MO435 — Probabilistic Machine Learning
INSTITUTO DE COMPUTAÇÃO — UNICAMP

Probability Theory

Lesson No. 1

João Victor da Silva Guerra

1 Introduction

Given some data from a particular collection of observations, which can be weight of individuals within a group of people, number of heads in a series of coin tosses and number of monkeys in a zoo. Such observations are called samples and can be analyzed by two different statistical methods: descriptive and inferential statistics. The first method provide a detailed description of a sample using, for example, mean and variation. The inferential method use a random sample of data, that are subject of a random variation, to describe and make generalizations about the population [1].

There are two general interpretations of inferential statistics: frequentist and Bayesian. Both approaches provide tools to evaluate about competing hypothsesis. The frequentist interpretation argues that probabilities represent a long term frequencies of events. On the other hand, the Bayesian interpretation use propabilities to quantify uncertainty about an event [2]. When comparing frequentist and Bayesian statistics, the last can be applied to model uncertainty about events that do not have long run frequencies.

In machine learning applications, the idea of repeated trials does not make sense; however, the Bayesian approach is valid since we have uncertainty about our predictions and want to make generalizations about our population from a sample.

2 Core concepts

In this section, we briefly review core concepts of probability theory, which are useful for machine learning applications.

2.1 Random variables

A random variable, usually denoted as X, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of variables: discrete and continuous. A discrete variable is a quantitative random variable that can assume a countable number of values, e.g., number of students in a school and number of heads when flipping coins. A continuous variable is a quantitative random variable that can assume an uncountable number of values, e.g., height of students, BMI of individuals in a population and time required to run a mile. However, most of the fundamental rules are interchangable between those types.

The expression $p(A)$ denotes the probability of an event A, in which must satisfy that $0 \leq p(A) \leq 1$. With $p(A) = 0$, the event A will not occur, and with $p(A) = 1$, the event will not occur.

2.2 Fundamental rules

In this topic, we review the fundamental rules of probability theory. Given two events, A and B, we will define the probabilities of two events, the Bayes' theorem, and independence.

2.2.1 Probability addition rule

The union of sets is the set the set that contains all the elements of each of the composing sets. The definition of the union is $A \cup B = \{x : x \in A \vee x \in B\}$. The probability of the union of events is defined as follows:

$$p(A \cup B) = p(A) + p(B) + p(A \cap B) \tag{1}$$

where $p(A \cap B)$ is the joint probability.

If events A and B are mutually exclusive, the probability is:

$$p(A \cup B) = p(A) + p(B) \tag{2}$$

2.2.2 Joint probability

The intersection of sets is the set of elements that belongs to all of the sets. The definition of the intersection is $A \cap B = \{x : x \in A \land x \in B\}$. The joint probability, also known as product rule, of events is defined as follows:

$$p(A,B) = p(A \cap B) = p(A|B) \times p(B) \tag{3}$$

where $p(A|B)$ is the conditional probability of event A, given event B is true.

2.2.3 Marginal probability

Given the joint probabilities of events A and B, we may be interested in the probability of just one event, irrespective of the outcome of another event. The marginal distribution, also known as sum rule or rule of total probability, of an event is defined as follows:

$$p(A) = \sum_{b}^{B} p(A,B) = \sum_{b}^{B} p(A|B=b) \times p(B=b) \tag{4}$$

and $p(B)$ can be defined similarly.

2.2.4 Conditional probability

Given the occurence of an event (for example, event A, so $p(A) > 0$), the conditional probability of another event (for example, event B) is defined as follows:

$$p(B|A) = \frac{p(B,A)}{p(A)} \tag{5}$$

2.2.5 Bayes' theorem

Bayes' theorem, also known as Bayes' rule, relates the probability of an "direct" conditional event to the "inverse" conditional event and the priors of each event [3]. The defintion is as follows:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(B|A) \times p(A)}{p(B)} \tag{6}$$

However, combining the sum rule described above, the Bayes' theorem can be rewritten as:

$$p(A|B) = \frac{p(B|A) \times p(A)}{\sum_{b}^{B} p(A|B=b)p(B=b)} \tag{7}$$

2.2.6 Independence and Conditional independence

If two events, such as events A and B, are unconditionally independent, the joint probability ($p(A,B)$) is equivalent to the product of the marginal probability of each as follows:

$$A \perp B \iff p(A \cap B) = p(A) \times p(B) \tag{8}$$

Since variables usually influence each other, the above case is rare. However, the influence of two events may be mediated by a third event, instead of being direct. Then, two events, such as A and B, are conditionally independent, given a thrid event, such as C (so $p(C) > 0$), if and only if the joint conditional probability ($p(A,B|C)$) is equivalent to the product of the conditional marginal probabilities as follows:

$$A \perp B|C \iff p(A \cap B|C) = p(A|C) \times p(B|C) \tag{9}$$

2.3 Discrete random variables

Suppose that discrete random variable X may take on any value from a finite or countably infinite set χ. The probability, also knwon as **probability mass function (pmf)**, of a event $X = x_\text{i}$ is defined as $p(X = x_\text{i})$ or $p(x_\text{i})$, which also satisfies the following property:

$$\sum_{x_\text{i} \in \chi} p(x_\text{i}) = 1 \tag{10}$$

Discrete and continuous random variables have a **cumulative distribution function (cdf)**, which is the probability of $X \leq x$ and defined as $F(X = x)$ or $F(x)$. Then, for a discrete random variable, the cdf of X will be discontinuous at points x_i and defined as follows:

$$F(x) = p(X \leq x) = \sum_{x_\text{i} \leq x} p(X = x_\text{i}) = \sum_{x_\text{i} \leq x} p(x_\text{i}) \tag{11}$$

2.4 Continuous random variables

On the other hand, a continuous random variable is not defined at specific values but over an interval of values. Hence, in this topic, we extend probability to cover continuous quantities.

Suppose a random variable X may take all values over an interval $a \leq X \leq b$. Given the events $A = (X \leq a)$, $B = (X \leq b)$ and $C = (a < X \leq b)$. Assuming A and C are mutually exclusive events, we note that $B = A \cup C$. The probability addition rule yields

$$p(B) = p(A) + p(C) \tag{12}$$

and rearranging

$$p(C) = p(B) - p(A) \tag{13}$$

Using the previously defined cdf notation $(F(x))$, we can rewrite the previous equation as:

$$P(a < x \leq b) = F(b) - F(a) \tag{14}$$

Given the cdf $(F(x))$ is the area under the **probability density function (pdf)**, we define the pdf $(f(x))$ as follows:

$$f(x) = \frac{d}{dx} F(x) \tag{15}$$

Hence, combining equations above, the probability of a continuous random variable in a finite interval is computed as:

$$P(a < x \leq b) = F(b) - F(a) = \int_a^b f(x) dx \tag{16}$$

References

[1] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[2] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, pp. 27–33.

[3] J. Joyce, “Bayes’ theorem,” in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Spring 2019, Metaphysics Research Lab, Stanford University, 2019.