Bayesian Inference

TM Chapter 6

Outline

- Bayes Theorem
- MAP & ML hypotheses
- MAP learners
- Minimum description length (MDL) principle
- Bayes-optimal, Gibbs, and Naive Bayes classifiers
- Expectation Maximization (EM) algorithm

Why Study Bayesian Inference?

Provides practical learning algorithms:

- Naive Bayes classifiers & Bayesian belief networks
- Allows prior knowledge (in the form of prior belief) to be combined with observed data to give probabilistic prediction
- Allows new input instance to be classified by combining predictions of multiple hypotheses weighted by their beliefs
- Incrementally updates belief of hypothesis with each training example

Provides useful conceptual framework:

- Provides "gold standard" to evaluate other learning algorithms
- Additional insight into Occam's razor

Bayes' Theorem/Belief Update

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

- $P(h)$: prior belief of hypothesis h
- $P(D|h)$: likelihood of data D given h
- $P(D) = \sum_{h \in H} P(D|h)P(h)$: marginal likelihood/evidence of D
- $P(h|D)$: posterior belief of h given D

Limitations.

- Requires specifying probabilities and underlying distributions
- Often prohibitively expensive to compute evidence

How to Choose Hypothesis?

We generally want the most probable hypothesis given the training data, i.e., maximum *a posteriori* hypothesis:

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

If $P(h) = P(h')$ for any $h, h' \in H$, then we can further simplify and choose the maximum likelihood hypothesis:

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

Example: Medical Diagnosis

Does the patient have cancer or not?

- A patient takes a lab test and the result comes back +ve.
- The test returns a correct +ve result in only 98% of the cases in which cancer is actually present,
- and a correct –ve result in only 97% of the cases in which cancer is not present.
- Furthermore, 0.008 of the entire population have this cancer.

$P(cancer) =$ $\qquad$ $P(\neg cancer) =$

$P(+ \mid cancer) =$ $\qquad$ $P(- \mid cancer) =$

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

Basic Probability Formulas

Chain rule for probability. Joint probability $P(A_1, \ldots, A_n)$ of a conjunction of n events $A_1, \ldots, A_n$:

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P(A_i | A_1, \ldots, A_{i-1}) \ .$$

Inclusion-exclusion principle. Probability of a disjunction/ union of n events $A_1, \ldots, A_n$:

$$\begin{aligned} P(\textstyle\bigcup_{i=1}^{n} A_i) = & \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i, A_j) \\ & + \sum_{1 \leq i < j < k \leq n} P(A_i, A_j, A_k) - \ldots \\ & + (-1)^{n-1} P(A_1, \ldots, A_n) \ . \end{aligned}$$

Marginalization. If events $A_1, \ldots, A_n$ are mutually exclusive s.t. $\sum_{i=1}^{n} P(A_i) = 1$, then $P(B) = \sum_{i=1}^{n} P(B|A_i) P(A_i)$.

BRUTE-FORCE MAP HYPOTHESIS LEARNER

1. For each hypothesis $h \in H$, compute posterior belief

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

2. Output hypothesis h_{MAP} with highest posterior belief

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

Relation to Concept Learning

Consider our usual concept learning task:

- Input instance space X, hypothesis space H, unknown target concept/function $c : X \rightarrow \{0, 1\} \in H$, noise-free training examples $D = \{d\}$ where $d = \langle \mathbf{x}_d, c(\mathbf{x}_d) \rangle$

- FIND-S outputs most specific hypothesis from version space $VS_{H,D}$

What would Bayes rule produce as the MAP hypothesis?

Does FIND-S output a MAP hypothesis?

Relation to Concept Learning

- Assume that input instances $\mathbf{x}_d$ for $d \in D$ are fixed
- Choose $P(D \mid h)$:

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

- Choose $P(h)$ to be uniform distribution:

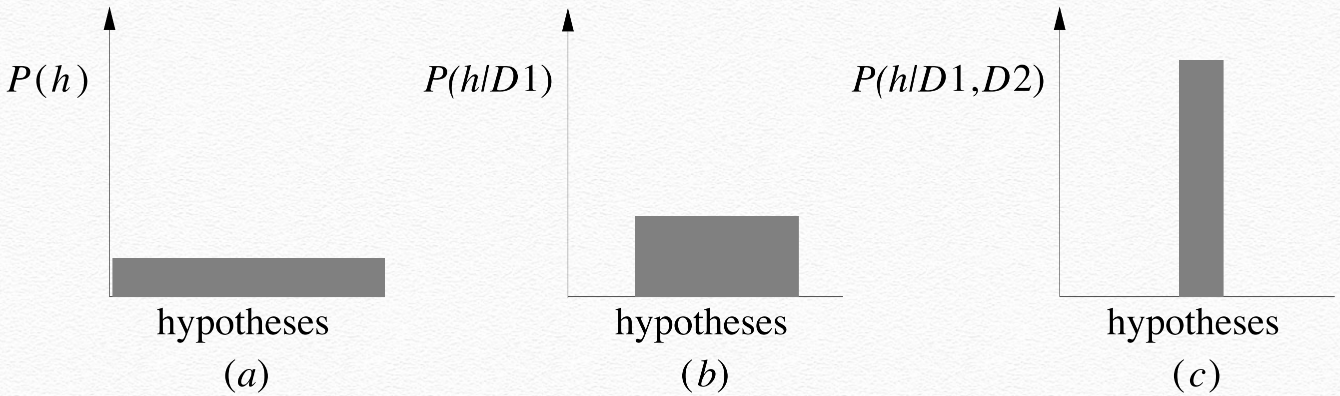
$$P(h) = \frac{1}{|H|} \text{ for all } h \in H.$$

Then,

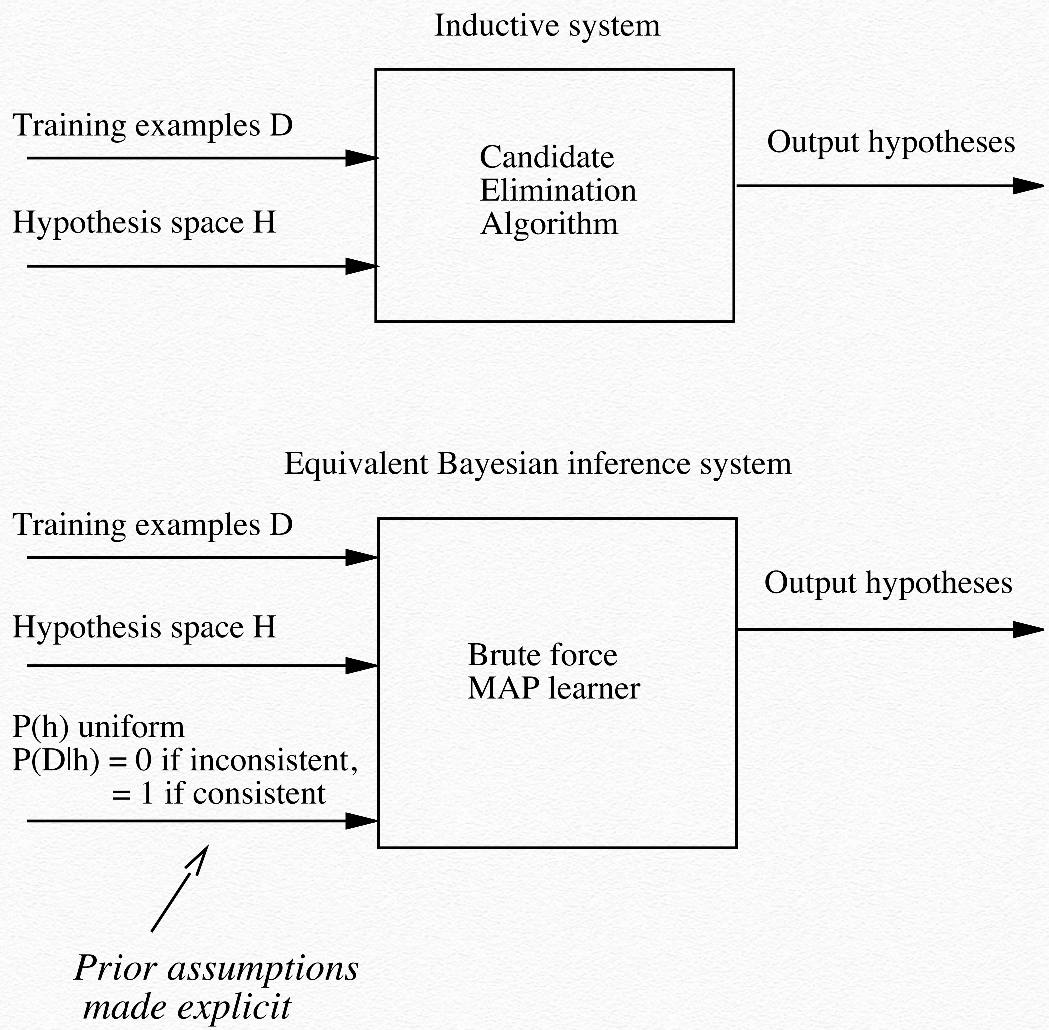
$$P(h|D) = \begin{cases} \dfrac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D, \\ 0 & \text{otherwise.} \end{cases}$$

Every consistent hypothesis is a MAP hypothesis!

Belief Update

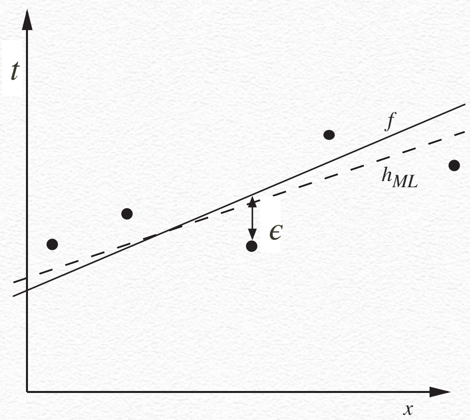


Characterizing Learning Algorithms by Equivalent MAP Learners



Learning a Continuous-Valued Function

Consider any real-valued target function f & training examples $D = \{\langle \mathbf{x}_d, t_d \rangle\}$ where t_d is a noisy target output for training example d:



- $t_d = f(\mathbf{x}_d) + \epsilon_d$
- ϵ_d is a random noise variable drawn independently for each $\mathbf{x}_d$ according to $\epsilon_d \sim \mathcal{N}(0, \sigma^2)$

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

$$h_{\mathrm{ML}} = \arg\min_{h \in H} \frac{1}{2} \sum_{d \in D} (t_d - h(\mathbf{x}_d))^2 .$$

Learning a Continuous-Valued Function

$$
\begin{aligned}
h_{\mathrm{ML}} &= \operatorname*{arg\,max}_{h \in H} p(D|h) \\
&= \operatorname*{arg\,max}_{h \in H} \prod_{d \in D} p(t_d | h, \mathbf{x}_d) \\
&= \operatorname*{arg\,max}_{h \in H} \prod_{d \in D} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t_d - h(\mathbf{x}_d))^2}{2\sigma^2} \right) \\
&= \operatorname*{arg\,max}_{h \in H} \sum_{d \in D} \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(t_d - h(\mathbf{x}_d))^2}{2\sigma^2} \\
&= \operatorname*{arg\,max}_{h \in H} \sum_{d \in D} -\frac{(t_d - h(\mathbf{x}_d))^2}{2\sigma^2} \\
&= \operatorname*{arg\,max}_{h \in H} \frac{1}{2} \sum_{d \in D} -(t_d - h(\mathbf{x}_d))^2 \\
&= \operatorname*{arg\,min}_{h \in H} \frac{1}{2} \sum_{d \in D} (t_d - h(\mathbf{x}_d))^2
\end{aligned}
$$