**Chapter 3**

**PROBABILITY AND INFORMATION THEORY**

In artificial intelligence applications, we use probability theory in two major ways.

1. First, the laws of probability tell us how AI systems should reason, so we design our algorithms to compute or approximate various expressions derived using probability theory.
2. Second, we can use probability and statistics to theoretically analyze the behavior of proposed AI systems.

Another resource focused exclusively on teaching linear algebra, such as Shilov(1977).

* 1. **Why Probability?**
* In many cases, it is more practical to use a simple but uncertain rule rather than a complex but certain one. For example, the simple rule “Most birds fly” is cheap to develop and is broadly useful, while a rule of the form, “Birds fly, except for very young birds that have not yet learned to fly, sick or injured birds that have lost the ability to fly, flightless species of birds including the cassowary, ostrich and kiwi. . .” is expensive to develop, maintain and communicate, and after all this effort is still very brittle and prone to failure.
* Probability theory was originally developed to analyze the frequencies of events.
* When we say that an outcome has a probability p of occurring, it means that if we repeated the experiment (e.g., draw a hand of cards) infinitely many times, then proportion p of the repetitions would result in that outcome.
* If a doctor analyzes a patient and says that the patient has a 40% chance of having the flu, this means something very different—we cannot make infinitely many replicas of the patient, nor is there any reason to believe that different replicas of the patient would present with the same symptoms yet have varying underlying conditions. In the case of the doctor diagnosing the patient, we use probability to represent a degree of belief, with 1 indicating absolute certainty that the patient has the flu and 0 indicating absolute certainty that the patient does not have the flu.
* The former kind of probability, related directly to the rates at which events occur, is known as **frequentist probability**, while the latter, related to qualitative levels of certainty, is known as **Bayesian probability**.
  1. **Random Variables**

A random variableis a variable that can take on different values randomly.

Random variables may be discrete or continuous.

A **discrete random variable** is one that has a finite or countably infinite number of states. Note that these states are not necessarily the integers; they can also just be named states that are not considered to have any numerical value.

A **continuous random variable** is associated with a real value.

* 1. **Probability Distributions**

A probability distributionis a description of how likely a random variable or set of random variables is to take on each of its possible states. The way we describe probability distributions depends on whether the variables are discrete or continuous

* + 1. **Discrete Variables and Probability Mass Functions**
* The probability mass function maps from a state of a random variable to the probability of that random variable taking on that state. The probability that x = x is denoted as P (x), with a probability of 1 indicating that x = x is certain and a probability of 0 indicating that x = x is impossible.
* Probability mass functions can act on many variables at the same time. Such a probability distribution over many variables is known as a **joint probability distribution**. P (x = x, y = y) denotes the probability that x = x and y = y simultaneously. We may also write P (x, y) for brevity.
* To be a probability mass function on a random variable x, a function P must satisfy the following properties:

1. The domain of P must be the set of all possible states of x.
2. ∀x ∈ x,0 ≤ P(x) ≤ 1.

An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.

1. .

We refer to this property as being **normalized**. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring.

* + 1. **Continuous Variables and** **Probability Density Functions**
* When working with continuous random variables, we describe probability distributions using a probability density function(PDF)rather than a probability mass function. To be a probability density function, a function p must satisfy the following properties:

1. The domain of must be the set of p all possible states of x.
2. ∀x ∈ x, p(x) ≥ 0. Note that we do not require p(x) ≤ 1.
3. ∫p(x)dx = 1.

* A probability density function p(x) does not give the probability of a specific state directly, instead the probability of landing inside an infinitesimal region with volume δx is given by p(x)δx.
* We can integrate the density function to find the actual probability mass of a set of points. Specifically, the probability that x lies in some set S is given by the integral of p (x) over that set. In the univariate example, the probability that x lies in the interval [a, b] is given by

∫[a,b] p(x)dx

* 1. **Marginal Probability**

Sometimes we know the probability distribution over a set of variables and we want to know the probability distribution over just a subset of them. The probability distribution over the subset is known as the marginal probabilitydistribution.

For example, suppose we have discrete random variables x and y, and we know P(x, y). We can find P(x) with the *sum rule*:

For continuous variables, we need to use integration instead of summation:

* 1. **Conditional Probability**

In many cases, we are interested in the probability of some event, given that some other event has happened. This is called a *conditional probability*.

* 1. **The Chain Rule of Conditional Probabilities**

Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable:

For example:

P (b, c) = P (b | c) P(c)

P (a, b, c) = P (a | b, c) P (b, c)

P (a, b, c) = P (a | b, c) P (b | c) P(c).

* 1. **Independence and Conditional Independence**
* Two random variables x and y are **independent**if their probability distribution can be expressed as a product of two factors, one involving only x and one involving only y:

∀x ∈ x, y ∈ y, p(x = x, y = y) = p(x = x) p(y = y).

* Two random variables x and y are **conditionally independent**given a random variable z if the conditional probability distribution over x and y factorizes in this way for every value of z:

∀x ∈ x, y ∈ y, z ∈ z, p(x = x, y = y | z = z) = p(x = x | z = z) p(y = y | z = z).

x⊥y means that x and y are independent, while x⊥y | z means that x and y are conditionally independent given z.

* 1. **Expectation, Variance and Covariance**
* The **expectation or expected value**of some function f(x) with respect to a probability distribution P(x) is the average or mean value that f takes on when x is drawn from P. For discrete variables, this can be computed with a summation:

while for continuous variables, it is computed with an integral:

Expectations are linear, for example:

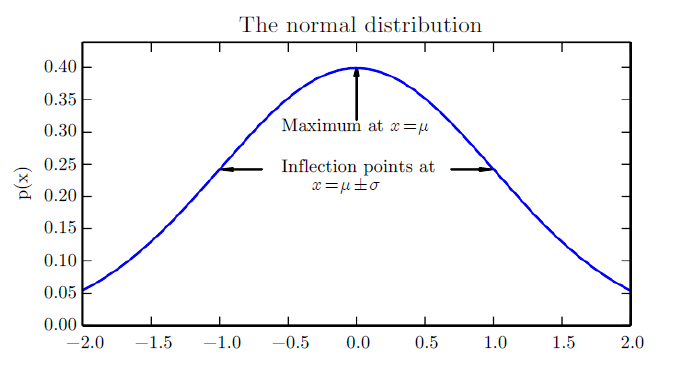
when α and β are not dependent on x.

* The **variance**gives a measure of how much the values of a function of a random variable x vary as we sample different values of x from its probability distribution:

When the variance is low, the values of f(x) cluster near their expected value. The

square root of the variance is known as the **standard deviation**.

* The **covariance**gives some sense of how much two values are linearly related to each other, as well as the scale of these variables:
* High absolute values of the covariance mean that the values change very much and are both far from their respective means at the same time. If the sign of the covariance is positive, then both variables tend to take on relatively high values simultaneously. If the sign of the covariance is negative, then one variable tends to take on a relatively high value at the times that the other takes on a relatively low value and vice versa.
* It is possible for two variables to be dependent but have zero covariance. For example, suppose we first sample a real number x from a uniform distribution over the interval [−1, 1]. We next sample a random variable s. With probability 1/2 , we choose the value of s to be 1. Otherwise, we choose the value of s to be −1. We can then generate a random variable y by assigning y = sx. Clearly, x and y are not independent, because x completely determines the magnitude of y. However, Cov(x, y) = 0.
  1. **Common Probability Distributions**
     1. **Bernoulli Distribution**
* The Bernoullidistribution is a distribution over a single binary random variable.
* It is controlled by a single parameter φ ∈ [0, 1], which gives the probability of the random variable being equal to 1.
* It has the following properties:
  + 1. **Multinoulli Distribution**
* The multinoullior categoricaldistribution is a distribution over a single discrete variable with k different states, where k is finite.1
* The multinoulli distribution is parametrized by a vector p ∈ [0, 1]k−1, where pi  gives the probability of the ith state.
* The final, kth state’s probability is given by 1 – 1Tp. Note that we must constrain 1T p ≤ 1.
  + 1. **Gaussian Distribution**
* The most commonly used distribution over real numbers is the normal distribution, also known as the Gaussian distribution:
* mean of the distribution is E[x] = μ.
* The standard deviation of the distribution is given by σ, and the variance by σ2.



**Figure 1:** *The normal distribution N(x; μ, σ2) exhibits a classic “bell curve” shape, with the x coordinate of its central peak given by μ, and the width of its peak controlled by σ.*

*In this example, we depict the standard normal distribution, with μ = 0 and σ = 1.*

* When we evaluate the Probability Density Function, we need to square and invert σ, which is computationally expensive.
* A more efficient way of parametrizing the distribution is to use a parameter β ∈ (0, ∞) to control the precisionor inverse variance of the distribution:
* The normal distribution is a good default choice for two major reasons.

1. The central limit theoremshows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many complicated systems can be modeled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behavior.
2. Out of all possible probability distributions with the same variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers. We can thus think of the normal distribution as being the one that inserts the least amount of prior knowledge into a model.

* The normal distribution generalizes to Rn , in which case it is known as the multivariate normal distribution. It may be parametrized with a positive definite symmetric matrix Σ:
* The parameter μ still gives the mean of the distribution, though now it is vector-valued. The parameter Σ gives the covariance matrix of the distribution.
* the covariance is not a computationally efficient way to parametrize the distribution, since we need to invert Σ to evaluate the Probability Density Function. We can instead use a precision matrixβ:
  + 1. **Exponential and Laplace Distribution**
* In the context of deep learning, we often want to have a probability distribution with a sharp point at x = 0. To accomplish this, we can use the exponential distribution:
* The exponential distribution uses the indicator function 1x≥0 to assign probability zero to all negative values of x.
* A closely related probability distribution that allows us to place a sharp peak of probability mass at an arbitrary point μ is the Laplace distribution
  + 1. **The Dirac Distribution and Empirical Distribution**
* In some cases, we wish to specify that all the mass in a probability distribution clusters around a single point. This can be accomplished by defining a Probability Density Function using the Dirac delta function, δ(x):

p(x) = δ(x − μ).

* The Dirac delta function is defined such that it is zero-valued everywhere except 0, yet integrates to 1.
* By defining p( x) to be δ shifted by −μ we obtain an infinitely narrow and infinitely high peak of probability mass where x = μ.
* A common use of the Dirac delta distribution is as a component of an empirical distribution,
* which puts probability mass 1/m on each of the m points x(1), . . . ,x(m) forming a given data set or collection of samples. The Dirac delta distribution is only necessary to define the empirical distribution over continuous variables. For discrete variables, the situation is simpler: an empirical distribution can be conceptualized as a multinoulli distribution, with a probability associated to each possible input value that is simply equal to the empirical frequency of that value in the training set.
  + 1. **Mixtures of Distributions**

A mixture distribution is made up of several component distributions. On each trial, the choice of which component distribution generates the sample is determined by sampling a component identity from a multinoulli distribution:

where P(c) is the multinoulli distribution over component identities.

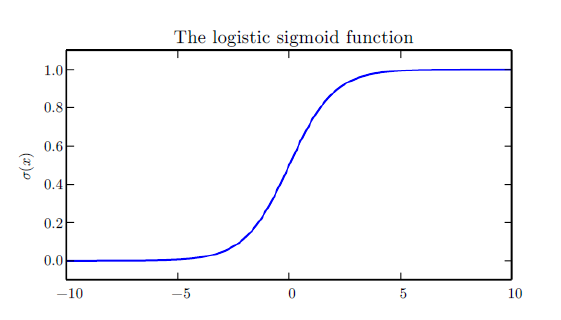
A latent variableis a random variable that we cannot observe directly. The component identity variable c of the mixture model provides an example. Latent variables may be related to x through the joint distribution, in this case, P(x, c) = P(x | c) P (c).

The distribution P (c) over the latent variable and the distribution P(x | c) relating the latent variables to the visible variables determines the shape of the distribution P (x) even though it is possible to describe P (x) without reference to the latent variable.

A very powerful and common type of mixture model is the Gaussian mixturemodel, in which the components p(x | c = i) are Gaussians. Each component has a separately parametrized mean μ(i) and covariance Σ(i).

In addition to the means and covariances, the parameters of a Gaussian mixture specify the prior probability αi = P ( c = i) given to each component i. The word “prior” indicates that it expresses the model’s beliefs about c **before** it has observed x. By comparison, P( c | x) is a posterior probability, because it is computed **after** observation of x. A Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific, non-zero amount of error by a Gaussian mixture model with enough components.

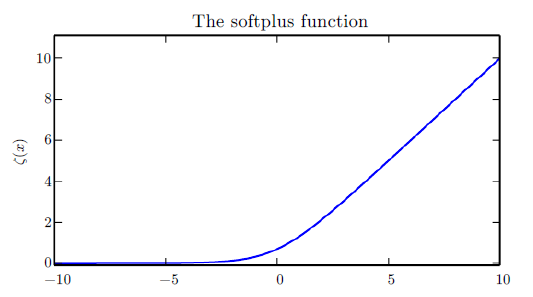
* 1. **Useful Properties of Common Functions**
* The logistic sigmoid is commonly used to produce the φ parameter of a Bernoulli distribution because its range is (0,1), which lies within the valid range of values for the φ parameter.
* sigmoid function saturateswhen its argument is very positive or very negative, meaning that the function becomes very flat and insensitive to slight changes in its input.



**Figure 2:** *Logistic Sigmoid function*

* The softplus function can be useful for producing the β or σ parameter of a normal distribution because its range is (0, ∞).

ζ(x) = log (1 +ex)



**Figure 2:** *softplus function*

The following properties are all useful enough:

* 1. **Bayes’ Rule**

We often find ourselves in a situation where we know P(y | x) and need to know P(x | y). Fortunately, if we also know P(x), we can compute the desired quantity using *Bayes’ rule*:

Note that while P (y) appears in the formula, it is usually feasible to compute P(y) =∑x P(y | x)P(x), so we do not need to begin with knowledge of P(y).

* 1. **Technical Details of Continuous Variables**