

هنتكلم عن ال statistical learning و هنشوف ازاى نقدر ن explain الجوردم معين من خلال ال statistical perspective .. معني ال statistical learning هو ان في كونسبت انت مش متأكد منو و عملية ال learning هنا هدفها انها تقلل ال uncertainty .. ال idea ان لو في شوية uncertainty يبقى احنا ممكن نستخدم distribution عشان يقفش ال uncertainty و احنا بنتعلم و احنا بنلاحظ الداتا .. نقدر نقلل ال uncertainty عن طريق انك تحسب revised distribution .. اللي هي correspond to the process of learning

انهارده السؤال بتاعنا احنا لو هنستخدم الديستريبيشنز عشان ن capture و ن quantify ال uncertainty بتاعتنا ايه هي الطريقه الصح اننا ن update الديستريبيشنز اللي عندنا ... هنشوف دلوقت الحوار ده بيتم ازاى ...

Statistical Learning

- View: we have uncertain knowledge of the world
- Idea: learning simply reduces this uncertainty

اللي احنا هنعملو برضو اننا هن review شوية علي ال Probabilities .. أول حاجه اليه هي ال probability distribution .. المقصود بيها اننا عندنا dice و هنرميه و هيطلع عندنا 6 احتمالات لل outcome .. وال outcomes ديه ليها احتماليات انها تظهر .. هو ده ال distribution .. احنا هنستخدم ال probabilities عشان encode ال uncertainty عن العالم .. و هحتاج راندم فاريلز كثير فشاخه .. زي مثلاً لما هنتكلم علي الجو . هنتكلم علي درجة الحرارة و الرطوبة و فبالثالي عندي أكثر من راندم فاريلز .. لما بيبقا عندي أكثر من واحد .. فانا بتكلم علي Joint probability distribution علي أكثر من راندم فاريلز .. وهنا انت تقدر تتكلم علي $P(\text{temp} = 20)$ او ان الرطوبة عند لفل معين .. فانا عندي أكثر من ايفنت ..

لما بيبقا عندنا Joint distribution نقدر ن charectize ال quantities اللي عاوزنها في دومين معين . و احنا مش بنهتم بال combinations بتاعت كل ال random variables ... بس احنا فعلاً مهتمين ب اننا نعمل prediction ل راندم فاريلز 1 بس .. فانا مهتم بال Temp .. بس عندي joint distribution ماسكه درجة الحرارة و الرطوبة و المطر .. وحاجات كثير .. فالسؤال هنا .. ازاى أ extract ال distribution بالنسبه ل 1 variable only زي الحرارة مثلاً .. هنا نقدر نستخدم ال Marginalisation rule اللي هي اسمها sumout rule .. والفكره هنا ان لو عندنا Joint distribution علي A and B ... وانا محتاج بس ال marginal of A .. في الحاله ديه اقدر احسبها عن طريق اني اجمع كل ال possible values ل variable B .. فهنا ال Capital letter هو ال random variable انما ال small letter هي ال values بتاعت الراندم فاريلز دول .. فلو انت عاوز المارجن بتاع A .. اجمع B و العكس ..

Joint distribution

- Given two random variables A and B :
- Joint distribution:

$$\Pr(A = a \wedge B = b) \text{ for all } a, b$$

- Marginalisation (sumout rule):**

$$\Pr(A = a) = \sum_b \Pr(A = a \wedge B = b)$$

$$\Pr(B = b) = \sum_a \Pr(A = a \wedge B = b)$$

تعال نحل الأسئلة الي في الصورة الجايه ..

Example: Joint Distribution

	sunny			~sunny	
	cold	~cold		cold	~cold
headache	0.108	0.012	headache	0.072	0.008
~headache	0.016	0.064	~headache	0.144	0.576

$$P(\text{headache} \wedge \text{sunny} \wedge \text{cold}) = \quad P(\sim \text{headache} \wedge \text{sunny} \wedge \sim \text{cold}) =$$

$$P(\text{headache} \vee \text{sunny}) =$$

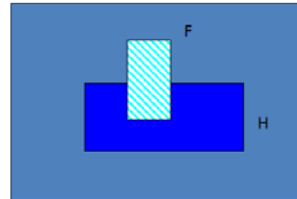
$$P(\text{headache}) =$$

marginalization

ايه هي احتمالية ال headache and sunny and cold بتساوي 0.108 ... تقدر كمان تبص علي احتمالية انك مش مصدع و في شمس و معندكش برد . هتلاقى ب 0.064 ايه احتمالية انك مصدع أو في شمس .. محتاج تجمع .. خذ الصف بتاع ال headache كلو .. و ال sunny خذ كل ال table .. 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28 و آخر واحده هي ال P(headache) .. هو يعتبر ال marginalization .. فانت هتجمع كل الاحتماليات بتاعت كل الراندم فاريبلز التانييين ... 0.2 = 0.008 + 0.072 + 0.012 + 0.108

Conditional Probability

- $\Pr(A|B)$: fraction of worlds in which B is true that also have A true



H="Have headache"
F="Have Flu"

$$\begin{aligned}\Pr(H) &= 1/10 \\ \Pr(F) &= 1/40 \\ \Pr(H|F) &= 1/2\end{aligned}$$

Headaches are rare and flu is rarer, but if you have the flu, then there is a 50-50 chance you will have a headache

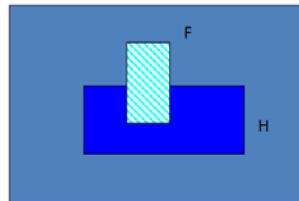
University of Waterloo

CS480/680 Spring 2019 Pascal Poupart

6

انت محتاج تبص علي $P(A|B)$.. هنا ال notation اننا عندنا vertical bar .. وعلي اليمين و الشمال شوية فاربيلز تانيين ... دلوقت انا ببص علي جزء من العالم اللي فيه ال B بتبقا true و كمان ببص في ال A تبص true .. فانت ببص علي ال B انو هو الرفرنس بتاعك .. و دلوقت انت بتحاول تلاقي ال A كمان تبصا true ...

Conditional Probability



H="Have headache"
F="Have Flu"

$$\begin{aligned}\Pr(H) &= 1/10 \\ \Pr(F) &= 1/40 \\ \Pr(H|F) &= 1/2\end{aligned}$$

$\Pr(H|F)$ = Fraction of flu inflicted worlds in which you have a headache

= (# worlds with flu and headache) / (# worlds with flu)

= (Area of "H and F" region) / (Area of "F" region)

= $\Pr(H \wedge F) / \Pr(F)$

University of Waterloo

CS480/680 Spring 2019 Pascal Poupart

7

احفظ السلايد الجايه ..

Conditional Probability

- Definition:
$$\Pr(A|B) = \Pr(A \wedge B) / \Pr(B)$$
- Chain rule:
$$\Pr(A \wedge B) = \Pr(A|B) \Pr(B)$$

Memorize these!

University of Waterloo

CS480/680 Spring 2019 Pascal Poupart

8

Example: Joint Distribution

	sunny		~sunny	
	cold	~cold	cold	~cold
headache	0.108	0.012	0.072	0.008
~headache	0.016	0.064	0.144	0.576

$$\begin{aligned}
 \text{Pr(headache} \wedge \text{cold} \mid \text{sunny}) &= \frac{0.108}{0.108 + 0.012 + 0.016 + 0.064} = 0.54 \\
 \text{Pr(headache} \wedge \text{cold} \mid \sim \text{sunny}) &= \frac{0.072}{0.072 + 0.008 + 0.144 + 0.576} = 0.09
 \end{aligned}$$

خد بالك different context بيكا مفيش سبب يخلي المجموع بتاع الاحتماليات بيكا ب واحد ..

Bayes Rule

- Note

$$\text{Pr}(A|B)\text{Pr}(B) = \text{Pr}(A \wedge B) = \text{Pr}(B \wedge A) = \text{Pr}(B|A)\text{Pr}(A)$$

- Bayes Rule

$$\text{Pr}(B|A) = \frac{[\text{Pr}(A|B)\text{Pr}(B)]}{\text{Pr}(A)}$$

$$\frac{p(B)p(A|B)}{p(A)}$$

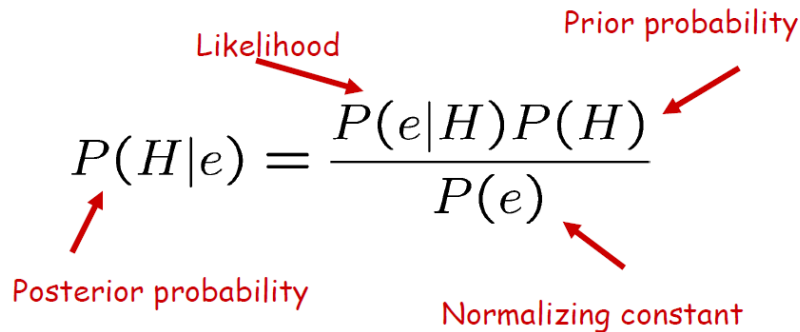
Memorize this!

بايز روول مهمه جداً .. وفي فرع من فروع الماشين ليرننج اسمو Bayesian learning مبني علي القاعده اللي فوق ديه .. في هنا حاجه مهمه تعرفها . تخيل ان عندك H hypothesis space .. واحنا قلنا في ال statistical learning .. احنا هنفكر في ال learning ك process of reducing uncertainty ففي البدايه انا عندي ال hypothesis space بس انا مش عارف انه في واحد هي الاحسن .. فانا عندي uncertainty كتيره .. يستحسن لو عرفت اعملها ب Pior distribution .. ديه اللي هت capture my uncertainty .. و بعدين هنقوم جايبين شوية داتا اللي هي ال evidence .. فانا لو عندي specific hypothesis .. أقدر اقيس ال Likelihood انك تجيب evidence وانت عندك ال hypothesis space هنا bayes rule .. بتسمحلك انك تروح من ال conditional ل conditional ثانيه ..

Using Bayes Rule for inference

- Often we want to form a hypothesis about the world based on what we have observed
- Bayes rule is vitally important when viewed in terms of stating the belief given to hypothesis H , given evidence e

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$



فدلوقت فكَر في الماشين ليرننج من ال statistical prespective انها reduction of uncertainty اللي captured by bayes rule فاحنا بنبدأ ب initial distribution in H ... اللي هو بي capture the uncertainty .. وبعدين ن observe some data ... أو evidence . وبعدين ن condition علي الداتا ديه .. وبعدين ن revise our uncertainty .. وبعدين ن obtain posterior distribution

Bayesian Learning

- **Prior:** $\Pr(H)$
- **Likelihood:** $\Pr(e|H)$
- **Evidence:** $e = \langle e_1, e_2, \dots, e_N \rangle$
- **Bayesian Learning** amounts to computing the posterior using Bayes' Theorem:
$$\Pr(H|e) = k \Pr(e|H)\Pr(H)$$

قال Bayesian learning بنبدأ ب Prior .. بشكل عام يعني هيقا برضو عندنا ال likelihood distribution اللي بيقلنا how likely it is to obtain a certain dataSet or evidence given each hypothesis .. ودلوقت لو عندنا اداتا سيت نقدر نستخدم بايز رول عشان نحسب ال posterior و ده اللي هي encode للي احنا اتعلمناه .. لما احنا كتبنا ال bayes rule ثاني في الصوره اللي فوق ديه علي طول ... انت عندك ال K ... ده انت حطيطو بدل انك تقسم علي ال $P(\text{evidence})$.. عشان ده طلع انو مجرد normalization يعني .. ليه لإن اللي انت بتحاول تعملو هنا هو ان عندي ال hypothesis space و عندي شوية hypotheses وانا بحاول احدد انهي واحده هي أحسن واحده او هي الصح ... و دلوقت هحدد ايه هي ال Posterior probability في كل واحده من ال hypothesis .. و زي منت شايك ان ال Posterior probability هي بالنسبه لنفس ال context اللي هو ال evidence او الداتا سيت .. وديه هي هي لكل ال hypothesis .. فده معناه ان الترم ده بيعتمد علي ال e بس .. فبالتالي ده مجرد كونستانت .. الهدف منو انو ياخذ ال numerator و يعمل ال normalization عشان يجيب ارقام من صفر ل 1 .. فال process of bayesian learning .. بتديك posterior distribution .. بس علي ارض الواقع .. ال poseterior distribution .. هو مش اللي احنا بندور عليه .. انت عاوز تجيبا hypothesis وتستخدمها عشان تجيب prediction .. مثلاً يعني

انا مهتم اني اعمل weather prediction .. انا ممكن يكون عندي موديل عن الجو ليه شوية بارمترز ... كل الموديلز اللي كان ممكن استخدمها هم دول ال hypothesis space اللي عندي .. سبب من كل الموديلز الي ممكن استخدمها .. فهنا في الصوره الجايه ديه :

Bayesian Prediction

- Suppose we want to make a prediction about an unknown quantity X
- $$\Pr(X|e) = \sum_i \Pr(X|e, h_i)P(h_i|e)$$
$$= \sum_i \Pr(X|h_i)P(h_i|e)$$
- Predictions are weighted averages of the predictions of the individual hypotheses
- Hypotheses serve as “intermediaries” between raw data and prediction

في الصورة اللي فوق ديه عندك h_i .. ديه هتعتبر عن a possible model ... و بعدين X هتبقا قيمه لل temperature .. يعني مثلاً انا ه predict درجة الحرارة بكرة ... وانت عندك model for the weather .. والمشكله اللي بحاول اتعلمها هي :

What is the right model that I should use to make predictions

فأول ما بيقا معنا ال posterior distribution ل hypothesis ... انت لسه محتاج انك تستخدمو عشان تعمل Prediction عن متغير اسمو X .. والمشكله ان انا عندي هنا distribution over hypothesis .. لو كان عندي single hypothesis .. تقدر تستخدمها وخلاص... بس انت عندك كتير جدا من ال hypothesis .. و انا حالياً قمت عامل اني expressed some probability with respect to each one of them

فازاي اقدر اعمل بريدكشن بالنسبه لكل ال hypothesis دول .. ففي حل من الحلول .. انك تاخذ ال weighted combination .. فاللي الاكسبرشن ده بيقولو ان كل hypothesis اقدر اعمل prediction about the temperature بس انا كمان اتعلمت من الداتا ايه اللي ممكن يكون a good power للموديل الي عندي اللي هو $P(h|e)$.. فبالتالي اقدر بيقا عندي distribution علي كل ال possible weather models واللي فعلياً بعملو هو اني ب cmbine كل ال predictions عن طريق اني استخدم weight بناءً علي ال posterior probability ..

....

فانت ممكن تستخدم a posterior distribution عشان تعرف weights for each hypothesis و دلوقت بيقا عندك weighted combination of the prediction of all these hypotheses to obtain my final prediction

تعال نشوف شوية أمثله ..

أول واحد اللي هو ال Candy example ... عندنا نوع مش بنحبو .. ونوع بنحبو .. الحاجات اللي ممكن نشترىها مش بتقولك النوع اللي انت بتشتريه هو نوعو ايه .. بس الحلوي ديه بنتباع في شنت .. في شنته فيها 100% cherry .. وفي 75% تشيري ل 25% لايم .. و 50 50 و 25 75 و 100%

بص ع السلايد

Candy Example

- Favorite candy sold in two flavors:
 - Lime (hugh)
 - Cherry (yum)
- Same wrapper for both flavors
- Sold in bags with different ratios:
 - 100% cherry
 - 75% cherry + 25% lime
 - 50% cherry + 50% lime
 - 25% cherry + 75% lime
 - 100% lime

دلوقت عشان نخلي الدنيا أرخم حاجه .. الشنط ديه surprise bags .. انت في المحل تقدر تشتري اي شنطه من دول . بس انت لا عارف الشنطه ديه انهبي واحده من الشنط .. ولا عارف كمان نسبة ال cherry ل ال lime كام في الشنطه اللي هتشتريها ..

طيب تعال نقول اننا هنشتري واحده من الشنط ديه .. وانت جاي تسأل ايه هي نسبة ال flowers بتاعت الشنطه اللي انت ماسكها ف ايدك ناووو .. فالدكتور قالك تعال نعمل تجربه .. كل شوية K من الحلويات الي في الشنطه .. وبعدين حاول ت estimate من ال candies الي انت كلت منها .. ايه هي النسبه اللي جوا الشنطه .. و كمان توقع الطعم بتاع اللي جايه ..

Candy Example

- You bought a bag of candy but don't know its flavor ratio
- After eating k candies:
 - What's the flavor ratio of the bag?
 - What will be the flavor of the next candy?

Statistical Learning

- **Hypothesis H:** probabilistic theory of the world

- h_1 : 100% cherry
- h_2 : 75% cherry + 25% lime
- h_3 : 50% cherry + 50% lime
- h_4 : 25% cherry + 75% lime
- h_5 : 100% lime

5 types of bags

- **Examples E:** evidence about the world

- e_1 : 1st candy is cherry
- e_2 : 2nd candy is lime
- e_3 : 3rd candy is lime
- ...

evidence = data

eating candy = obtaining data

بدء احنا نقدر ن formulate ال prior distribution .. وبعدين نحسب ال posterior distribution اللي هت correspond ل reducing ال uncertainty وتتعلم حاجه عن نسبة ال flavors ..

في Bayesian learning ال prior distribution اللي بنعملو هو تخمين بنعملو في البدايه خالص .. فهنا عندنا 5 ارقام .. دول الي هم احتمالية

$$P(h_1) = 0.1$$

$$P(h_2) = 0.2$$

$$P(h_3) = 0.4$$

$$P(h_4) = 0.2$$

$$P(h_5) = 0.1$$

الفيكتور اللي عندك في الصوره الجايه ده هو اللي انت بتختارو .. ده اللي بي correspond to the initial belief ... قول مثلاً انت عند ال store و بتشتري شطه قبل ما تاكل أي من الحلوي .. الفيكتور ده بيعبر عن اول نظره للشطه اللي موجوده و أول وهله للشطه اللي موجوده حاجه عشوائيه يعني .. حاجه educated guess ...

Candy Example

- Assume prior $\Pr(H) = \langle 0.1, 0.2, 0.4, 0.2, 0.1 \rangle$
- Assume candies are **i.i.d. (identically and independently distributed)**

$$\Pr(\mathbf{e}|h) = \prod_n P(e_n|h)$$

- Suppose first 10 candies all taste lime:

$$\Pr(\mathbf{e}|h_5) =$$

$$\Pr(\mathbf{e}|h_3) =$$

$$\Pr(\mathbf{e}|h_1) =$$

دلوقت احنا هناكل candies و قام ايه طلع كل ال 10 اللي كلناهم كلهم lime .. لو انت كلت ال 10 و كلهم طلعو الطعم ده .. بيقا طبيعي انك تقول ان الشنطه ديه أكيد من الشنط اللي فيها lime بس .. بس تعال نشوف ايه اللي بيحصل لما بنستخدم Bayesian learning in terms of estimating what is the ratio of the flavor

احنا محتاجين a likelihood distribution .. فالتالي ده عبارته عن احتمالية ل given hypothesis اللي هي given ratio of flavor ... هن observe certain flavors زي الحاله اللي عندنا هنا 10 limes ..

عشان الداتا سبت هيبقا فيها داتا بوينتس كتيره .. بص علي الجوينت بروبابلتي بتاعت الداتا سبت .. فهتلاقىها بتساوي انها بقت حاصل ضرب ال $P(e|h)$

$$P(10 \text{ limes} | h_5) = 1^{10} = 1$$

$$P(10 \text{ limes} | h_3) = 0.5^{10} = 0.00097$$

$$P(10 \text{ limes} | h_1) = 0^{10} = 0$$

دلوقت انت عندك ال prior وعندك ال likelihood .. اضربهم ف بعض و هات ال posterior distribution ... فينأء علي بايز ثيوريم .. انت تقدر تحسب ال posterior distribution اللي هي $P(h | e)$.. عن طريق ال hypothesis space و given a dataset .. اضرب ال prior في ال likelihood و هيطلعك ال posterior distribution

Bayesian Learning

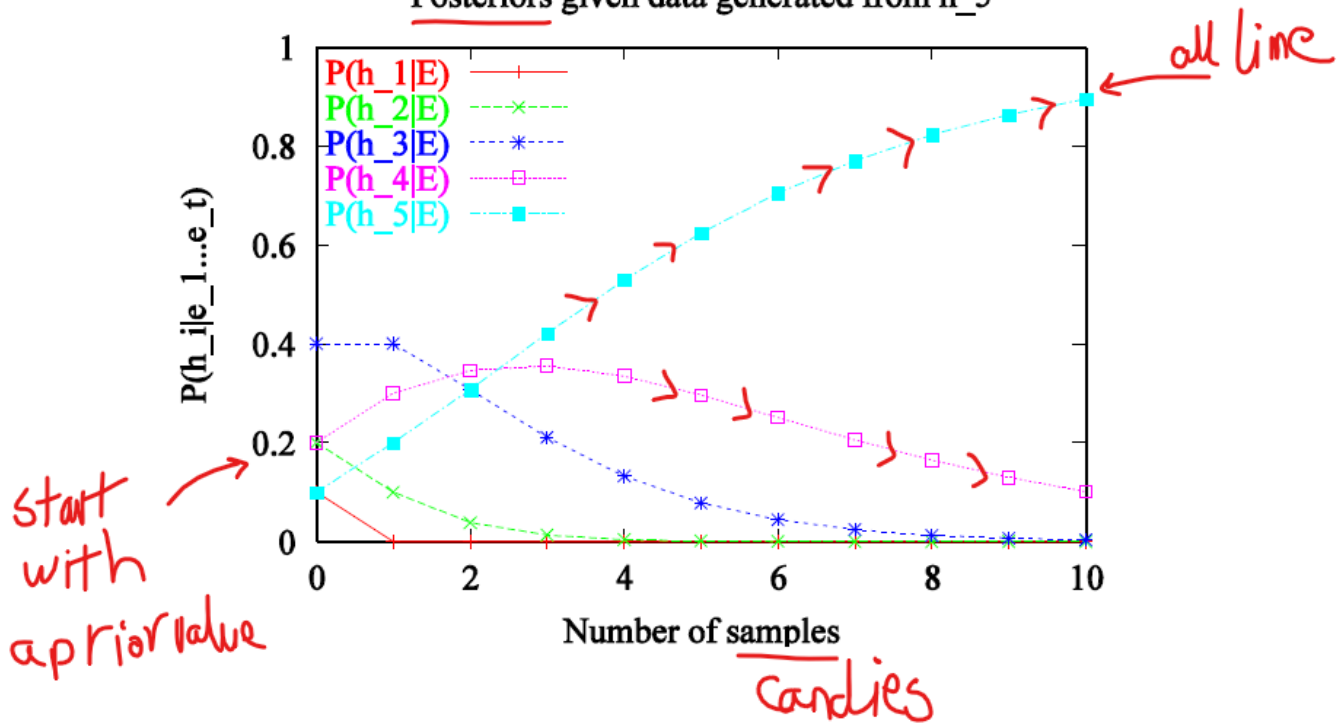
- **Prior:** $\Pr(H)$
- **Likelihood:** $\Pr(e|H)$
- **Evidence:** $e = \langle e_1, e_2, \dots, e_N \rangle$
- **Bayesian Learning** amounts to computing the posterior using Bayes' Theorem:

$$\Pr(H|e) = k \Pr(e|H) \Pr(H)$$

normalize
Likelihood
prior

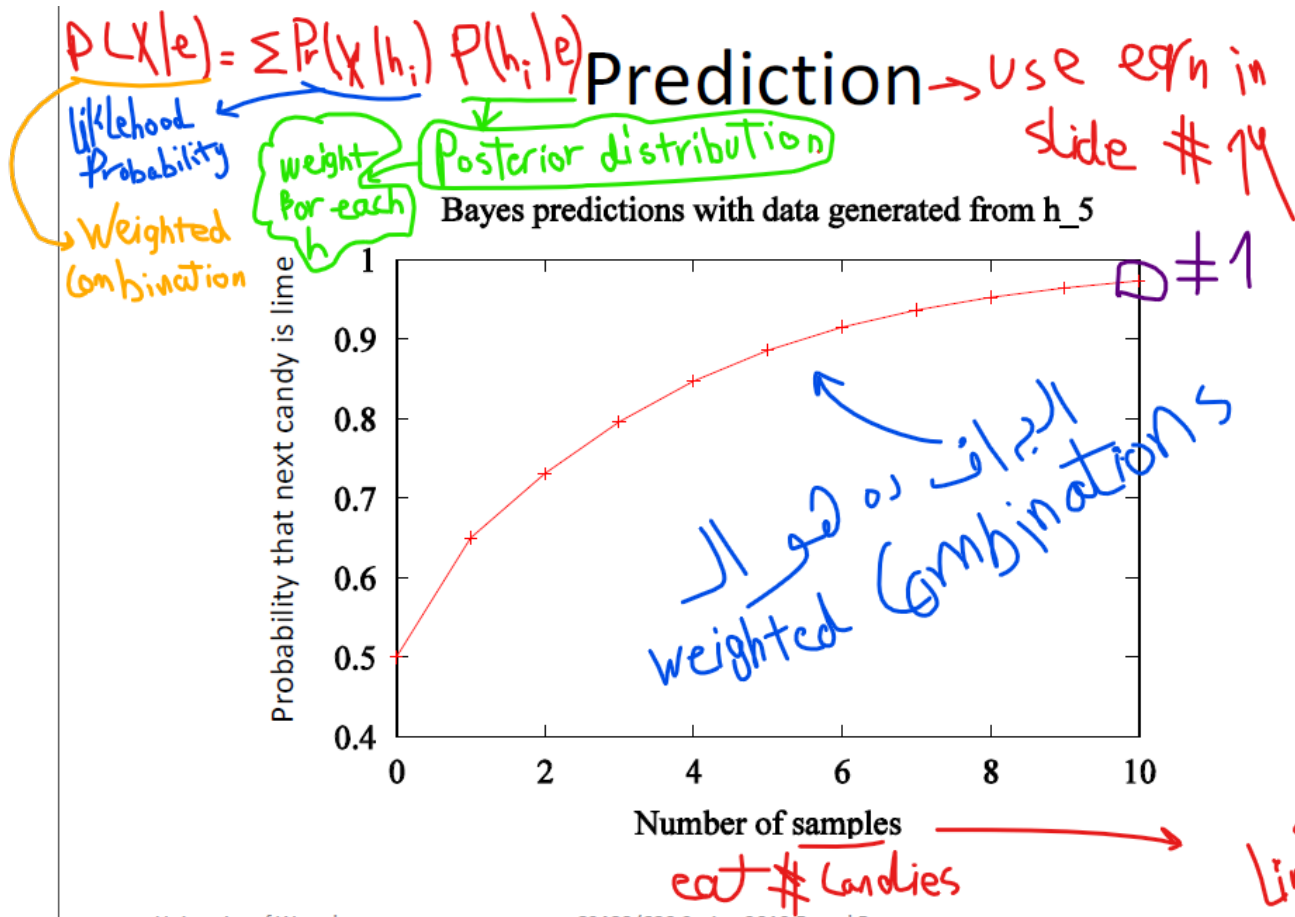
Probability Posterior for each hypothesis

Posteriors given data generated from h_5



ال prior hypothesis هي حابه subjective ... اللي هو ك domain expert انت هتختارها .. ولازم تكون بت reflect ال initial belief بناء علي أي نوع من المعلومات عندك .. فبالتالي ال initial probability عند الصفر هم دول ال prior ..

تعال نشوف الجراف الجاي ... ده بيص علي اننا نعمل a prediction .. عشان زي ما قلنا قبل كذا .. احنا مش مهتمين بحساب ال POSTERIOR DISTRIBUTION مش ديه اللي بندور عليها .. ديه بيبقى معناها اني عندي DISTRIBUTION OVER POSSIBLE MODELS اقدر استخدمهم عشان اعمل PREDICTION .. بس في الأبلكيشن انا مش مهتم اني أدني الناس ديه ال DISTRIBUTION ده .. هم مش هيعرفو يعملو بيه ايه بالظبط هتلاقيه بيقال ما تقولي بس ايه هي درجة الحرارة بكرا وخلص ليه مديني حابه انا مش فاهم هي ايه .. فهنا بص ع الجراف



تعال نلخص الدنيا ... احنا شفنا ازاى ال BAYESIAN LEARNING بيشتغل .. تعال نشوف السلايد الجايه

Bayesian Learning

- Bayesian learning properties:
 - Optimal (i.e. given prior, no other prediction is correct more often than the Bayesian one)
 - No overfitting (all hypotheses considered and weighted)
- There is a price to pay:
 - When hypothesis space is large, Bayesian learning may be intractable
 - i.e. sum (or integral) over hypothesis often intractable
- Solution: approximate Bayesian learning

الدكتور بيقولك ان BAYESIAN LEARNING عنده مناعه لل OVERFITTING .. لانيو بيستخدم كل ال HYPOTHESES اللي موجوده .. اللي احنا هنا بنقولو كالاتي:

How much each hypothesis fits the data essentially allows me to compute a posterior polity, use it as a weight to combine many hypotheses together, so even if there is 1 hypothesis that could overfit, I'm still considering other hypotheses that will counterbalance this hypothesis. As a result, there is no overfitting

في تمن بيتدفع لحاجه زي كدا .. المشكله هنا ان لو عندنا hypothesis space very large .. لو عندك neural network ليها شوية بارمترز كل بارمتر يقدر ياخذ قيمه علي ال Real line اللي هو بياخد ارقام حقيقيه يعني .. فده هيقا عندك انفتت سبيس من البارمترز و بالتالي ال hypothesis space هيقا كبير فشاخه .. وديه مشكله للبرديشكن و لل computation for the posterior distribution ...

البرديشكن عشان احنا هنحتاج ناخذ weighted combinations فلو كان infinite space .. انت كدا عندك weighted integral وممكن مبيقاش متاح في ال closed form .. في أرض الواقع انت بتستخدم approximations .. خذ الهايبو اللي ليها اعلي احتمالية .. واعمل prediction بناء عليها هي .. بس انت ممكن تبقا غلطان وانت بتختار .. فهي مش هتطلع نواتج كويسه زي اللي قبل كدا يعني

Maximum a posteriori (MAP)

- Idea: make prediction based on **most probable hypothesis** h_{MAP}

$$h_{MAP} = \operatorname{argmax}_{h_i} \Pr(h_i|e)$$

$$\Pr(X|e) \approx \Pr(X|h_{MAP})$$

- In contrast, Bayesian learning makes prediction based on **all** hypotheses weighted by their probability

انا الحقيقه فصلت ف آخر 10 دقائق ... من أول ال MAP و ال Max likelihood .. ابقا اقراهم ولو مفهمتش .. اسمعهم تاني ..