Computational learning and discovery

CSI 873 / MATH 689

Instructor: I. Griva

Wednesday 7:20 - 10 pm

Bayesian learning

- Provides a probabilistic approach to learning
- Can calculate explicit probabilities for hypotheses
- Perform well on practice
- Help understand better other learning algorithms

Features of Bayesian Learning

- Each training example either increase or decrease the probability that some hypothesis is correct

- Capable of probabilistic predictions

- Prior knowledge (such as probability for a candidate hypothesis) can be used

- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities

The goal of Bayesian Learning

To determine the best hypothesis from H, given the observe training data D and the prior knowledge about the quality of the hypotheses from H!

best hypothesis = most probable hypothesis

Bayes Theorem

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

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D (posterior prob.)
- $P(D|h)$ = probability of D given h

Maximum a posteriori hypothesis (MAP)

Maximum a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

If we assume that *P*(*h*) = const for any *h* then we Are choosing the maximum likelihood (ML) hypothesis:

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Example

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

Summary of basic probability formulas

- *Product Rule*: probability $P(A \wedge B)$ of a conjunction of two events A and B:

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

- *Sum Rule*: probability of a disjunction of two events A and B:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- *Theorem of total probability*: if events $A_1, \ldots, A_n$ are mutually exclusive with $\Sigma_{i=1}^{n} P(A_i) = 1$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Brute-Force Bayes MAP Learning

1. For each hypothesis h in H, calculate the posterior probability

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

2. Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

MAP and Concept Learning

Consider our usual concept learning task

- instance space X, hypothesis space H, training examples D
- consider the FINDS learning algorithm (outputs most specific hypothesis from the version space $VS_{H,D}$)

What would Bayes rule produce as the MAP hypothesis?

Does $FindS$ output a MAP hypothesis??

MAP and Concept Learning

Assume fixed set of instances $\langle x_1, \ldots, x_m \rangle$

Assume D is the set of classifications

$D = \langle c(x_1), \ldots, c(x_m) \rangle$

Choose $P(D|h)$

- $P(D|h) = 1$ if h consistent with D
- $P(D|h) = 0$ otherwise

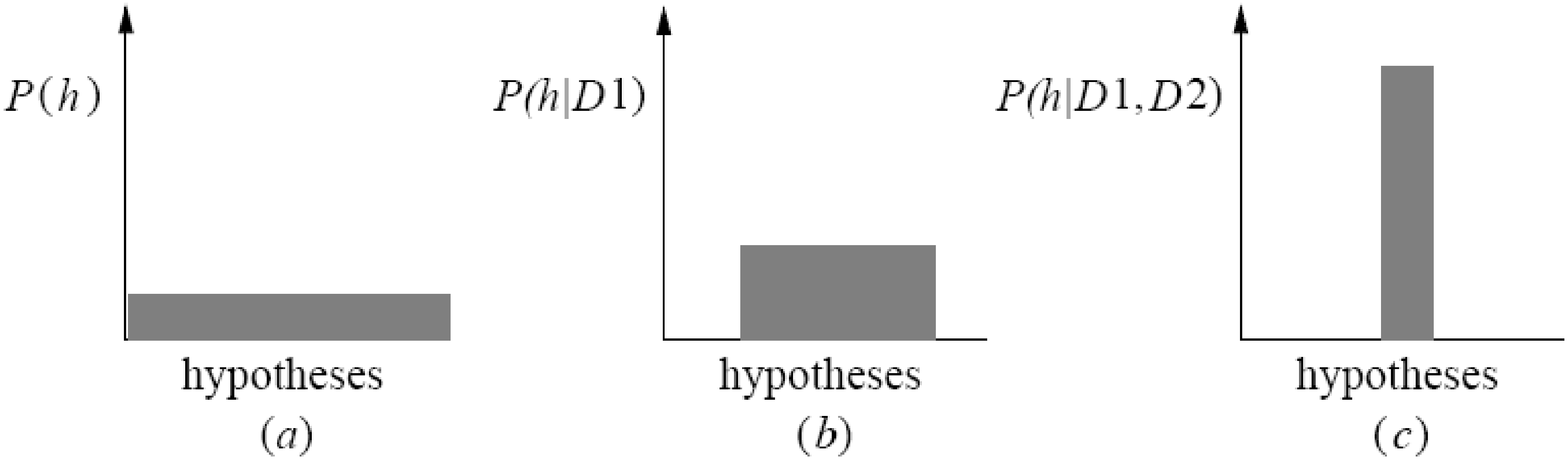
Choose $P(h)$ to be *uniform* distribution

- $P(h) = \frac{1}{|H|}$ for all h in H

Then,

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases}$$

MAP and Concept Learning



Characterizing Learning algorithms by Equivalent MAP systems