

本文件之意義在於(後面的原話): The point here is we've now used Bayesian learning to derive a bunch of different things that we've actually been using all along. I think it is important because it told us something that we were thinking and tells us in fact we were right.

From 概率書(modified some wordings):

$P(A|B)$ 表示事件 B 發生的條件下 事件 A 發生的概率.

設樣本空間的一個劃分為 A_1, A_2, \dots, A_n , 則對事件 B 有: $P(B) = \sum_i P(B|A_i)P(A_i)$.

Bayes 公式:

設 A_1, A_2, \dots, A_n 為樣本空間的一個劃分, 則有:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

在 Bayes 公式中, 往往把 B 理解為「結果」, 樣本空間的劃分 A_1, A_2, \dots, A_n 理解為「原因」.

Tao: Bayes 公式求的就是已知「結果」, 求「原因」發生之概率.

Bayesian Learning

- LEARN THE BEST Hypothesis Given Data & some domain knowledge
- LEARN THE MOST PROBABLE H GIVEN Data & domain knowledge

$$\text{argmax}_{h \in H} P_r(h|D)$$

D 俱體是甚麼意思? 看了第 2 段就清楚了.

1. HI Michael. >> Hey how's it going? >> So I want to talk about something today Michael. I want to talk about Bayesian Learning, and I've been inspired by our last discussion on learning theory to think a little bit more about what it is exactly that we're trying to do. I'm in the mood beyond specific algorithms to just think more generally The sort that learning people want us to do, learning theory people want us to do and I think Bayesian Learning is a nice place to start. Sound fair? >> Yeah, that sounds really cool, I think that might be a nice formal framework for thinking about some of these problems. >> Good. Good. So, I'm going to start out. By making a few assertions, which I hope you

will agree with, and if you agree with this then we'll be able to kind of move forward and ask some pretty cool questions okay? So Bayesian learning, so the kind of idea here behind Bayesian learning is this sort of fundamental Underlying assumption about what we're trying to do with the machine learning. So, I've written it down here, here's what I'm going to claim we're trying to do. We are trying to learn the best hypothesis we can given some data and some domain knowledge. Do you buy that as an assertion? >> Yeah, it's, and pretty much everything we've talked about so far has had a form kind of like that. We're searching through a hypothesis base and As you've pointed out on multiple occasions there's this kind of extra domain knowledge that comes into play for example when you pick a like a similarity metric first thing like [INAUDIBLE] >> Right and of course we always have the data because we're machine learning people and we always have data. So this is what we've been trying to do and I'm going to suggest that we can be a little bit more precise about what we mean by best and I'm going to try to do that and see if you agree with me. Okay, so I'm going to rewrite what I've written already except I'm replacing best with most probable. Okay. So what I'm going to claim is that what we've really been trying to do with all these algorithms we're doing is we're trying to learn the most likely or the most probable hypothesis given the data and whatever domain knowledge we bring [UNKNOWN]. You buy that? >> Interesting. I'm not sure yet. I mean, so is it the hypothesis that it's most likely to be returned by the algorithm? >> No, it's the hypothesis that we think is most likely, given the data that we've seen. Given the training set and given whatever domain knowledge that we bring to bear on the problem, the best hypothesis is the one that is most likely, that is Most probable. Or most l, probably the correct one. >> Interesting. Well, are we going to be able to connect that to what we were talking before? Which is generally we were selecting hypotheses based on things like their error. >> Yes. Exactly. We are going to be able to connect that. We are definitely going to be able to connect that. But. >> Okay. >> I can;t go forward unless I can convince you that it's reasonable to at least start out thinking about best being the same thing as most probable. Yeah, I'm willing to go forward with this. It sounds interesting. >> So if you're willing to move forward with this, then I want to write one more thing down and then we can sort of dive into it. So if you buy that we're trying to learn the most probable hypothesis, the most likely one, the one that has the highest chance of being correct given the data, and our domain knowledge, then we can write that down in math speak pretty simply. It's the probability of, some particular hypothesis h , drawn from some hypothesis class. Given some amount of data which I'm just going to refer to as D from now on. Okay? And that's just, exactly what we said just above when we talk about the most probable age, given the data. Okay? >> Well wait. Two things. One is so D is not distribution which we've had in the past. >> That's true. >> So I guess as long as we keep that straight. And the other one is No that's, you're just telling me the probability of some particular hypothesis h . >> That's right. So, we want to somehow, given this quantity we want to find, the best or the, most likely, of the hypothesis given this. Does that make sense? >> Yes. >> So we want to find the argmax (argmax 意思見下), of h , drawn from your hypothesis class. That is we want to find the hypothesis drawn from the hypothesis class that has the highest probability given the data. >> Perfect. >> Okay, good. So we're going to spend the next 45 hours. >> [LAUGH] >> Exploring this expression. >> Okay so that's like what, like 6 hours per letter. >> [LAUGH] Yeah and that's fine because its important.

From online:

What is the difference between argmax and max ?

To use a mathematical example, consider the function $f(x) = 1 - x^2$.

Then $\text{Max } f(x) = 1$

but $\text{Argmax } f(x) = 0$

as $x = 0$ is the (unique) value for which $f(x) = \text{max } f(x) = 1$.

Bayesian Learning

$$\underset{h \in H}{\operatorname{argmax}} \Pr(h|D)$$

$$\Pr(h|D) = \frac{\Pr(D|h) \Pr(h)}{\Pr(D)}$$

prior on the data

data gives the hypothesis

$$D = \{(x_i, d_i)\}$$

Bayes' Rule

$$\Pr(a,b) = \Pr(a|b)\Pr(b)$$

$$\Pr(a,b) = \Pr(b|a)\Pr(a)$$

$$h(x) = \{x \geq 10\}$$

$$x = 7, \begin{matrix} T \\ F \end{matrix} \begin{matrix} 0 \\ 1 \end{matrix}$$

(xi, di)的意思 和 $h(x)=\{x \geq 10\}$ 的意思 見文中
注意{(xi, di)}就是 D

2. Alright Michael. So like I said, we're going to spend all this time trying to, to unpack this particular equation. And the first thing we need to do is we need to come up with another form of it that we might have some chance of actually understanding of actually getting through. So I want to use something called **Bayes' rule**. Do you remember Bayes' rule? >> I do. >> Okay, what's Bayes' Rule? >> The man with the Bayes makes the rule. Oh wait, no, that's the golden rule. >> That's right, no. >> The Bayes Rule, is, it relates, it, I don't know. I think of it as just letting you switch which thing is on which side of the bar. >> Okay, so. >> Do you want me to give the whole expression? >> Yeah, give me the whole expression. >> So if we're going to apply Bayes' Rule to the probability of h given D. We can move, turn it around and make it equal to the probably of D given H. And it would be great if we could just stop with that, but we can't. We have to now kind of put them in the same space. So, we multiply by the probability of H, and then we divide by the probability of D. And sometimes that's just a normalization and we don't have to worry about it too much. But that's, that's the bay, that's Bayes' rule right there. >> So this is Bayes' rule. And it actually is really easy to derive. It falls it follows directly from the chain rule in probability theory. Do you think it's worthwhile? Showing people that or just they should just accept it. >> Well, I mean, you could just, you might be able to just see it. Just, the, the thing on top of the, the normalization, the probability of D given h times probability of h. That's actually the probability of D and h together. Right. So the probability of h times the probability of d over h as you say also the chain rule basically the definition of conditional probability in conjunctions and if you move the probability of d over to the left hand side you can see we're really just saying the same thing two different ways. It's just the probability of h and d. So then we're done. >> No, that's right. So I can write down what you just said. And use different letters just to make it more confusing, so >> Oh good. >> You can point out that the probability of A and B, by the chain rule, is just the probability of A given B, times the probability of B. But because order doesn't matter, it's also the case that the probability of A and B. Is the probability of b given a times the probability of a. And that's just the chain rule. And so if these two quantities equal to one another's exactly what you say, I could say well, the probability of a given b is just the probability of b given a times the probability a divided by the probability of b. And

that's exactly what we have over here. >> Good. So now that we've mastered that all your Bayes are belong to us. [LAUGH] >> How long have you been saying that? >> The...just, only about 3 or 4 minutes. >> [LAUGH] Fair enough. Okay, so we have Bayes's rule. And what's really nice about Bayes's rule is that while it's a very simple thing, it's also true. It follows directly from probability theory. But more importantly for machine learning, it gives us a handle to talk about. What it is we're exactly trying to do when we say we're trying to find the most probable hypothesis, given the data. So let's just take a moment to think about what all these terms mean. We know what this term here means. The, it's just the probability of some hypothesis given the data. But what do all these other terms mean? I want to start with this term, the probability of the data. It's really nothing more than your prior belief of seeing some particular set of data. Now, and as you point out, Michael, often it just ends up to be a normalizing term and typically does not matter, though we'll see a couple of cases where it does matter, helps us to, to sort of think about a few things. But generally speaking, whatever it is Since the only thing that we care about is the hypothesis, we're trying to find that, the probability of the data doesn't depend on the hypothesis, so typically we ignore it, but it's nice to just be clear about what it means. The other terms are a bit more interesting. They matter a little bit more. This term here, the probability is the probability of the data given the hypothesis right? >> Mm. Seems like learning backwards. >> It does seem like learning backwards but [what's really nice about this quantity \(\$\Pr\(D|h\)\$ \) is that unlike the other quantity, the probability of the hypothesis given the data, it's actually, turns out to be pretty easy to think about the likelihood that we would see some data given that we were in a world where some hypothesis, \$h\$, is true. So there is a little bit of subtlety there and I, let me, let me unpack that subtlety a little bit. So we've been talking about the data if its sort of a thing that is floating out in air, but \[we know that the data is actually our training data. And it's a set of inputs \\(\\$x_i\\$ \\) and lets just say for the sake of argument we are going to do classification learning, it's a set of labels \\(\\$d_i\\$ \\) that are associated with those inputs.\]\(#\) So just to drive the point home, I'm going to call those \$d\$'s, little \$d\$'s. And so our data is made up of a bunch of these training examples. \[And these training examples are whatever input that we get\]\(#\) coming from a teacher, coming from ourselves, coming from nature, coming from somewhere \[and the associated label that goes along with them. So when you talk about the probability of the data given the hypothesis, what you're talking about, well, what's the likelihood that given that I've got all of these \\$x_i\\$'s and given that I'm living in a world where this particular hypothesis that I would see these particular labels.\]\(#\) Does that make sense Michael? >> I see. Yeah, so, so I can imagine a more complicated kind of notation where, we're, we're kind of accepting the \$X\$ s as given. But the labels is what we are actually saying is something that we want to assigned probability to. >> Right so its not really that the \$x\$'s matter in the sense that we are trying to understand those. What really mattes re the labels that are associated with them. And we will see an example of that in a moment. But I wanted to make sure that you get this subtled. >> So in a sense then I guess you're saying that the probability of \$D\$ given \$H\$ component, or, or quantity, is really like running the hypothesis. It's like, It's like labeling the data. >> Okay Michael, just to make sure we get this. \[Let's imagine we're in a universe, where the following hypothesis is true. It returns true, in exactly the cases where some input number \\$X\\$, is greater than or equal to 10 And it returns false otherwise.\]\(#\) Okay? >> Yup. >> Okay. So here's a question for you. \[Let's say that our data was made up of exactly one point. And that value set \\$x\\$ equal to 7. Okay? What is the probability that 「 the label associated with 7 would be true 」 ?\]\(#\) >> Huh. So you're saying we're in a world where \$h\$ is holding and that the \$h\$, \$h\$ is being used to generate labels. So it wouldn't do that right? So, \[the probability ought to be zero.\]\(#\) >> That's exactly right and \[what's the probability that it would be false?\]\(#\) 1 minus 0 \[LAUGH\] which \[we'll call 1.\]\(#\) >> Which we'll call 1. That's exactly right. So it's, it's just that simple. That, the probability of the data given the hypothesis, is really about, given a set of \$x\$'s, what's the probability that I would see some particular label. Now, what's nice about that is, is, as you point out, is that, it's as if we're running the hypothesis. Well, given a hypothesis, it's really easy, or at least it's easier usually, to compute the probability of us seeing some labels. So, \[this quantity \\(\\$\Pr\\(D|h\\)\\$ \\) is a lot easier to figure out than the original quantity \\(\\$\Pr\\(h|D\\)\\$ \\) that we're looking for.\]\(#\) The](#)

expect to change, go up or go down, or stay the same, that would influence whether the probability of a hypothesis goes up. >> So the probability of the hypothesis given the data, what could make that combined quantity go up, so one is looking at the right hand side, the probability of the hypothesis, so, so if you have a hypothesis that has a higher prior, has, is more likely to be a good one. Before you see the data then that would raise it after you see the data too. >> Right. >> And I guess the probability of the data given the hypothesis should go up. Oh, which is kind of like accuracy. It's kind of like saying that if you pick a hypothesis that does a better job of labeling the data, then also your probability of the hypothesis will go up. >> Right. Anything else? >> I guess the probability of the data going down. But that's not really a change from the hypothesis. >> Right. But it is true that if those goes down, then the probability in the hypothesis can and the data will go up. But as you point out, it's not connected to the hypothesis directly. And I'll write in equation for you in, in just a moment that'll kind of make that, I think, a little bit clearer. Okay, but you got all this, right? So I think you understand it. So we got Bayes' Rule. And, notice what we've done. We've gone from this sort of general notion of saying we need to find the best hypothesis, to actually coming up with an equation, that sort of makes explicit what we mean by that. That what we care about is the probability of some hypothesis given the data. That's what we mean by best. And that, that can be further thought as, the probability of us seeing, some labels on some data, given hypothesis. Times the probability of the hypothesis, even without any data whatsoever, normalized by the probability of the data. So let's play around with Bayes' rules a little bit and make certain that we all, we all kind of get it. Okay? >> Sure. >> Okay.

4. Okay Mike, are you ready for a quiz? >> Uh-huh. Okay, so, here, let me, let me set up the, the situation for you. So a man goes to see his doctor, okay, because his back hurts or something. >> Aww. >> And she gives him a, I know, it's really sad. It's his, the left side of his lower back, he's been playing too much racquetball. Anyway, so a man goes to see a doctor, and she gives him a lab test. Now this test is pretty good, okay? It returns a correct positive. That is, if you have the thing that this lab test is testing for, it will say you have it 98 percent of the time, okay? So it only gives you a false positive two percent of the time. And at the same time, it will return a correct negative, that is if you don't have what the lab test is testing for, it will say you don't have it. 97% of the time, so it has a false negative rate of only 3%. >> Wait, hang on. So, just, what's his problem? >> Oh, that's the question. So, the test looks for a disease. So, give me a disease. >> Spleen? >> Okay, I like that. So the test looks for spleentitis. Now spleentitis is such a rare disease that nobody's ever heard of it, And it turns out that it's so rare that only about this fraction of the population has it. Okay? >> Mm-hm. >> That make sense? So we're looking for spleentitis. It's a very rare disease, but this test is really good at determining whether you have it or determining whether you don't have it >> Can I tell you that, its, spleentitis appeared zero times in google. [LAUGH] So it really is quite rare. >> It really is quite rare. But what does google know? OK, so you got it all Michael? >> Yeah. So its a really rare disease and we have a very accurate test for it. >> Good. Man goes to see the doctor. She gives him a lab test. Its a pretty good lab test. Its checking for spleentitis, relatively rare disease and the test comes back positive. >> Oh. >> Yes. So, test is positive. So, here is the quiz question. >> Should we be net, notifying his next of kin? >> Yes. Does he have spleentitis? >> You said, just said he had spleentitis. >> No, I said the test says he had spleentitis. Or the test looks for spleentitis, and the test came back positive. So, does he have spleentitis? Yes or no? Alright, before I try to answer that can I just, ask for clarification, can I get a clarification? >> Please. >> So the 98 is a percentage and the 97 is a percentage, is .008 also a percentage? >> No it's not. So if I wanted to convert it to a percentage it would be .8%. >> Got it. Alright, now I think I have, what I need. >> Okay, alright, so, you think about it. Go.

Quiz : BAYES' RULE

Prior's matter

A man goes to see a doctor. She gives him a lab test. The test returns correct positive 98% of the time, and a correct negative 97% of the time. The test looks for spleentitis ~ .008 has it. 50%
50%
5%
.8%

TEST IS POSITIVE!

Does he have spleentitis?

NO

$$P(s|+) = P(+|s)P(s)/P(+)$$

$$= .98 \cdot .008 = .00784 \sim 21\%$$

$$P(s|+) = P(+|s)P(s)/P(+)$$

$$= .05 \cdot .992 = .02976 \sim 79\%$$

題目的意思就是, 如果那人實際得了 spleentitis 病, 則「實驗結果顯示得了 spleentitis 病」的概率為 98%, 如果那人實際沒得 spleentitis 病, 則「實驗結果顯示沒得 spleentitis 病」的概率為 97%. 現在已知「實驗結果顯示得了 spleentitis 病」, 則那人實際得 spleentitis 病的可能性有多大? 上圖中的解法既不好懂, 還算錯了 (21%應為 26%). 所以不要看上圖的解法, 看下面我的就可以了:

Bayes 公式:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

已知:

P(測出得了 | 實際得了) = 98%

P(測出沒得 | 實際沒得) = 97%

要求:

P(實際得了 | 測出得了),

即 A1 = 實際得了, A2 = 實際沒得, B = 測出得了, 故:

P(實際得了 | 測出得了)=

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_i P(B|A_i)P(A_i)} = \frac{98\% \times 0.8\%}{98\% \times 0.8\% + (1 - 97\%) \times (1 - 0.8\%)} = 26.3\%,$$

26.3%比較小, 所以可以斷定他沒得 spleentitis 病.

5. >> Okay Michael, what's the answer? >> Does he have spleentitis? >> Yes, does he have spleentitis? >> I don't think we know, for sure. >> Mm? What do mean by that? >> Well, I mean. It's a noisy and probabilistic world right. So the test told us that things look like he has spleentitis and the test is usually right. But the test is sometimes wrong and it can give the wrong answer and that's really all we know, so we can't be sure. >> Okay but if you had to pick one. If you had to yes or no, like our students they did when they took the quiz. Which one would you pick? Yes or no. >> So, I guess C the pants. I would just say, yes because the test says, yes but if I guess I was trying to be more precise, I may go through and work out the probability and I guess if it's more likely to have than not to have, then I'd say and otherwise I'd say, no. >> Okay. So how would you go about doing that? Walk me through it. >> Based on the name of the quiz, I think I'd go with Bayes' Rule. >> Okay. So [LAUGH] I like that. So Bayes' Rule, is everyone recall, is the probability of the hypothesis given the data is equal to the probability of the data given the hypothesis times the probability of the hypothesis divided by the probability of the data. So, >> [LAUGH] >> Let's write all that out. So what is the probability of spleentitis, which I'm just going to write as an s. Given. >> We're making jokes about spleentitis, but we don't want that to be confused with splenitis, which is a real thing and probably not very pleasant. So apologies to anyone out there with splenitis. But this is spleentitis, which is really totally different. >> Is splenitis a real thing? >> Yeah. >> :Really what is it? >> Enlargement and inflammation in the spleen and the spleen as a result of infection or possibly a parasite infestation or cysts. >> So what you're saying is that's gross and we don't want to think about it. OK good so Woo okay, so the probability of getting splentitis and probably isn't even real. >> Totally, its totally different, its definitely not real >> Yea definitely not. Given that we gotten a positive result and you say that we should use Baye's rules so that would be in this case what? >> So it's the same as the probability of the positive result given that you have spleentitis. >> Mm-hm. >> Times the probability, the prior probability of having spleentitis. >> Mm-hm. >> And I want to say normalize, but like divided by the probability of a positive test result. >> And what would be, the probabili. The other option is that you don't have spleentitis. >> Mm-hm. >> Even though you got a positive result. And that would be equal to? >> The probability of a positive result given you don't have spleentitis. >> Mm-hm. >> Times the prior probability of not having spleentitis. >> huh. >> Divided by the, again the same thing. The probability of the test results. So that's, those two things added together, needed to be one. >> Right. But as you point out. If we just want to figure out which one is bigger than the other. We don't actually have to know this. >> Hm, good point. >> So we can ignore it, okay. Okay, so, let's compute this. So, what is in fact, the probability of me getting a plus, given that I have spleenitis? >> Right. So it says in the setup, the test results correct positive 98% of the time. So, I, I think that's what it means. It means that if you really do have it, it's going to say that you have it with that probability. >> Okay, so That's just point nine eight. OK? And that's times the prior probability of having spleentitis which is? >> Right. .008. And what's that equal to? >> It is equal to. 0.0078. >> Okay, fine. We can do the same thing over here. So what's the probability of getting a positive if you don't have spleentitis >> So, the probability of a correct negative is 97%. That means if you really don't have it, it's going to say you don't have it, so probability of positive result given that you don't have it, that should be the 3%. >> That's exactly right. Times the prior probability of not having spleentitis which is? >> That's right, and that is equal to? >> So, which number is bigger? >> The one that has the larger significant digit. >> Which one of those two is that? >> I mean, obviously, the one that's bigger is the, you don't have it. >> That's right. So the answer would be no. >> And in fact the probability is almost 80%. >> Yeah. >> Which is crazy. So, it's like, you go into the doctor, you've run a test, the doctor says congratulations, you don't have speentitis, because the test says you do. >> That's right. [LAUGH] >> So, what does that tell you? >> That seems stupid. >> That does seem stupid, but what does it tell you About Bayes' Rule. What is Bayes' Rule capture. What is thing that make the answer no, despite the fact, you have a high reliability test that says yes. >> I. Okay. So I guess, I guess the way to think

about it is, a random person showing up in the doctors office, is very unlikely to have this particular disease. And even the tiny, little, small percentage probability that the test would give a wrong answer is completely swamped by the fact that you probably don't have the disease. But I guess this isn't really factoring in the idea that, you know, presumably this lab test was run for some other reason. There was some other evidence that there was concern. >> Or the doctor just really wanted some more money, because She needs a new boat. >> Yeah, I know a lot of doctors. >> I do too. >> And most of them don't work like that. >> Yeah most, well most of them have PhD's not MD's. So, another way of summarizing what you just said Michael, I think, is that priors (即 $Pr(h)$ 和 $Pr(D)$, 見第 3 段圖) matter. >> I want to say the thing that I got out of this is tests don't matter. >> Well, tests matter. >> Like what's the purpose of running a test if it's going to come back and say. Well it used to be that I was pretty sure you didn't have it and now I am still pretty sure you don't have it. >> Well the point of running a test is you run a test when you have a reason to believe that the test might be useful. So what is the one thing, if I could only change one thing without getting completely ridiculous, what's the easy well, I don't know what's easy, what's the easiest thing for me to change about this setup. I have three numbers here. This one, this one and this one. What would be the easiest number to change? >> Well, in some sense none of them seem that easy to change but I guess maybe what you're trying to get me to say is that if we look at a different population of people then we can change that .008 number to something else, like if we only give the test to people who have other signs of splentitis. Then then it, it would probably be a much bigger number. >> Right, so changing the test, making the test better might be hard, presumably you know, billion of dollars of research have gone into that, but if you don't give the test to people who you don't have any reason to believe have Splentitis, just walking off the street, as you put it, a random person walking off the street, then you can change the priors, so some other evidence. That you might have splentitis might lead the prior to change, and then the test would suddenly be useful. So this, by the way, is an argument for why you don't want to just require that everyone take tests for certain things. Because if the prior probability is low, then the test isn't very useful. On the other hand, as soon as you have any reason to believe we have strong evidence that someone might have some condition, then it makes sense to test them for it. >> So it's like a stop and frisk situation. >> It's exactly like a stop and frisk situation. I'm looking at you [INAUDIBLE]. Okay But in some sense, your use of the word prior is a little confusing there. So it's not that we're changing the prior, it's that we're...we have some additional evidence that we can factor in. And I guess we can imagine that that's part of the prior, but it seems like it's post-ilia. >> Yeah, it does. And it, but... One way to think about it, you actually, I think you just captured it in what you just said, right? Which is you can think of as a prior. Well, a prior to what? So it's your prior belief over a set of hypotheses, given the world you happen to be in. If you're in a world where random people walk in to take a test for splentitis, then there's a low prior probability that they have it. If you're in a world where the only people who come in are people who are from a population where the prior probability is significantly higher, then you would have a different prior. It's really a question about where you are in the process when you actually formulate your question. >> So would it be worth asking people how, how likely would it have to be that you have splentitis to make this test at all useful? Right, that would change a positive, a positive result would actually change your mind about whether someone has it. >> yeah, actually that, I think that's something that I, I'll leave for the for the, for the interested reader, where would that prior probability have to change so that getting a positive result, I would be more likely to believe that you actually have it than not. That does bring up a philosophical question, though, which is So what, just because the priors have changed, doesn't mean that suddenly the test is useful, or that the test is going to give you an answer that somehow distinguishes and is this positive. And from a mathematical point of view, the question of whether this number is 0.008 or, or 0.8, you know, 8 10ths of a percent, where does it change? Does it change at 5%? Or does it change at 50%? Or does it change at 500%? It probably changes at 500%. You know, what, where is the place in which suddenly a positive result would make you believe they actually had splentitis or whatever disease you're looking

for. Okay? >> Okay.

Bayesian Learning

For each $h \in H$

calculate $P(h|D) = P(D|h) P(h) / P(D)$

OUTPUT :

$$h = \underset{h}{\operatorname{argmax}} P(h|D)$$

h : candidate hypothesis,

H : hypothesis space

$h = \operatorname{argmax} P(h | D)$ 意思是: simply output whichever hypothesis has maximum probability

6. Okay, Michael, so we've gotten through that quiz and you see that Bayes' rule actually gives you some information. It actually helps you make a decision. So I'm going to suggest that, that whole exercise we went through was actually our way of walking through an algorithm. So here's a particular algorithm that follows from what we just did. And let me just write that down for you. All right, so here's the algorithm, Michael, so it's very simple. For each H in H , that is, each candidate hypothesis in our hypothesis space, simply calculate the probability of that hypothesis given the data W which we know is equal to the probability of the data given that hypothesis times the prior probability of the hypothesis, divided by the probability of the data. And then simply output whichever process has maximum probability. Does that make sense? >> Yeah.

Bayesian Learning

For each $h \in H$
calculate

$$P(h|D) \doteq P(D|h) P(h)$$

OUTPUT :

$$h_{\text{map}} = \underset{h}{\operatorname{argmax}} P(h|D)$$

$$h_{\text{ml}} = \underset{h}{\operatorname{argmax}} P(D|h)$$

uniform

MAP =
maximum
a posteriori

ML =
maximum
likelihood

NOT PRACTICAL

h_{MAP} 完整的表達式是(後面會用到):

$h_{\text{MAP}} = \operatorname{argmax}_h P(h | D) = \operatorname{argmax}_h P(D | h) P(h)$. h_{MAP} 就是我們要求的那個。

>> Okay, so I do want to point out that (下句話的意思是: 由於我們要找的是 hypothesis that has maximum probability, 所以實際上不用算 $P(D)$) since all we care about is computing the argmax, as before, we don't actually ever have to compute that little bit so, and that's a good thing because we don't always know what the prior probability on the data is, so we can ignore it for the purposes of finding the maximal hypothesis. >> So the place you removed it from, it seems like that's not actually valid, because it's not the case that the probability of h given d equals, it's the probability of d given h times the probability of h . It just means that we don't care what the probability is when we go to compute the argmax. That's right, so, in fact, it's probably better to say that I'm going to approximate the probability hypothesis given the data by just calculating the probability of the data given the hypothesis times the probability of the hypothesis and just go ahead and ignore the denominator. Precisely because it doesn't change the maximal age. >> Yeah, so it's, it's nice that that goes away. >> Right, because it's hard to know, often what the prior, what the prior probability over the data is. >> It would be nice if we didn't have to worry about the other one, either. >> Which other one? >> The probability of h , where's that coming from? >> right, so where does that come from? So that's a deep philosophical question. Sometimes it's just something you believe, and you can write down. And sometimes it's a little harder. And that's actually good that you bring that up. When we compute, our probabilities this way (即只算 $P(D|h) P(h)$) so it's actually got a name, it's the MAP or the maximum a posteriori hypothesis and that makes sense, it's the biggest posterior given all of your priors. But you're right Michael that often it's just as hard to say anything particular about your prior over the hypothesis (即 $P(h)$) as it is to say something about your prior of the data (即 $P(D)$) and, so it is very common to drop that. And, in dropping that, we're actually computing the argmax over the probability of the data given the hypothesis (即 $P(D|h)$). And, that is known as the maximum likelihood hypothesis. >> I guess you can't call it the maximum A priori hypothesis, because then it would also be MAP. >> Exactly, although I've never thought about that before. By the way, just just to be clear, we're not really

dropping this (即 $\Pr(h)$), in this case, what we really said, is that, our prior belief is that all hypotheses are equally likely. So we have a uniform prior that is, the probability of any given hypothesis is exactly the same as the probability as any other given hypothesis. >> I see, so you're saying if, if we assume that they all are equally likely, then, the choice of hypothesis doesn't change that term at all, the p of h term, so it really is equivalent to just ignoring it. >> Exactly, in some constant, we don't even have to know what the constant is. But whatever it is, it's the same everywhere and therefore it doesn't affect the other terms or, in particular, affect the argmax computation. >> So that's actually pretty cool right? Once you think about what we just did. We just took something that was very hard. Computing the probability of a hypothesis given the data and turned it into something much easier that is... Computing the probability of you seeing the data labels given a particular hypothesis and it turns out that those are effectively the same thing if you don't have a strong prior. So that's really cool, so we're done right? We now know how to find the best hypothesis. You're just finding the most likely hypothesis or the most probable one and that turns out to be the same thing as just simply finding the hypothesis that best matches the data. We're done its all, its easy. Everything's good. >> So, the math seems very nice and pretty and easy but isn't it hiding a lot of work to actually do these computations? >> Well, sure well well look you know how to do multiplication that's pretty easy right? >> [LAUGH]. >> So I guess the only hard part is we have to look at every single hypothesis. >> Yeah, that's just a slight, little, you know, issue. >> So, mathematically meaningful, but computationally questionable. >> Hm. >> So, the big point there, is that it's not practical. Well, unless the number of hypotheses is really, really small. But as we know, a lot of the hypotheses spaces that we care about, like, for example, linear separators, are actually infinite. And so it's going to be very difficult to use this algorithm directly. But despite all that, I think that there's still something important that we get out of thinking about it this way in just the same way that we get something important out of thinking about vc dimension. Even if we're not entirely sure how to compute it in some particular case. This really gives us a gold standard, right? We have an algorithm, at least a conceptual algorithm, that tells us what the right thing to do would be if we're capable of computing it directly. So, that's good because we can maybe prove things about this and compare results that we get from some Real live algorithms to what we might expect to get but also it turns out it's pretty cute because it helps us to say other things about what it is we actually expect to learn. And I'm going to give you a couple examples of those just to sort of prove my point, sound good? >> Yeah. >> Okay.

Bayesian Learning in Action!

① Given $\{ \langle x_i, d_i \rangle \}$ as noise-free examples of c

② $c \in H$

③ uniform prior

$$Pr(h) = \frac{1}{|H|}$$

$$Pr(D|h) = \begin{cases} 1 & \text{if } d_i = h(x_i) \quad \forall x_i, d_i \in D \\ 0 & \text{otherwise} \end{cases}$$

$$Pr(D) = \sum_{h \in H} \underbrace{Pr(D|h)}_{1 \text{ if } h \in VS_{H,D}} \underbrace{Pr(h)}_{\frac{1}{|H|}} = \sum_{h \in VS_{H,D}} 1 \cdot \frac{1}{|H|} = \frac{|VS|}{|H|}$$

$$Pr(h|D) = \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS|}{|H|}} = \frac{1}{|VS|} \quad h \in |VS|$$

$$Pr(h|D) = \frac{Pr(D|h)Pr(h)}{Pr(D)}$$

x_i 和 d_i 表示甚麼意思? 見第 8 段前的藍字.

第 1 點最右是 examples of c

$Pr(D|h)$ 式中右邊是 1 if... 和 0 otherwise

From previous note:

version space is essentially the space of all the hypotheses, that are consistent with the data (即 training set)

最後一行, $Pr(h|D) = 1 * 1/|H| / (|VS| / |H|)$, 為何是 $1 * \dots$? 為何沒有 $0 * \dots$?

答: 實際上有 $0 * \dots$, 只是這裡沒寫. 下面的文字說了: And by the way, if it's not consistent with it, then it's zero. 注意此式後面有個 $h \in |VS|$.

7. >> Okay Michael, so let's see if we can actually use this as a way of deriving something maybe that we already knew. So I'm going to go through a couple of these because I actually think, well, frankly I just think it's kind of cool. But, I'm hoping I can convince you it's sort of cool too and that we get something out of it. Okay, so let me set up the word, I'm going to set up a problem, and it's going to be a kind of generic problem, and I'm going to see what we can get out of it, okay? So this is machine learning, so we're going to be given a bunch of data, so there are three assumptions that I'm going to make here. The first is that we're going to be given a bunch of labeled training data, which I'm writing here as x_i and d_i , so x_i is whatever the input space is, and d_i are these labels. And let's say, it doesn't actually even matter what the labels are, but let's say that the labels are classification labels. Okay? >> Hm. >> All right. And furthermore, not only we're given this data as examples drawn from some underlying concept c , but they're, in fact, noise-free. Okay? So they're true examples that tell you what c is. Okay? >> Mm-hm. >> So I'm going to say, in fact, let me write that down because I think it's important. They're noise-free examples. Okay. >> Like $d_i = c(x_i)$. >> That's right, for all x_i . So,

the second assumption, is that the true concept c is actually in our hypothesis space, whatever that hypothesis space is. And finally, we have no reason to believe that any particular hypothesis in our hypothesis space is more likely than any other. And so, we have a uniform prior over our hypotheses.

>> So it's like the one thing we know is that we don't know anything. >> That's right. So, sometimes people called this an uninformative prior because you don't know anything. Except of course I've always thought that's a terrible name because it's a completely informative prior. In fact it's equally as informative as every other prior in that it tells you something that all hypothesis are equally likely. But that's >> I thought it was called an uninformed prior. >> Is it? So it's just an ignorant prior is what you're telling me. Yeah. >> Okay. Well, then maybe that's the problem. I just always had a problem with it because people keep calling it uninformative and the really mean uninformed. Okay. In any case, so these are our, these are our assumptions. We've got a bunch of data, it's noise free, the concept is actually in the hypothesis base we care about and we have a uniform prior. So we need to compute the best hypothesis. So given that we want to somehow compute the probability of some hypothesis given the data, right? That's just Bay's Rule. So, Michael, you've got the problem right? >> Yes. >> [LAUGH] okay. So in order to compute the probability of a hypothesis given the data, we just need to figure out all of these other terms. So let me just write down some of the terms and you can tell me what you think the answer. Okay. >> Well, what was the question? >> The question is, well we want to compute some kind of expression for the probability of a hypothesis given the data. So given some particular hypothesis, I want to know what's the probability of that hypothesis given the data, okay? >> Yeah. >> Okay, you got the setup. So, we're going to compute that by figuring out these three terms over here. So, let's just pick, one of them to do. Let's try the prior probability. So Michael, what's the prior probability on H ? >> Did we say that it was a finite hypothesis class? >> It is a finite hypothesis class. >> Then it's like, one over the size of that hypothesis class because it's uniform. >> Exactly right, uniform means Exactly that. Okay so we've got one of our terms, good job. Let's pick another term. How about the probability of data given the hypothesis. What's that? >> The probability, so I guess noise free, and we know that it's noise free so it's always, so they're always going to be zeros and ones (跟 noise free 甚麼關係? 見第 8 段前的藍字). >> Mm-hm. >> So, and it's going to be a question of whether or not the data is consistent with that hypothesis. Right, if the labels all match. >> Right. >> What we expect them to be if that really were the hypothesis, then we get a one, otherwise we get a zero. That's exactly right. So let me see if I can write down what I think you just said. The probability of the data, given the hypothesis, is, therefore one if it's the case, that the labels and the hypothesis agree for every single one of the training exercises. Right? >> Yep >> Is that what you said? Good. And if any of them disagree, then the probability is zero. So that's actually very important. It's important to, to understand exactly what it means for, to have the probability to get a hypothesis, as we mentioned before. That the English version of this is, $\Pr(D | h)$ 即: what's the probability that I would see data with these labels in a universe where h is actually true. Which is different from saying that H is true or H is false. It's really a common about the labels that you see on a data. In a universe, where H happens to be true. >> Okay, but you know, it's occurring to me now that you wrote that down, that we've talked about this idea before. >> When? >> Well, so, like there's a shorter way of writing that. Which is D if H equals one if H is in the version space of D . >> Huh, that's exactly right, that's exactly right. So, in fact, that will help us to compute the final term that we need, which is the probability of seeing the data labels. So, how do we go about computing that? Well, it's exactly going to boil down to the version space as you say, let me just write out a couple of steps so that it's pretty Kind of easy to see. It's sometimes easier in these situations to kind of break things up. So, the probability of the data sort of formally, is equal to just this. So we can write the probability of the data as being, basically, a marginalized version of the probability of the data given each of the hypotheses times the probability of the hypotheses. Now, this is only true in a world where our hypotheses are mutually exclusive. Okay so let's assume we are in that world because frankly that's what we always assume and this little trick is going to workout for us because we are going to get to take advantage of two terms that we already

computed naming the probability that the data given the hypothesis and the probability of a particular hypothesis so we know that prior probability of a hypothesis is right, its just one over the size of the hypothesis space and how am I going to substitute in this equation for the probability of the data given the hypothesis? >> So, I don't know. I would write that differently. I mean, it's basically it's like the indicator function on whether or not H_i is in the version space of D . >> Right, that's exactly right. So in fact this is not a good way to have written it. Let's see if I can come up with a, a good notational way of doing it. Let's say, for every hypothesis that is in the version space of the hypothesis space given the labels that we've got. Okay? How's that count? >> Okay. >> So rather than having to come up with an indicator function, I'm just going to define V as the subset of all those hypotheses that are consistent with the data. >> Yeah exactly >> Okay, and so what's the probability of those? >> One It's one and it's zero otherwise, so then, we can simplify the sum and it's simply what? ? >> The sum of the one, ooh! The one of each doesn't even depend on the hypothesis. >> mm-mh! >> I see wait I don't see oh yes I do, I do its one over the size of version space. No its the size of the version space over the size of the hypothesis space. >> That's exactly right. Basically for every single hypothesis in the version space we're going to add one and how many of those are? Well the size of the version space number of those. And multiply all that by one over the size hypothesis space, and so the probability the data is that term. So now we can just substitute all of that, into our handy dandy equation up there, and let's just do that. So the probability of the hypothesis given the data, is the probability of the data given the hypothesis Which we know is one for all those that are consistent, zero otherwise. The probability of the prior probability over the hypothesis is just one over the size of the hypothesis space, and the probability of the data is the size of the version space Over the size of the hypothesis base which, when we divide everything out, is simply this. Got it? >> Got it. >> So, what does that all say? It says that, given a bunch of data, [your probability of a particular hypothesis being correct, or being the best one or the right one, is simply uniform over all of the hypotheses that are in the version space. That is, are consistent with the data that we see.](#) >> Nice. >> It is kind of nice. [And by the way, if it's not consistent with it, then it's zero.](#) So, this is only true for hypotheses that are still in A version space and zero otherwise. Now notice that all of this sort of works out only in a world where you really do have noise free examples, and you know that the concept is actually in your hypothesis space and, just as crucially that you have a uniform prior for all the hypotheses. Now this is exactly the algorithm that we talked about before right. We talked about before what would we do. To kind of decide whether a hypothesis was good enough in this sort of noise-free world. And the answer we came up with is you should just pick one of them that's in the version space. And what this says is there's no reason to pick one over the other from the version space. They're all equally as good or rather equally as likely to be correct. >> Yeah, that follows. >> Yeah. So there you go. So it turns out you can actually do something with this. Notice by the way that we did not pick a particular hypothesis space, we did not pick a particular form of our instance space, we did not actually say anything at all about exactly what the labels were other than that they were labels of some sort. The strongest assumption that we made was a uniform prior, so this is always the right thing to do. At least in a Bayesian sense in a world where you've got noise free data, you have to find that hypothesis space, and you have uniform priors. Just pick something from the consistent set of hypotheses.

Quiz: Noisy Data

$$\langle x_i, d_i \rangle$$

$$d_i = k \cdot x_i \sim \Pr\left(\frac{1}{2^k}\right)$$

$$k = \{1, 2, 3, \dots\}$$

x	d	
1	5	$\frac{1}{32}$
3	6	$\frac{1}{4}$
11	11	$\frac{1}{2}$
12	36	$\frac{1}{8}$
20	100	$\frac{1}{32}$

$$h(x) = x$$

$$k = \frac{d_i}{x_i}$$

$$\frac{1}{2^{d_i/x_i}}$$

$$\Pr(D|h) = \frac{1}{65536}$$

$$\Pr(D|h) = \prod_i \Pr(d_i|h)$$

$$= \prod_i \frac{1}{2^{(d_i/x_i)}}$$

$$\text{if } d_i \bmod x_i = 0 \quad \forall d_i, x_i$$

右上角的式子的意思是 $\Pr(d_i = k \cdot x_i) = 1/2^k$

右下角是 if $d_i \bmod x_i = 0$, for all (d_i, x_i) . 意思就是 $d_i = k \cdot x_i$.

题目的意思是:

x_i 是 input, 每個 x_i 都可以弄出一個 output 來(即 d_i), 一個例子就是 x_i 代表一輛車(比如 x_i 為此車的車牌號), d_i 為此車能私下賣多少錢. $\Pr(d_i = k \cdot x_i) = 1/2^k$ 的意思就是, 對每一輛車 x_i , 它能賣 $k \cdot x_i$ 元錢的概率都為 $1/2^k$. 現在我有一個假設 h , 此假設為: 每輛車賣的錢等於它的車牌號(即 $h(x)=x$).

現在我們對這些車的出售情況進行觀察, 對這些車的實際售價作了記錄, 列在了左下的那個 (x, d) 表中. 要問的是: 「在假設 h 的情況下, 這些車賣出該表中價格」的概率是多少? 即 $\Pr(D|h)$ 為多少?

可以看出 $\Pr(D|h)$ 的值跟 h 沒甚麼關係, 這是因為 h 不能表示車應當賣多少錢, 而每輛車的售價都有個客觀的概率分佈, 車的售價是由這個概率分佈決定的, 而不是由 h 決定的. 在第 10 段中, 我們會看到一個例子, in which 車的售價概率分佈跟 h 有關.

前面第 7 段中 $\Pr(D|h)$ 等於 1 或 0, 是因為 noise free, 該處 $\langle x_i, d_i \rangle$ 表示車 x_i 就應當賣 d_i 元錢, 即觀察到的售價就是 應當賣的售價, 沒有賣錯了的(即沒有 noise), 只有 h 跟 D (即 $\langle x_i, d_i \rangle$) 相符(即 $d_i = h(x_i)$)時, $\Pr(D|h)$ 才等於 1, 否則 $\Pr(D|h)$ 等於 0. 而本段中的是有 noisy data, 且每輛車的售價都有個客觀的概率分佈, 故 $\Pr(D|h)$ 是由這個概率分佈決定的, 而不是由 h 決定的.

8. Alright, Michael, I got a quiz for you, okay? >> Sure. >> So, in the last example we had noise free data. So I want to think a little bit about what happens if we have some noisy data. And so I'm going to come up with a really weird, noisy model. But hopefully it illustrates the point. Okay. >> Sure. >> Okay so I got a bunch of training data, it's $\langle x_i, d_i \rangle$ and here's how the true underlying process sort of works. So give us some particular x_i , you get a label which is d_i which is equal to $k \cdot x_i$ where k is some number So one of the counting numbers, one, two, three, four, five, six, seven, eight, and so on and so forth. And the probability that you actually get anyone of those multiples of x_i is equal to

$1/2^k$. Now why did I choose $1/2^k$? Because it turns out that the sum of all those two to the k's from one through infinity happens to equal to one. So it's a true probability distribution. >> Hmm, okay. >> So it's just a neat little geometric distribution. So, you understand the setup so far? >> I think so, so before hypothesis were producing answers then we looked for them to be exactly in the data. Now we're saying that the hypothesis produces an answer, and it gets kind of smooshed around a little bit before it reappears in the table, that's the noisy part. >> Right, so you, you're not going to be in a case now, that if the hypothesis disagrees with the label it sees. That in fact that means no **it can't possibly be the right hypothesis because there's some stochastic process going on** that might corrupt your output label, if you want to think of it as corruption, since it's noisy. Okay? >> Okay, yeah sure. >> Alright? >> Okay, so here's a set of data that you got. Here's a bunch of x's that, that make up our training data one, three, 11, 12, and 20. For some reason they're in ascending order. And the labels that go along with them are five, six, 11, 36, and 100. So you'll notice that they're all multiples of some sort of the input x. Okay? >> Alright. >> Now I have a candidate hypothesis. H of x which just returns x. That's kind of neat. So it's the identity function. So, what I want you to do is to compute the probability of seeing this particular data set in a world where that hypothesis, the identity function, is in fact true. >> The identity function plus this noise process. >> Yes. >> **And one other question quickly this, this noise process is supplied independently to each of these inputs, outputs, pairs?** >> **Yes, absolutely.** >> Okay, then, yeah, I think I can do that. Uh-huh. >> Okay, go.

9. Okay, Michael. You got the answer? >> Yeah, I think, well I can work through it, I don't actually have the number yet. >> Okay, let's do that. >> So, alright, so in a world where. >> In a world where. >> Where this is the hypothesis that actually matters. We're saying that X comes in, the hypothesis spits that same X out. And then this noise process causes it to become a multiple. And the probability of a multiple is this one over two to the case. So, the probability that that would happen from this hypothesis. for the very first data item. The one to five, would be $1/32$. That's the probability that a one would produce a five by this process. >> Okay. How do you, how'd you figure that out? >> Cause the k that we would need the multiplier would have to be five. And so the probability for that multiplier is exactly one over two to the five which is $1/32$. >> Okay. >> And so then I would use that same thought process on the next one which says that it is doubled and the way that this particular process would have produced a doubling would be if with, with probability a quarter. >> Uh-hm. >> And, the next data element would have been produced by this process with probability at half, because it's k will be 1, and $1/2$ to the k would be half, >> Okay, I like this. >> Right? The next one will be an 8th, because it's tripled, >> Uh-hm. >> And the last one is also a multiplier of 5, just like the first one, so that will be $1/32$ as well, >> Mm-hm. >> Alright but now we need to assign a probability to the whole data set, and because you told me it was okay to think about these things happening independently, the probability that all these things would happen is exactly the product. >> Right. >> So I'll multiply a $1/32$ and a quarter and $1/2$ and an 8th and a $1/32$, so that's like a factor of $5 \times 2 \times 7 \times 1 \times 8 \times 11 \times 16 \times 65,536$. So it should be 1 over, oh you already wrote it. 65,536. Yea that. >> Yes that's absolutely correct Michael. Well done. Okay so, that's right, but you did it with a bunch of specific numbers. Is there a more generic Is there a general form that we could write down? >> Yeah, I think so, we're doing something pretty regular once I fell into a pattern. So, I took the D, and divided by X, so D over X tells me that the multiplier that was used, so that's like, the K. >> So. D over x gave you the k. >> And it was one over 2 to the that. >> Okay, so one over 2 to the that. >> And it was then the product of, of that quantity for all of the data elements, so all the i's. So product over all the i's of that. >> Okay. >> But **we have to be careful because if it was the case that for any of our xi's the d wasn't a multiple of it, that can't happen under this hypothesis and the whole probability needs to go to zero.** >> Right. >> So they all have to be divisible otherwise all bets are off. >> Okay, so in other words if d of i mod x of i is equal to zero and this formula holds and it's zero otherwise. >> Exactly. >> Okay. Sounds good. Okay, great Michael.

So that's right and that was exactly the right way of thinking about it. And now, what we're going to do next, is we're going to take what we've just gone through. This sort of process of thinking about, how to generate data labels. for, you know, noisy cases and we're going to apply to it what I think you will find will be a pretty cool derivation. Sound good? >> Awesome! >> Excellent.

BAYESIAN LEARNING

- Given: $\{ \langle x_i, d_i \rangle \}$
 - $d_i = \underbrace{f(x_i)}_{\text{u.d.}} + \epsilon_i \leftarrow \text{error}$
 - $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$x \leftarrow \text{height}$
 $d \leftarrow \text{weight in } \epsilon$

$\sum_i \text{ squared errors}$

$$\begin{aligned}
 h_{ML} &= \operatorname{argmax}_h P(\underline{h} | \underline{D}) \\
 &= \operatorname{argmax}_h P(\underline{D} | \underline{h}) \\
 &= \operatorname{argmax}_h \prod_i P(d_i | h) \\
 &= \operatorname{argmax}_h \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(d_i - h(x_i))^2 / \sigma^2} \\
 &= \operatorname{argmax}_h \sum_i -\frac{1}{2} (d_i - h(x_i))^2 / \sigma^2 \\
 &= \operatorname{argmax}_h - \sum_i (d_i - h(x_i))^2 \\
 &= \operatorname{argmin}_h \sum_i (d_i - h(x_i))^2
 \end{aligned}$$

右下角由 \prod 到 \sum , 是取了一個 log, 只是沒把 log 寫出來. 即式子沒寫完, 但可以理解.
 右下角最後一行是 argmin, 而不是 argmax.

10. Okay Michael, so that was pretty good with the quiz. I want to do another derivation and I want you to help me with it, okay? >> Hm. Cool. >> Okay, so Michael, we have a similar setup to what we've had before. We're given a bunch of training data, XI inputs and DI outputs. But this time we're dealing with real valued functions. **So the way d_i are constructed is there's some deterministic function f , that we pass the f s through. And that gives us some value. And that's really what we're trying to figure out. What is the underlying f ? But to make our job a little bit harder, we have noisy outputs. So, **for every single d_i that is generated, there's some small error, epsilon that is added to it. Now, this particular error term, is in fact drawn from a normal distribution with zero mean and some variance.** We don't know what the variance is. It's going to turn out. It doesn't actually matter. There's some variance going on here. **The important thing is that there's zero mean.** So, you got it? >> **And it's important that it's probably the same variance for all the data.** >> That's right, in fact, each of these epsilon sub i 's are drawn iid. >> And is that f , are we assuming it's linear? >> Nope, we're not assuming that it's linear. >> Okay. >> It's just some function. >> All right, I'm with you. >> Okay, so you got it? >> Yep. >> All right. So, here's my question to you. What is The maximum likelihood hypothesis. >> Do we know f ? Can I just say f ? >> No, we don't know f . All we see are x sub i 's and d sub i 's. But we know there is some underlying f . And we know that it's noisy, according to some normal distribution. >> I don't know how I would find that. >> Well let's try to walk it through. So. We know how to find the maximum likelihood hypothesis, at least we know an equation for it. The maximum likelihood hypothesis is simply the one that maximizes this expression. >> Right. That was when we assumed a uniform prior on the hypotheses. >> Exactly. And so we, this is sort of the easiest case to think about**

Where it turns out that finding the hypothesis that best fits the data is the same as finding a hypothesis that describes the data the best. If you make an assumption about a uniform distribution, or a uniform prior. Okay, so. This is all we have to do now is figure out what we're going to do to expand this expression. So what do you think we should do first? The probability of the data given the hypothesis. Right. So each we assumed IID. >> Mm-hm. >> You actually helpfully even wrote that down. So we can expand that into the product over all the data elements of the probability of that data element given the hypothesis. And x. >> Okay, so let's do that, Michael. Let's write that out. So, finding the hypothesis that maximizes the data that we see, as you point out, is just a product over each of the independent data that we see. Or datums. So that's good. That's one nice step. So we've gone from talking about all of the data together to each of the individual training data that we see. So what do we do next? What is the probability of seeing one particular $P_{sub i}$, given that we're in a world where H is true. >> So okay, given that H is true that means whatever the corresponding x_i is, if we push that through the f function, then the d_i is going to be F of X_i plus some error term so I guess if we took d_i minus $F(X_i)$, that would tell us what the error term is and then we just need an expression for saying how likely it is that we get that much error. >> Right, so, what is the expression that tells us that? >> I'm guessing it's something that uses the normal distribution, it probably has an E in it. >> [LAUGH] I think that's absolutely right. So, let's be particular about what you said. So, when you say that we should push it through $f(x)$, let's be clear that that's basically what h is supposed to be. Our goal here is, given all of this training data, let's recover what the true $f(x)$ is. And that's what our H is. Each of our hypotheses a guess about what the true underlying deterministic function f is. So, if we have some particular labels, some particular value $D_{sub i}$ that is at variance with that. What's the probability of us seeing something that far away from the true underlying F . Well, it's completely determined by, the noise model. And the noise is a Gaussian. So, we can actually write down Gaussian. Do you remember what the equation for a Gaussian is? >> Yes. It's exactly something that has an E in it. >> That's right. So I'll see, I'm going to start writing it and you see if you remember any of what I'm writing down. >> E to the... >> No. >> Okay, good. >> It's 1 over. >> E to the. >> No. >> Okay. >> Square root of, it's, it's coming back to you now. $2\pi\sigma^2$. >> Okay. >> Times... >> I was going to put that in after. >> Oh, okay. So now you get your E , so E to the what? >> It's going to be the value, which, in our case, is, like, H of X_i minus D_i . >> Yeah. And then I feel like, we probably square that? >> Yep. >> And then we divide by σ^2 ? >> right. >> Really? >> Yeah. >> Sweet! >> And your missing one tiny thing. >> There needs to be another two. >> Yes. And in fact it's minus one half. >> Got it. >> So, this is exactly the form of the Gaussian in the normal distribution. And what it basically says is the probability of seeing some particular point, in this case D_i . Given that the mean is H of X . Which is to say that's the underlying function. Is exactly this expression. E to the minus one half, of the distance from the mean, squared, divided by the variance. Okay. And that's just, you either remember that or you don't. But that's just the definition of a Gaussian. So that means the probability of us seeing the data is the product of the probability of us seeing each of the data items. And that's just the product of this expression here. Good. Now, we need to simplify this. We could stop here because this is true, but we really need to simplify this and I think it's pretty... Not too hard to do. It's pretty easy. >> Mm..hm. >> What kind of trick do you think we would do here to simplify this? >> So, first thing I would do is, noticed that the 1 over $\sqrt{2\pi\sigma^2}$ doesn't depend on i at all, and maybe move it outside the product but then realize, well, actually since we're doing an argmax anyway, it's not going to have any impact at all. [CROSSTALK] I would just like cross that baby out. >> I like that. No point in keeping it. All right, now I'm hoping that the other σ^2 we can make that go away too. So I'm tempted to just cross it out, but I'd rather, I'd be much more happy if I had a good explanation for why that's okay. >> Well, so what's the normal trick, so we're trying to maximize the function, right? What you just said is we can get rid of this particular constant expression because it doesn't affect the max. What's making it hard for you to get rid of the σ^2 here is that it's being passed through some exponential and you can't remember off the top of your head what clever

work you can do with constants inside of exponentials. So it would be nice if we could get rid of the exponential. >> Very good. So because log is concave. >> No, because it's monotonic. >> um-hm. We can take the log of the whole shabang. So this is going to be equal to the argmax of the sum of the log of that expression, which is going to move the thing to the outside and the log of E, so that's going to be good, so it's going to be the sum of the superscript thing, the power. >> Right. So let's write that down. Okay, so just to be sure that that was clear to everybody, let's just point out that we basically took the log of both, the natural log of both sides, and so we said, instead of trying to find the maximum hypothesis or the maximum likelihood hypothesis by evaluating this expression directly, we instead evaluated the log of that expression. And as you'll recall from intermediate algebra, the log of a product is the same as the sum of the logs, and the log of E to something is just that thing. >> As long as we do natural log. >> As long as we do natural log when we have E. If we were doing something to the, 2 to the power of something, we'd want to do log base 2. Okay. >> Got it. And you said to do it to both sides but we really didn't need to do it to both sides we just needed to do it inside the things we taking the argmax. >> That's correct. Okay, so we've got here. So, is there any other simplifying that we can do. >> Yeah, yeah now it seems much clearer so the. The negative one half divided by sigma squared all can move outside the sum cause it doesn't depend on I at all. >> Right. And then the sigma squared you said that before you said that that wasn't going to turn out to matter. Both sigma squares ended up, you know, getting tossed into the rubbish heap. >> That's right. >> And I want to be careful with the negative sign. Like I feel like the half can go and the sigma square can go but the negative has to stay. >> You're right. The half can go. And the sigma squared can go. So that leaves us with this expression. So I've taken, gotten rid of the one half, like you suggested. Got rid of the sigma squared like you suggested, and I moved the minus sign outside of the summation. And I'm left with this expression. >> I have a thought about getting rid of that minus sign. >> Well how would you get rid of a minus sign? >> So the max of a negative is the min. Right, so we can get rid of the minus sign by just simply minimizing instead of maximizing that expression. We end up with this expression. >> Nice. That's much simpler than where we started. The e is gone. >> It's much simpler. We got rid of a bunch of e's. We got rid of a bunch of turns out extraneous constants. We got rid of multiplication.

We did a bunch of stuff, and we ended up with this. >> You know, we got rid of two pis. It's kind of sad I would like some pie. >> Mm, I wonder what kind of pie it was? >> Pecan pie? >> [LAUGH] >> Okay, so we got this expression, and that's kind of nice on your own you say, but actually it's even nicer than that. >> What? >> What does this expression remind you of Michael? >> The Sum of Squares. >> This is exactly it. This is, in fact. The sum of squared error, which is awesome. >> Yeah, whoever decided it would be a good idea to model noise as a Gaussian was really on to something. >> Mm-hm. Now, think about what this means, Michael. We just took, using Bayesian Learning, a very simple idea of maximizing a likelihood. We did nothing but substitution, here and there. With the noise model. We got rid of a bunch of things that we didn't have to get rid of. We cleverly used the natural log. Notice that the minus sign can be taken away with the min. And, [we ended up with some of squared error. Which suggests that all that stuff we were doing with back propagation. And, all these other kinds of things we're doing with receptrons is the right thing to do.](#) Minimizing the sum of square error, which we've just been doing before. Is in fact the right thing to do according to Bayesian learning. >> Right in this case meaning meaning what a Bayesian would say. >> Meaning what a Bayesian would say which I believe is sort of right by definition. More importantly here it is. >> They certainly believe it. >> Well, they, they do frequently. >> Oh! I see what you did there. >> No one will get that but, but us. Anyway, the thing is this is the thing you should minimize if you're trying to find the maximum likelihood hypothesis. Now, I just want to say something. That is beautiful. Absolutely beautiful. That you do something very simple like finding the maximum [UNKNOWN] hypothesis and you end up deriving some of squared errors. >> So, just to make sure that I'm understanding. because I see some beauty here, but maybe not all of it. We didn't talk about what the hypothesis class here was. Right, so, if you don't know what the hypothesis class is... You're, you're

kind of stalled at this point, but if we say the hypothesis class is say linear functions. >> Mm-hm. >> Then, what we're saying is we can do linear regression, because linear regression is exactly about minimizing the sum of the squares, right? So linear regression comes popping out of this kind of Bayesian perspective just like that, so is, is that part of what makes it so cool? >> That is part of what makes it cool, but I just think more generally about gradient descent right? The way gradient descent works is you take a derivative by stepping in this, in this space of the air function, which is sum of squared error. >> I see, so you get gradient descend too. >> Yes, you get all of the stuff that people have been doing. Now, there's a piece of beauty there, which is that we derived things like gradient descent and linear regression, all of the stuff we were talking about before and we have an argument. For why it's the right thing to do at least in a Bayesian sense. But there's an even deeper beauty here, which is tied in with ugliness, which is the reason this is the right thing to do, is because of the specific assumptions that we've made. So what were the assumptions that we made? We assumed that there was some True deterministic function that was mapping our x's to our y's and that they were corrupted say transmission error or line noise or however you want to think about it. They are corrupted by some noise that has a very particular form. Uncorrelated, independently drawn, Gaussian noise, with mean zero. So the less pretty way of thinking about it is. Whenever you're trying to minimize the sum of squared error, you are in fact assuming that the data that you have has been corrupted by Gaussian noise. And if it's corrupted by some other noise, or you're actually not trying to model determinance function, of this sort. And then you are in fact, possibly, in fact most likely doing the wrong thing. >> I mean are there other noise models that lead to some other kinds of learning. >> Sure, pick any other model in here that doesn't look Gaussian at all, and you would end up with something else. I don't know what you would end up with because. You know, you couldn't do all these cute tricks with natural logs but yes, you would end up with something different. And one question you might ask yourself is well, if I try to do minimizing the sum of the squared errors, or something for which this model was not the right one, what sort of bad things might happen? Here let me give you an example, let's imagine that we're looking at this here, and our X's are, I don't know measurements of people. Okay? So height and weight. Something like that. >> Mm-hm. >> And in fact let's make it, let's make it let's make it even simpler than that. Let's imagine that our x is our height. And our outputs, our y's, are say weight. And what we're trying to learn is some kind of function from height to weight. Now, this doesn't make a lot of sense to have a true [INAUDIBLE], but I'm trying to make a point here. So what we're saying here is that we, we measure our height and then we measure weight. That there's some simple relationship between them that's captured by f. But, when we measure the weight, we get a sort of noisy version of that weight. Okay? That seems reasonable. But what's not reasonable is we're saying. Our measurement of the weight is noisy, but our measurement of height is not. >> Because if the x's are noisy, then this is not a valid assumption. >> I see. >> So, it seems to work a lot of the time and we have an argument for when it will work, but it's not clear that this particular assumption actually makes a lot of sense in the real world. Even though in practice it seems to do just fine. Okay, got it? >> I think so though I feel like if the error if you put an error term inside the f along with the x and f is say linear. >> Mm-hm. >> Then maybe it pops out and it just becomes another part of the noise term and, and it all still goes through. Like I feel lines are still pretty happy even with that. >> No I think you're right. Lines would be happy here because linear, I mean linear functions are very nicely behaved in that way. But of course, they'd have to be the same noise model in order for it to work the way you want it to work. >> Yeah. >> They'd have to both be Gaussian. They have to both have zero mean, right? And they'd have to be independent of one another. So your measuring device that gives you an error for your height would also have to give you an independent normal error for the weight. Yeah. Though I feel like my scale and my yardstick actually are fairly independent. And they're Gaussian? . >> Oh mine is clearly Gaussian. >> Yeah. >> Yeah. Well at least they're normal. >> They're normally are. >> Mm-hm. >> Okay good. So let's move on to the next thing Michael. Let's try one more example of this and, and then I hope that means you got it, okay? >> Sure. >> Beautiful.

Quiz

$$\begin{aligned} & - \langle x_i, d_i \rangle \\ & - d_i = \underline{f(x_i)} + \epsilon_i \end{aligned}$$

x	d
1	1
3	0
6	5
10	2
11	1
13	4

Insert code

BEST h ?

12 ☒ $h(x) = x \bmod 9$

15.8 19.4 ☐ $h(x) = x/3$ $.16\bar{6}x + .9\bar{6}$

18.8 19 ☐ $h(x) = 2$ 2.2

11. So before we go on to the next example, Michael, I wanted to do a quick [quiz](#), just to make certain you really get what's going on here. The, the sort of power of looking, using Bayesian learning. The, the main insight, I think, I, I want to drive home here, is something you said. Which is that, when we were doing regression before, when we were talking about the perceptrons, we actually had in our head a particular kind of function, a particular hypothesis class. In here with what been talking about with Bayesian learning, the answer to finding to sum of squared errors was independent of the hypothesis class and only dependent upon the key assumptions that we were making, mainly that we had labeled the data with certain form, and that that data was generated by a process that took deterministic function and added some Gaussian noise to it. So, here's the quiz. [Here's your training data. You've got a bunch of xs and a bunch of ds. These are the values that you have to learn. And I want you to tell me, which of these three functions over here, these three hypotheses is the best one under this assumption.](#) Got it? >> mod? Are we allowed to do that? >> We are allowed to do that because >> It's just a function. >> It's just a function, man. >> Interesting. >> It's just a function. >> So we've got a linear, a constant function, a linear function, and we've talked about those, but we've also got a mod function. alright, and [we've got a uniform prior over these three hypotheses.](#) >> Yup. >> Okay. Yeah, I think I can do that. >> Okay. Go.

12. Alright, we're back, what's the answer Michael? >> So, you want me to work it through? >> Sure. >> So what I did first is I made it to, I extended the table that you had. >> Okay. >> To include each of these, the output for each of these three functions. What I'm basically, [what I like to do is compute the squared error for each of these three functions on that data and then choose the one that has the lowest squared error.](#) >> Make total sense to me. Sounds good enough to be an algorithm. Aren't you going to write out the table? >> Well, I mean, I decided to do that, and then there was like one too many steps, and I just threw out my hand and just wrote this all down. Okay, so, we'll just say "Insert code here", because that's what you did, that was the step. >> And, what did your code tell you? >> Well, let me start with the constant function, because that's the easiest piece of code. So, I'm saying what's the difference between each of the D values and two. Squaring it all and summing it up and I get 19. And I can do the same thing, instead of using two I use x over three take the difference of that to the D values and square that and I get 19 point [INAUDIBLE] four, four, four, four, four. Then I can do,

right, so now at this point I'm rock and rolling. I can actually just substitute in my nine, and I get 12? >> Not, not something-odd 12? >> No, just 12, so the error's 12. So that has the smallest error. So even though that's sort of a crazy, like, stripy function right. Like, it increases linearly and then it resets at 9. >> Mmhmm. >> It actually fits the state of the best. >> That is correct. Your code is correct, Michael. Well done. Well actually, looking at this data that sort of makes sense to me, right. Because if you look at the first three examples. Of the data, the outputs are very close. But the outputs of the next three are much bigger, and by doing a mod nine, what you effectively do is say, this is the identity function above this line. And then below the line, it's as if I'm sort of subtracting nine from all of them, and that makes them closer together. >> Hm. >> And so it just happens to work out here. But surely that's just because we came up with a bad constant function and a bad linear function. Do you think there's a better linear function? >> So I mean because it's the squared error, we're really just talking about linear regression. >> Right. >> So I can just run linear regression. So I get an intercept of 0.9588 And a, and a slope of 0.1647. >> Okay >> So that's, so that's my linear function of choice. >> Okay, so that's, what was, what was that again? >> So x times, you know, it's like a six, I guess, like 0.165 probably. >> 0.165x >> Plus >> Mm hm. >> Plus 0.959. So that's our function, that's our best linear function, the function that minimizes greater. So it better end up being, it better end up being less than 19.4, or I'm a liar. >> Mm-hm. >> And now I need to take the difference between that and di square it, and sum. 15.7. >> Hm. So that gives you 15, I'm going to say 15.8. So that is better. >> Yeah, so it's better than the X over three, but it's also worse than the mod 9. >> Hm. >> and the best constant function, has to be worse, because the linear functions include the constant functions as a subset, so this is, that 15.8 is. Is better than the best constant function too. Oh its easy to do though right? Because the best constant function is just the mean of the data. >> What is the mean of the data? >> huh two is pretty close. >> Yeah that's interesting. >> Well that's >> Kind of in the middle of the pack I guess. >> That sort of works right because two the error for two was actually lower than the error for x divided by three. And for what it's worth the error for 2.17 as constant function to 2.2 is 8.8, 18.8, 18.8 sorry. >> Yeah, you're not the [INAUDIBLE]. >> Yeah, eight would have been less than everything. >> Okay. So, what have we learned here? >> That sometimes you want to use mod. >> Yeah. >> If your data is weird. [LAUGH] >> You have you have definitely modified my box. >> Well I'm glad you found it mod. Hmm. [LAUGH] PUNS. Okay, good, so I think that was a good, that was, that was a good exercise. So I'm going to give you one more example of deriving something and then we're going to move on. >> Cool. >> Okay.

Bayesian Learning

$$\begin{aligned} h_{\text{MAP}} &= \operatorname{argmax} P(D|h) P(h) \\ &= \operatorname{argmax} [\lg P(D|h) + \lg P(h)] \\ &= \operatorname{argmin} \left[\underbrace{-\lg P(D|h)}_{\text{event w probability } p} \quad \underbrace{-\lg P(h)}_{\sim \text{has length } -\lg p} \right] \\ &\quad \operatorname{argmin} \quad \text{length}(D|h) + \text{length}(h) \end{aligned}$$

13, 14, 15 這三段是幹甚麼的? 答: 這三段的内容是: 我們很容易就推出了圖中 那個含負號的式子, 然後剩下的內容都是在討論怎麼用 information theory 來理解這個式子。

\lg 即 \log_2

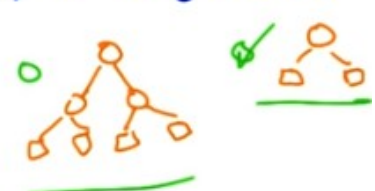
左下角那一堆是 information theory 裡的, 看不懂的話 記住即可

13. >> All right, Michael. So I, all I have written up here for you is, are a maximum a posteriori equation, right? So the best hypothesis is the one that maximizes this expression. Nothing new, right? So I want to do a little trick, the same trick that you did before. So, you noticed that when we had E to the something, that we could use the natural log on E to get rid of everything. So I am going to try to do the same thing here. In the nat, why did the natural log work again? >> Well, it's the inverse of the E, but it let us turn products into sums. >> Right. And the other reason it worked is because it's a. >> Oh, it's monotonic. >> Right, it's a monotonic function and so it doesn't change the argmax. So, I'm going to do the log of both sides here. But this time I'm going to do log base 2, for no particular reason other than it'll turn out to help later. So, I'm just going to take the log of this entire expression, which, because it turns products into sums, gives me this. And by the way for those of you who haven't noticed, I drew in a little bit of notation here. When you write just LG, it's just log base 2. Okay, so, we agree that the answer to this equation and the answer to this equation is the same. And now I'm going to do one other little trick, exactly the trick that you used before. I'm going to change my max into a min, by simply multiplying everything by minus 1. >> Okay, I don't quite see where you're going here. >> But you agree that we haven't changed the answer. >> I agree that we haven't changed the answer. >> Okay. Do a log in there, do a minus sign in there that took us from a max to a min, but I haven't changed the answer. Now, do you recognize anything about these expressions? I'll give you a hint. Information theory. >> Okay. So, information theory is usually entropy, which is like some of $P \log P$ stuff. >> Right. >> I'm not seeing that. >> Well, there you, there's your log and there's your P. >> Sure. [LAUGH] >> [LAUGH] >> It's not P times that though. >> That's true. But we know from information theory, based exactly on this notion of entropy, that the optimal code for some event with probability P has length $-\lg P$. So, that just comes straight out of information theory. That's where all the

entropy stuff comes from. Okay. So, if we have some event that has some particular probability P of happening, the best code for it has this structure, minus log of P . >> Okay. >> So, if we take this fact that we know, and we apply it to here, what is this actually saying? This is saying that, in order to find the maximum a posteriori hypothesis, we want to somehow minimize two terms (那兩個負的) that can be described as lengths. >> Okay. I can see that. >> So my question to you is, given that this definition over here, that an event with probability P has some length minus log P , what is this the length of? >> So that would be the length of the probability of the data given the hypothesis. >> Mm-hm. >> And the length of the hypothesis, or the probability of the hypothesis. >> Well no, it's just the length of that hypothesis. >> Oh, because the event is what has the length. Oh, I see. So it's the length of 'the data given the hypothesis', and the length of 'the hypothesis'. >> Right. So let's write that out. >> But I was just doing, like, pattern matching there. It's not clear to me what a length of a hypothesis is. Hypotheses are functions. I don't know how to take a tape measure to a function. >> That's fair. So this is the length of the hypothesis. Hypothesis. Right? >> Yep. >> So, you said you don't know what that means. But, let's think about that out loud for a moment. What does it mean to have a length of a hypothesis? That's really sort of the number of bits you need to describe a particular hypothesis, right? >> Okay. >> Okay. And in fact, that's exactly what it means. That's why we use log base 2. So, if we want to minimize the length of a hypothesis, what does that mean, the number of bits that we need to represent the hypothesis? >> The number of bits that we need to represent the hypothesis is, I guess, in some representation, or, so in this case I guess it would be some optimal representation. We are taking all the different hypotheses and writing them out. The ones that are more likely have a higher P of H , because that's the prior. And those are going to have smaller lengths than the optimal code. And the ones that are less common are going to have longer codes. >> Well, let's make it more concrete.

Bayesian Learning

$$\begin{aligned}
 h_{\text{MAP}} &= \operatorname{argmax} P(D|h) P(h) \\
 &= \operatorname{argmax} [\lg P(D|h) + \lg P(h)] \\
 &= \operatorname{argmin} \left[\underbrace{-\lg P(D|h)}_{\text{event w probability } P \sim \text{has length } -\lg P} + \underbrace{-\lg P(h)}_{\text{argmin length}(h)} \right]
 \end{aligned}$$



14. So here are two decision trees, which one has the smallest length? Go.

15. Okay, Michael. Which of these two decision trees is smaller? >> [LAUGH] The one on the right is smaller. >> That's exactly right because it's easier, it's, it's easier to represent it in sort of almost any

obvious way that you could think of. It has fewer nodes, so smaller decision trees, trees with fewer nodes, less depth, whatever you need to make it smaller, have smaller lengths than bigger decision trees. So that means that if all we cared about was the second term ($\text{length}(h)$) here. We would prefer smaller decision trees, over bigger decisions trees. >> Which we do. >> Which we do. Now what about this over here? The, what does it mean? So this is pretty straight forward. You got this right? >> That the length of. Well, I mean guess what's weird that you, you're kind of moving back and forth between a notion of a prior, which is where the p of h came from and a notion of Well, you know, if we're going to actually have to describe the hypothesis you're going to have to write it down in some way, and this gives you a way of measuring how long it takes to write it down. But I guess what this whole derivation is doing is linking those two concepts, so that you can think about our bias for shorter decision trees as actually being the prior, right? Actually being the thing that says the smaller ones are more likely And vica versa, that when we think about things that are priors, that are assigning higher probability to certain things, it's kind of like giving them a shorter description. >> Right, so infact if you were to take this example literally here, that we prefer smaller trees to bigger trees, this kind of a bayesian argument for occam's razor. >> And pruning. >> And pruning. Well, you, often use razors to prune, so it makes perfect sense.

Bayesian Learning : Minimum Description Length

$$\begin{aligned}
 h_{\text{MAP}} &= \text{argmax} P(D|h) P(h) \\
 &= \text{argmax} [\lg P(D|h) + \lg P(h)] \\
 &= \text{argmin} \left[\underbrace{-\lg P(D|h)}_{\substack{\text{event w probability } p \\ \sim \text{has length} \\ -\lg p}} \quad \underbrace{-\lg P(h)}_{\substack{\text{argmax length}(D|h) + \text{length}(h) \\ \text{misclassification or "error"} \\ \text{"size of } h\text{"}}} \right]
 \end{aligned}$$

右下角是 size of h

>> Ok, so this is kind of straight forward, that basically smaller trees are smaller than bigger trees. It sort of makes sense. Now, what about this over here? What does it mean to talk about the length of the data given a particular hypothesis. >> Uh...I could think of one interpretation there. So like, if the hypothesis generates the data really well, then you don't really need the data at all, right? You just have...you already have the hypothesis. The data is free. Right? But if it deviates a lot from the hypothesis, then you're going to have to have a long description of where the deviations are. So maybe it's kind of capturing this sort of notion of how well it fits. >> Right, that's exactly right. So I like that

explanation so let me write it down. So here we literally just mean something like size of h . But over here we are talking about, well sort of error right? if the hypo, if just exactly what you said if the hypothesis perfectly describes the data, then you don't need any of the data. But let's imagine that the hypothesis gets all of the data labels wrong. Then when you send the hypothesis over to this person. This, this sort of person we're making up who, trying to understand the Daden hypothesis. And you're also going to have to send over what all the correct answers were. So, what this really is, is a notion of miss-classification error, or just error in general. If we're thinking about regression. So, basically, what we're saying is, if we're trying to find the maximum Imposter Hypothesis. We want to maximize this expression. We want to find the age that maximizes this expression. That's the same as finding the age that maximizes the log of that expression, which gives you this. Which is the same as minimizing this expression, which is just maximizing this expression but throwing a minus one in front. But these terms (那兩個負的) actually have meanings in information theory: the best hypothesis, the hypothesis with the maximum ~eapaspiron probability is the one that minimizes error and the size of your hypothesis. You want the most simple hypothesis that minimizes your error. That is pretty much Occam's razor. What is important here in reality is that these are often traded off of one another. If I give a more complicated or bigger hypothesis, I can typically drive down my error. Or I can have a little bit of error for a smaller hypothesis. But this is the sort of fundamental tradeoff here. You want to find The simplest hypothesis that still explains your data, that is, minimizes your error. >> Hm. >> So this actually has a name, and that is the minimum description, and there have been many algorithms over the years that have tried to do this directly by simply trading off some notion of error, and some notion of size. And finding the tradeoff between them that actually works. Now, if you think about it for a little while Michael you'll realize that yea this sort of makes sense at the hand wavy level at which I just talked about it. But, you do have some real issues here about for example units. So, I don't know if the units of the size of the hypothesis are directly comparable to the counts of errors or you know sum of squared errors or something like that and so you have to come up with some way of translating between them... And some way of making the decision whether you would rather minimize this or you'd rather minimize that if you were forced to make a decision. But the basic idea is still the same here. That the best hypothesis is the one that minimizes error without paying too much of a price for the complexity of the hypothesis. >> Wow. So I've been sitting here thinking about, so with decision trees, this notion of length feels... Like you could translate it directly into bits right like you actually had to write it down and transmit it, it makes a lot of sense. But then I was thinking about neural networks. And, and, and given that a fixed neural network architecture it's always the same number of weights and they're just numbers. So you just transmit those numbers. So I thought, hmmm, this isn't really helping us understand ~neural-nets-art-al and then it occurred to me that those weights, if they get really you're going to need more bits to express those big weights. And in fact that's exactly when we get over fitting with neural nets if we let the weights get too big. So like this gives a really nice story for understanding neural nets as well. >> Right. That the complexity is not in the number of parameters directly but in what you need to represent the value of the parameters. >> Wow. >> So I could have ten parameters that are all binary, in which case I need ten bits. Or they could be arbitrary real numbers, in which case I might need, well, an arbitrary number of bits. That's really weird. >> Yeah, but the point here, Michael, I want to wrap this up. The point here is we've now used Bayesian learning to derive a bunch of different things that we've actually been using all along, and so again the beauty of Bayesian learning is that it gives you a sort of handle on why you might be making some of the decisions that you're making. >> It seems like this raises the theory question that you threw at me in a previous unit. Right. Which is like well so if it doesn't really tell us anything we didn't already know, how important is it? >> Well in this case, I think it is important because it told us something that we were thinking and tells us in fact we were right. So now we can comfortably go out in the world minimizing some of squared error when we're in a world where there is some kind of [UNKNOWN] transmission noise. We can go about trying to Believe Occam's Razor because Bayes told us so. [LAUGH] Thanks to Shannon.

And so on and so forth. We can do these things and know that in some sense, they're the right things to do, at least in a Bayesian sense. >> Neat. >> Okay, good. Now one more thing, Michael, I'm going to show you. Which is that everything I've told you so far is a lie. >> Ah...

Bayesian Classification

$h(x)$	$Pr(h D)$
h_1 +	.4
h_2 -	.3
h_3 -	.3

BEST LABEL for x?

○ +
☒ -

- for $h \in H$ compute $Pr(h|D)$
 output argmax

$$\underline{V_{MAP}} = \underset{v}{\operatorname{argmax}} \sum_h P(v|h) P(h|D) \quad - \quad \text{weighted vote } h \in H, \quad Pr(h|D)$$

右下角第一個“-”是 the way to find the best hypothesis, 第二個“-”是 the way to find the best label

16. Okay, Michael, so here's a quiz. Now it's a pretty straightforward quiz. I just want you to use everything that you've learned so far. Okay, so you have three hypotheses. Let's call them h_1 , h_2 , and h_3 , okay? >> Mm-hm. >> For the sake of argument. (左邊那列) Here's what each of these hypotheses outputs or some particular x . h_1 says +. h_2 says -. h_3 says -. Okay? >> Mm-hm. >> Now here we've, already made it easy for you and we've computed the probability of, a particular hypothesis given some set of data. I'm not showing you the data but I'm showing you the answer for it. Okay? So the probability of h_1 given the data is 0.4, h_2 is 0.3, and h_3 is 0.3. Got it? Wait hang on, so, okay, I see that corresponds here about this given data, what's x ? >> x is some input, it doesn't matter, just like it doesn't matter what the data is. >> [LAUGH] 'Kay. >> Just call it so that, that, x is x , okay? It's just some object out in the world and each hypothesis labels it. Plus or minus. >> 'Kay. >> Or can label it plus or minus. And H_1 decides if for that X , it's positive, and the other two hypotheses decide that it is, in fact, negative and H_1 has probability that is the maximum a posterior probability h_1 is .4, h_2 is .3 and h_3 is .3. So my question is, very simple. Using all of the magic we've done, this is just to make sure you've got it, Michael, I dunno, we've done a lot of derivations, we've walked away from some things [LAUGH] we gotta make sure we get back to basics here. What is the best label for x ? Is it plus, or is it minus? I see why this is tricky. Okay. >> And go.

17. And we're back. What's the answer, Michael? >> Okay, so it depends. >> What does it depend on? I've given you everything. This is straightforward. >> Well, so, okay, I guess. The here, so here's what I'm seeing. So I'm, what I'm seeing is that hypothesis 1 is the most likely hypothesis. >> It's not just

the most likely, it's the most a posteriori. >> Well, that's what I mean by likely. Right, is the map hypothesis? It's the maximum a posteriori hypothesis. So if we say, [what is the label according to the map hypothesis? Boom, it's plus.](#) >> Yes. >> But, if we're saying what's the most likely label. So the most likely label is, is, we have to actually look over all the hypotheses and in a sense, let them vote. So the probability that the label is minus is actually 0.6, which is greater than 0.4, so if I had to pick (the answer), I would go with minus. >> And you would be correct. So I did a little tricky thing here for you Michael. You've been complaining about my titles, because everyone said Bayesian learning and I changed the title here to Bayesian Classification. >> Ohhh. >> Because in fact the problem here, we've been talking about all along is, what's the best hypothesis. But here. I ask you what's the best label. >> Hm. And exactly as you point out, [finding the best hypothesis is a, is a very simple algorithm.](#) Here I'll write it for you because we did it before. For every H in hypothesis set, simply compute the probability that it is the best one, and then simply output max. That's how you find the best hypothesis, but that's not how you find the best label. [The way you find the best label is you basically do a weighted vote for every single hypothesis in the hypothesis set, according to the weight being the probability of that hypothesis given the data.](#) >> Okay. >> So the best, if you can only output hypothesis and use that hypothesis, in fact, you would say plus. But if you asked everyone to vote, just like we did with boosting, just like we did effectively with KNN and all these other kind of. Weighted regression techniques we've used before, you need to do the voting. >> And I, and I feel like I could probably derive that using rules of probability. Right, because really what we want is we're trying to maximize the probability of the label, given the data, and I think the probability laws would tell us that's equal to sum of all hypotheses of the hypothesis and the label given the data, which is, like, the probability of the hypothesis given the data, times the probability of the label given the hypothesis, and that's what we did, we summed up. You know, the probability of the label given the hypothesis is either one or zero. That's your left column. And then we're summing up the probabilities that corresponding to the pluses. And we're summing up the probabilities corresponding to the minuses and choosing the largest one. >> So, this is what you just said written down as an equation. basically, the most likely value. Is the one that maximizes this expression. And this follows directly from Bayesian's rule where (左下角 Vmap 一式, V 表示 value) now instead of trying to maximize the hypothesis given the data, you're trying to maximize the value given the data. And I think it's pretty straightforward to derive that but I'd like to leave it up to the students to do it on their own. Okay, so Michael, [in some sense everything that I've told you before is a lie, in that I've led you down this path that somehow, finding the best hypothesis is the right thing to do. But the truth is, finding the value is what we actually care about. Finding a hypothesis is just a means to an end. And if we have a way of actually computing the probabilities for all the hypotheses, then we should let them both in order to find the best actual label or the best value for it.](#) >> Got it. >> All right. Good.

Bayesian Learning Classification

- Bayes' rule! Swap "causes" & "effects"
$$Pr(h|D) \sim Pr(D|h) P(h)$$
- priors matter
- h_{MAP} , h_{ML}
- derived rules we've used
- voting h \leadsto Bayes optimal Classifier
optimality & gold standards.

18. Okay Michael so this wraps up all this Bayesian Learning stuff. What have we learned today? >> We did Bayes rule. >> We learned Bayes rule. We even learned how to derive Bayes rule. >> And it was super useful because it lets you swap, kind of, causes and effect. >> So I like the way you put that, Michael that we're swapping causes and effects. Sort of mathematically when we think about Bayes rule, what that really lets us do is. Instead of having to compute the probability of a hypothesis given the data We instead view to compute the probability of the data given the hypothesis, which is typically a much easier thing to do. And what makes it of course Bayes rule in general is that you wait that by the prior probability over the hypothesis. Which in fact is one of the important things that we learned which is that priors matter. So anything else we learned? >> Yep, we did the MAP hypothesis, Maximum a posteriori. Right. We learned about HMap, and we also learned about HML. >> ML, right. The maximum likelihood hypothesis. > Right. And what's the maximum likelihood hypothesis? How's it relate to the maximum a posteriori hypothesis? >> It's the map that you get when the prior is uniform. >> Right. Alright. And we, oh, we connected up maximum uposteriory(?) and lease(?) squares. >> Yeah, that was pretty, I really liked that. So, we basically der, we deroved. We derived a bunch of things we'd been doing before. And short of showed that there's actually a good argument for them. At least if you're Bayesian. There are good arguments for doing some doing some squares. There are good arguments for Achem's(?) Razor(?). We'd actually be able to give real justification for doing them other than, well sure it makes us one of them. >> Right so that includes the minimum description length story. >> Mm-hm. >> And then finally, you told me that was all a lie, and you said that really what you want to do is this other kind of way of picking that actually factors in the probability of all the different hypotheses and having them essentially vote. Right. What we really care about, is classification (想想也是, 前面幾個算法都是做 classification 的, 到 Bayesian 怎麼就變成了找 hypothesis, 原來目的還是 classification). We're learning in the end and so we also learned about base classifiers. So in fact, what we described before, which is voting of hypothesis, turns out to be the Bayes optimal classifier. I didn't say that, but it is very important to note. In fact, what you should be noting there is

not only is it the Bayes optimal classifier, it's the Bayes optimal classifier (重音在 optimal 上). And what that means is that on average you cannot do any better than basically doing a weighted vote of all the hypotheses according to the probability of the hypothesis given the data. You cannot do any better than this on average. So again, what Bayesian learning gives us and what Bayesian classification gives us is a way of talking about optimality... In gold standards. What'd you think, Michael? >> That's really neat. >> I like it. I mean, I have to tell you, I really think that this stuff is kind of cool. It's always nice to be able to take things that actually work and explain them according to some framework, some underlying theory. >> I wonder though, it seems like all these Bayesian equations lead us to the question, of how we actually infer probabilities from various different quantities and observations. So is there a way to do that? >> So I think the answer is yes. And maybe you should go figure it out and then tell me about it next time. >> Okay. [LAUGH] All right, as you wish. >> As you wish. >> Stay tuned. >> Anyway, this has been a lot of fun, Michael. I will talk to you later. >> Thanks. >> Bye.