BASIC PROBABILITIES

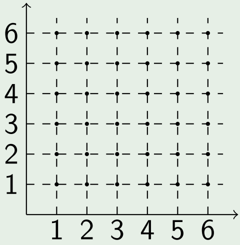
Prof. Dan A. Simovici

UMB

Suppose that $(\Omega, \mathcal{E}, P)$ 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.
In many cases, $\mathcal{E}$ consists of all subsets of Ω, and we will make this assumption unless a special statement says otherwise.

Example

Roling two dice is described by a finite probability space that consists of 36 elementary events: $(1,1),(1,2),\ldots,(6,6)$.



An event in the previous example is a subset of $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$, that is, a subset of the set of pairs $\{(u, v) \mid 1 \leqslant u \leqslant 6, 1 \leqslant v \leqslant 6\}$.

Example

- throws that have the same number of both dice:

$$S = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$$

- throws such that the sum of the numbers is greater than 8:

$$B = \{(2,6), (3,5), (3,6), (4,4), (4,5), (4,6), (5,3), (5,4), (5,6), (6,6)$$

Note that Ω consists of 36 elementary events and there are $2^{36} \approx 10^{12}$ events in this very simple probability space

Probability of an event V in this context is the number $P(V)$ given by

$$P(V) = \frac{|V|}{|\Omega|}.$$

Example

We have

$$P(S) = \frac{6}{36} = \frac{1}{6},$$

and

$$P(B) = \frac{18}{36} = \frac{1}{2}$$

Informally, Borel sets of $\mathbb{R}$ are the sets that can be constructed from open or closed sets by repeatedly taking countable unions and intersections.

Definition

Let $(\Omega, \mathcal{E}, P)$ be a probability space. A function $X : \Omega \longrightarrow \mathbb{R}$ is a random variable if $X^{-1}(U) \in \mathcal{E}$ for every Borel subset of $\mathbb{R}$.

Definition

A simple random variable is defined by a table:

$$X : \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix},$$

where $x_1, \ldots, x_n$ are the values assumed by X and $p_i = P(X = x_i)$ for $1 \leqslant i \leqslant n$. We always have $p_1 + \cdots + p_n = 1$.

The expected value of X is

$$E[X] = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n.$$

Example

A random variable X whose distribution is:

$$X : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix},$$

where $p + q = 1$ is said to have a Bernoulli distribution with parameter p. Note that

$$E[X] = p \text{ and } \mathrm{var}(X) = pq.$$

Example

Let $p, q \in [0,1]$ be two numbers such that $p+q=1$. Consider the random variable defined by

$$X : \begin{pmatrix} 0 & 1 & \cdots & k & \cdots & n \\ q^n & \binom{n}{1} q^{n-1} p & \cdots & \binom{n}{k} q^{n-k} p^k & \cdots & p^n \end{pmatrix},$$

We refer to a random variable with this distribution as a *binomial random variable*. Note that

$$q^n + \binom{n}{1} q^{n-1} p + \cdots + \binom{n}{k} q^{n-k} p^k + \cdots + p^n = (q+p)^n = 1.$$

Example cont'd

The expectation of a binomial variable is

$$E[X] = np.$$

The variance of a random variable X is

$$\text{var}(X) = E[(X - E(X))^2] = E[X^2] - (E[X])^2.$$

In the case of a binomial variable the variance is $\text{var}(X) = npq$.

The Characteristic Function of an Event

If A is an event, then the function $1_A : \Omega \longrightarrow \{0,1\}$ 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 $\operatorname{var}(1_A) = P(A)(1 - P(A))$.

The event $A \wedge B$ takes place when both A and B occur; the event $A \vee B$ takes place when at least one of A and B occur.

Example

The event $S \wedge B$ takes place when the result of throwing the dice results in a pair of numbers (n, n) whose sum is greater than 8 and consists of the pairs:

$$(4, 4), (5, 5), (6, 6)$$

Therefore, $P(S \wedge B) = \frac{3}{36} = \frac{1}{12}$.

Definition

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

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

Example

The probability of the event S conditioned on B is

$$P(S|B) = \frac{P(S \wedge B)}{P(B)} = \frac{\frac{1}{12}}{\frac{1}{2}} = \frac{1}{6},$$

and

$$P(B|S) = \frac{P(S \wedge B)}{P(S)} = \frac{\frac{1}{12}}{\frac{1}{6}} = \frac{1}{2}.$$

Definition

Two events A, B are independent if $P(A \wedge B) = P(A)P(B)$.

If A, B are independent events, then

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

and

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

Note that B and S are independent events because $P(B \wedge S) = \frac{1}{12} = P(B)P(S)$.

- 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 $A_1, \ldots, A_n$ 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 probabile hypothesis is called a maximum a posteriori hypothesis, MAP.
h_{MAP} is

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

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,

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

$P(S|h)$ 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 $\ominus$.

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

Prior knowlege: 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

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

$$\begin{aligned} P(\oplus|h_1)P(h_1) &= 0.98 * 0.008 = 0.0078, \\ P(\oplus|h_0)P(h_0) &= 0.03 * 0.992 = 0.0298. \end{aligned}$$

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 probablity:

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

- Output the hypothesis h_{MAP} with

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

Assumption for the Brute-Force Bayes Concept Learning:

- Training data is $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $y_i = f(x_i)$ for $1 \leqslant i \leqslant m$ 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

- $P(h) = \frac{1}{|H|}$;
- The probability of S given h is 1 if S is consistent with h and 0 otherwise:

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

Let $\mathrm{VS}_{H,S}$ 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.