MachineLearning-Lecture13

Instructor (Andrew Ng):Okay, good morning. For those of you actually online, sorry; starting a couple of minutes late. We're having trouble with the lights just now, so we're all sitting in the dark and they just came on. So welcome back, and what I want to do today is continue our discussions of the EM Algorithm, and in particular, I want to talk about the EM formulation that we derived in the previous lecture and apply it to the mixture of Gaussians model, apply it to a different model and a mixture of naive Bayes model, and then the launch part of today's lecture will be on the factor analysis algorithm, which will also use the EM. And as part of that, we'll actually take a brief digression to talk a little bit about sort of useful properties of Gaussian distributions.

So just to recap where we are. In the previous lecture, I started to talk about unsupervised learning, which was machine-learning problems, where you're given an unlabeled training set comprising m examples here, right? And then – so the fact that there are no labels; that's what makes this unsupervised or anything. So one problem that I talked about last time was what if you're given a data set that looks like this and you want to model the density PFX from which you think the data had been drawn, and so with a data set like this, maybe you think was a mixture of two Gaussians and start to talk about an algorithm for fitting a mixture of Gaussians model, all right? And so we said that we would model the density of XP of X as sum over Z PFX given Z times P of Z where this later random variable meaning this hidden random variable Z indicates which of the two Gaussian distributions each of your data points came from and so we have, you know, Z was not a nomial with parameter phi and X conditions on a coming from the JAFE Gaussian was given by Gaussian of mean mu J and covariant sigma J, all right?

So, like I said at the beginning of the previous lecture, I just talked about a very specific algorithm that I sort of pulled out of the air for fitting the parameters of this model for finian, Francis, phi, mu and sigma, but then in the second half of the previous lecture I talked about what's called the EM Algorithm in which our goal is that it's a likelihood estimation of parameters. So we want to maximize in terms of theta, you know, the, sort of, usual right matter of log likelihood – well, parameterized by theta. And because we have a later random variable Z this is really maximizing in terms of theta, sum over I, sum over Z, P of XI, ZI parameterized by theta. Okay? So using Jensen's inequality last time we worked out the EM Algorithm in which in the E step we would chose these probability distributions QI to the l posterior on Z given X and parameterized by theta and in the M step we would set theta to be the value that maximizes this. Okay? So these are the ones we worked out last time and the cartoon that I drew was that you have this long likelihood function L of theta that's often hard to maximize and what the E step does is choose these probability distribution production QI's. And in the cartoon, I drew what that corresponded to was finding a lower bounds for the log likelihood. And then horizontal access data and then the M step you maximize the lower boundary, right? So maybe you were here previously and so you jumped to the new point, the new maximum of this lower bound. Okay? And so this little curve here, right? This lower bound function here that's really the right-hand side of that augments. Okay? So this whole thing in the augments. If you view this thing as a function of theta, this function of theta is a lower bounds for the log likelihood of theta and so the M step we maximize this lower bound and that corresponds to jumping to this new maximum to lower bound.

So it turns out that in the EM Algorithm – so why do you evolve with the EM algorithm? It turns out that very often, and this will be true for all the examples we see today, it turns out that very often in the EM Algorithm maximizing the M Step, so performing the maximization the M Step, will be tractable and can often be done analytically in the closed form. Whereas if you were trying to maximize this objective we try to take this formula on the right and this maximum likely object, everyone, is to take this all on the right and set its derivatives to zero and try to solve and you'll find that you're unable to obtain a solution to this in closed form this maximization. Okay?

And so to give you an example of that is that you remember our discussion on exponential family marbles, right? It turns out that if X and Z is jointly, I guess, a line in exponential families. So if P of X, Z prioritized by theta there's an explanation family distribution, which it turns out to be true for the mixture of Gaussians distribution. Then turns out that the M step here will be tractable and the E step will also be tractable and so you can do each of these steps very easily. Whereas performing – trying to perform this original maximum likelihood estimation problem on this one, right? Will be computationally very difficult. You're going to set the derivatives to zero and try to solve for that. Analytically you won't be able to find an analytic solution to this. Okay?

So what I want to do in a second is actually take this view of the EM Algorithm and apply it to the mixture of Gaussians models. I want to take these E steps and M Steps and work them out for the mixture of Gaussians model, but before I do that, I just want to say one more thing about this other view of the EM Algorithm. It turns out there's one other way of thinking about the EM Algorithm, which is the following: I can define an optimization objective J of theta, Q are defined it to be this. This is just a thing in the augments in the M step. Okay? And so what we proved using Jensen's inequality is that the log likelihood of theta is greater and equal to J of theta Q. So in other words, we proved last time that for any value of theta and Q the log likelihood upper bounds J of theta and Q. And so just to relate this back to, sort of, yet more things that you all ready know, you can also think of covariant cause in a sense, right? However, our discussion awhile back on the coordinate ascent optimization algorithm. So we can show, and I won't actually show this view so just take our word for it and look for that at home if you want, that EM is just coordinate in a set on the function J. So in the E step you maximize with respect to Q and then the M step you maximize with respect to theta. Okay? So this is another view of the EM Algorithm that shows why it has to converge, for example. If you can – I've used in a sense of J of theta, Q having to monotonically increase on every iteration. Okay?

So what I want to do next is actually take this general EM machinery that we worked up and apply it to a mixture Gaussians model. Before I do that, let me just check if there are questions about the EM Algorithm as a whole? Okay, cool.

So let’s go ahead and work on the mixture of Gaussian’s EM, all right? MOG, and that’s my abbreviation for Mixture of Gaussian’s. So the E step were called those Q distributions, right? In particular, I want to work out – so Q is the probability distribution over the late and random variable Z and so the E step I’m gonna figure out what is these compute – what is Q of ZI equals J. And you can think of this as my writing P of ZI equals J, right? Under the Q distribution. That’s what this notation means. And so the EM Algorithm tells us that, let’s see, Q of J is the likelihood probability of Z being the value J and given XI and all your parameters. And so, well, the way you compute this is by Dave’s rule, right? So that is going to be equal to P of XI given ZI equals J times P of ZIJ divided by – right? That’s all the – by Dave’s rule. And so this you know because XI given ZI equals J. This was a Gaussian with mean mu J and covariant sigma J. And so to compute this first term you plug in the formula for the Gaussian density there with parameters mu J and sigma J and this you’d know because Z was not a nomial, right? Where parameters given by phi and so the problem of ZI being with J is just phi J and so you can substitute these terms in. Similarly do the same thing for the denominator and that’s how you work out what Q is. Okay? And so in the previous lecture this value the probability that ZI equals J under the Q distribution that was why I denoted that as WIJ. So that would be the E step and then in the M step we maximize with respect to all of our parameters. This, well I seem to be writing the same formula down a lot today. All right. And just so we’re completely concrete about how you do that, right? So if you do that you end up with – so plugging in the quantities that you know that becomes this, let’s see. Right. And so that we’re completely concrete about what the M step is doing. So in the M step that was, I guess, QI over Z, I being over J. Just in the summation, sum over J is the sum over all the possible values of ZI and then this thing here is my Gaussian density. Sorry, guys, this thing – well, this first term here, right? Is my P of XI given ZI and that’s P of ZI. Okay? And so to maximize this with respect to – say you want to maximize this with respect to all of your parameters phi, mu and sigma. So to maximize with respect to the parameter mu, say, you would take the derivative for respect to mu and set that to zero and you would – and if you actually do that computation you would get, for instance, that that becomes your update to mu J. Okay? Just so I want to – the equation is unimportant. All of these equations are written down in the lecture notes. I’m writing these down just to be completely concrete about what the M step means. And so write down that formula, plug in the densities you know, take the derivative set to zero, solve for mu J and in the same way you set the derivatives equal to zero and solve for your updates for your other parameters phi and sigma as well. Okay?

Well, just point out just one little tricky bit for this that you haven’t seen before that most of you probably all ready now, but I’ll just mention is that since phi here is a multinomial distribution when you take this formula and you maximize it with respect to phi you actually have an additional constraint, right? That the sum of I – let’s see, sum over J, phi J must be equal to one. All right? So, again, in the M step I want to take this thing and maximize it with respect to all the parameters and when you maximize this respect to the parameters phi J you need to respect the constraint that sum of J phi J must be equal to one. And so, well, you all ready know how to do constraint maximization, right? So I’ll throw out the method of the granjay multipliers and generalize the granjay when you talk about the support of X machines. And so to actually perform the maximization in terms of phi J you construct to the granjay, which is – all right? So that's the equation from above and we'll denote in the dot dot dot plus theta times that, where this is sort of the granjay multiplier and this is your optimization objective. And so to actually solve the parameters phi J you set the parameters of this so that the granjay is zero and solve. Okay? And if you then work through the math you get the appropriate value to update the phi J's too, which I won't do, but I'll be – all the full directions are in the lecture notes. I won't do that here.

Okay. And so if you actually perform all these computations you can also verify that. So I just wrote down a bunch of formulas for the EM Algorithm. At the beginning of the last lecture I said for the mixture of Gaussian's model – I said for the EM here's the formula for computing the WIJ's and here's a formula for computing the mud's and so on, and this variation is where all of those formulas actually come from. Okay? Questions about this? Yeah?

Student:[Inaudible]

Instructor (Andrew Ng):Oh, I see. So it turns out that, yes, there's also constrained to the phi J this must be greater than zero. It turns out that if you want you could actually write down then generalize the granjayn incorporating all of these constraints as well and you can solve to [inaudible] these constraints. It turns out that in this particular derivation – actually it turns out that very often we find maximum likely estimate for multinomial distributions probabilities. It turns out that if you ignore these constraints and you just maximize the formula luckily you end up with values that actually are greater than or equal to zero and so if even ignoring those constraint you end up with parameters that are greater than or equal to zero that shows that that must be the correct solution because adding that constraint won't change anything. So this constraint it is then caused – it turns out that if you ignore this and just do what I've wrote down you actually get the right answer. Okay? Great.

So let me just very quickly talk about one more example of a mixture model. And the perfect example for this is imagine you want to do text clustering, right? So someone gives you a large set of documents and you want to cluster them together into cohesive topics. I think I mentioned the news website news.google.com. That's one application of text clustering where you might want to look at all of the news stories about today, all the news stories written by everyone, written by all the online news websites about whatever happened yesterday and there will be many, many different stories on the same thing, right? And by running a text-clustering algorithm you can group related documents together. Okay?

So how do you apply the EM Algorithm to text clustering? I want to do this to illustrate an example in which you run the EM Algorithm on discreet valued inputs where the input – where the training examples XI are discreet values. So what I want to talk about specifically is the mixture of naïve Bayes model and depending on how much you remember about naïve Bayes I talked about two event models. One was the multivariant vanuvy event model. One was the multinomial event model. Today I'm gonna use the multivariant vanuvy event model. If you don’t remember what those terms mean anymore don’t worry about it. I think the equation will still make sense. But in this setting we’re given a training set X1 for XM. So we’re given M text documents where each XI is zero one to the end. So each of our training examples is an indimensional bit of vector, right? S this was the representation where XIJ was – it indicates whether word J appears in document I, right? And let’s say that we’re going to model ZI the – our latent random variable meaning our hidden random variable ZI will take on two values zero one so this means I’m just gonna find two clusters and you can generalize the clusters that you want.

So in the mixture of naïve Bayes model we assume that ZI is distributed and mu E with some value of phi so there’s some probability of each document coming from cluster one or from cluster two. We assume that the probability of XI given ZI, right? Will make the same naïve Bayes assumption as we did before. Okay? And more specifically – well, excuse me, right. Okay. And so most of us [inaudible] cycles one given ZI equals say zero will be given by a parameter phi substitute J given Z equals zero. So if you take this chalkboard and if you take all instances of the alphabet Z and replace it with Y then you end up with exactly the same equation as I’ve written down for naïve Bayes like a long time ago. Okay?

And I’m not actually going to work out the mechanics deriving the EM Algorithm, but it turns out that if you take this joint distribution of X and Z and if you work out what the equations are for the EM algorithm for maximum likelihood estimation of parameters you find that in the E step you compute, you know, let’s say these parameters – these weights WI which are going to be equal to your perceived distribution of Z being equal one conditioned on XI parameterized by your phi’s and given your parameters and in the M step. Okay? And that’s the equation you get in the M step. I mean, again, the equations themselves aren’t too important. Just sort of convey – I’ll give you a second to finish writing, I guess. And when you’re done or finished writing take a look at these equations and see if they make intuitive sense to you why these three equations, sort of, sound like they might be the right thing to do. Yeah?

Student:[Inaudible]

Instructor (Andrew Ng):Say that again.

Student:Y –

Instructor (Andrew Ng):Oh, yes, thank you. Right. Sorry, just, for everywhere over Y I meant Z. Yeah?

Student:[Inaudible] in the first place?

Instructor (Andrew Ng):No. So what is it? Normally you initialize phi’s to be something else, say randomly. So just like in naïve Bayes we saw zero probabilities as a bad thing so the same reason you try to avoid zero probabilities, yeah. Okay? And so just the intuition behind these equations is in the E step WI’s is you’re gonna take your best guess for whether the document came from cluster one or cluster zero, all right? This is very similar to the intuitions behind the EM Algorithm that we talked about in a previous lecture. So in the E step we’re going to compute these weights that tell us do I think this document came from cluster one or cluster zero. And then in the M step I’m gonna say does this numerator is the sum over all the elements of my training set of – so then informally, right? WI is one there, but I think the document came from cluster one and so this will essentially sum up all the times I saw words J in documents that I think are in cluster one. And these are sort of weighted by the actual probability. I think it came from cluster one and then I’ll divide by – again, if all of these were ones and zeros then I’d be dividing by the actual number of documents I had in cluster one. So if all the WI’s were either ones or zeroes then this would be exactly the fraction of documents that I saw in cluster one in which I also saw were at J. Okay? But in the EM Algorithm you don’t make a hard assignment decision about is this in cluster one or is this in cluster zero. You instead represent your uncertainty about cluster membership with the parameters WI. Okay?

It actually turns out that when we actually implement this particular model it actually turns out that by the nature of this computation all the values of WI’s will be very close to either one or zero so they’ll be numerically almost indistinguishable from one’s and zeroes. This is a property of naïve Bayes. If you actually compute this probability from all those documents you find that WI is either 0.0001 or 0.999. It’ll be amazingly close to either zero or one and so the M step – and so this is pretty much guessing whether each document is in cluster one or cluster zero and then using formulas they’re very similar to maximum likely estimation for naïve Bayes. Okay? Cool. Are there – and if some of these equations don’t look that familiar to you anymore, sort of, go back and take another look at what you saw in naïve Bayes and hopefully you can see the links there as well. Questions about this before I move on? Right, okay.

Of course the way I got these equations was by turning through the machinery of the EM Algorithm, right? I didn’t just write these out of thin air. The way you do this is by writing down the E step and the M step for this model and then the M step same derivatives equal to zero and solving from that so that’s how you get the M step and the E step.

So the last thing I want to do today is talk about the factor analysis model and the reason I want to do this is sort of two reasons because one is factor analysis is kind of a useful model. It’s not as widely used as mixtures of Gaussian’s and mixtures of naïve Bayes maybe, but it’s sort of useful. But the other reason I want to derive this model is that there are a few steps in the math that are more generally useful. In particular, where this is for factor analysis this would be an example in which we’ll do EM where the late and random variable – where the hidden random variable Z is going to be continued as valued. And so some of the math we’ll see in deriving factor analysis will be a little bit different than what you saw before and they’re just a – it turns out the full derivation for EM for factor analysis is sort of extremely long and complicated and so I won’t inflect that on you in lecture today, but I will still be writing more equations than is – than you’ll see me do in other lectures because there are, sort of, just a few steps in the factor analysis derivation so I'll physically illustrate it.

So it's actually [inaudible] the model and it's really contrast to the mixture of Gaussians model, all right? So for the mixture of Gaussians model, which is our first model we had, that – well I actually motivated it by drawing the data set like this, right? That one of you has a data set that looks like this, right? So this was a problem where n is two-dimensional and you have, I don't know, maybe 50 or 100 training examples, whatever, right? And I said maybe you want to give a label training set like this. Maybe you want to model this as a mixture of two Gaussians. Okay? And so a mixture of Gaussian models tend to be applicable where m is larger, and often much larger, than n where the number of training examples you have is at least as large as, and is usually much larger than, the dimension of the data.

What I want to do is talk about a different problem where I want you to imagine what happens if either the dimension of your data is roughly equal to the number of examples you have or maybe the dimension of your data is maybe even much larger than the number of training examples you have. Okay? So how do you model such a very high dimensional data? Watch and you will see sometimes, right? If you run a plant or something, you run a factory, maybe you have a thousand measurements all through your plants, but you only have five – you only have 20 days of data. So you can have 1,000 dimensional data, but 20 examples of it all ready. So given data that has this property in the beginning that we've given a training set of m examples. Well, what can you do to try to model the density of X? So one thing you can do is try to model it just as a single Gaussian, right? So in my mixtures of Gaussian this is how you try model as a single Gaussian and say X is intuitive with mean mu and parameter sigma where sigma is going to be done n by n matrix and so if you work out the maximum likelihood estimate of the parameters you find that the maximum likelihood estimate for the mean is just the empirical mean of your training set, right. So that makes sense. And the maximum likelihood of the covariance matrix sigma will be this, all right? But it turns out that in this regime where the data is much higher dimensional – excuse me, where the data's dimension is much larger than the training examples you have if you compute the maximum likely estimate of the covariance matrix sigma you find that this matrix is singular. Okay? By singular, I mean that it doesn't have four vanq or it has zero eigen value so it doesn't have – I hope one of those terms makes sense. And there's another saying that the matrix sigma will be non-invertible. And just in pictures, one complete example is if D is – if N equals M equals two if you have two-dimensional data and you have two examples. So I'd have two training examples in two-dimen – this is X1 and X2. This is my unlabeled data. If you fit a Gaussian to this data set you find that – well you remember I used to draw constables of Gaussians as ellipses, right? So these are examples of different constables of Gaussians. You find that the maximum likely estimate Gaussian for this responds to Gaussian where the contours are sort of infinitely thin and infinitely long in that direction. Okay? So in terms – so the contours will sort of be infinitely thin, right? And stretch infinitely long in that direction. And another way of saying it is that if you actually plug in the formula for the density of the Gaussian, which is this, you won't actually get a nice answer because the matrix sigma is non-invertible so sigma inverse is not defined and this is zero. So you also have one over zero times E to the sum inversive and non-inversive matrix so not a good model. So let’s do even better, right? So given this sort of data how do you model P of X?

Well, one thing you could do is constrain sigma to be diagonal. So you have a covariance matrix X is – okay? So in other words you get a constraint sigma to be this matrix, all right? With zeroes on the off diagonals. I hope this makes sense. These zeroes I’ve written down here denote that everything after diagonal of this matrix is a zero. So the massive likely estimate of the parameters will be pretty much what you’ll expect, right? And in pictures what this means is that the [inaudible] the distribution with Gaussians whose controls are axis aligned. So that’s one example of a Gaussian where the covariance is diagonal. And here’s another example and so here’s a third example. But often I’ve used the examples of Gaussians where the covariance matrix is off diagonal. Okay? And, I don’t know, you could do this in model P of X, but this isn’t very nice because you’ve now thrown away all the correlations between the different variables so the axis are X1 and X2, right? So you’ve thrown away – you’re failing to capture any of the correlations or the relationships between any pair of variables in your data. Yeah?

Student: Is it – could you say again what does that do for the diagonal?

Instructor (Andrew Ng):Say again.

Student: The covariance matrix the diagonal, what does that again? I didn’t quite understand what the examples mean.

Instructor (Andrew Ng):Okay. So these are the contours of the Gaussian density that I’m drawing, right? So let’s see – so post covariance issues with diagonal then you can ask what is P of X parameterized by mu and sigma, right? If sigma is diagonal and so this will be some Gaussian dump, right? So not in – oh, boy. My drawing’s really bad, but in two-D the density for Gaussian is like this bump shaped thing, right? So this is the density of the Gaussian – wow, and this is a really bad drawing. With those, your axis X1 and X2 and the height of this is P of X and so those figures over there are the contours of the density of the Gaussian. So those are the contours of this shape.

Student: No, I don’t mean the contour. What’s special about these types? What makes them different than instead of general covariance matrix?

Instructor (Andrew Ng):Oh, I see. Oh, okay, sorry. They’re axis aligned so the main – these, let’s see. So I’m not drawing a contour like this, right? Because the main axes of these are not aligned with the X1 and X-axis so this occurs found to Gaussian where the off-diagonals are non-zero, right? Cool. Okay. You could do this, this is sort of work. It turns out that what our best view is two training examples you can learn in non-singular covariance matrix, but you’ve thrown away all of the correlation in the data so this is not a great model.

It turns out you can do something – well, actually, we'll come back and use this property later. But it turns out you can do something even more restrictive, which is you can constrain sigma to equal to sigma squared times the identity matrix. So in other words, you can constrain it to be diagonal matrix and moreover all the diagonal entries must be the same and so the cartoon for that is that you're constraining the contours of your Gaussian density to be circular. Okay? This is a sort of even harsher constraint to place in your model. So either of these versions, diagonal sigma or sigma being the, sort of, constant value diagonal are the all ready strong assumptions, all right? So if you have enough data maybe write a model just a little bit of a correlation between your different variables. So the factor analysis model is one way to attempt to do that. So here's the idea. So this is how the factor analysis model models your data. We're going to assume that there is a latent random variable, okay? Which just means random variable Z. So Z is distributed Gaussian with mean zero and covariance identity where Z will be a D-dimensional vector now and D will be chosen so that it is lower than the dimension of your X's. Okay? And now I'm going to assume that X is given by – well let me write this. Each XI is distributed – actually, sorry, I'm just. We have to assume that conditions on the value of Z, X is given by another Gaussian with mean given by mu plus lambda Z and covariance given by matrix si. So just to say the second line in an equivalent form, equivalently I'm going to model X as mu plus lambda Z plus a noise term epsilon where epsilon is Gaussian with mean zero and covariant si. And so the parameters of this model are going to be a vector mu with its n-dimensional and matrix lambda, which is n by D and a covariance matrix si, which is n by n, and I'm going to impose an additional constraint on si. I'm going to impose a constraint that si is diagonal. Okay? So that was a form of definition – let me actually, sort of, give a couple of examples to make this more complete. So let's give a kind of example, suppose Z is one-dimensional and X is two-dimensional so let's see what this model – let's see a, sort of, specific instance of the factor analysis model and how we're modeling the joint – the distribution over X of – what this gives us in terms of a model over P of X, all right?

So let's see. From this model to let me assume that lambda is 2, 1 and si, which has to be diagonal matrix, remember, is this. Okay? So Z is one-dimensional so let me just draw a typical sample for Z, all right? So if I draw ZI from a Gaussian so that's a typical sample for what Z might look like and so I'm gonna – at any rate I'm gonna call this Z1, Z2, Z3 and so on. If this really were a typical sample the order of the Z's would be jumbled up, but I'm just ordering them like this just to make the example easier. So, yes, typical sample of random variable Z from a Gaussian distribution with mean of covariance one. So – and with this example let me just set mu equals zero. It's to write the – just that it's easier to talk about. So lambda times Z, right? We'll take each of these numbers and multiply them by lambda. And so you find that all of the values for lambda times Z will lie on a straight line, all right? So, for example, this one here would be one, two, three, four, five, six, seven, I guess. So if this was Z7 then this one here would be lambda times Z7 and now that's the number in R2, because lambda's a two by one matrix. And so what I've drawn here is like a typical sample for lambda times Z and the final step for this is what a typical sample for X looks like. Well X is mu plus lambda Z plus epsilon where epsilon is Gaussian with mean nu and covariance given by si, right? And so the last step to draw a typical sample for the random variables X I'm gonna take these non – these are really same as mu plus lambda Z because mu is zero in this example and around this point I’m going to place an axis aligned ellipse. Or in other words, I’m going to create a Gaussian distribution centered on this point and this I’ve drawn corresponds to one of the contours of my density for epsilon, right? And so you can imagine placing a little Gaussian bump here. And so I’ll draw an example from this little Gaussian and let’s say I get that point going, I do the same here and so on. So I draw a bunch of examples from these Gaussians and the – whatever they call it – the orange points I drew will comprise a typical sample for whether distribution of X is under this model. Okay? Yeah?

Student: Would you add, like, mean? Instructor:

Oh, say that again.

Student: Do you add mean into that?

Instructor (Andrew Ng):Oh, yes, you do. And in this example, I said you do a zero zero just to make it easier. If mu were something else you’d take the whole picture and you’d sort of shift it to whatever value of mu is. Yeah?

Student:[Inaudible] horizontal line right there, which was Z. What did the X’s, of course, what does that Y-axis corresponds to?

Instructor (Andrew Ng):Oh, so this is Z is one-dimensional so here I’m plotting the typical sample for Z so this is like zero. So this is just the Z Axis, right. So Z is one-dimensional data. So this line here is like a plot of a typical sample of values for Z. Okay? Yeah?

Student: You have by axis, right? And the axis data pertains samples.

Instructor (Andrew Ng):Oh, yes, right.

Student: So sort of projecting them into that?

Instructor (Andrew Ng):Let’s not talk about projections yet, but, yeah, right. So these beige points – so that’s like X1 and that’s X2 and so on, right? So the beige points are what I see. And so in reality all you ever get to see are the X’s, but just like in the mixture of Gaussians model I tell a story about what I would imagine the Gau—the data came from two Gaussian’s was is had a random variable Z that led to the generation of X’s from two Gaussians. So the same way I’m sort of telling the story here, which all the algorithm actually sees are the orange points, but we’re gonna tell a story about how the data came about and that story is what comprises the factor analysis model. Okay? So one of the ways to see the intrusion of this model is that we’re going to think of the model as one way just informally, not formally, but one way to think about this model is you can think of this factor analysis model as modeling the data from coming from a lower dimensional subspace more or less so the data X here Y is approximately on one D line and then plus a little bit of noise – plus a little bit of random noise so the X isn't exactly on this one D line. That's one informal way of thinking about factor analysis.

We're not doing great on time. Well, let's do this. So let me just do one more quick example, which is, in this example, let's say Z is in R2 and X is in R3, right? And so in this example Z, your data Z now lies in 2-D and so let me draw this on a sheet of paper. Okay? So let's say the axis of my paper are the Z1 and Z2 axis and so here is a typical sample of point Z, right? And so we'll then take the sample Z – well, actually let me draw this here as well. All right. So this is a typical sample for Z going on the Z1 and Z2 axis and I guess the origin would be here. So center around zero. And then we'll take those and map it to mu plus lambda Z and what that means is if you imagine the free space of this classroom is R3. What that means is we'll take this sample of Z's and we'll map it to position in free space. So we'll take this sheet of paper and move it somewhere and some orientation in 3-D space. And the last step is you have X equals mu plus lambda Z plus epsilon and so you would take the set of the points which align in some plane in our 3-D space the variable of noise of these and the noise will, sort of, come from Gaussians to the axis aligned. Okay? So you end up with a data set that's sort of like a fat pancake or a little bit of fuzz off your pancake. So that's a model – let's actually talk about how to fit the parameters of the model. Okay?

In order to describe how to fit the model I'm sure we need to re-write Gaussians and this is in a very slightly different way. So, in particular, let's say I have a vector X and I'm gonna use this notation to denote partition vectors, right? X1, X2 where if X1 is say an r-dimensional vector then X2 is an estimational vector and X is an R plus S dimensional vector. Okay? So I'm gonna use this notation to denote just the taking of vector and, sort of, partitioning the vector into two halves. The first R elements followed by the last S elements. So let's say you have X coming from a Gaussian distribution with mean mu and covariance sigma where mu is itself a partition vector. So break mu up into two pieces mu1 and mu2 and the covariance matrix sigma is now a partitioned matrix. Okay? So what this means is that you take the covariance matrix sigma and I'm going to break it up into four blocks, right? And so the dimension of this is there will be R elements here and there will be S elements here and there will be R elements here. So, for example, sigma 1, 2 will be an R by S matrix. It's R elements tall and S elements wide.

So this Gaussian over to down is really a joint distribution of a loss of variables, right? So X is a vector so XY is a joint distribution over X1 through X of – over XN or over X of R plus S. We can then ask what are the marginal and conditional distributions of this Gaussian? So, for example, with my Gaussian, I know what P of X is, but can I compute the modular distribution of X1, right. And so P of X1 is just equal to, of course, integrate our X2, P of X1 comma X2 DX2. And if you actually perform that distribution – that computation you find that P of X1, I guess, is Gaussian with mean given by mu1 and sigma 1, 1. All right. So this is sort of no surprise. The marginal distribution of a Gaussian is itself the Gaussian and you just take out the relevant sub-blocks of the covariance matrix and the relevant sub-vector of the mu vector – E in vector mu. You can also compute conditionals. You can also – what does P of X1 given a specific value for X2, right? And so the way you compute that is, well, the usual way P of X1 comma X2 divided by P of X2, right? And so you know what both of these formulas are, right? The numerator – well, this is just a usual Gaussian that your joint distribution over X1, X2 is a Gaussian with mean mu and covariance sigma and this by that marginalization operation I talked about is that. So if you actually plug in the formulas for these two Gaussians and if you simplify the simplification step is actually fairly non-trivial. If you haven't seen it before this will actually be – this will actually be somewhat difficult to do. But if you plug this in for Gaussian and simplify that expression you find that conditioned on the value of X2, X1 is – the distribution of X1 conditioned on X2 is itself going to be Gaussian and it will have mean mu of 1 given 2 and covariant sigma of 1 given 2 where – well, so about the simplification and derivation I'm not gonna show the formula for mu given – of mu of one given 2 is given by this and I think the sigma of 1 given 2 is given by that. Okay?

So these are just useful formulas to know for how to find the conditional distributions of the Gaussian and the marginal distributions of a Gaussian. I won't actually show the derivation for this.

Student: Could you repeat the [inaudible]?

Instructor (Andrew Ng):Sure. So this one on the left mu of 1 given 2 equals mu1 plus sigma 1,2, sigma 2,2 inverse times X2 minus mu2 and this is sigma 1 given 2 equals sigma 1,1 minus sigma 1,2 sigma 2,2 inverse sigma 2,1. Okay? These are also in the lecture notes. Shoot. Nothing as where I was hoping to on time. Well, actually it is. Okay?

So it turns out – I think I'll skip this in the interest of time. So it turns out that – well, so let's go back and use these in the factor analysis model, right? It turns out that you can go back and – oh, do I want to do this? I kind of need this though. So let's go back and figure out just what the joint distribution factor analysis assumes on Z and X's. Okay? So under the factor analysis model Z and X, the random variables Z and X have some joint distribution given by – I'll write this vector as mu ZX in some covariance matrix sigma. So let's go back and figure out what mu ZX is and what sigma is and I'll do this so that we'll get a little bit more practice with partition vectors and partition matrixes. So just to remind you, right? You have to have Z as Gaussian with mean zero and covariance identity and X is mu plus lambda Z plus epsilon where epsilon is Gaussian with mean zero covariant si. So I have the – I'm just writing out the same equations again. So let's first figure out what this vector mu ZX is. Well, the expected value of Z is zero and, again, as usual I'll often drop the square backers around here. And the expected value of X is – well, the expected value of mu plus lambda Z plus epsilon. So these two terms have zero expectation and so the expected value of X is just mu and so that vector mu ZX, right, in my parameter for the Gaussian this is going to be the expected value of this partition vector given by this partition Z and X and so that would just be zero followed by mu. Okay? And so that's a d-dimensional zero followed by an indimensional mu. That's not gonna work out what the covariance matrix sigma is. So the covariance matrix sigma – if you work out definition of a partition. So this is into your partition matrix. Okay? Will be – so the covariance matrix sigma will comprise four blocks like that and so the upper left most block, which I write as sigma 1,1 – well, that uppermost left block is just the covariance matrix of Z, which we know is the identity. I was gonna show you briefly how to derive some of the other blocks, right, so sigma 1,2 that’s the upper – oh, actually, excuse me. Sigma 2,1 which is the lower left block that’s E of X minus EX times Z minus EZ. So X is equal to mu plus lambda Z plus epsilon and then minus EX is minus mu and then times Z because the expected value of Z is zero, right, so that’s equal to zero. And so if you simplify – or if you expand this out plus mu minus mu cancel out and so you have the expected value of lambda – oh, excuse me. ZZ transpose minus the expected value of epsilon Z is equal to that, which is just equal to lambda times the identity matrix. Okay? Does that make sense? Cause this term is equal to zero. Both epsilon and Z are independent and have zero expectation so the second terms are zero.

Well, so the final block is sigma 2,2 which is equal to the expected value of mu plus lambda Z plus epsilon minus mu times, right? Is equal to – and I won’t do this, but this simplifies to lambda lambda transpose plus si. Okay? So putting all this together this tells us that the joint distribution of this vector ZX is going to be Gaussian with mean vector given by that, which we worked out previously. So this is the new ZX that we worked out previously, and covariance matrix given by that. Okay? So in principle – let’s see, so the parameters of our model are mu, lambda, and si. And so in order to find the parameters of this model we’re given a training set of m examples and so we like to do a massive likely estimation of the parameters. And so in principle one thing you could do is you can actually write down what P of XI is and, right, so P of XI XI is actually – the distribution of X, right? If, again, you can marginalize this Gaussian and so the distribution of X, which is the lower half of this partition vector is going to have mean mu and covariance given by lambda lambda transpose plus si. Right? So that’s the distribution that we’re using to model P of X.

And so in principle one thing you could do is actually write down the log likelihood of your parameters, right? Which is just the product over of – it is the sum over I log P of XI where P of XI will be given by this Gaussian density, right. And I’m using theta as a shorthand to denote all of my parameters. And so you actually know what the density for Gaussian is and so you can say P of XI is this Gaussian with E mu in covariance given by this lambda lambda transpose plus si. So in case you write down the log likelihood of your parameters as follows and you can try to take derivatives of your log likelihood with respect to your parameters and maximize the log likelihood, all right. It turns out that if you do that you end up with sort of an intractable atomization problem or at least one that you – excuse me, you end up with a optimization problem that you will not be able to find and in this analytics, sort of, closed form solutions to. So if you say my model of X is this and found your massive likely parameter estimation you won’t be able to find the massive likely estimate of the parameters in closed form. So what I would have liked to do is – well, so in order to fit parameters to this model what we’ll actually do is use the EM Algorithm in with the E step, right? We’ll compute that and this formula looks the same except that one difference is that now Z is a continuous random variable and so in the E step we actually have to find the density QI of ZI where it’s the, sort of, E step actually requires that we find the posterior distribution that – so the density to the random variable ZI and then the M step will then perform the following maximization where, again, because Z is now continuous we now need to integrate over Z. Okay? Where in the M step now because ZI was continuous we now have an integral over Z rather than a sum. Okay?

I was hoping to go a little bit further in deriving these things, but I don’t have time today so we’ll wrap that up in the next lecture, but before I close let’s check if there are questions about the whole factor analysis model. Okay. So we’ll come back in the next lecture; I will wrap up this model and because I want to go a little bit deeper into the E and M steps, as there’s some tricky parts for the factor analysis model specifically. Okay. I’ll see you in a couple of days.

[End of Audio]

Duration: 75 minutes