Deep learning by Ian goodfellow

▼ Chapter 3 - Probability and Information Theory

While probability theory allows us to make uncertain statements and to reason in the presence of uncertainty, information theory enables us to quantify the amount of uncertainty in a probability distribution.

▼ 3.1 - Why Probability?

▼ Three possible sources of uncertainty?

- Inherent stochasticity in the system being modeled. For example, most interpretations of quantum mechanics describe the dynamics of subatomicparticles as being probabilistic. We can also create theoretical scenarios that we postulate to have random dynamics, such as a hypothetical card game where we assume that the cards are truly shuffled into a random order.
- Incomplete observability. Even deterministic systems can appear stochastic when we cannot observe all the variables that drive the behavior of the system. For example, in the Monty Hall problem, a game show contestant is asked to choose between three doors and wins a prize held behind the chosen door. Two doors lead to a goat while a third leads to a car. The outcome given the contestant's choice is deterministic, but from the contestant's point of view, the outcome is uncertain.
- Incomplete modeling. When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions. For example, suppose we build a robot that can exactly observe the location of every object around it. If the robot discretizes space when predicting the future location of these objects, then the discretization makes the robot immediately become uncertain about the precise position of objects: each object could be anywhere within the discrete cell that it was observed to occupy.

▼ 3.2 - Random Variables

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

Random variables may be discrete or continuous. A discrete random variableis one that has a finite or countably infinite number of states. Note that thesestates are not necessarily the integers; they can also just be named states thatare not considered to have any numerical value. A continuous random variable isassociated with a real value.

▼ 3.3 - Probability Distributions

A probability distribution is 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.

▼ 3.3.1 - Discrete Variables and Probability Mass Functions

▼ What is a probability mass function(PMF)?

A probability distribution over discrete variables may be described using aproba-bility mass function(PMF).

▼ What properties a function should satisfy to be categorised as PMF?

- The domain of P must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$. 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.
- $\sum_{x \in \mathrm{x}} P(x) = 1$. 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.

For example, consider a single discrete random variable x with k different states. We can place a uniform distribution on x—that is, make each of its states equally likely—by setting its PMF to P (x = xi) =1/k for all i. We can see that this fits the requirements for a probability mass function.The value 1/k is positive because k is a positive integer. We also see that Summation i P (x = xi) = Summation i 1/k=k/k= 1, so the distribution is properly normalised.

▼ 3.3.2 - Continuous Variables and Probability Density Functions

▼ What properties a function should satisfy to be a PDF?

- The domain of p must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
- $\int p(x)dx = 1$.

▼ Define PDF?

A probability density function p(x) gives the probability of landing inside an infinitesimal region with volume δx is given by p(x)δx.

▼ PDF example

For an example of a PDF corresponding to a specific probability density over a continuous random variable, consider a uniform distribution on an interval of the real numbers. We can do this with a functionu(x;a, b), where a and b are the end points of the interval, with b > a. The “;” notation means “parametrized by”;we consider x to be the argument of the function, while a and b are parameters that define the function. To ensure that there is no probability mass outside the interval, we say u(x;a, b) = 0 for all x not belonging to [a, b]. Within [a, b],u(x;a, b) =1b−a. We can see that this is non-negative everywhere. Additionally, it integrates to 1. We often denote that x follows the uniform distribution on [a, b] by writing x ∼ U(a, b).

▼ 3.4 - Marginal Probability

▼ What is 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 probability distribution.

▼ 3.5 - Conditional Probability

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

denote the conditional probability that y=y given x=x as P(y=y | x=x). This conditional probability can be computed with the formula.

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}.$$

The conditional probability is only defined whenP(x=x)>0.We cannot computethe conditional probability conditioned on an event that never happens.It is important not to confuse conditional probability with computing whatwould happen if some action were undertaken. The conditional probability thata person is from Germany given that they speak German is quite high, but ifa randomly selected person is taught to speak German, their country of origindoes not change. Computing the consequences of an action is called making anintervention query. Intervention queries are the domain ofcausal modeling,which we do not explore in this book.

▼ 3.6 - The Chain Rule of Conditional Probabilities

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

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}).$$

This observation is known as the chain rule, or product rule, of probability

For example, applying the definition twice, we get

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}). \end{aligned}$$

▼ 3.7 - 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:

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y).$$

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:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z).$$

We can denote independence and conditional independence with compact notation: x⊥y means that x and y are independent, while x⊥y | z means that x and y are conditionally independent given z.

▼ 3.8 - Expectation, Variance and Covariance

▼ What is expectation in probability?

The expectation, or expected value, of some function f(x) with respect to a probability distributionP(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:

$$\mathbb{E}_{\mathrm{x} \sim P}[f(x)] = \sum_x P(x)f(x),$$

$$\mathbb{E}_{\mathrm{x} \sim p}[f(x)] = \int p(x)f(x)dx.$$

▼ What is variance in probability?

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:

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right].$$

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

▼ What is covariance in probability?

The covariance gives some sense of how much two values are linearly related to each other, as well as the scale of these variables:

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right].$$

▼ 3.9

▼ 3.10 - Useful Properties of Common Functions

▼ What is a logistic sigmoid?

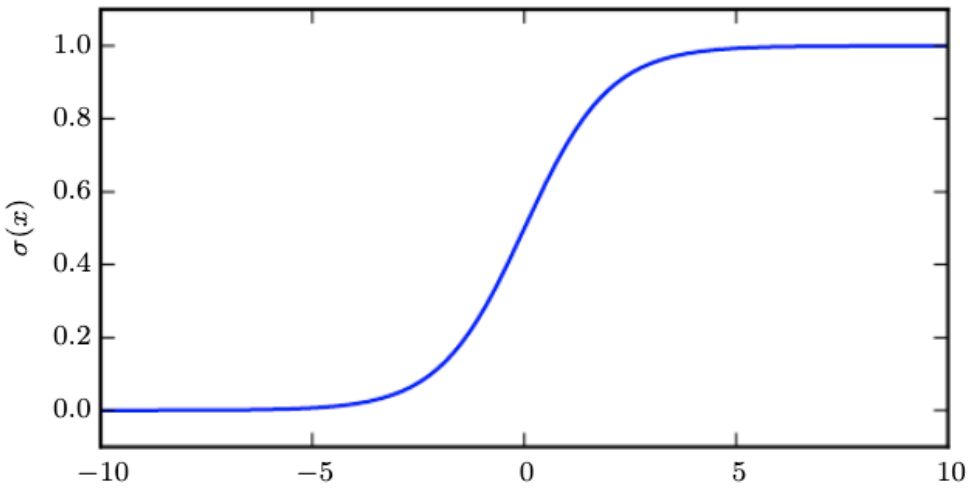

Figure 3.3: The logistic sigmoid function.

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

▼ What is softplus function?

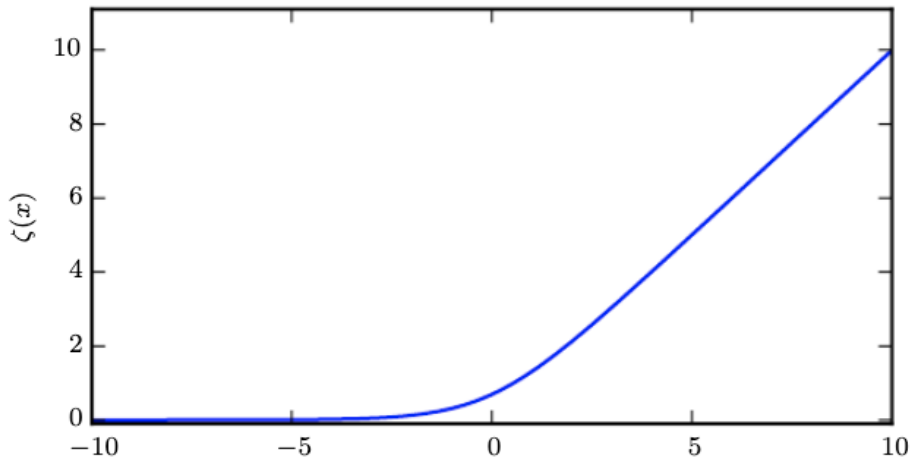

Figure 3.4: The softplus function.

$$\zeta(x) = \log\left(1 + \exp(x)\right).$$

▼ 3.11 - Bayes' Rule

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

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}.$$

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

▼ 3.12 - Technical Details of Continuous Variables

▼ What is measure theory?

A proper formal understanding of continuous random variables and probability density functions requires developing probability theory in terms of a branch of mathematics known as measure theory.

For our purposes, measure theory is more useful for describing theorems that apply to most points in Rn but do not apply to some corner cases. Measure theory provides a rigorous way of describing that a set of points is negligibly small. Such a set is said to have measure zero. We do not formally define this concept in this textbook. For our purposes, it is sufficient to understand the intuition that a set of measure zero occupies no volume in the space we are measuring. For example,withinR2, a line has measure zero, while a filled polygon has positive measure.Likewise, an individual point has measure zero. Any union of countably many sets that each have measure zero also has measure zero (so the set of all the rational numbers has measure zero, for instance)

▼ What does 'almost everywhere' mean in measure theory?

A property that holds almost everywhere holds throughout all space except for on a set of measure zero.

- ▼ Study the relation between Jacobian matrix and continous random variable

▼ 3.13 - Information Theory

- ▼ What is information theory?

  Information theory is a branch of applied mathematics that revolves around quantifying how much information is present in a signal. In this context, information theory tells how to design optimal codes and calculate the expected length of messages sampled from specific probability distributions using various encoding schemes.

- ▼ How do we quantify our intuition?

  We would like to quantify information in a way that formalizes this intuition.

  •Likely events should have low information content, and in the extreme case,events that are guaranteed to happen should have no information contentwhatsoever.

  • Less likely events should have higher information content.

  •Independent events should have additive information. For example, findingout that a tossed coin has come up as heads twice should convey twice asmuch information as finding out that a tossed coin has come up as headsonce.

- ▼ What is self-information of an event?

  To satisfy all three of these properties, we define the **self-information** of an event $\mathrm{x} = x$ to be

  $$I(x) = -\log P(x). \tag{3.48}$$

- ▼ **Cover differential entropy/shannon entropy and divergence**

▼ 3.14 - **Structured Probabilistic Models**