

Bayesian Learning

Prof. Dan A. Simovici

UMB

Suppose that (Ω, \mathcal{E}, P) is a probability space, \mathcal{E} is a family of subsets of Ω known as **events**, and P is a probability. The elements of Ω are **elementary events**.

If B is an event such that $P(B) > 0$ one can define the **probability of an event A conditioned on B** as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Note that if A, B are independent events, then $P(A|B) = P(A)$.

The Characteristic Function of an Event

If A is an event, then the function $1_A : \Omega \rightarrow \{0, 1\}$ defined by

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise,} \end{cases}$$

is a random variable,

$$1_A : \begin{pmatrix} 0 & 1 \\ 1 - P(A) & P(A) \end{pmatrix}$$

Note that $E(1_A) = P(A)$ and $\text{var}(1_A) = P(A)(1 - P(A))$.

Recapitulation of Conditional Probabilities

- The **product rule** or the **Bayes theorem**:

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

- The **sum rule**:

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

- The **total probability rule**: if A_1, \dots, A_n are mutually exclusive and $\sum_{i=1}^n P(A_i) = 1$, then

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

In ML we are often interested in determining the best hypothesis from some space H given the observed data S .

“Best” means in this context, the most probable hypothesis given

- the data S , and
- any initial knowledge of **prior** probabilities of hypotheses in H .

- “Prior probabilities” (or a priori probabilities) mean probabilities of hypotheses **before** seeing the data S .
- “Posterior probabilities” mean probabilities of hypotheses **after** seeing the data S .

If no prior knowledge exist all hypotheses have the same probability. In ML we are interested to compute $P(h|S)$ that h holds given the observed training data S .

Bayes' Theorem in ML

For a sample S and a hypothesis h we have

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

Note that:

- $P(h|S)$ increases with $P(h)$ and with $P(S|h)$.
- $P(h|S)$ decreases with $P(S)$ because the more probable is that S will be observed independent of h , the less evidence S provides for h .

Learning Scenario

Consider some set of candidate hypotheses H and seek the most probable hypothesis given the observed data S .

Any such maximally probable hypothesis is called a **maximum a posteriori** hypothesis, **MAP**.

h_{MAP} is

$$\begin{aligned} h_{MAP} &= \operatorname{argmax}_{h \in H} P(h|S) \\ &= \operatorname{argmax}_{h \in H} \frac{P(S|h)P(h)}{P(S)} \\ &= \operatorname{argmax}_{h \in H} P(S|h)P(h) \end{aligned}$$

because $P(S)$ is a constant.

Maximum Likelihood Hypothesis

In some cases we assume that every hypothesis of H is **apriori equally probable**, that is, $P(h_i) = P(h_j)$ for all $h_i, h_j \in H$.

Now,

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

$P(S|h)$ is known as the **likelihood** of S given h .

Example

A medical diagnosis problem:

The hypothesis space contains two hypotheses:

- h_0 : patient has no cancer;
- h_1 : patient has cancer.

An imperfect diagnosis test that has two outcomes; \oplus and \ominus .

$$\begin{aligned} P(\oplus|h_1) &= 0.98 & P(\oplus|h_0) &= 0.03 \\ P(\ominus|h_1) &= 0.02 & P(\ominus|h_0) &= 0.97 \end{aligned}$$

Prior knowledge: Only 0.08% of population has cancer; 99.2% does not.

Example (cont'd)

The test returns \oplus . Should we conclude that the patient has cancer?

The **MAP** hypothesis is obtained as

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

$$P(\oplus|h_1)P(h_1) = 0.98 * 0.008 = 0.0078,$$

$$P(\oplus|h_0)P(h_0) = 0.03 * 0.992 = 0.0298.$$

The **MAP** hypothesis is h_0 ; the patient has no cancer.

Brute-Force Bayes Concept Learning

- For each hypothesis $h \in H$ calculate the posterior probability:

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

- Output the hypothesis h_{MAP} with

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

Assumption for the Brute-Force Bayes Concept Learning:

- Training data is $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, where $y_i = f(x_i)$ for $1 \leq i \leq m$ and it is noise-free.
- The target hypothesis is contained in H .
- We have no apriori reason to believe that any hypothesis is more probable than the other

Consequences

- $P(h) = \frac{1}{|H|};$



$$P(S|h) = \begin{cases} 1 & \text{if } y_i = h(x_i) \text{ for } 1 \leq i \leq m \\ 0 & \text{otherwise;} \end{cases}$$

The probability of S given h is 1 if S is consistent with h and 0 otherwise.

Let $VS_{H,S}$ be the subset of hypotheses of H that is consistent with S .

- If S is inconsistent with h then $P(h|S) = \frac{0 \cdot P(h)}{P(S)} = 0$.
- If S is consistent with h then

$$P(h|S) = \frac{1 \cdot \frac{1}{|H|}}{P(S)} = \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,S}|}{|H|}} = \frac{1}{|VS_{H,S}|}$$

Since the hypotheses are mutually exclusive (that is, $P(h_i \wedge h_j) = 0$ if $i \neq j$), by the total probability law:

$$\begin{aligned}
 P(S) &= \sum_{h_i \in H} P(S|h_i)P(h_i) \\
 &= \sum_{h \in VS_{H,S}} 1 \cdot \frac{1}{|H|} + \sum_{h \notin VS_{H,S}} 0 \cdot \frac{1}{|H|} \\
 &= \sum_{h \in VS_{H,S}} 1 \cdot \frac{1}{|H|} = \frac{|VS_{H,S}|}{|H|}.
 \end{aligned}$$

Note that under this setting every consistent hypothesis is a MAP hypothesis.