Probability review

James Chuang

February 27, 2017

Contents

My notes reviewing probability and information theory, combining information from the Stanford CS229 probability theory notes and the Chapter 3 of the Deep Learning Book. Again, most of the material is directly transcribed from these sources. Note that the two sources use slightly different notation, and I haven't been bothered to make everything consistent.

Probability theory

- a mathematical framework for representing uncertain statements
- a means of quantifying uncertainty, and axioms for new uncertain statements

Information allows us to quantify the amount of uncertainty in a probability distribution

why probability? (DL 3.1)

Machine learning must always deal with uncertain quantities, and sometimes may need to deal with stochastic quantities. There are three possible sources of uncertainty:

- Inherent stochasticity in the system being modeled. (e.g. quantum mechanics or randomly shuffled cards.)
- Incomplete observability. Deterministic systems can appear stochastic when we cannot observe all of the variables that drive the behavior of the system.
- Incomplete modeling. Models necessarily discard some of the observed information, resulting in uncertainty in the model's predictions.

Two ways of thinking about probability:

- ***frequentist*** probability- related directly to the rates at which events occur. E.g., flip a fair coin an infinite number of times and half of the time it will be heads.
- ***Bayesian*** probability- related to qualitative levels of certainty, i.e. a *degree of belief* that an event will happen. (E.g., the patient has a 40% chance of having the flu.)

elements of probability (CS229-1)

The three ***axioms of probability*** define probability on a set:

1. ***sample space*** Ω: The set of all possible outcomes of a random experiment. Each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment. For example, the sample space of rolling a 6-sided die is the set $\Omega = \{1, 2, 3, 4, 5, 6\}$.
2. ***set of events*** (or ***event space***) $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called ***events***) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment). An event can be a singleton set, e.g. in the die example, rolling a 1 is an event.
3. ***probability measure***: The probability measure describes how likely each event in the sample space will occur. More formally, it is a function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties:

- $P(A) \geq 0$, for all $A \in \mathcal{F}$
- $P(\Omega) = 1$
- If $A_1, A_2, ...$ are disjoint events (i.e. $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

properties:

- If $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min(P(A), P(B))$
- (union bound) $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega \backslash A) = 1 - P(A)$
- (law of total probability) If $A_1, \ldots, A_k$ are a set of disjoint events such that $\underset{i=1}{\cup} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$

random variables (CS229-2, DL-3.2)

The deep learning book says, “a ***random variable*** is a variable that can take on different values randomly,” which I guess is technically correct but sounds suspiciously like what I would write on an exam if I didn’t know the real answer. More formally, a random variable X is a function $X : \Omega \to \mathbb{R}$, denoted $X(\omega)$ or simply X. The value that a random variable may take on can be denoted in lowercase: x. For example, $X(\omega)$ can be the number of heads occurring in a sequence of coin tosses, ω.

Random variables may be discrete or continuous. A ***discrete random variable*** is one that has a finite or countably infinite number of states. The probability of the set associated with a random variable X taking on a specific value k is:

$$P(X = k) \triangleq P(\{\omega : X(\omega)\})$$

A ***continuous random variable*** is associated with a real value, i.e. it takes on an infinite number of possible values. The probability that X takes on a value between $a, b \in \mathbb{R}$, where $a < b$ is:

$$P(a \leq X \leq b) \triangleq P(\{\omega : a \leq X(\omega) \leq b\})$$

probability distribution functions (DL-3.3, CS229 2.1-2.3)

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. In order to specify the probability measure used when dealing with random variables, it is often convenient to specify alternative functions, i.e. cumulative distribution functions (CDFs), probability mass functions (PMFs), and probability density functions (PDFs).

A ***cumulative distribution function (CDF)*** is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as,

$$F_X(x) \triangleq P(X \leq x)$$

Properties of the CDF:

- $0 \leq F_X(x) \leq 1$

- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- $x \leq y \Rightarrow F_X(x) \leq F_Y(y)$

Probability distributions over discrete random variables are described using a ***probability mass function*** (PMF), a function $p_X : \Omega \to \mathbb{R}$ such that:

$$p_X(x) \triangleq P(X = x)$$

For discrete random variables, the notation Val(X) is used to indicate the set of possible values that the random variable X may assume (Val(X) is the domain of $p_X(x)$). For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten coin tosses, then Val$(X) = \{0, 1, 2, \ldots, 10\}$.

Properties of PMFs:

- $0 \leq p_X(x) \leq 1$
- $\sum_{x \in \text{Val}(X)} p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = P(X \in A)$

Probability distributions over continuous random variables (for which the CDF is differentiable everywhere) are described using a ***probability density function*** (PDF), a function $f_X(x) \triangleq \frac{dF_X(x)}{dx}$.

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 ∂x is given by $p(x)\partial x$.

joint and marginal distributions (CS229 3.1, DL 3.4)

When working with multiple random variables, we sometimes want to know the values that they assume simultaneously during outcomes of a random experiment. For example, consider an experiment where a coin is flipped ten times, where the two random variables of interest are $X(\omega) =$ number of heads and $Y(\omega) =$ the length of the longest run of consecutive heads. In this case of two random variables, the ***joint cumulative distribution function*** of X and Y, is defined by

$$F_{XY}(x, y) \triangleq P(X \leq x, Y \leq y)$$

The joint CDF $F_{XY}(x, y)$ and the distribution functions $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y) dy F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y) dx$$

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*** distibution. 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***:

$$\forall x \in X, P(X = x) = \sum_y P(X = x, Y = y)$$

For continuous variables, summation is replaced with integration:

$$p(x) = \int p(x, y) dy$$

conditional probability (DL 3.5, CS229 1.1, 3.4)

In the discrete case, the ***conditional probability mass function*** of Y given X is:

$$P(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{P(X = x)}$$

From the above definition, the conditional probability is only defined when $P(X = x) > 0$, i.e. the conditional probability cannot be computed on an event that never happens. Two events are ***independent*** iff $P(y, x) = P(y)P(x)$, or equivalently, $P(y \mid x) = P(y)$. That is, independence is equivalent to saying that observing x does not have any effect on the probability of y.

In the continuous case, the situation is technically more complicated because the probability that a continuous random variable X takes on a specific value x is zero. Ignoring this technical point, we define, by analogy to the discrete case, the ***conditional probability density*** of Y given $X = x$ to be:

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

provided $f_X(x) \neq 0$.

chain rule of conditional probabilities (DL 3.6)

Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable. This follows directly from the above definition of conditional probabilities:

$$\begin{aligned}
&P(X_1, \ldots, X_n) \\
&= P(X_n | X_1, X_2, \ldots, X_{n-1}) P(X_1, X_2, \ldots, X_{n-1}) \\
&= \ldots \\
&= P(X_1) \prod_{i=2}^{n} P(X_i | X_1, \ldots, X_{i-1})
\end{aligned}$$

For example,

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

Bayes' rule (CS229 3.5, DL 3.11)

For discrete random variables X and Y:

$$P_{Y|X}(y \mid x) = \frac{P_{XY}(x, y)}{P_X(x)} = \frac{P_{X|Y}(x \mid y)}{\sum_{y' \in \mathrm{Val}(Y)} P_{X|Y}(x \mid y') P_Y(y')}$$

For continuous random variables X and Y:

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x \mid y') f_Y(y') dy'}$$

independence and conditional independence (DL 3.7)

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:

$$\forall x \in X, y \in Y, \quad p(X = x, Y = y) = p(X = x) p(Y = y)$$

Two random variables 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:

$$\forall x \in X, y \in Y, z \in Z, \quad p(X = x, Y = y, Z = z) = p(X = x \mid Z = z) p(Y = y \mid Z = z)$$

Independence can be denoted shorthand: $X \perp Y$ indicates that X and Y are independent, while $X \perp Y \mid Z$ means that X and Y are conditionally independent given Z.

expectation (DL 3.8, CS229 2.4)

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. The expectation can be thought of as a weighted average of the values that $f(x)$ can take on for different values of x, where the weights are given by $P(X)$. As a special case, the expectation of a random variable itself is found by letting $f(x) = x$; this is the ***mean*** of the random variable X. For discrete variables:

$$\mathbf{E}_{X\sim P}[f(x)] = \sum_x P(x)f(x)$$

while for continuous variables:

$$\mathbf{E}_{X\sim p}[f(x)] = \int p(x)f(x)dx$$

When the identity of the distribution is clear from the context, we can omit it and simply write the name of the random variable that the expectation is over, i.e. $\mathbf{E}_X\left[f(x)\right]$. If it is clear which random variable the expectation is over, we can omit the subscript entirely, i.e. $\mathbf{E}[f(x)]$.

Properties of the expected value:

- $\mathbf{E}[a] = a$ for any constant $a \in \mathbb{R}$
- Expectations are linear, i.e.
$$\mathbf{E}_X[\alpha f(x) + \beta g(x)] = \alpha\mathbf{E}_X[f(x)] + \beta\mathbf{E}_X[g(x)]$$
, when α and β are not dependent on x.
- for a discrete random variable X, $\mathbf{E}[\mathbf{1}\{X = k\}] = P(X = k)$, where $\mathbf{1}\{\}$ is the indicator function.

variance and covariance (DL 3.8, CS229 2.5, 3.7)

The ***variance*** gives a measure of how concentrated the distribution of a random variable X is around its mean:

$$\text{Var}(f(x)) = \mathbf{E}\left[(f(x) - \mathbf{E}[f(x)])^2\right]$$

Using the properties of the expected value (namely that expectations are linear and that $\mathbf{E}[X]$ is a constant wrt an outer expectation), we can derive an alternate expression for the variance:

$$\begin{aligned}\text{Var}[X] &= \mathbf{E}[(X - E[X])^2] \\ &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}^2[X]] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}\left[X\mathbf{E}[X]\right] + \mathbf{E}\left[\mathbf{E}^2[X]\right] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}^2[X] \\ &= \mathbf{E}[X^2] - \mathbf{E}^2[X]\end{aligned}$$

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***.

Properties of the variance:

- $\text{Var}[a] = 0$ for any constant $a \in \mathbb{R}$
- $\text{Var}[af(x)] = a^2\text{Var}[f(x)]$ for any constant $a \in \mathbb{R}$

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

$$\text{Cov}(f(x), g(y)) = \mathbf{E}\left[(f(x) - \mathbf{E}[f(x)])\,(g(y) - \mathbf{E}[g(y)])\right]$$

Using an argument similar to that for variance (linearity of expectations and the fact that expectations of constants are just the constants), this can be rewritten as:

$$\begin{aligned}\text{Cov}[X,Y] &= \mathbf{E}[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])] \\ &= \mathbf{E}[XY - X\mathbf{E}[Y] - Y\mathbf{E}[X] + \mathbf{E}[X]\mathbf{E}[Y]] \\ &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] - \mathbf{E}[Y]\mathbf{E}[X] + \mathbf{E}[X]\mathbf{E}[Y] \\ &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]\end{aligned}$$

Properties of expectation and covariance for two random variables:

- (linearity of expectation) $\mathbf{E}[f(X,Y)+g(X,Y)] = \mathbf{E}[f(X,y)] + \mathbf{E}[g(X,Y)]$
- $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X,Y]$
- if X and Y are independent, then $\text{Cov}[X,Y] = 0$
- if X and Y are independent, then $\mathbf{E}[f(X)g(Y)] = \mathbf{E}[f(X)]\mathbf{E}[g(Y)]$

Correlation normalizes the covariance in order to measure how much the variables are related, without being affected by the scale of the separate variables. Covariance and dependence are related, but distinct. Two variables that are independent have zero covariance, and two variables with non-zero covariance are dependent. However, independence is a stronger requirement than zero covariance, because it excludes nonlinear relationships in addition to the linear relationships specified by zero covariance. Thus, it is possible for two variables to be dependent but have zero covariance.

The ***covariance matrix*** Σ of a random vector $X : \Omega \to \mathbb{R}^n$ is an $n \times n$ matrix, such that

$$\Sigma_{i,j} = \text{Cov}[x_i, x_j]$$

From the definition of covariance, we have:

$$\begin{aligned}\Sigma &= \begin{bmatrix}\text{Cov}[X_1,X_1] & \cdots & \text{Cov}[X_1,X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n,X_1] & \cdots & \text{Cov}[X_n,X_n]\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{E}[X_1^2] - \mathbf{E}^2[X_1] & \cdots & \mathbf{E}[X_1X_n] - \mathbf{E}[X_1]\mathbf{E}[X_n] \\ \vdots & \ddots & \vdots \\ \mathbf{E}[X_nX_1] - \mathbf{E}[X_n]\mathbf{E}[X_1] & \cdots & \mathbf{E}[X_n^2] - \mathbf{E}^2[X_n]\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{E}[X_1^2] & \cdots & \mathbf{E}[X_1X_n] \\ \vdots & \ddots & \vdots \\ \mathbf{E}[X_nX_1] & \cdots & \mathbf{E}[X_n^2]\end{bmatrix} - \begin{bmatrix}\mathbf{E}[X_1]\mathbf{E}[X_1] & \cdots & \mathbf{E}[X_1]\mathbf{E}[X_n] \\ \vdots & \ddots & \vdots \\ \mathbf{E}[X_n]\mathbf{E}[X_1] & \cdots & \mathbf{E}[X_n]\mathbf{E}[X_n]\end{bmatrix} \\ &= \mathbf{E}[XX^T] - \mathbf{E}[X]\mathbf{E}[X]^T \\ &= \ldots \\ &= \mathbf{E}[(X-\mathbf{E}[X])(X-\mathbf{E}[X])^T]\end{aligned}$$

Properties of the covariance matrix:

- $\Sigma \succeq 0$, i.e. Σ is positive semidefinite
- $\Sigma = \Sigma^T$, i.e. Σ is symmetric

The diagonal elements of the covariance give the variances:

$$\Sigma_{i,i} = \text{Var}(x_i)$$

common discrete probability distributions (DL 3.9, CS229 2.6)

Bernoulli distribution

The Bernoulli distribution is a distribution over a single binary random variable, parameterized by $\phi \in [0,1]$, which gives the probability of the random variable being equal to 1. It can be thought of as a coin toss where the probability of heads is given by ϕ. It is a special case of the binomial distribution with $n = 1$. The Bernoulli distribution has the following properties:

- $P(x = 1) = \phi$
- $P(x = 0) = 1 - \phi$
- $P(X = x) = \phi^x (1 - \phi)^{1-x}$
- $\mathbf{E}_X[X] = \phi$
- $\text{Var}_X(X) = \phi(1 - \phi)$

binomial distribution

$X \sim$ binomial(n, p) (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p:

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

geometric distribution

$X \sim$ geometric(p) (where $p > 0$): the number of flips of a coin with heads probability p until the first heads:

$$p(x) = p(1 - p)^{x-1}$$

Poisson distribution

$X \sim$ Poisson(λ) (where $\lambda > 0$): a probability distribution over the nonnegatve integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

common continuous probability distributions (DL 3.9, CS229 2.6)

uniform distribution

$X \sim$ uniform(a, b) (where $a < b$): equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

exponential distribution

$X \sim$ exponential(λ) (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

normal (Gaussian) distribution

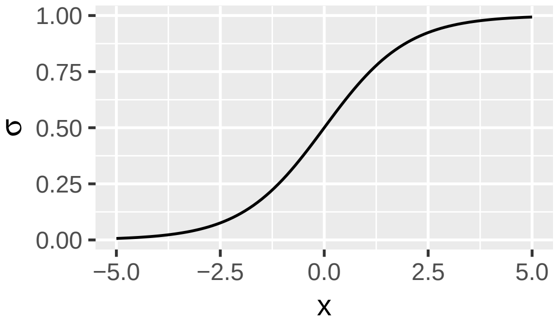
$X \sim$ normal(μ, σ^2):

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

useful properties of common functions (DL 3.10)

logistic sigmoid

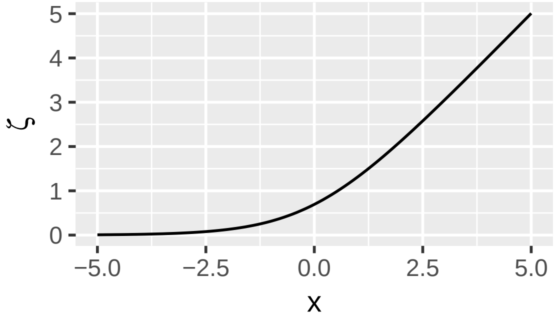
$$\sigma(x) = \frac{1}{1+\exp(-x)}$$



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. The sigmoid function saturates when its argument is very positive or very negative, meaning that the function becomes very flat and insensitive to small changes in its input.

softplus

$$\zeta(x) = \log(1+\exp(x))$$



The softplus function can be useful for producing the β or σ parameter of a normal distribution because its range is $0, \infty$. It also arises commonly when manipulating expressions involving sigmoids. The name 'softplus' comes from the fact that it is a smoothed vesion of $x^+ = \max(0, x)$.

Useful properties of the logistic sigmoid and softplus functions:

- $\sigma(x) = \frac{\exp(x)}{\exp(x)+\exp(0)}$
- $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$
- $1-\sigma(x) = \sigma(-x)$
- $\log\sigma(x) = -\zeta(-x)$
- $\frac{d}{dx}\zeta(x) = \sigma(x)$
- $\forall x \in (0,1), \sigma^{-1} = \log(\frac{x}{1-x})$
- $\forall x > 0, \zeta^{-1}(x) = \log(\exp(x)-1)$
- $\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy$
- $\zeta(x) - \zeta(-x)x$

The function $\sigma^{-1}(x)$ is called the ***logit*** in statistics.

information theory

Information theory- a branch of applied math revolving around quantifying how much information is present in a signal.

The basic intuition: learning that an unlikely event has occurred is more informative than learning that a likely event has occurred. (E.g., "The sun rose this morning" is uninformative, but "There was a solar eclipse this morning" is informative.)

Quantify information in a way that formalizes this intuition:

- Likely events should have low information content. Events that are guaranteed to happen should have no information content.
- Less likely events should have greater information content.
- Independent events should have additive information. E.g., finding out that a tossed coin has come up as heads twice should convey twice as much information as finding out that a tossed coin has come up heads once.

A function which satisfies these properties is the ***self-information*** of an event $X = x$:

$$I(x) = -\log P(x)$$

Using the base e logarithm, as above, information is in units of ***nats***. Using the base 2 logarithm, information is in units of ***bits*** or ***shannons***.

Self-information deals only with a single outcome. The amount of uncertainty in an entire probability distribution is quantified using the ***Shannon entropy***:

$$H(x) = \mathbf{E}_{X\sim P}[I(x)] = -\mathbf{E}_{X\sim P}[\log P(x)]$$

The Shannon entropy is the expected amount of information in an event drawn from that distribution. It gives a lower bound on the information needed on average to encode symbols drawn from a distribution P. When X is continuous, the Shannon entropy is known as the ***differential entropy***.

If we have two separate probability distributions $P(X)$ and $Q(X)$ over the same random variable X, we can measure how different these two distributions are using the ***Kullback-Leibler (KL) divergence***:

$$D_{KL}(P \parallel Q) = \mathbf{E}_{X\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbf{E}_{X\sim P}[\log P(x) - \log Q(x)]$$

The KL divergence is 0 iff P and Q are the same distribution. However, it is not a true distance metric because it is not symmetric: In general, $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$.

A quantity that is closely related to the KL divergence is the ***cross-entropy*** $H(P, Q) = H(P) + D_{KL}(P \parallel Q)$:

$$H(P, Q) = -\mathbf{E}_{X\sim P} \log Q(x)$$

Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence, because Q does not participate in the omitted term.