Outline

- Bayes Theorem
- MAP hypotheses
- MAP learners

Bayesian Learning

Features:

- Training examples incrementally increase/decrease probability of each hypothesis
- Considers both prior knowledge (prior probabilities) and observed data
- Accomodate probabilistic hypotheses (e.g., "83% of students will pass the test")
- Classify new instances by combining predictions of multiple hypotheses
- Often inefficient
- Framework for characterizing algs that don't explicitly use probabilities

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
- $P(D|h)$ = probability of D given h

Choosing Hypotheses

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

Generally want the most probable hypothesis given the training data

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 $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

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

Bayes Theorem

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.

$$P(cancer) = \qquad P(\neg cancer) =$$
$$P(+|cancer) = \qquad P(-|cancer) =$$
$$P(+|\neg cancer) = \qquad P(-|\neg cancer) =$$

Basic Formulas for Probabilities

- *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 MAP Hypothesis Learner

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

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

Relation to Concept Learning

Consider our usual concept learning task

- instance space X, hypothesis space H, training examples D
- consider $FindS$ learning algorithm (picks a 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??

Relation to 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 $P(D|h)$:

Relation to 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

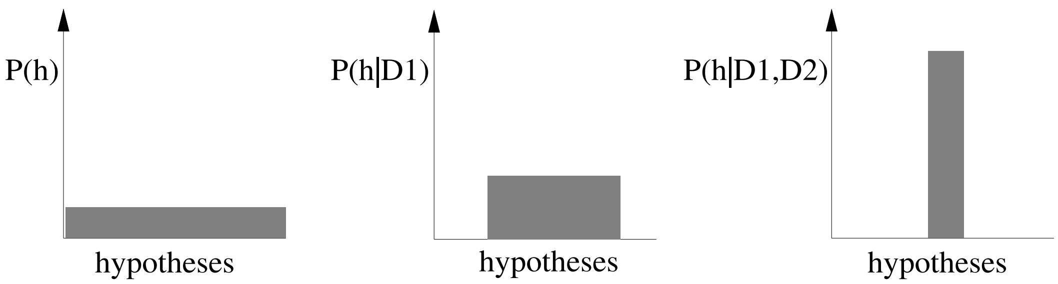
Choose $P(h)$ to be *uniform* distribution

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

Then,

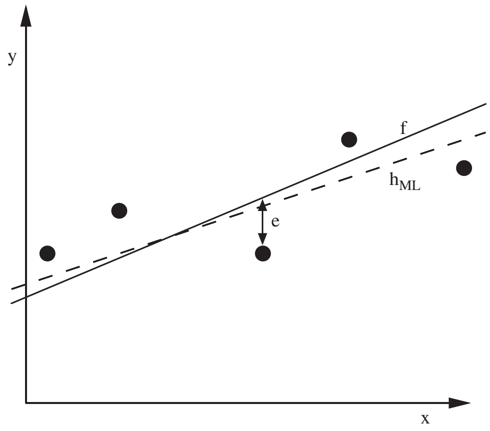
$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D \\ \\ 0 & \text{otherwise} \end{cases}$$

Evolution of Posterior Probabilities



(a) (b) (c)

Learning A Real Valued Function



Consider any real-valued target function f
Training examples $\langle x_i, \hat{f}(x_i) \rangle$, where

- $\hat{f}(x_i) = f(x_i) + e_i$
- e_i is random variable (noise) drawn independently for each x_i according to some Gaussian distribution with mean=0

Then the maximum likelihood hypothesis h_{ML} is the one that minimizes the sum of squared errors:

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} \left(\hat{f}(x_i) - h(x_i) \right)^2$$

Why isn't $h_{ML} = f$?