

CS480/680

Lecture 4: May 15, 2019

Statistical Learning

[RN]: Sec 20.1, 20.2, [M]: Sec. 2.2, 3.2

Statistical Learning

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

Terminology

- **Probability distribution:**
 - A specification of a probability for each event in our sample space
 - Probabilities must sum to 1
- Assume the world is described by two (or more) random variables
 - **Joint probability distribution**
 - Specification of probabilities for all combinations of events

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):**

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

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

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

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

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

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

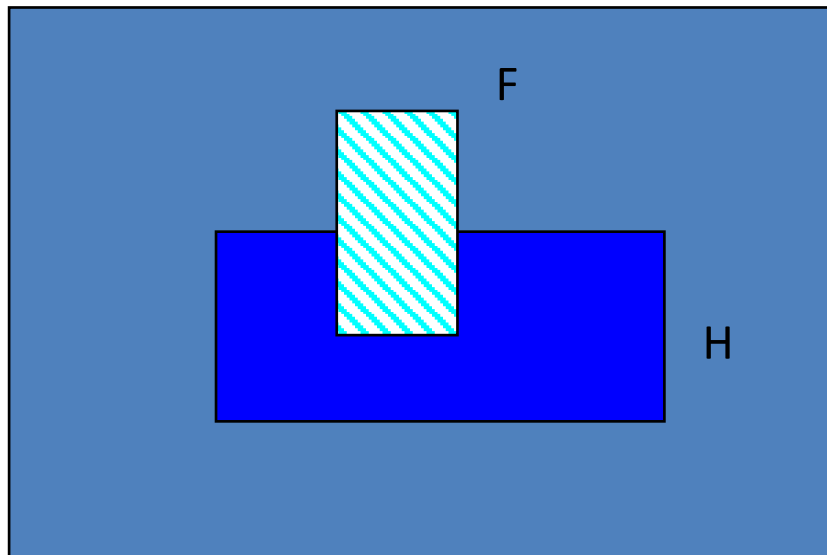
$$P(\text{headache}) =$$



marginalization

Conditional Probability

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



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

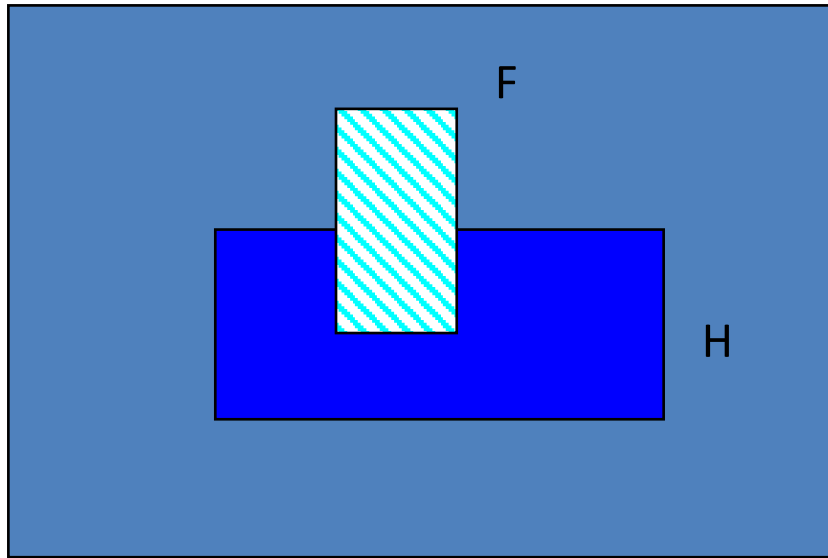
$$\Pr(H) = 1/10$$

$$\Pr(F) = 1/40$$

$$\Pr(H|F) = 1/2$$

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

Conditional Probability



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

$$\Pr(H) = 1/10$$

$$\Pr(F) = 1/40$$

$$\Pr(H|F) = 1/2$$

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

$$= (\# \text{ worlds with flu and headache}) / (\# \text{ worlds with flu})$$

$$= (\text{Area of "H and F" region}) / (\text{Area of "F" region})$$

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

Conditional Probability

- Definition:

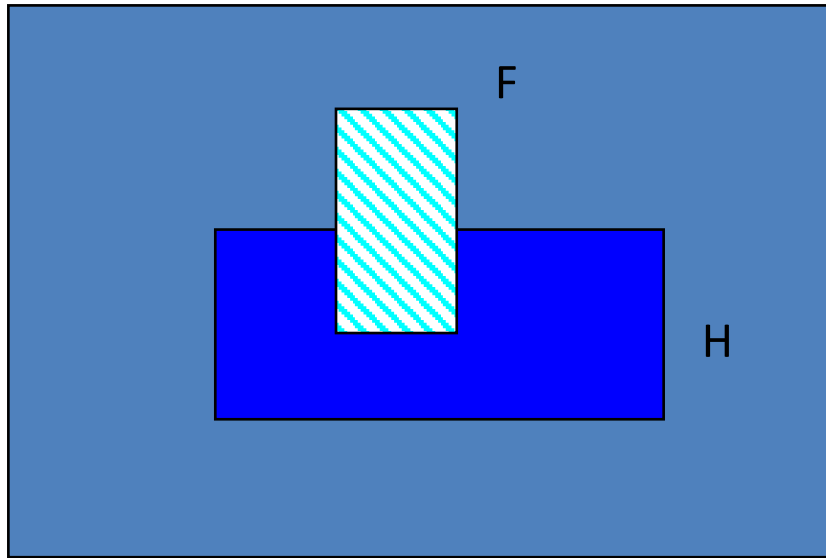
$$\Pr(A|B) = \Pr(A \wedge B) / \Pr(B)$$

- Chain rule:

$$\Pr(A \wedge B) = \Pr(A|B) \Pr(B)$$

Memorize these!

Inference



One day you wake up with a headache. You think "Drat! 50% of flues are associated with headaches so I must have a 50-50 chance of coming down with the flu"

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

$$\Pr(H) = 1/10$$

$$\Pr(F) = 1/40$$

$$\Pr(H|F) = 1/2$$

Is your reasoning correct?

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

$$\Pr(F|H) =$$

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

$$\Pr(\text{headache} \wedge \text{cold} \mid \text{sunny}) =$$

$$\Pr(\text{headache} \wedge \text{cold} \mid \sim \text{sunny}) =$$

Bayes Rule

- Note

$$\Pr(A|B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \wedge A) = \Pr(B|A)\Pr(A)$$

- Bayes Rule

$$\Pr(B|A) = [(\Pr(A|B)\Pr(B))]/\Pr(A)$$

Memorize this!

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

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

Likelihood (red text) points to $P(e|H)$

Prior probability (red text) points to $P(H)$

Posterior probability (red text) points to $P(H|e)$

Normalizing constant (red text) points to $P(e)$

Bayesian Learning

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

Bayesian Prediction

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

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

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
- **Examples E:** evidence about the world
 - e_1 : 1st candy is cherry
 - e_2 : 2nd candy is lime
 - e_3 : 3rd candy is lime
 - ...

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:

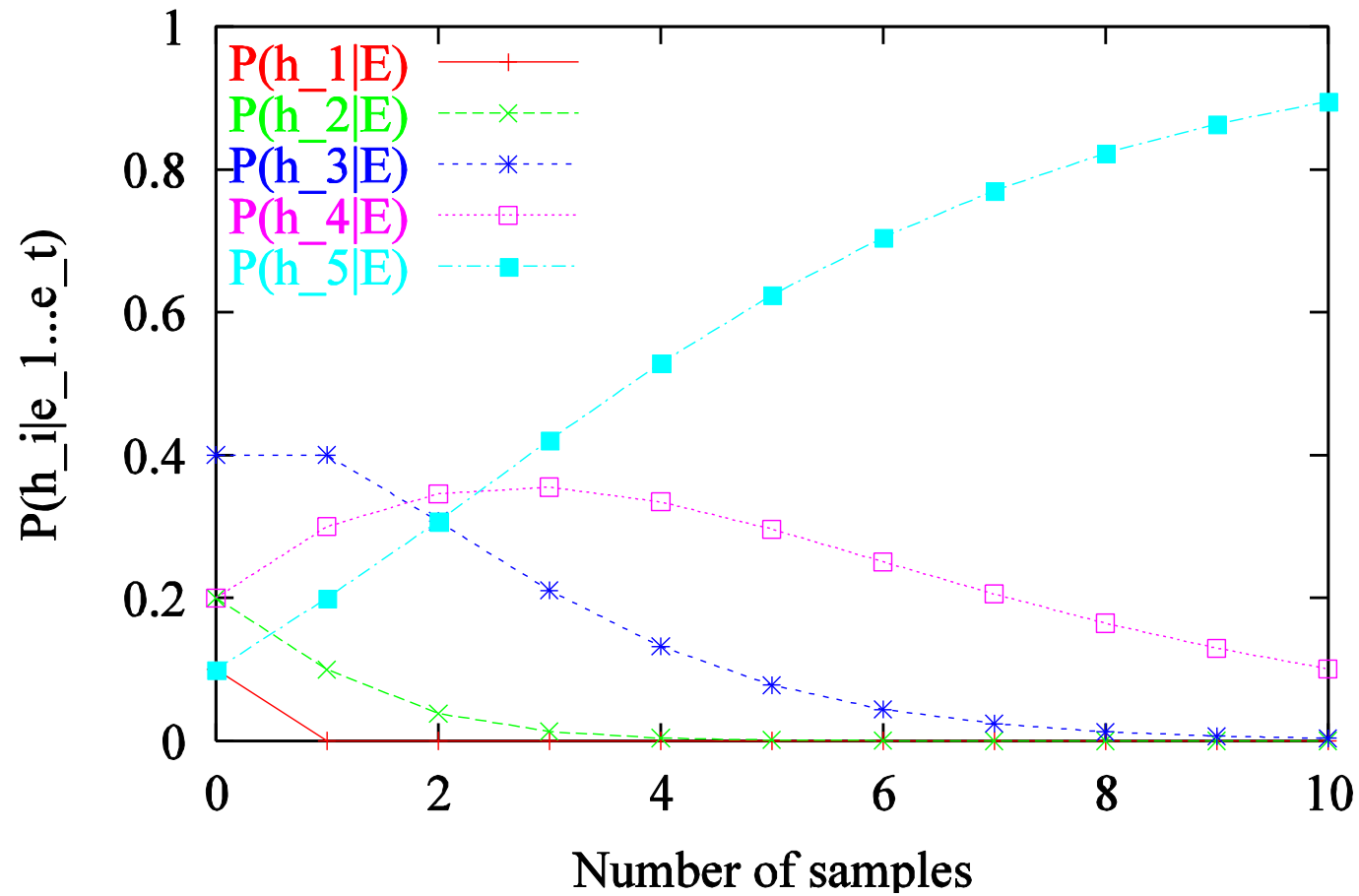
$$\Pr(\mathbf{e}|h_5) =$$

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

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

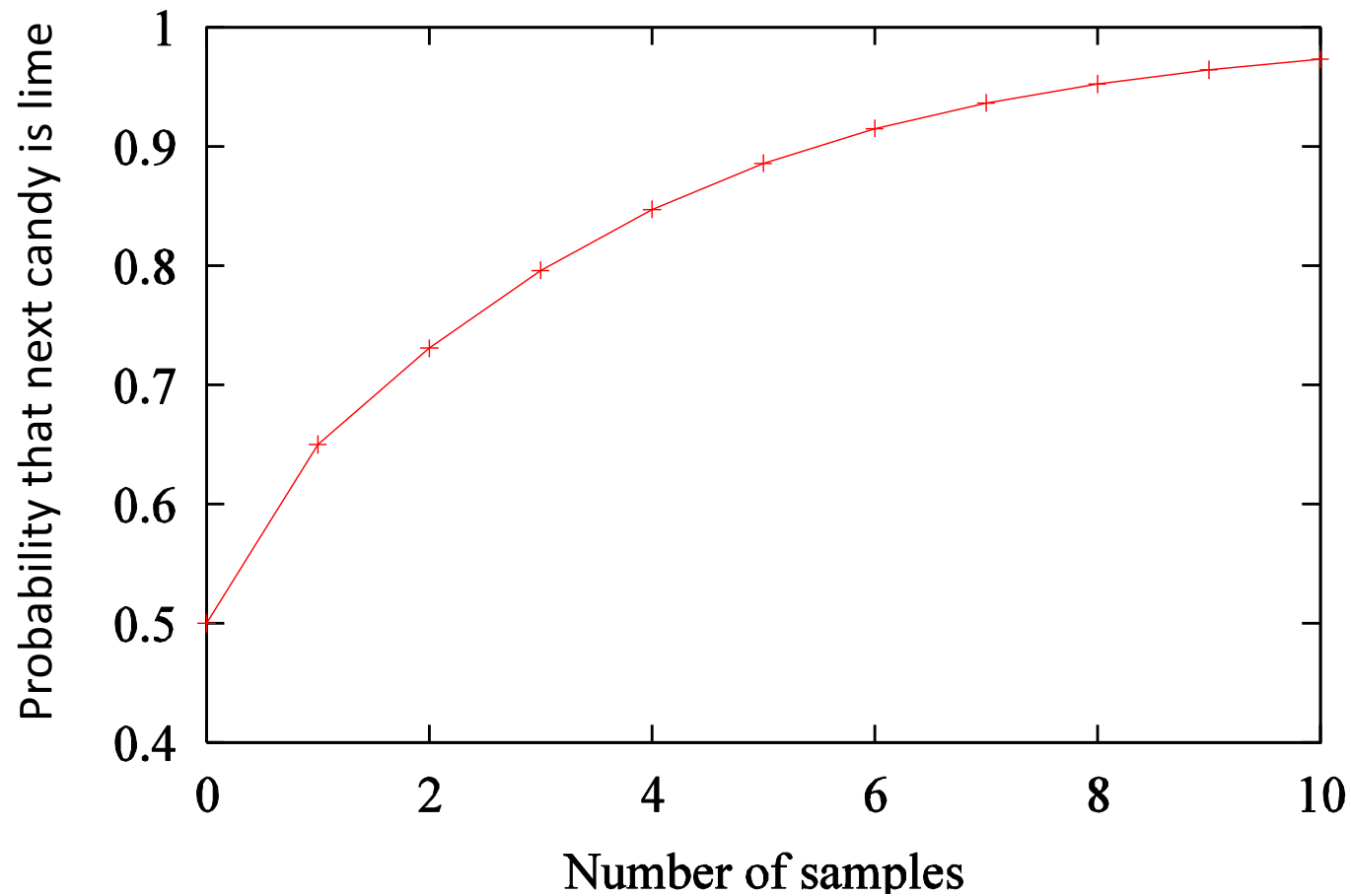
Posterior

Posteriors given data generated from h_5



Prediction

Bayes predictions with data generated from h_5



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

Maximum a posteriori (MAP)

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

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

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

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

MAP properties

- MAP prediction **less accurate** than Bayesian prediction since it relies only on **one** hypothesis h_{MAP}
- But MAP and Bayesian predictions converge as data increases
- **Controlled overfitting** (prior can be used to penalize complex hypotheses)
- **Finding h_{MAP} may be intractable:**
 - $h_{MAP} = \operatorname{argmax}_h \Pr(h|\mathbf{e})$
 - Optimization may be difficult

Maximum Likelihood (ML)

- Idea: simplify MAP by assuming uniform prior (i.e., $\Pr(h_i) = \Pr(h_j) \forall i, j$)

$$h_{MAP} = \operatorname{argmax}_h \Pr(h) \Pr(\mathbf{e}|h)$$

$$h_{ML} = \operatorname{argmax}_h \Pr(\mathbf{e}|h)$$

- Make prediction based on h_{ML} only:

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

ML properties

- ML prediction **less accurate** than Bayesian and MAP predictions since it ignores prior info and relies only on **one** hypothesis h_{ML}
- But ML, MAP and Bayesian predictions converge as data increases
- Subject to **overfitting** (no prior to penalize complex hypothesis that could exploit statistically insignificant data patterns)
- Finding h_{ML} is often easier than h_{MAP}
$$h_{ML} = \operatorname{argmax}_h \sum_n \log \Pr(e_n|h)$$