Probability for Machine Learning

StartOnAI Programming and Mathematics Tutorial 4

1 Topic Summary

Probability is the study of uncertainty, and is used in both data analysis and machine learning models. It gives us the language and tools we need to quantify the uncertainty of events; Machine learning is centered around developing predictive models from uncertain (imperfect/incomplete) data. Thus, the notion of uncertainty is essential to the field of machine learning.

This tutorial will provide you with only the essentials of probability necessary to embark further on your machine learning journey through our higher-level tutorials, provided that you have a sufficient mathematical background (up to and including calculus). Rereading chapters is also advised due to the concise nature of this tutorial.

2 Main Concepts

2.1 Construction of a Probability Space

2.1.1 Probability Spaces

The idea of a *probability space* allows us to quantify the idea of a probability, though we will usually not work directly with probability spaces. Instead, we usually work with *random variables*, which are functions that map outcomes of random experiments (such as choosing a card from a deck) to a set of properties that we are interested in. Furthermore, we can also have a distribution or law associated with a random variable (expanded upon in the next subchapter).

Modern probability is based on a set of axioms proposed by Andrey Kolmogorov, who made significant contributions to probability theory, that introduce the three concepts of *sample space*, *event space*, and *probability measure*:

- Sample Space (Ω):

  The sample space is the set of all possible outcomes of the experiment (e.g. the sample space of a coin toss would be {hh, ht, th, tt}, where "h" denotes heads and "t" denotes tails)

- Event Space ($\mathcal{A}$)

  Any subset (set contained in the sample) of the sample space is considered to be an *event*. The event space is the set of all events, and is also known as the *power set* of Ω.

- Probability (P)

  With each event $E \in \mathcal{A}$ (each event in the event space), we associate a number $P(A)$ that represents the probability (or likelihood) that the event will occur. $P(A)$ is called the probability of A.

The probability of a single event E must lie in the interval $[0, 1]$, and the total probability over all outcomes in the sample space Ω must be 1 ($P(\Omega) = 1$). A probability space is defined by the triple $(\Omega, \mathcal{A}, P)$.

2.1.2 Random Variables

In machine learning, we often avoid explicitly referring to the probability space, but instead refer to probabilities on quantities of interest, which we denote by $\mathcal{T}$. In this tutorial, we will refer to $\mathcal{T}$ as the *target space* and refer to elements of $\mathcal{T}$ as states. We introduce a function $X : \Omega \to \mathcal{T}$ that takes an element of Ω (an outcome) and returns a particular quantity of interest x, a value in $\mathcal{T}$, which we refer to as a *random variable*. (Note that X is neither random nor a variable; it is a function)

For instance, consider the process of throwing a dice. Let X be a random variable that depends on the outcome of the throw. A natural choice for X would be to map the outcome i to the value i (that is, mapping the event of throwing an "one" to the value of 1).

In a sense, random variables allow us to abstract away from the formal notion of an event space, as we can define

random variables that capture the appropriate events. Note that in the above example, we could have defined a random variable that takes on value 1 if outcome i is odd and 0 otherwise. These types of binary random variables are very common in practice, and are known as *indicator variables*.

From here on out, we will be discussing probability with respect to random variables, as random variables will allow us to provide a more uniform treatment of probability theory. Notation wise, the probability of a random variable X taking on the value of a will be denoted as one of the following:

$$P(X = a) \text{ or } P_X(a) \tag{1}$$

The range of a random variable X will be denoted with $Val(X)$.

2.1.3 Distributions, Joint Distributions, and Marginal Distributions

The *distribution* of a random variable refers to the probability of a random variable taking on certain values. For instance, let random variable X be defined on the outcome space Ω of a dice throw. If the dice throw is fair, the distribution of X will be

$$P_X(1) = P_X(2) = \cdots = P_X(6) = \frac{1}{6} \tag{2}$$

Notation wise, we will use $P(X)$ to denote the distribution of random variable X.

We call a distribution a *joint distribution* when the distribution involves more than one variable, as the probability is determined jointly by the involved variables.

For instance, let X be a random variable defined on the outcome space of a dice throw, and let Y be a random variable defined on the outcome space of a coin toss. The joint distribution of these two random variables is denoted by $P(X, Y)$, with the probability of X taking value a and Y taking value b being denoted by $P_{X,Y}(a, b)$. Each of the possible outcome combinations ((1, 0), (1, 1), (2, 0), ...) compose the joint distribution of X and Y.

Given a joint distribution over random variables X and Y, the *marginal distribution* refers to the probability distribution of a random variable on its own. To find the marginal distribution of a random variable, we "sum out" all of the other variables from the distribution:

$$P(X) = \sum_{b \in Val(Y)} P(X, Y = b) \tag{3}$$

The name of marginal distribution stems from the fact that if we add up all the entries of a row/column of a joint distribution and write the answer at the end (margin) of it, that will be the probability of the random variable taking on that value.

2.1.4 Conditional Distributions

Conditional distributions specify the distribution of a random variable (X in this case) when the value of another random variable (Y in this case) is known/another event is known to be true, and is defined as:

$$P(X = a|Y = b) = \frac{P(X = a|Y = b)}{P(Y = b)} \tag{4}$$

To conceptualize this in an example, suppose that we know that the result of a dice throw was odd. Let X be the random variable of the dice throw, and Y be an indicator variable that takes on the value of 1 if the dice throw turns up odd. Our probability would be expressed as

$$P(X = 1|Y = 1) = \frac{P(X = 1|Y = 1)}{P(Y = 1)} = \frac{1/6}{1/2} = \frac{1}{3} \tag{5}$$

In the case of additional random variables, the conditional probability extends as shown:

$$P(X = a|Y = b, Z = c) = \frac{P(X = a|Y = b, Z = c)}{P(Y = b, Z = c)} \tag{6}$$

Notation wise, $P(X|Y = b)$ denotes the distribution of random variable X when $Y = b$, and $P(X|Y)$ denotes a set of distributions of X, one for each of the different values that Y can take.

2.1.5 Independence

When we say that a random variable X is *independent* from another random variable Y, we mean that the distribution of X does not change upon learning the value of Y:

$$P(X) = P(X|Y) \tag{7}$$

Likewise, if X is independent of Y, then Y is independent of X. We express this with the notation $X \perp Y$. An equivalent statement about the independence of random variables X and Y is:

$$P(X, Y) = P(X)P(Y) \tag{8}$$

Another important concept to note is *conditional independence*. With conditional independence, if we know the value of one/a set of random variable(s), then we know that some other random variables will be independent of each other. We say that X and Y are *conditionally independent* given Z if

$$P(X, Y|Z) = P(X|Z)P(Y|Z) \tag{9}$$

For instance, if we were to train a learning algorithm to classify emails as spam or not spam, we could assume that the probability of a word x appearing in the email is conditionally independent of a word y appearing given whether the email is spam or not.

2.1.6 Chain Rule and Bayes' Rule

The *chain rule* is used to evaluate the joint probability of some random variables, and is especially useful when there are (conditional) independence across random variables. The chain rule is given by:

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots P(X_n|X_1, X_2, \ldots X_{n-1}) \tag{10}$$

Bayes' rule allows us to compute the conditional probability $P(X|Y)$ from $P(Y|X)$, which essentially "inverts" the conditions, and is given by:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{11}$$

If $P(Y)$ is not given, we can apply the *law of total probability* to find $P(Y)$:

$$P(Y) = \sum_{a \in Val(X)} P(X = a, Y) = \sum_{a \in Val(X)} P(Y|X = a)P(X = a) \tag{12}$$

2.2 Discrete and Continuous Probabilities

2.2.1 Discrete Distributions: Probability Mass Functions

By a *discrete distribution*, we mean that the random variable of the underlying distribution can take on only *finitely many* different values/the sample space is finite.

To define a discrete distribution, we can simply enumerate the probability of the random variable taking on each of the possible values, which is known as the *probability mass function*, as it divides up a unit mass (the total probability) and places them on the different values a random variable can take, and can also be extended analogously to joint distributions and conditional distributions.

Let X be a discrete random variable with the finite sample space $\Omega = x_1, x_2, \ldots$. The following would be considered the probability mass function (p.m.f) of X:

$$P_X(x_k) = P(X = x_k), \text{for } k = 1, 2, 3, \ldots \tag{13}$$

For instance, let X be defined as the number of heads observed after flipping a coin twice. Our sample space is given by $\Omega = \{HH, HT, TH, TT\}$; Therefore, all of the possible outputs of random variable X are represented by $R_X = \{0, 1, 2\}$. Since this is a finite set, X is a discrete random variable. Using this information, we can find the probability mass function as shown:

$$P_X(0) = P(X = 0) = P(TT) = \frac{1}{4} \tag{14}$$

$$P_X(1) = P(X = 1) = P(\{HT, TH\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \tag{15}$$

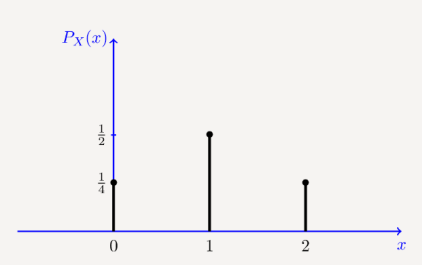
$$P_X(2) = P(X = 2) = P(HH) = \frac{1}{4} \tag{16}$$

Thus, we can, in general, write

$$P_X(x) = P(X = x) \tag{17}$$

We can also plot this PMF to better visualize it:

Figure 1: Visualization of the PMF of random variable X representing the number of heads observed after two coin flips (Pishro-Nik)



2.2.2 Continuous Distributions: Probability Density Functions

By a *continuous distribution*, we mean that the random variable of the underlying distribution can take on *infinitely many* different values/that the sample space is infinite. Defining a continuous distribution is tricky, as we cannot assign a non-zero value to a value due to the infinitely large sample space.

Thus, we will use a *probability density function* (PDF) f to define a continuous distribution. f is a non-negative, integrable function such that

$$\int_{Val(X)} f(x)dx = 1 \tag{18}$$

where X is a continuous random variable.

The probability of a random variable X distributed according to a PDF f is computed as follows:

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{19}$$

Note that this implies that the probability of a continuously distributed random variable taking on any given single value is 0.

Another concept to note is the *cumulative distribution function*, which is a function F that gives the probability of a random variable being smaller than some value:

$$F(b) = P(X \leq b) = \int_{-\infty}^b f(x)dx \tag{20}$$

where f is a probability density function related to the overlying cumulative distribution function F.

2.3 Expectation and Variance

2.3.1 Expectation

One common operation that we perform on a random variable is to compute its *expectation* (also known as its mean, expected value, or first moment). The expectation of a random variable X, denoted by $E(X)$ is given by

$$E(X) = \sum_{a \in Val(X)} aP(X = a) \text{ or } E(X) = \int_{a \in Val(X)} x\ f(x)dx \tag{21}$$

We may also be interested in computing the expected value of some function Y of a random variable X, which can be done by defining Y as $f(X)$ and computing the expected value of Y.

When working with the sums of random variables, one of the most important rules is the *linearity of expectations*. Let $X_1, X_2, \ldots, X_n$ be (possibly dependent) random variables:

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) \tag{22}$$

Likewise, if we assume X and Y are independent random variables, the product of the expections of X and Y follow the same patterns as the above:

$$E(XY) = E(X)E(Y) \tag{23}$$

2.3.2 Variance

The *variance* of a distribution is a measure of the "spread" (or "second moment") of a distribution, and is defined as:

$$Var(X) = E((X - E(X))^2) \tag{24}$$

The variance of a distribution is denoted σ^2, as we often want to find σ, the *standard deviation*. To find the variance of the random variable X, it is often easier to compute using the following formula:

$$Var(X) = E(X^2) - (E(X))^2 \tag{25}$$

Note that, unlike expectation, variance is not a linear function of a random variable X.

If random variables X and Y are independent, then

$$Var(X + Y) = Var(X)Var(Y) \tag{26}$$

Sometimes we also discuss the *covariance* of two random variables, which is a measure of how "closely related" two random variables are, and is defined by:

$$Cov(X, Y) = E((X - E(X))(Y - E(Y))) \tag{27}$$

2.4 Essential Distributions

2.4.1 Bernoulli

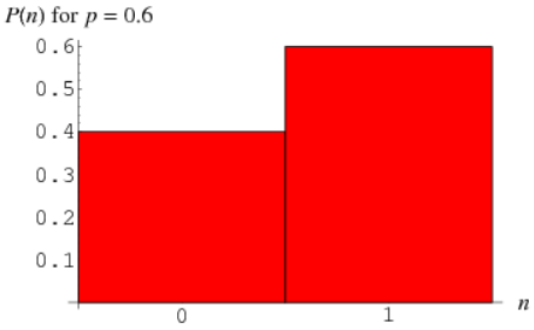
The *Bernoulli distribution* is a discrete distribution having two possible outcomes labelled $n = 0$ and $n = 1$, in which $n = 1$ occurs with probability p ($P(X = 1)$) and $n = 0$ occurs with probability $1 - p$, with $0 \leq p \leq 1$.

The probability distribution of a Bernoulli random variable X is therefore

$$P(X) = p^x(1 - p)^{1-x} \tag{28}$$

We can also visualize the Bernoulli distribution by plotting it:

Figure 2: Visualization of the Bernoulli distribution when $p = 0.6$ (Wolfram Alpha)



2.4.2 Poisson

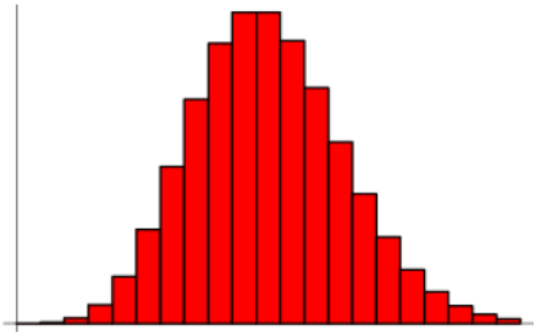
The *Poisson distribution* is a very useful distribution that deals with the arrival of events by measuring probability of the number of events happening over a fixed period of time, given a fixed average rate of occurrence, and that the events take place independently of the time since the last event. It is parametrized by the average arrival rate λ.

The probability mass function is given by

$$P(X = k) = \frac{exp(-\lambda)\lambda^k}{k!} \tag{29}$$

The mean value of a Poisson random variable is λ, and its variance is also λ. We can visualize the Poisson distribution by plotting it:

Figure 3: Visualization of a Poisson distribution (Wolfram Alpha)



2.4.3 Gaussian

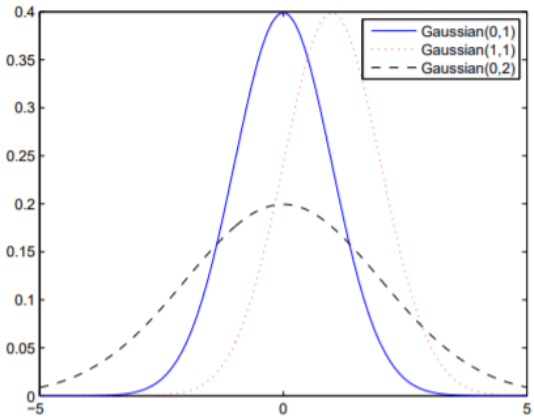
The *Gaussian distribution* (also known as the normal distribution) is the most well-studied probability distribution for continuous-valued random variables. There are many areas of machine learning that benefit from using a Gaussian distribution, including linear regression, density estimation, and reinforcement learning.

The Gaussian distribution is determined by two parameters: the mean μ and the variance σ^2. The probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{30}$$

To get a better sense of how the distribution changes with respect to the mean and the variance, we have plotted three different Gaussian distributions in Figure 4, shown below:

Figure 4: Visualization of three different Gaussian distribution, each with parameters (μ, σ^2) (Ieong)



In later tutorials, we may work with *multivariable Gaussian distributions*. A k-dimensional Gaussian distribution is parameterized by (has parameters) (μ, Σ), where μ is a k-dimensional vector of means, and Σ is a $k \times k$-dimensional *covariance matrix*. The probability density function is now defined over vectors of input, and is given by:

$$f(x) = \frac{1}{\sqrt{2\pi^k|\Sigma|}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{31}$$

Working with a multi-variate Gaussian distribution can be daunting at times, so one way we can make working with such easier is to assume that the covariances are zero when we first attempt a problem. When the covariances are zero, the determinant $|\Sigma|$ will simply be the product of the variances, and the inverse Σ^{-1} can be found by taking the inverse of the diagonal entries of Σ.

3 Applications to Machine Learning

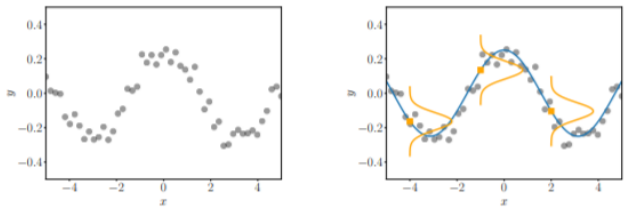
3.1 Regression

In regression, we aim to find a continuous function f that maps inputs x to corresponding function values $f(x)$. We assume that we are given a set of training inputs x_n and corresponding observations (training outputs) $y_n = f(x_n) + \epsilon$,

where ϵ is an independent and identically distributed random variable that describes measurement/observation noise.

We want to find a function that not only models the training data, but generalizes well to predicting function values that are not part of the training data. A typical example of a regression problem is shown below:

Figure 5: Visualization of a typical regression problem and one possible solution. (Deisenroth, Faisal, Ong; 2020)



Regression is a fundamental problem in machine learning, and regression problems appear in a diverse range of research areas and applications, including time-series analysis, control systems, and deep learning

Given training data, we want to infer the function that generated said data. Finding this function requires solving multiple subproblems, including choosing a model & parameterization (what model best suits this data? what parameters should we use?), finding good parameters (optimization), overfitting (does our model fit the data *too* closely?), and uncertainty modelling (how well can our model generalize to data that it's never seen before?).

To address the uncertainty modelling subproblem, we can use a probabilistic approach and explicitly model noise using a likelihood function. Within that likelihood function is ϵ, an independent, identically distributed Gaussian measurement noise (statistical noise having a PDF equal to that of the Gaussian distribution) with mean 0 and variance σ^2. If we assume we know σ^2, we can modify our likelihood function to be parametric with parameters θ, which we can find with *maximum likelihood estimation*, where we find parameters that maximize the likelihood. Intuitively, maximizing the likelihood means maximizing the predictive distribution of the training data given the model parameters.

TO find the desired parameters that maximize the likelihood, we typically perform gradient ascent (or gradient descent on the negative likelihood). In the case of linear regression we consider here, however, a closed-form solution exists, which makes iterative gradient descent unnecessary. In practice, instead of maximizing the likelihood directly, we apply the log-transformation to the likelihood function and minimize the negative log-likelihood.

3.2 Density Estimation

When we apply machine learning to data, we often aim to represent data in some form. A straightforward approach would be to just take the data points themselves as the representation of the data. However, this approach is lacking when we have a large dataset or if we are interested in representing characteristics of the data.

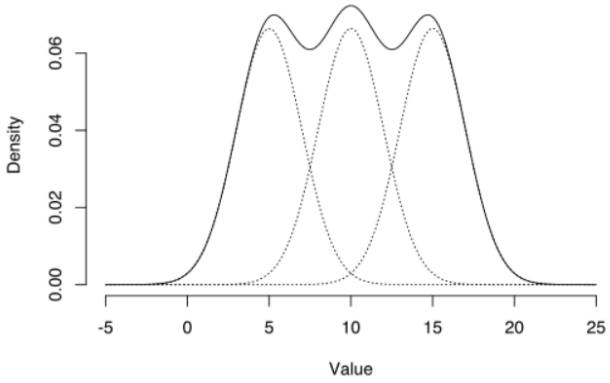
In density estimation, we represent the data compactly using a density from a parametric family (e.g. a Gaussian distribution). For instance, we may be interested in representing the data compactly via a Gaussian distribution by utilizing the mean and variance of a dataset, which can be found via maximum likelihood estimation.

In practice, the Gaussian distribution, along with many other distribution, have limited modelling capabilities. However, we can use *mixture models*, a more expressive family of distributions, for density estimation.

A *Gaussian mixture model* is a density/probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. These parameters can be optimized via the *expectation–maximization*.

Each distribution has different levels of importance, as determined by the previously mentioned parameters, and is reflected in the fitted mixture model, an example of which is shown below:

Figure 6: A Gaussian mixture model. (Glen, 2020)



4 References

Deisenroth, M., Faisal, A., & Ong, C. (2020). Probability and Distributions. In Mathematics for Machine Learning (pp. 120-151). Cambridge: Cambridge University Press. doi:10.1017/9781108679930.007

Ieong, S. (2006) *Probability Theory Review for Machine Learning.* Personal Collection of S. Ieong, Stanford University, Stanford CA.