CS489/698
Lecture 4: Jan 15, 2018

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:

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

- **Marginalisation (sumout rule):**

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

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

Example: Joint Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

P(headache∧sunny∧cold) =

P(~headache∧sunny∧~cold) =

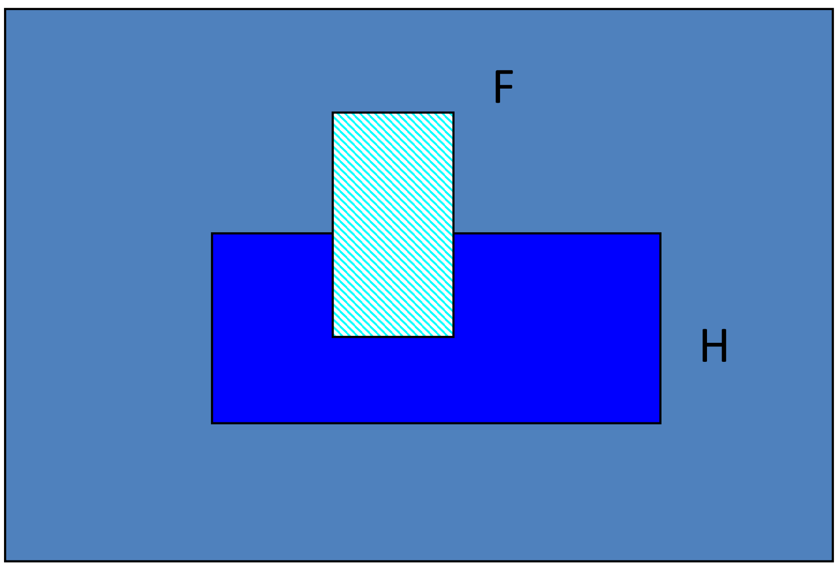
P(headache∨sunny) =

P(headache) =

marginalization

Conditional Probability

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

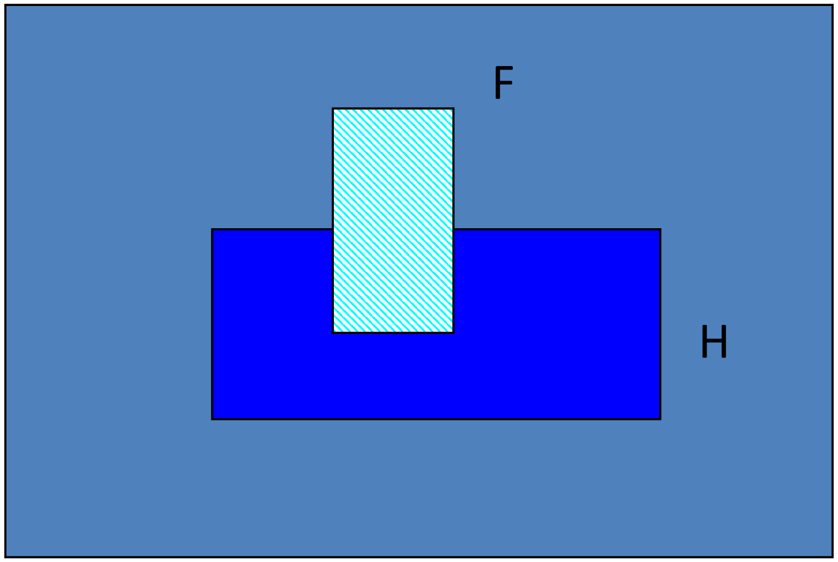


H="Have headache"
F="Have Flu"

$$\Pr(H) = 1/10$$
$$\Pr(F) = 1/40$$
$$\Pr(H|F) = 1/2$$

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

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

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

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

Conditional Probability

- Definition:

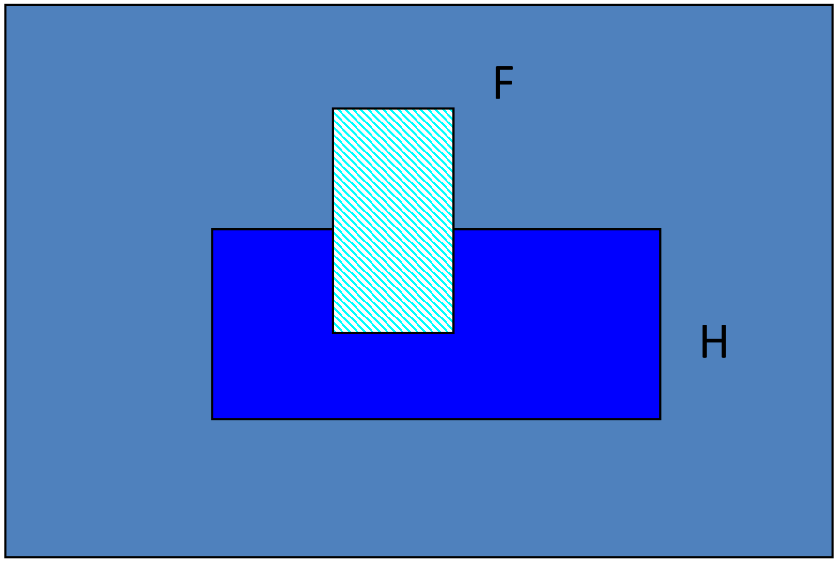
$$\Pr(A|B) = \Pr(A \wedge B) / \Pr(B)$$

- 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?

$\Pr(F \Lambda H) =$

$\Pr(F|H) =$

Example: Joint Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

$\Pr(headache \land cold \mid sunny) =$

$\Pr(headache \land cold \mid \sim sunny) =$

Bayes Rule

- Note

$$\Pr(A|B)\Pr(B) \;=\; \Pr(A \Lambda B) \;=\; \Pr(B \Lambda A) = \Pr(B|A)Pr(A)$$

- 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

Likelihood

Prior probability

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

Posterior probability

Normalizing constant

Bayesian Learning

- **Prior:** $\Pr(H)$
- **Likelihood:** $\Pr(e|H)$
- **Evidence:** $\boldsymbol{e} \ = < e_1, e_2, \dots, e_N >$

- **Bayesian Learning** amounts to computing the posterior using Bayes' Theorem:
$$\Pr(H|\boldsymbol{e}) \ = \ k \Pr(\boldsymbol{e}|H)\Pr(H)$$

Bayesian Prediction

- Suppose we want to make a prediction about an unknown quantity X

- $\Pr(X|\boldsymbol{e}) = \Sigma_i \Pr(X|\boldsymbol{e}, h_i)P(h_i|\boldsymbol{e})$
  $= \Sigma_i \Pr(X|h_i)P(h_i|\boldsymbol{e})$

- 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 $\Pr(H) \; = < 0.1, 0.2, 0.4, 0.2, 0.1 >$
- Assume candies are **i.i.d. (identically and independently distributed)**

$$\Pr(\boldsymbol{e}|h) \; = \; \Pi_n \, P(e_n|h)$$

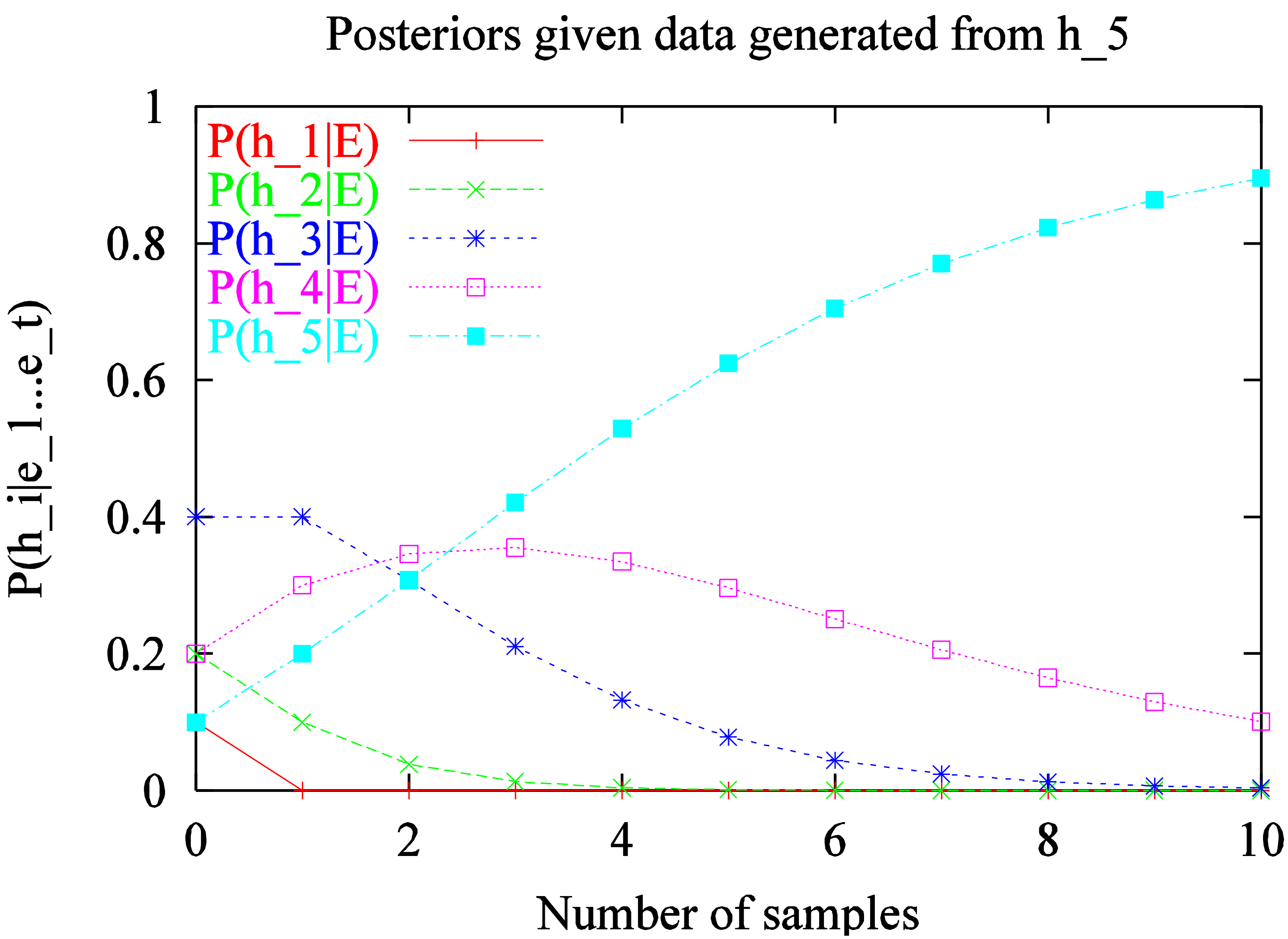
- Suppose first 10 candies all taste lime:

$$\Pr(\boldsymbol{e}|h_5) \; =$$

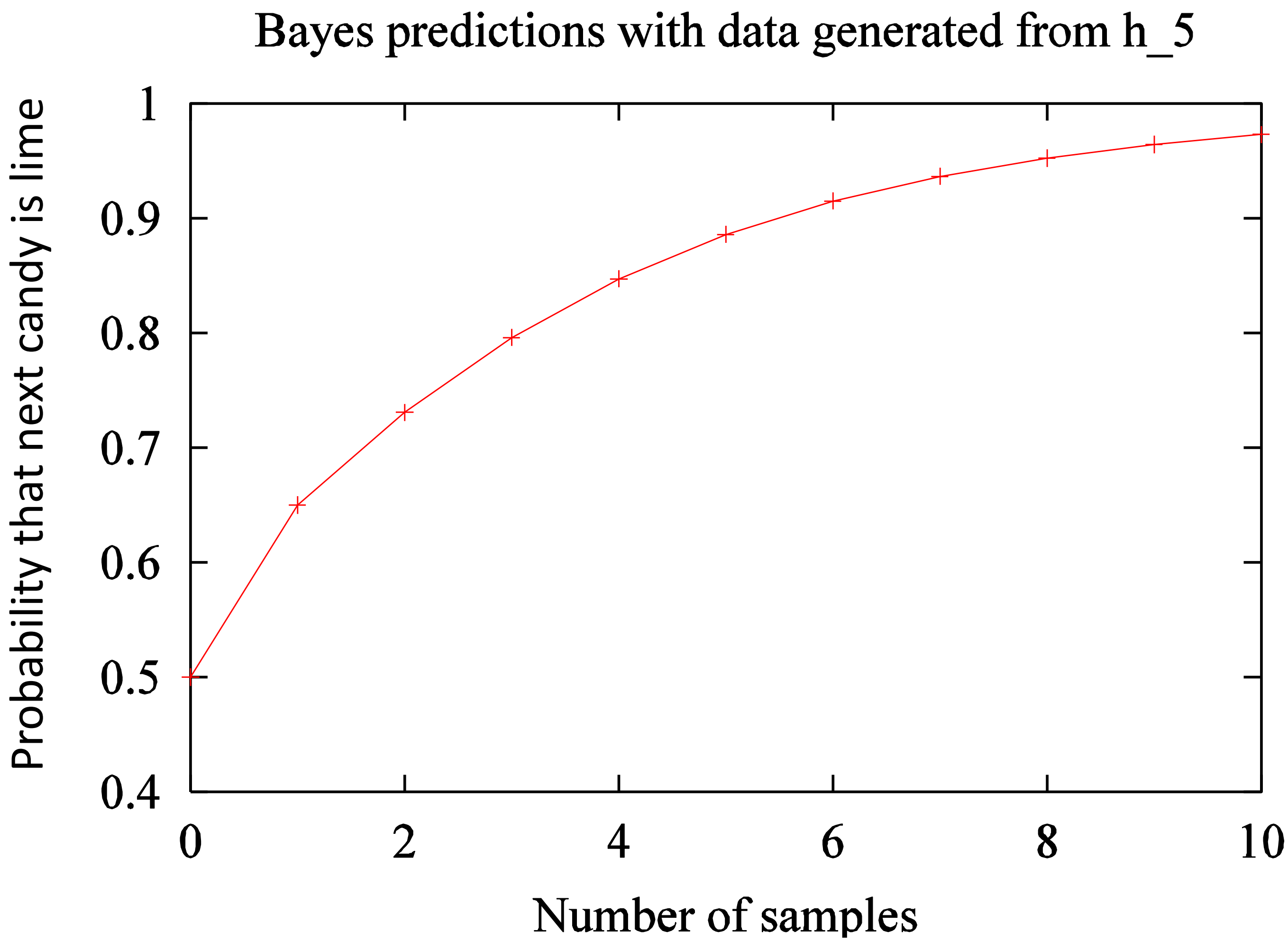
$$\Pr(\boldsymbol{e}|h_3) \; =$$

$$\Pr(\boldsymbol{e}|h_1) \; =$$

Posterior


Posteriors given data generated from h_5
P(h_1|E)
P(h_2|E)
P(h_3|E)
P(h_4|E)
P(h_5|E)
1
0.8
0.6
0.4
0.2
0
P(h_i|e_1...e_t)
0
2
4
6
8
10
Number of samples

Prediction



Bayes predictions with data generated from h_5
Probability that next candy is lime
Number of samples
1
0.9
0.8
0.7
0.6
0.5
0.4
0
2
4
6
8
10

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}

$$h_{MAP} = argmax_{h_i} \Pr(h_i|\boldsymbol{e})$$
$$\Pr(X|\boldsymbol{e}) \approx \Pr(X|h_{MAP})$$

- 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:**
  - $h_{MAP} = argmax_h \Pr(h|\boldsymbol{e})$
  - Optimization may be difficult

Maximum Likelihood (ML)

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

$$h_{MAP} = argmax_h \Pr(h) \Pr(\boldsymbol{e}|h)$$

$$h_{ML} = argmax_h \Pr(\boldsymbol{e}|h)$$

- Make prediction based on h_{ML} only:

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

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}

$$h_{ML} = argmax_h \Sigma_n \log \Pr(e_n|h)$$