CS 5806 Machine Learning II

Lecture 8 - Statistical Learning 1: Bayesian Learning for Classification
September 18th, 2023
Hoda Eldardiry

Recommended Reading

- [7] Sec 20.1, 20.2
- [8] Sec. 2.2, 3.2
- References are listed on canvas /pages/textbook-resources

Lecture Objectives

- Learn how to justify & explain certain algorithms from a statistical perspective

Statistical Learning

- **Learning**
  — Uncertain knowledge of the world (uncertain about some concepts)
  — Learning reduces this uncertainty

- **Capture & quantify uncertainty**
  — Statistics & probability theory
  — Use a distribution to capture uncertainty

- **Learning reduces uncertainty**
  — By updating the distribution

- **Today**
  — Update distributions
  — Compute results of learning

Probability distribution

- Characterize the world using random variables

- Quantify uncertainty in the world using probability

- **Probability distribution**
  — Specifies **probability for each event** in a sample set
  — [Uncertainty when rolling a dice: 6 possible outcomes]
  — Probabilities must sum to 1

- **Joint probability distribution**
  — Assume world is described by 2 (or more) random variables
  — [Weather: temperature, wind, humidity]
  — Specifies **probabilities for all combinations of events**
  — [Probability that temp, humidity, wind speed take certain values]

- What is the process of having different values of those random values?

Joint distribution

- Given two random variables A & B (quantities of interest)
- **Joint distribution** $P(A = a \wedge B = b) \forall_{a,b}$
- To make a prediction for only one random variable (perhaps we don't care about the combination)
- Given join prob of temp, humidity, wind. Want to extract distribution of temp
- **Marginalization (sumout rule)**

$$P(A = a) = \sum_b P(A = a \wedge B = b)$$ Sumout all possible values for B

random variable

$$P(B = b) = \sum_a P(A = a \wedge B = b)$$ Sumout all possible values for A

values

Joint Distribution - Example

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

$$P(headache \wedge sunny \wedge cold) =$$

$$P(\sim headache \wedge sunny \wedge \sim cold) =$$

$$P(headache \vee sunny) =$$

$$P(headache) =$$

Joint Distribution - Example

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

$P(headache \wedge sunny \wedge cold) =$ 0.108

$P(\sim headache \wedge sunny \wedge \sim cold) =$ 0.064

$P(headache \vee sunny) =$ 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28

$P(headache) =$ 0.108 + 0.012 + 0.072 + 0.008 = 0.2

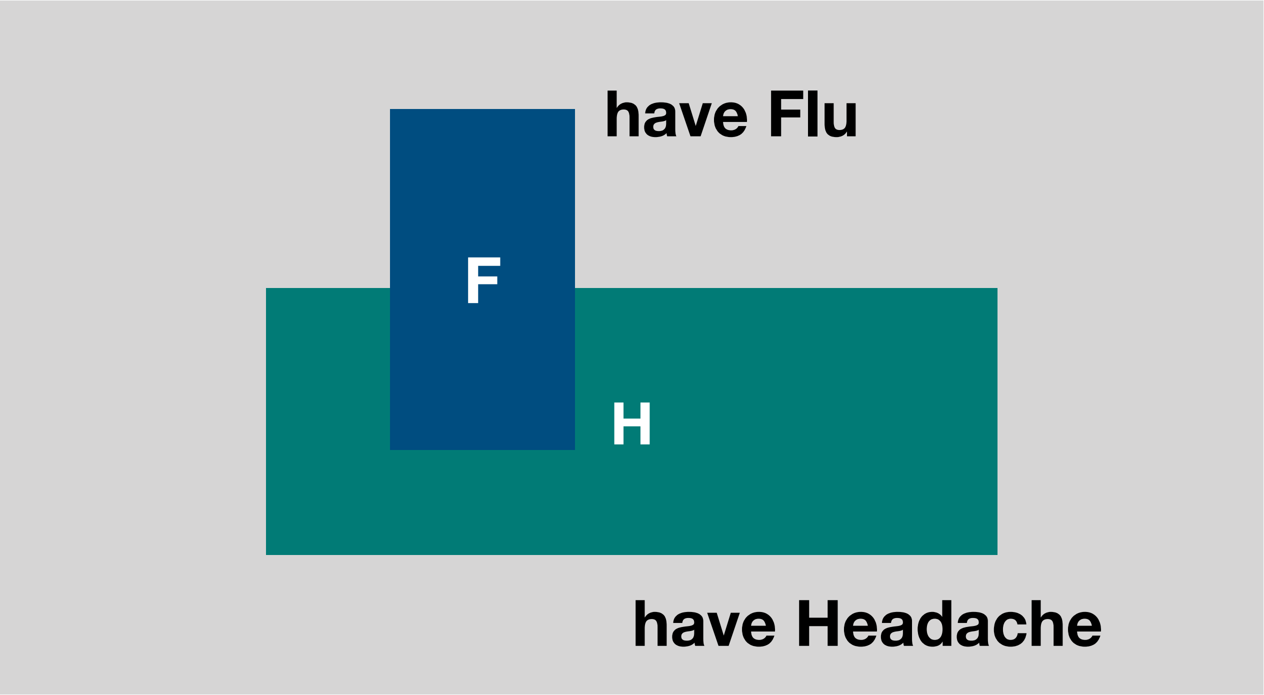
marginalization

Conditional Probability

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

- Headaches are rare,
  Flu is rarer

- But if you have flu
  50-50 chance
  you will have headache



P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

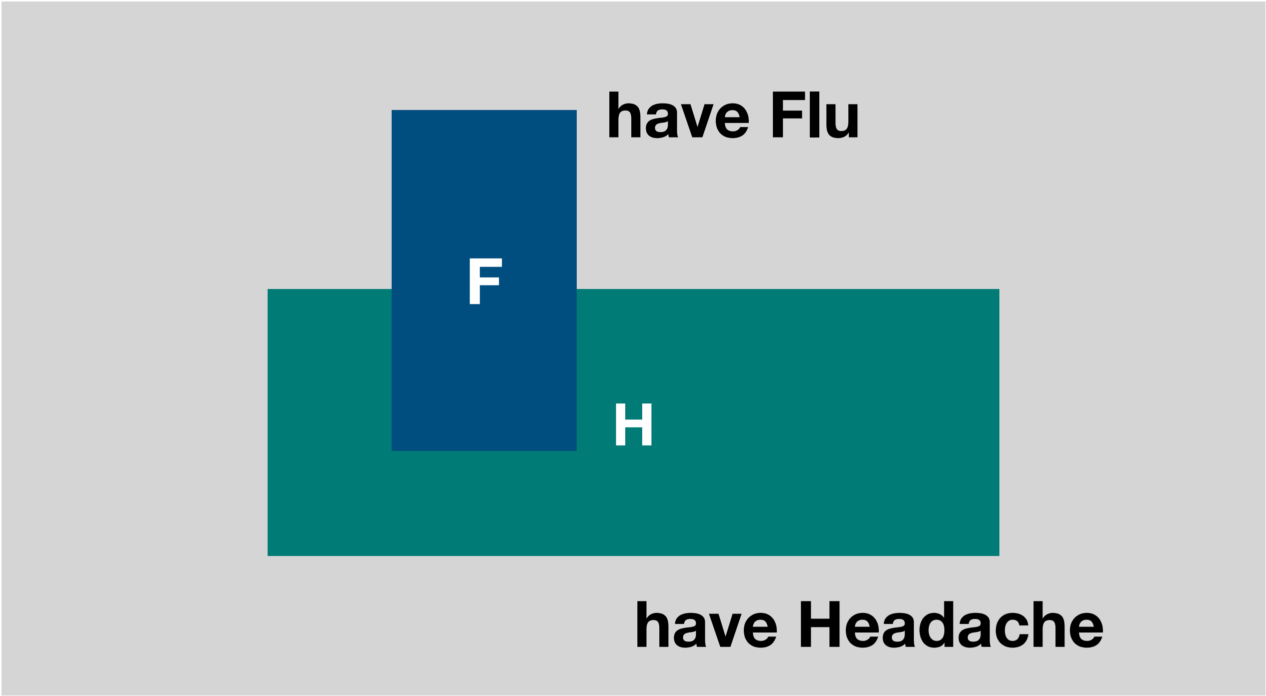
Conditional Probability

P(H|F) = fraction of flu inflicted worlds in which one has a headache

=#worlds (flu & headache) / #worlds (flu)

= area (H&F) / area (F)

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



***P*(H) = 1/10**
***P*(F) = 1/40**
***P*(H|F) = 1/2**

Conditional Probability

- **Definition**

$$P(A \mid B) = P(A \wedge B)/P(B)$$

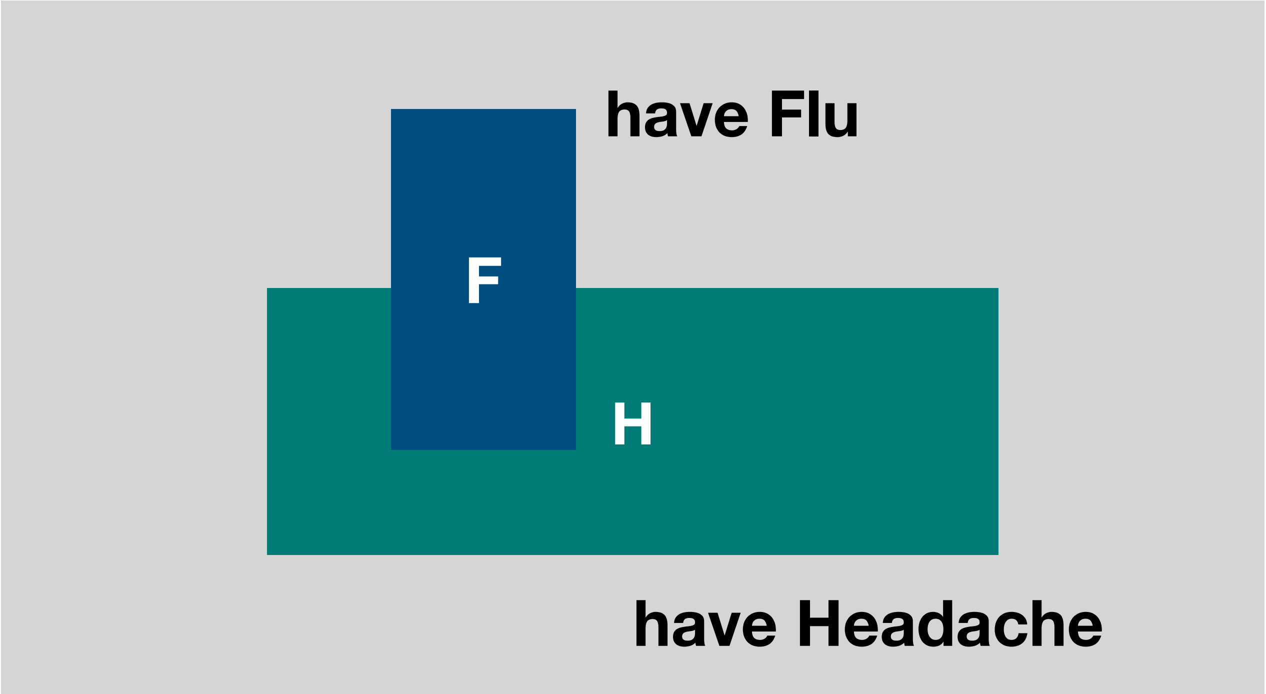
- **Chain rule**

$$P(A \wedge B) = P(A \mid B)P(B)$$

Inference

— you wake up with a headache
— you think..“50% of Flu is associated with headaches”
— so ..“50% chance I have Flu”

— is your reasoning correct?



$$P(F \wedge H) = P(H|F)P(F) = 1/2 * 1/40 = 1/80$$

$$P(F|H) = P(F \wedge H)/P(H) = (1/80)/(1/10) = 1/8$$

***P*(H) = 1/10**
***P*(F) = 1/40**
***P*(H|F) = 1/2**

Joint Distribution - Example

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

$$P(headache \wedge cold \,|\, sunny) = \quad P(headache \wedge cold \wedge sunny)/p(sunny)$$

= 0.108 / (0.108 + 0.012 + 0.016 + 0.064) = 0.54

$$P(headache \wedge cold \,|\, \sim sunny) = \quad P(headache \wedge cold \wedge \sim sunny)/p(\sim sunny)$$

= 0.072 / (0.072 + 0.008 + 0.144 + 0.576) = 0.09

Summary

- Probability distributions quantify uncertainty about the world
- Learning reduces uncertainty
- Probability distributions: joint, marginal, conditional
- Conditional probability $P(A \mid B) = P(A \wedge B)/P(B)$
- Chain rule $P(A \wedge B) = P(A \mid B)P(B)$

Bayes Rule

- **Note**

$$P(A \mid B)P(B) = P(A \wedge B) = P(B \wedge A) = P(B \mid A)P(A)$$

- **Bayes rule**

$$P(B \mid A) = [P(A \mid B)P(B)] \ / \ P(A)$$

Using Bayes Rule for inference

- Form a **hypothesis** about the world based on what we **observe**
- **Bayes rule** enables stating ..
- .. the belief given to **hypothesis *H***, given **evidence *e***

Prior probability

Likelihood

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

Posterior probability

Normalizing constant

Bayesian Learning

Likelihood

Prior probability

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

Posterior probability

Normalizing constant

- **Prior:** P(H)
- **Likelihood:** P(e|H)
- **Evidence: e** = <e_1, e_2, …, e_N>
- **Bayesian Learning =** compute posterior using Baye's Theorem
- $P(H\,|\,\mathbf{e}) = k\ P(\mathbf{e}\,|\,H)P(H)$

Bayesian Prediction

- To make a prediction about unknown quantity X

$$P(X \mid \mathbf{e}) = \sum_i P(X \mid \mathbf{e}, h_i) P(h_i \mid \mathbf{e})$$

predictions

$$= \sum_i P(X \mid h_i) P(h_i \mid \mathbf{e})$$

weighted averages of predictions of individual hypotheses

predictions of individual hypothesis   weight: posterior/belief

- Predictions: weighted averages of predictions of individual hypothesis h_i: model

- Hypotheses serve as “intermediaries” between raw data and prediction

Candy Example

- Candy sold in two flavors: Lime, Cherry
- 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
- You bought a bag of candy but don’t know its flavor ratio
- We can run an experiment: eat ***k*** candies: then try to estimate:
  – What’s the flavor ratio of the bag?
  – What will be the flavor of the next candy?
- What is the hypothesis?
- What is the evidence?

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
  - …

Statistical Learning

- Assume prior P(H) = <0.1,0.2,0.4,0.2,0.1>
- Assume candies are **i.i.d. (identically and independently distributed)**

Likelihood distribution: probability of observing a flavor e given a hypothesis h

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

- Suppose first 10 candies all taste lime:

$$P(\mathbf{e} \mid h_5) = \quad 1^{10} \quad =1$$

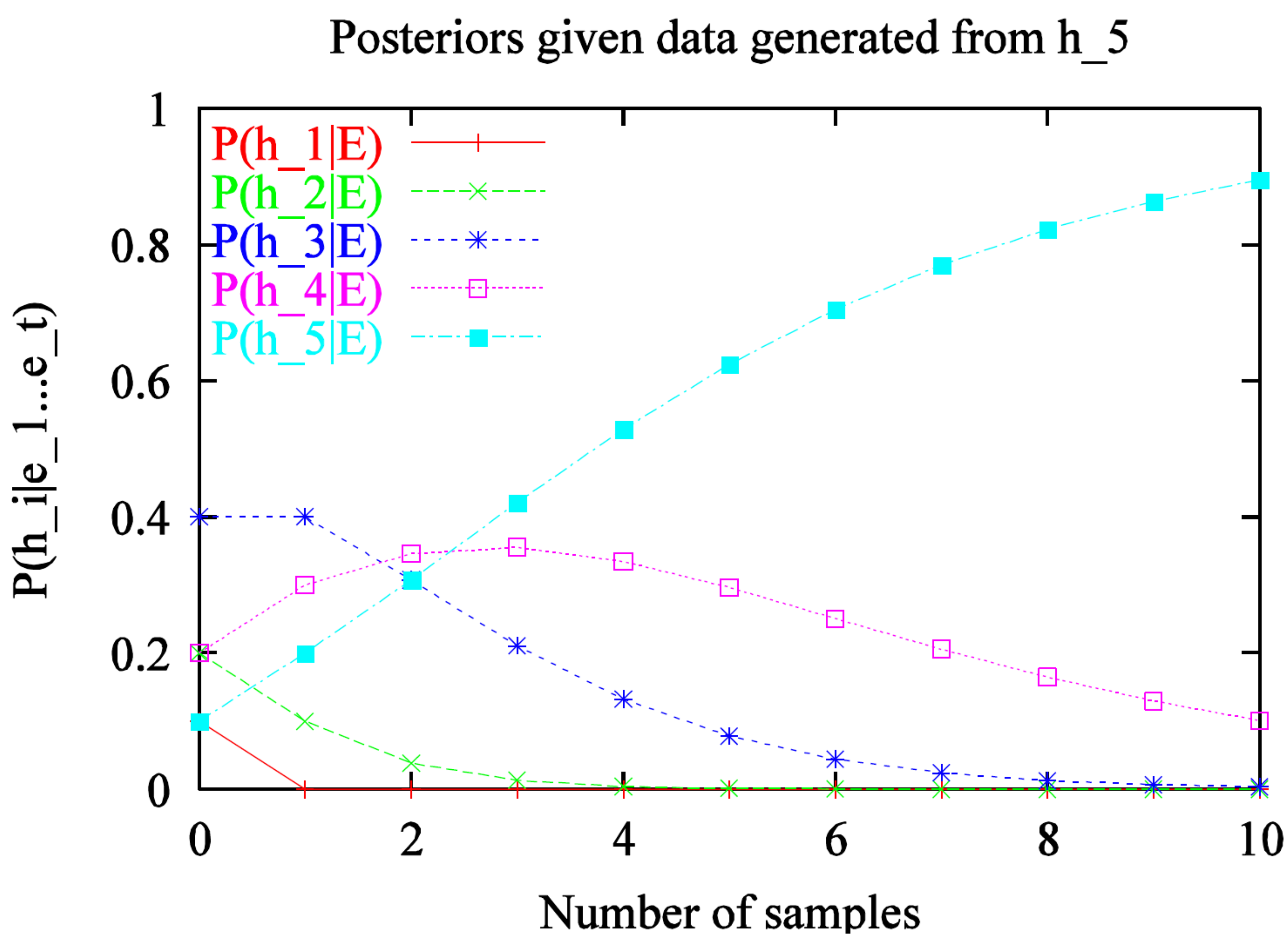
$$P(\mathbf{e} \mid h_3) = \quad (1/2)^{10} \quad = 0.00097$$

$$P(\mathbf{e} \mid h_1) = \quad (0)^{10} \quad = 0$$

- 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

Posterior

$$P(H \mid \mathbf{e}) = k\ P(\mathbf{e} \mid H)P(H)$$

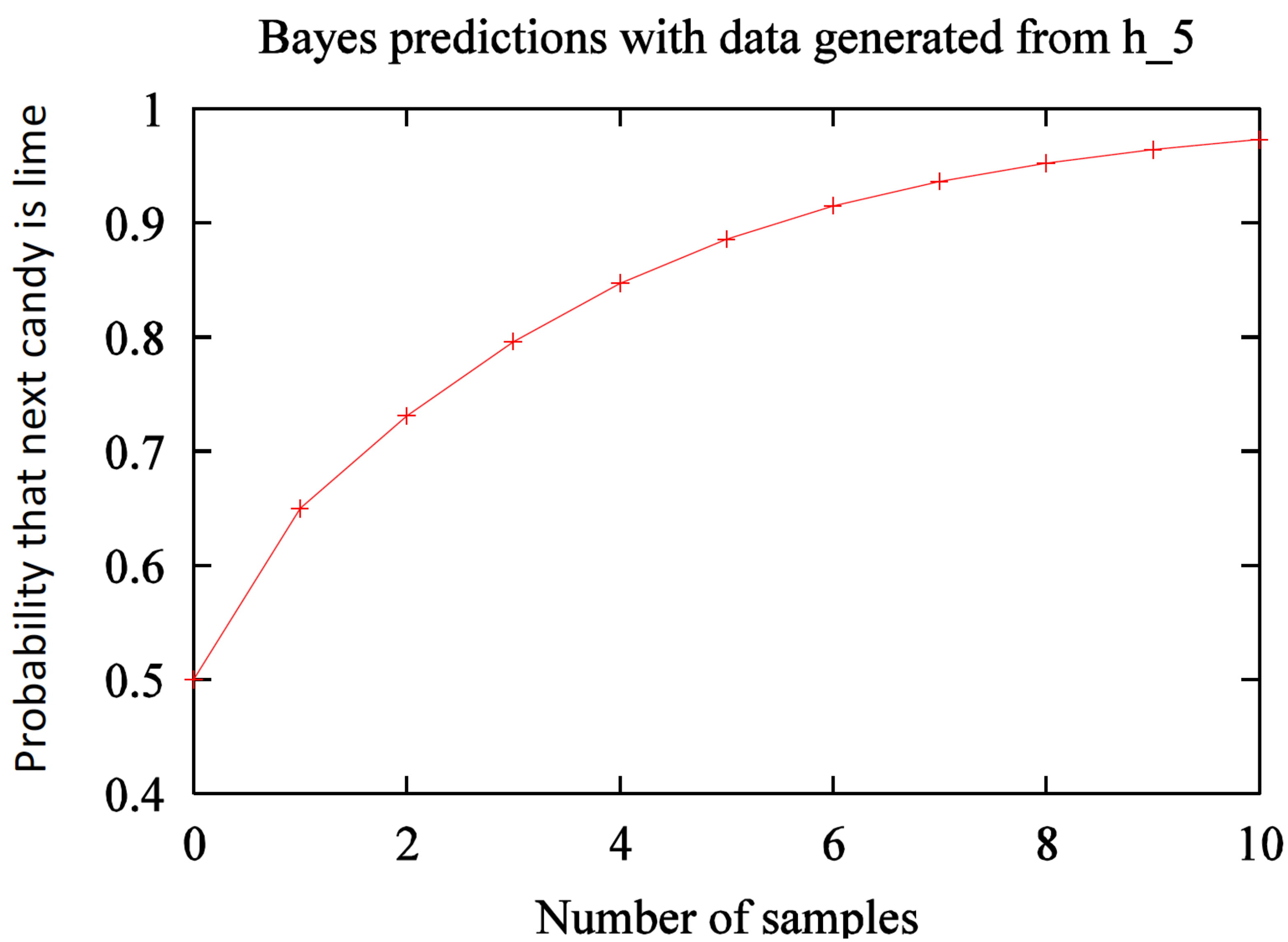


Prediction

Likelihood probability

$$P(X \mid \mathbf{e}) = \sum_i P(X \mid h_i) P(h_i \mid \mathbf{e})$$

Posterior distribution provides weights for each hypothesis



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)

- **Limitation**
  – When hypothesis space is large, Bayesian learning may be intractable
  – i.e. sum (or integral) over hypotheses often intractable

- **Solution**
  – "approximate" Bayesian learning