi, which is really [inaudible] is the function of phi. Just invert this formula. You find that phi is one over one plus [inaudible] minus [inaudible]. And so somehow the logistic function magically falls out of this. We'll take this even this even further later.

Again, copying definitions from the board on – from the previous board, A of [inaudible] I said is minus log of one minus phi. So again, phi and [inaudible] are function of each other, all right. So [inaudible] depends on phi, and phi depends on [inaudible].

So if I plug in this definition for [inaudible] into this – excuse me, plug in this definition for phi into that, I'll find that A of [inaudible] is therefore equal to log one plus [inaudible] to [inaudible]. And again, this is just algebra. This is not terribly interesting. And just to complete – excuse me. And just to complete the rest of this, T of Y is equal to Y, and B of Y is equal to one, okay.

So just to recap what we've done, we've come up with a certain choice of functions A, T and B, so then my formula for the exponential family distribution now becomes exactly the formula for the distributions, or for the probability mass function of the [inaudible] distribution. And the natural parameter [inaudible] has a certain relationship of the original parameter of the [inaudible]. Question?

Student:[Inaudible]

Instructor (Andrew Ng):Let's see. [Inaudible].

Student: The second to the last one.

Instructor (Andrew Ng):Oh, this answer is fine.

Student: Okay.

Instructor (Andrew Ng):Let's see. Yeah, so this is – well, if you expand this term out, one minus Y times log Y minus phi, and so one times log – one minus phi becomes this. And the other term is minus Y times log Y minus phi. And then – so the minus of a log is log one over X, or is just log one over whatever. So minus Y times log one minus phi becomes sort of Y times log, one over one minus phi. Does that make sense?

Student: Yeah.

Instructor (Andrew Ng):Yeah, cool. Anything else? Yes?

Student:[Inaudible] is a scaler, isn't it? Up there –

Instructor (Andrew Ng):Yes.

Student:– it's a [inaudible] transposed, so it can be a vector or –

Instructor (Andrew Ng):Yes, [inaudible]. So let's see. In most – in this and the next example, [inaudible] will turn out to be a scaler. And so – well, on this board. And so if [inaudible] is a scaler and T of Y is a scaler, then this is just a real number times a real number. So this would be like a one-dimensional vector transposed times a one-dimensional vector. And so this is just real number times real number.

Towards the end of today's lecture, we'll go with just one example where both of these are vectors. But for main distributions, these will turn out to be scalers.

Student:[Inaudible] distribution [inaudible]. I mean, it doesn't have the zero probability or [inaudible] zero and one.

Instructor (Andrew Ng):I see. So – yeah. Let's – for this, let's imagine that we're restricting the domain of the input of the function to be Y equals zero or one. So think of that as maybe in implicit constraint on it. [Inaudible]. But so this is a probability mass function for Y equals zero or Y equals one. So write down Y equals zero one. Let's think of that as an [inaudible].

So – cool. So this takes the [inaudible] distribution and invites in the form and the exponential family distribution. [Inaudible] do that very quickly for the Gaussian. I won't do the algebra for the Gaussian. I'll basically just write out the answers.

So with a normal distribution with [inaudible] sequence squared, and so you remember, was it two lectures ago, when we were dividing the maximum likelihood – excuse me,

oh, no, just the previous lecture when we were dividing the maximum likelihood estimate for the parameters of ordinary [inaudible] squares. We showed that the parameter for [inaudible] squared didn't matter.

When we divide the [inaudible] model for [inaudible] square [inaudible], we said that no matter what [inaudible] square was, we end up with the same value of the parameters.

So for the purposes of just writing lesson, today's lecture, and not taking account [inaudible] squared, I'm just going to set [inaudible] squared to be for the one, okay, so as to not worry about it.

Lecture [inaudible] talks a little bit more about this, but I'm just gonna – just to make [inaudible] in class a bit easier and simpler today, let's just say that [inaudible] square equals one. [Inaudible] square is essentially just a scaling factor on the variable Y.

So in that case, the Gaussian density is given by this, [inaudible] squared. And – well, by a couple of steps of algebra, which I'm not going to do, but is written out in [inaudible] in the lecture now so you can download. This is one root two pie, E to the minus one-half Y squared times E to E. New Y minus one-half [inaudible] squared, okay. So I'm just not doing the algebra.

And so that's B of Y, we have [inaudible] that's equal to [inaudible]. P of Y equals Y, and – well, A of [inaudible] is equal to minus one-half – actually, I think that should be plus one-half. Have I got that right? Yeah, sorry. Let's see – excuse me. Plus sign there, okay. If you minus one-half [inaudible] squared, and because [inaudible] is equal to [inaudible], this is just minus one-half [inaudible] squared, okay.

And so this would be a specific choice again of A, B and T that expresses the Gaussian density in the form of an exponential family distribution. And in this case, the relationship between [inaudible] and [inaudible] is that [inaudible] is just equal to [inaudible], so the [inaudible] of the Gaussian is just equal to the natural parameter of the exponential family distribution.

Student: Minus half.

Instructor (Andrew Ng):Oh, this is minus half?

Student:[Inaudible]

Instructor (Andrew Ng):Oh, okay, thanks. And so – guessing that should be plus then. Is that right? Okay. Oh, yes, you're right. Thank you. All right.

And so [inaudible] result that if you've taken a look in undergrad statistics class, turns out that most of the "textbook distributions," not all, but most of them, can be written in the form of an exponential family distribution.

So you saw the Gaussian, the normal distribution. It turns out the [inaudible] in normal distribution, which is a generalization of Gaussian random variables, so it's a high dimension to vectors. The [inaudible] normal distribution is also in the exponential family.

You saw the [inaudible] as an exponential family. It turns out the [inaudible] distribution is too, all right. So the [inaudible] models outcomes over zero and one. They'll be coin tosses with two outcomes. The [inaudible] models outcomes over K possible values. That's also an exponential families distribution.

You may have heard of the Parson distribution. And so the Parson distribution is often used for modeling counts. Things like the number of radioactive decays in a sample, or the number of customers to your website, the numbers of visitors arriving in a store. The Parson distribution is also in the exponential family.

So are the gamma and the exponential distributions, if you've heard of them. So the gamma and the exponential distributions are distributions of the positive numbers. So they're often used in model intervals, like if you're standing at the bus stop and you want to ask, "When is the next bus likely to arrive? How long do I have to wait for my bus to arrive?" Often you model that with sort of gamma distribution or exponential families, or the exponential distribution. Those are also in the exponential family.

Even more [inaudible] distributions, like the [inaudible] and the [inaudible] distributions, these are probably distributions over fractions, are already probability distributions over probability distributions. And also things like the Wisha distribution, which is the distribution over covariance matrices. So all of these, it turns out, can be written in the form of exponential family distributions.

Well, and in the problem set where he asks you to take one of these distributions and write it in the form of the exponential family distribution, and derive a generalized linear model for it, okay.

Which brings me to the next topic of having chosen and exponential family distribution, how do you use it to derive a generalized linear model? So generalized linear models are often abbreviated GLM's. And I'm going to write down the three assumptions. You can think of them as assumptions, or you can think of them as design choices, that will then allow me to sort of turn a crank and come up with a generalized linear model.

So the first one is – I'm going to assume that given my input X and my parameters theta, I'm going to assume that the variable Y, the output Y, or the response variable Y I'm trying to predict is distributed exponential family with some natural parameter [inaudible].

And so this means that there is some specific choice of those functions, A, B and T so that the conditional distribution of Y given X and parameterized by theta, those

exponential families with parameter [inaudible]. Where here, [inaudible] may depend on X in some way.

So for example, if you're trying to predict – if you want to predict how many customers have arrived at your website, you may choose to model the number of people – the number of hits on your website by Parson Distribution since Parson Distribution is natural for modeling com data. And so you may choose the exponential family distribution here to be the Parson distribution.

[Inaudible] that given X, our goal is to output the effective value of Y given X. So given the features in the website examples, I've given a set of features about whether there were any proportions, whether there were sales, how many people linked to your website, or whatever. I'm going to assume that our goal in our [inaudible] problem is to estimate the expected number of people that will arrive at your website on a given day.

So in other words, you're saying that I want H of X to be equal to – oh, excuse me. I actually meant to write T of Y here. My goal is to get my learning algorithms hypothesis to output the expected value of T of Y given X.

But again, for most of the examples, T of Y is just equal to Y. And so for most of the examples, our goal is to get our learning algorithms output, T expected value of Y given X because T of Y is usually equal to Y. Yes?

Student:[Inaudible]

Instructor (Andrew Ng):Yes, same thing, right. T of Y is a sufficient statistic. Same T of Y.

And lastly, this last one I wrote down – these are assumptions. This last one you might – maybe wanna think of this as a design choice. Which is [inaudible] assume that the distribution of Y given X is a distributed exponential family with some parameter [inaudible].

So the number of visitors on the website on any given day will be Parson or some parameter [inaudible]. And the last decision I need to make is was the relationship between my input teachers and this parameter [inaudible] parameterizing my Parson distribution or whatever.

And this last step, I'm going to make the assumption, or really a design choice, that I'm going to assume the relationship between [inaudible] and my [inaudible] axis linear, and in particular that they're governed by this – that [inaudible] is equal to theta, transpose X.

And the reason I make this design choice is it will allow me to turn the crank of the generalized linear model of machinery and come off with very nice algorithms for fitting say Parson Regression models or performed regression with a gamma distribution outputs or exponential distribution outputs and so on.

So let's work through an example. [Inaudible] equals theta transpose X works for the case where [inaudible] is a real number. For the more general case, you would have [inaudible] I equals theta I, transpose X if [inaudible] is a vector rather than a real number. But again, most of the examples [inaudible] will just be a real number.

All right. So let's work through the [inaudible] example. You'll see where Y given X parameterized by theta – this is a distributed exponential family with natural parameter [inaudible]. And for the [inaudible] distribution, I'm going to choose A, B and T to be the specific forms that cause those exponential families to become the [inaudible] distribution. This is the example we worked through just now, the first example we worked through just now.

So – oh, and we also have – so for any fixed value of X and theta, my hypothesis, my learning algorithm will make a prediction, or will make – will sort of output [inaudible] of X, which is by my, I guess, assumption [inaudible].

Watch our learning algorithm to output the expected value of Y given X and parameterized by theta, where Y can take on only the value zero and one, then the expected value of Y is just equal to the probability that Y is equal to one. So the expected value of a [inaudible] variable is just equal to the probability that it's equal to one.

And so the probability that Y equals one is just equal to phi because that's the parameter of my [inaudible] distribution. Phi is, by definition, I guess, is the probability of my [inaudible] distribution [inaudible] value of one.

Which we worked out previously, phi was one over one plus E to the negative [inaudible]. So we worked this out on our previous board. This is the relationship – so when we wrote down the [inaudible] distribution in the form of an exponential family, we worked out what the relationship was between phi and [inaudible], and it was this. So we worked out the relationship between the expected value of Y and [inaudible] was this relationship.

And lastly, because we made the design choice, or the assumption that [inaudible] and theta are linearly related. This is therefore equal to one over one plus E to the minus theta, transpose X.

And so that's how I come up with the logistic regression algorithm when you have a variable Y – when you have a [inaudible] variable Y, or also response variable Y that takes on two values, and then you choose to model variable [inaudible] distribution. Are you sure this does make sense? Raise your hand if this makes sense. Yeah, okay, cool.

So I hope you get the ease of use of this, or sort of the power of this. The only decision I made was really, I said Y – let's say I have a new machine-learning problem and I'm trying to predict the value of a variable Y that happens to take on two values. Then the only decision I need to make is I chose [inaudible] distribution.

I say I want to model – I want to assume that given X and theta, I'm going to assume Y is distributed [inaudible]. That's the only decision I made. And then everything else follows automatically having made the decision to model Y given X and parameterized by theta as being [inaudible].

In the same way you can choose a different distribution, you can choose Y as Parson or Y as gamma or Y as whatever, and follow a similar process and come up with a different model and different learning algorithm. Come up with a different generalized linear model for whatever learning algorithm you're faced with.

This tiny little notation, the function G that relates G of [inaudible] that relates the natural parameter to the expected value of Y, which in this case, one over one plus [inaudible] minus [inaudible], this is called the canonical response function. And G inverse is called the canonical link function.

These aren't a huge deal. I won't use this terminology a lot. I'm just mentioning those in case you hear about – people talk about generalized linear models, and if they talk about canonical response functions or canonical link functions, just so you know there's all of this.

Actually, many techs actually use the reverse way. This is G inverse and this is G, but this notation turns out to be more consistent with other algorithms in machine learning. So I'm going to use this notation. But I probably won't use the terms canonical response functions and canonical link functions in lecture a lot, so just – I don't know. I'm not big on memorizing lots of names of things. I'm just tossing those out there in case you see it elsewhere.

Okay. You know what, I think in the interest of time, I'm going to skip over the Gaussian example. But again, just like I said, [inaudible], Y is [inaudible], different variation I get of logistic regression. You can do the same thing with the Gaussian distribution and end up with ordinary [inaudible] squares model.

The problem with Gaussian is that it's almost so simple that when you see it for the first time that it's sometimes more confusing than the [inaudible] model because it looks so simple, it looks like it has to be more complicated. So let me just skip that and leave you to read about the Gaussian example in the lecture notes.

And what I want to do is actually go through a more complex example. Question?

Student:[Inaudible]

Instructor (Andrew Ng):Okay, right. So how do choose what theory will be? We'll get to that in the end. What you have there is the logistic regression model, which is a [inaudible] model that assumes the probability of Y given X is given by a certain form.

And so what you do is you can write down the log likelihood of your training set, and find the value of theta that maximizes the log likelihood of the parameters. Does that make sense? So I'll say that again towards the end of today's lecture.

But for logistic regression, the way you choose theta is exactly maximum likelihood, as we worked out in the previous lecture, using Newton's Method or gradient ascent or whatever. I'll sort of try to do that again for one more example towards the end of today's lecture.

So what I want to do is actually use the remaining, I don't know, 19 minutes or so of this class, to go through the – one of the more – it's probably the most complex example of a generalized linear model that I've used. This one I want to go through because it's a little bit trickier than many of the other textbook examples of generalized linear models.

So again, what I'm going to do is go through the derivation reasonably quickly and give you the gist of it, and if there are steps I skip or details omitted, I'll leave you to read about them more carefully in the lecture notes.

And what I want to do is talk about [inaudible]. And [inaudible] is the distribution over K possible outcomes. Imagine you're now in a machine-learning problem where the value of Y that you're trying to predict can take on K possible outcomes, so rather than only two outcomes.

So obviously, this example's already – if you want to have a learning algorithm, or to magically send emails for you into your right email folder, and you may have a dozen email folders you want your algorithm to classify emails into. Or predicting if the patient either has a disease or does not have a disease, which would be a [inaudible] classification problem.

If you think that the patient may have one of K diseases, and you want other than have a learning algorithm figure out which one of K diseases your patient has is all.

So lots of multi-cause classification problems where you have more than two causes. You model that with [inaudible]. And eventually – so for logistic regression, I had [inaudible] like these where you have a training set and you find a decision boundary that separates them.

[Inaudible], we're going to entertain the value of predicting, taking on multiple values, so you now have three causes, and the learning algorithm will learn some way to separate out three causes or more, rather than just two causes.

So let's write [inaudible] in the form of an exponential family distribution. So the parameters of a [inaudible] are phi one, phi two [inaudible] phi K. I'll actually change this in a second – where the probability of Y equals I is phi I, right, because there are K possible outcomes.

But if I choose this as my parameterization of the [inaudible], then my parameter's actually redundant because if these are probabilities, then you have to sum up the one. And therefore for example, I can derive the last parameter, phi K, as one minus phi one, up to phi K minus one. So this would be a redundant parameterization from [inaudible]. The result is over-parameterized.

And so for purposes of this [inaudible], I'm going to treat my parameters of my [inaudible] as phi one, phi two, up to phi K minus one. And I won't think of phi K as a parameter. I'll just – so my parameters are just – I just have K minus one parameters, parameterizing my [inaudible].

And sometimes I write phi K in my derivations as well, and you should think of phi K as just a shorthand for this, for one minus the rest of the parameters, okay.

So it turns out the [inaudible] is one of the few examples where T of Y – it's one of the examples where T of Y is not equal to Y. So in this case, Y is on of K possible values.

And so T of Y would be defined as follows; T of one is going to be a vector with a one and zeros everywhere else. T of two is going to be a zero, one, zero and so on. Except that these are going to be K minus one-dimensional vectors.

And so T of K minus one is going to be zero, zero, zero, one. And T of K is going to be the vector of all zeros. So this is just how I'm choosing to define T of Y to write down the [inaudible] in the form of an exponential family distribution. Again, these are K minus one-dimensional vectors.

So this is a good point to introduce one more useful piece of notation, which is called indicator function notation. So I'm going to write one, and then curly braces. And if I write a true statement inside, then the indicator of that statement is going to be one. Then I write one, and then I write a false statement inside, then the value of this indicator function is going to be a zero.

For example, if I write indicator two equals three [inaudible] that's false, and so this is equal to zero. Whereas indicator [inaudible] plus one equals two, I wrote down a true statement inside. And so the indicator of the statement was equal to one. So the indicator function is just a very useful notation for indicating sort of truth or falsehood of the statement inside.

And so – actually, let's do this here. To combine both of these, right, if I carve out a bit of space here – so if I use – so TY is a vector. Y is one of K values, and so TY is one of these K vectors. If I use TY as [inaudible] to denote the [inaudible] element of the vector TY, then TY – the [inaudible] element of the vector TY is just equal to indicator for whether Y is equal to I.

Just take a – let me clean a couple more boards. Take a look at this for a second and make sure you understand why that – make sure you understand all that notation and why this is true.

All right. Actually, raise your hand if this equation makes sense to you. Most of you, not all, okay. [Inaudible].

Just as one kind of [inaudible], suppose Y is equal to one – let's say – let me see. Suppose Y is equal to one, right, so TY is equal to this vector, and therefore the first element of this vector will be one, and the rest of the elements will be equal to zero.

And so – let me try that again, I'm sorry. Let's say I want to ask – I want to look at the [inaudible] element of the vector TY, and I want to know is this one or zero. All right. Well, this will be one. The [inaudible] element of the vector TY will be equal to one if, and only if Y is equal to I.

Because for example, if Y is equal to one, then only the first element of this vector will be zero. If Y is equal to two, then only the second element of the vector will be zero and so on. So the question of whether or not – whether the [inaudible] element of this vector, TY, is equal to one is answered by just asking is Y equal to I.

Okay. If you're still not quite sure why that's true, go home and think about it a bit more. And I think I – and take a look at the lecture notes as well, maybe that'll help. At least for now, only just take my word for it.

So let's go ahead and write out the distribution for the [inaudible] in an exponential family form. So PFY is equal to phi one. Indicator Y equals one times phi two. Indicator Y equals to up to phi K times indicator Y equals K. And again, phi K is not a parameter of the distribution. Phi K is a shorthand for one minus phi one minus phi two minus the rest.

And so using this equation on the left as well, I can also write this as phi one times TY one, phi two, TY two, dot, dot, dot. Phi K minus one, TY, K minus one times phi K. And then one minus [inaudible]. That should be K.

And it turns out – it takes some of the steps of algebra that I don't have time to show. It turns out, you can simplify this into – well, the exponential family form where [inaudible] is a vector, this is a K minus one-dimensional vector, and – well, okay.

So deriving this is a few steps of algebra that you can work out yourself, but I won't do here. And so using my definition for TY, and by choosing [inaudible] A and B this way, I can take my distribution from [inaudible] and write it out in a form of an exponential family distribution.

It turns out also that – let's see. [Inaudible], right. One of the things we did was we also had [inaudible] as a function of phi, and then we inverted that to write out phi as a function of [inaudible]. So it turns out you can do that as well.

So this defines [inaudible] as a function of the [inaudible] distributions parameters phi. So you can take this relationship between [inaudible] and phi and invert it, and write out phi as a function of [inaudible]. And it turns out, you get that phi I is equal to [inaudible] – excuse me. And you get that phi I is equal to [inaudible] I of one plus that.

And the way you do this is you just – this defines [inaudible] as a function of the phi, so if you take this and solve for [inaudible], you end up with this. And this is – again, there are a couple of steps of algebra that I'm just not showing.

And then lastly, using our assumption that the [inaudible] are a linear function of the [inaudible] axis, phi I is therefore equal to E to the theta I, transpose X, divided by one plus sum over J equals one, to K minus one, E to the theta J, transpose X. And this is just using the fact that [inaudible] I equals theta I, transpose X, which was our earlier design choice from generalized linear models.

So we're just about down. So my learning algorithm [inaudible]. I'm going to think of it as [inaudible] the expected value of TY given X and [inaudible] by theta. So TY was this vector indicator function. So T one was indicator Y equals one, down to indicator Y equals K minus one. All right. So I want my learning algorithm to output this; the expected value of this vector of indicator functions.

The expected value of indicator Y equals one is just the probability that Y equals one, which is given by phi one. So I have a random variable that's one whenever Y is equal to one and zero otherwise, so the expected value of that, of this indicator Y equals one is just the probability that Y equals one, which is given by phi one.

And therefore, by what we were taught earlier, this is therefore [inaudible] to the theta one, transpose X over – well – okay. And so my learning algorithm will output the probability that Y equals one, Y equals two, up to Y equals K minus one. And these probabilities are going to be parameterized by these functions like these.

And so just to give this algorithm a name, this algorithm is called softmax regression, and is widely thought of as the generalization of logistic regression, which is regression of two classes. Is widely thought of as a generalization of logistic regression to the case of K classes rather than two classes.

And so just to be very concrete about what you do, right. So you have a machine-learning problem, and you want to apply softmax regression to it. So generally, work for the entire derivation [inaudible]. I think the question you had is about how to fit parameters.

So let's say you have a machine-learning problem, and Y takes on one of K classes. What you do is you sit down and say, "Okay, I wanna model Y as being [inaudible] given any

X and then theta." And so you chose [inaudible] as the exponential family. Then you sort of turn the crank. And everything else I wrote down follows automatically from you have made the choice of using [inaudible] distribution as your choice of exponential family.

And then what you do is you then have this training set, X, I, Y, I up to X, M, Y, M. So you're doing the training set. We're now [inaudible] the value of Y takes on one of K possible values.

And what you do is you then find the parameters of the model by maximum likelihood. So you write down the likelihood of the parameters, and you maximize the likelihood.

So what's the likelihood? Well, the likelihood, as usual, is the product of your training set of P of YI given XI parameterized by theta. That's the likelihood, same as we had before. And that's product of your training set of – let me write these down now. YI equals one times phi two of indicator YI equals two, dot, dot, dot, to phi K of indicator YI equals K.

Where, for example, phi one depends on theta through this formula. It is E to the theta one, transpose X over one plus sum over J – well, that formula I had just now. And so phi one here is really a shorthand for this formula, and similarly for phi two and so on, up to phi K, where phi K is one minus all of these things. All right.

So this is a –this formula looks more complicated than it really is. What you really do is you write this down, then you take logs, compute a derivative of this formula [inaudible] theta, and apply say gradient ascent to maximize the likelihood.

Student: What are the rows of theta? [Inaudible] it's just been a vector, right? And now it looks like it's two-dimensional.

Instructor (Andrew Ng):Yeah. In the notation of the [inaudible] I think have theta one through theta K minus one. I've been thinking of each of these as – and N plus one-dimensional vector. If X is N plus one-dimensional, then I've been – see, I think if you have a set of parameters comprising K minus one vectors, and each of these is a – you could group all of these together and make these, but I just haven't been doing that. [Inaudible] the derivative of K minus one parameter vectors.

Student:[Inaudible], what do they correspond to?

Instructor (Andrew Ng):[Inaudible]. We're sort of out of time. Let me take that offline. It's hard to answer in the same way that the logistic regression – what does theta correspond to in logistic regression? You can sort of answer that as sort of –

Student: Yeah. It's kind of like the [inaudible] feature –

Instructor (Andrew Ng):Yeah. Sort of similar interpretation, yeah. That's good. I think I'm running a little bit late. Why don't I – why don't we officially close for the day, but you can come up if you more questions and take them offline. Thanks.

[End of Audio]

Duration: 76 minutes