Statistical Machine Learning
Part 5

Horacio Gómez-Acevedo
Department of Biomedical Informatics

February 15, 2021

Probability refresher

Formal definition. A probability space is a triplet $(A, \mathfrak{A}, P)$, where A is a (non-empty) set, $\mathfrak{A}$ contains subsets of A called *events* (those events form a σ-algebra of A) and a function $P : \mathfrak{A} \to [0, 1]$ (the so-called *probability function*).

Example. Let's consider the experiment of rolling a die. A will be the outcomes $A = \{1, 2, \ldots, 6\}$ and $\mathfrak{A}$ will be the subsets of A. For instance, the event E consisting of the even number output is $\{2, 4, 6\}$. The probability function will be defined as

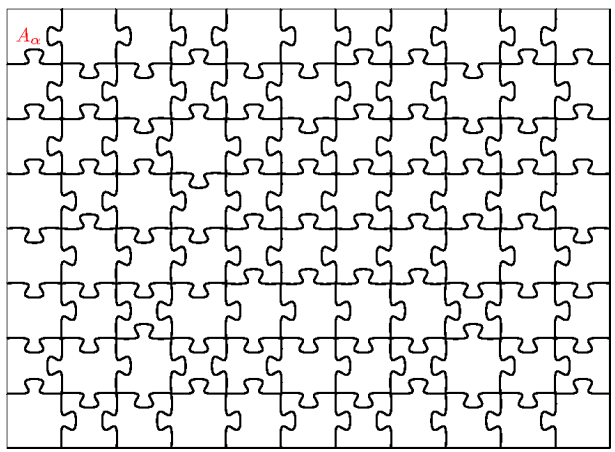
$$P(E) = \frac{\text{number of elements of } E}{\text{number of elements in } A} = \frac{3}{6}$$

Complete system of events

We called a *complete system of events* a collection $\{A_\alpha\}_{\alpha \in I}$ that satisfy:

- $A_\alpha \cap A_\beta = \emptyset$
- $\cup_{\alpha \in I} A_\alpha = A$.

A jigsaw puzzle is a simple visualization of this. A is the whole rectangle and each piece is one of the A_α.



Conditional Probabilities

If B and C are events and $P(C) > 0$, we say that the probability that the event B has occurred given that C has occurred is given by

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

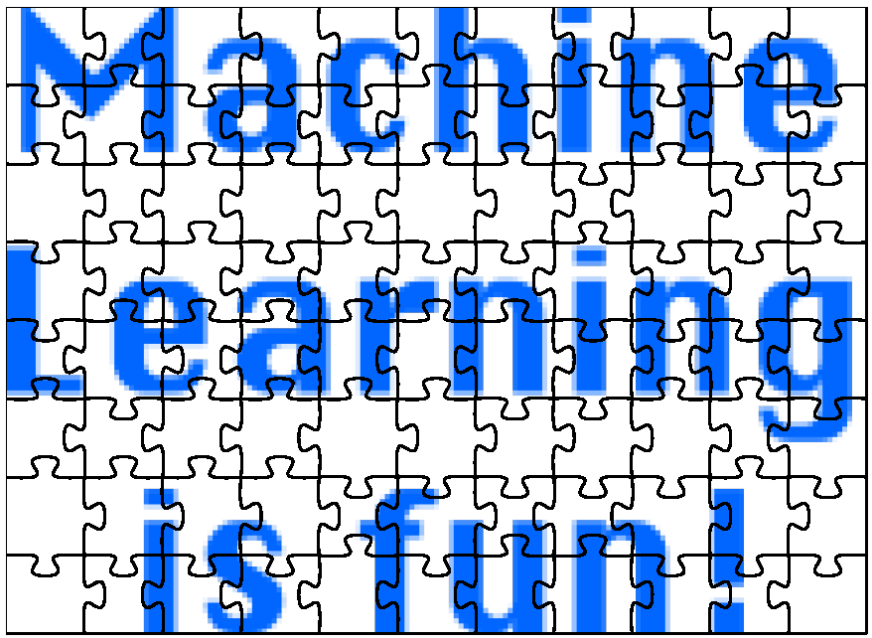
Total Probability Theorem. If we have a complete system of events $\{A_\alpha\}_\alpha \in I$ with $P(A_\alpha) > 0$ for each $\alpha \in I$, then for an arbitrary event E it holds

$$P(E) = \sum_\alpha P(E|A_\alpha)P(A_\alpha)$$

It is clear that every term $P(E|A_\alpha)P(A_\alpha) = P(A_\alpha \cap E)$. Thus, the probability of E is the sum of every part of E in each A_α.

Total Probability Theorem

Suppose our space is again the jigsaw puzzle rectangle and each piece is A_α as before. The probability measure will be the area cover by an event. In this case our "event" represents the blue letters. Visually the probability of the whole event will be the sum of the blue area in each of the pieces, thus the name "total probability"



Bayes Theorem

Following the same framework as before, we have one of the most useful results in probability

Bayes theorem. If E is an event with $P(E) > 0$, then

$$P(A_\beta|E) = \frac{P(E|A_\beta)P(A_\beta)}{\sum_{\alpha \in I} P(E|A_\alpha)P(A_\alpha)} \tag{1}$$

Note that the numerator of (1) is $P(E \cap A_\beta)$ whereas the denominator is $P(E)$ according to the total probability theorem!

Bayesian Perspective

Unlike the traditional (frequentist) inference, the Bayesian approach tries to determine conclusions about a parameter θ or unobserved data $\overline{y}$ as probability statements. For example, given the observed data y, Bayesian statements are based on the conditionals such as $P(\overline{y}|y)$ for *new data* $\overline{y}$, or $P(\theta|y)$ for a parameter θ. So following our Bayes theorem,

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)} \tag{2}$$

where $P(y) = \sum_\theta P(\theta)P(y|\theta)$ (or in the continuous case $P(y) = \int P(\theta)P(y|\theta)d\theta$).
The expression $P(\theta)$ is called the **prior**, whereas $P(\theta|y)$ is called **posterior**, and $P(y|\theta)/P(y)$ is called **likelihood**.
Note that in (2) the denominator does not depend on θ. So, the unnormalized posterior

$$P(\theta|y) \propto P(\theta) \cdot P(y|\theta)$$

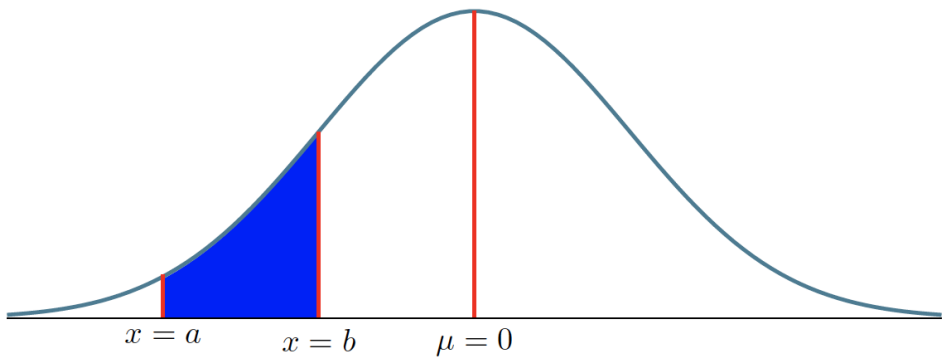
Density functions

When we have a continuous random variable X, the probability is calculated by integrals instead of sums. The **probability density function** $f(x)$ describes the probability of an event.
For instance, let's suppose we have a random variable X distributed as a $N(0, 1)$. The density function in this case is defined as

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

Then, the probability that our random variable lies in the interval $[a, b]$ is given by

$$P(a \leq X \leq b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx$$



Linear Discriminant Analysis

We need to classify an observation into one of K distinct classes ($K \geq 2$). The response variable Y can take on K possibilities. Let π_k denote the *prior* probability that a randomly chose observation comes from the kth class. Let $f_k(X) = P(X = x|Y = k)$ denote the *density function* of X for an observation that comes from the kth class. We know by the Bayes' theorem

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

Let's simplify *posterior probability* , so $p_k(X) = P(Y = k|X)$
What have we achieved thus far?

- We have a potential classifier