

07-280 Notes

Probability Background

Carnegie Mellon University
School of Computer Science

Contents

1	Probability	1
1.1	Vocabulary	1
1.2	Notation	2
1.2.1	Probability (of an event)	2
1.2.2	Distributions	3
1.2.3	Statistics	4
1.3	Notes	4
1.3.1	Random Variables	4
1.3.2	Continuous vs discrete random variables	4
1.3.3	Basic Rules	4
1.3.4	Law of Total Probability (marginalization)	6
1.3.5	Independence	6
1.3.6	Gaussian Distribution	6
1.3.7	Law of Total Expectation	7
1.3.8	Variance	7
1.3.9	Independent and identically distributed (i.i.d.)	7

1 Probability

1.1 Vocabulary

outcome: a specific result, i.e., heads, tails.

sample space: the set of all possible outcomes of the random experiment, Ω is the non-empty, finite set. In the case of flipping a single coin, we let $\Omega = \{ \text{heads}, \text{tails} \}$

events: a set of outcomes, which describe which outcomes correspond to the “event” happening. If you throw the die twice, then your event could be, for example, $\{2, 3\}$

random variables: A random variable is a function $X : \Omega \rightarrow \mathbb{R}$, which is just a labeling of the elements in Ω , the sample space, with some real numbers. This is a desirable transformation because then we can take the expectation of the random variable and other interesting concepts. i

distributions: Probability distributions may be associated with a random variable as a means to define the relationship between the range of a random variable and probability or probability density. Discrete random variables have discrete probability distributions that can be defined by a probability mass function. Likewise, continuous random variables have continuous probability distributions that can be defined by a probability density function.

probability mass function: Let $X : \Omega \rightarrow \mathbb{R}$, be a random variable. The probability mass function (pmf) of X is a function $p_X : \mathbb{R} \rightarrow [0, 1]$ such that for any $x \in \mathbb{R}$, $p_X(x) = P[X = x]$. The pmf must sum to 1, $\sum_{x \in \text{range}(X)} p_X(x) = 1$.

probability density function: The probability density function (pdf) of a continuous random variable X over some interval of values, S is an integrable function $f(x)$ satisfying the following:

1. $f(x)$ is positive everywhere in the interval of the domain, ie, $f(x) > 0$ for all $x \in S$
2. The area under the curve $f(x)$ in the interval is 1, ie, $\int_S f(x)dx = 1$
3. If $f(x)$ is the pdf of x , then the probability that x belongs to some interval A is given by the integral of $f(x)$ over that interval, that is, $P(X \in A) = \int_A f(x)dx$

parameters: Variables that when set, define a specific probability distribution. E.g., the parameters of the normal distribution are the mean and standard deviation, μ and σ , respectively.

1.2 Notation

Ω : Ω is typically used to represent the sample space, the set of all possible outcomes.

Capital letter, e.g A, B, X, Y : Capital letters (in probability) are usually events or random variables.

1.2.1 Probability (of an event)

$P(A)$ or $Pr(A)$ or $\mathbb{P}(A)$ or $\mathbf{P}\{A\}$: The probability of event A occurring

$P(X = x)$: The probability of random variable X taking on value x . Note that while extremely common, this is a notational shortcut. More formally, it would be written as $\mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$, the “probability of the set of outcomes where the random variable X maps the outcome to the value x ”.

$P(X = 1)$: The probability of random variable X taking on the specific value 1 (using the same notational shortcut as $P(X = x)$ described above.

$P(X = \text{heads})$: Likely an abuse of notation for a discrete random variable that maps the outcome “heads” one-to-one to a value. For example, if $X(\text{heads}) = 1$ and nothing else maps to 1, this shorthand notation could be more precisely written as $P(X = 1)$. Writing just $P(\text{heads})$ would have been correct from a notation standpoint, but it probably loses the connection to a specific random variable X .

$P(A \cap B)$ or $P(A, B)$: The probability of both A and B, jointly. We tend to use the comma notation for the joint probabilities.

$P(A | B)$: Conditional probability of A given B . The conditional probability symbol has operator precedence, so $P(A, B | C, D)$ is the probability of A and B given both C and D . (Other groupings don't really make sense.)

$A \perp B$: A and B are independent

$A \perp B | C$: A and B are independent given C . (Note the precedence of the conditional symbol. One might initially try to read this as $A \perp (B | C)$, but that doesn't quite make sense.)

1.2.2 Distributions

$p(x)$ or $p_X(x)$: probability mass function for discrete random variable X . We almost always drop the X subscript, relying on context to understand the implied random variable.

$f(x)$ or $f_X(x)$: probability density function for continuous random variable X . Also, often dropping the X subscript.

$F(x)$ or $F_X(x)$: cumulative distribution function for random variable X .

Note: In the rest of this section (and throughout the course), we'll just use p to indicate either a pmf or a pdf.

$P(X)$: If X is a random variable, this is a bit of an abuse of notation, likely meant to mean $P(X = x) \forall x \in \text{range}(X)$, which of course is just the pmf or pdf, $p(x)$.

$p(x = 1)$ or $P(x = 1)$: A fairly common abuse of notation in machine learning. It could be written slightly more precisely as $p_X(1)$, $p(1)$, or $P(X = 1)$ (which is already shorthand notation). Writing $p(x = 1)$ can be a convenient way of writing $p(x)$ when $x = 1$. Another example used quite often is $p(x | y = 1)$, which could be written more precisely as $p_{X,Y}(x | 1)$.

$p(x; \theta)$ or $p(x | \theta)$ or $p(x)$ or $p_\theta(x)$: (marginal) distribution (pmf/pdf) defined by parameter(s) θ . The $|$ is required when we are considering the variable that is conditioned upon to also be a random variable value (Bayesian reasoning), e.g., if we also have prior distribution over θ , $p(\theta)$. The colon notation represents a [Frequentist's](#) view of probability (as opposed to Bayesian). The colon and $|$ are often used interchangeably when we don't explicitly model distributions for the given variable.

$p(x, y)$ or $p_{X,Y}(x, y)$: pmf/pdf of joint distribution related to both random variables X and Y .

$p(x | y)$ or $p_{X|Y}(x | y)$: pmf/pdf of the conditional distribution of random variable X conditioned on the random variable Y taking on value y . Note that it's easy to confuse y as needing to be a known value; however, y is also an input to this pmf/pdf, which allows us to compute the mass/density for any combination of x and y .

$X \sim \text{DistributionName}(\theta)$: Random variable X is modeled with a 'enter distribution name' distribution defined by parameter(s) θ .

$x \sim \text{DistributionName}(\theta)$: The value x is sampled from (drawn from) a random variable with a 'enter distribution name' distribution defined by parameter(s) θ . There are a few distributions that have a common symbol for their distribution, e.g., $\mathcal{N}$ for a Gaussian distribution.

$x \sim \mathcal{N}(\mu, \sigma)$: The value x is drawn from a Gaussian distribution with mean μ and standard deviation σ .

1.2.3 Statistics

$\hat{\theta}$: The hat notation can be used to indicate a specific estimated value as opposed to a variable of the same name. For example, let $\hat{\theta}$ be the estimate of the parameter that minimizes the function $J(\theta)$.

$\mathcal{D}$: Dataset notation: a dataset is typically represented by calligraphic uppercase letters, often $\mathcal{D}$. Technically, a dataset is a multiset rather than a set because we can have repeated entries. We often index a dataset by labels in parenthesized superscripts $\mathcal{D} = \{z^{(1)}, z^{(2)}, \dots, z^{(n)}\}$, where $n = |\mathcal{D}|$, the number of elements in the set. This will become helpful when we get into a set of training examples, $\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$. Note: These parenthesized superscripts are not exponents! However, some people will not include these helpful parentheses, so you'll have to determine from context whether they are powers or not.

1.3 Notes

Note many of the rules in these notes are written for discrete probabilities. The associated rule can be converted to continuous distributions with either no changes or simple changes, such as converting the summation to an integral.

1.3.1 Random Variables

In this class, we will be using random variable notation. A random variable is a mapping of events to values, and then the associated pmf (or pdf) maps those values to probabilities (or densities). For example, if Z takes on the value 1 when the roll of a six-sided fair dice is even, the probability of rolling an even number will be denoted as the following:

$$P(Z = 1) = \frac{1}{2}$$

1.3.2 Continuous vs discrete random variables

Discrete random variables can only take a countable number of values (e.g., values we can roll in a dice) while continuous can take on infinitely many values. Our definitions for marginalization and the law of total probability assumed Y was a discrete random variable. If Y is not:

$$P(A = a) = \int_b P(A = a, B = b) * db$$

1.3.3 Basic Rules

Definition of Conditional Probability:

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

Chain Rule:

$$\begin{aligned}
 P(X, Y) &= P(X | Y)P(Y) \\
 &= P(Y | X)P(X) \\
 P(X_1, X_2, X_3) &= P(X_1, X_2 | X_3)P(X_3) \\
 &= P(X_1 | X_2, X_3)P(X_2, X_3)
 \end{aligned}$$

Chain Rule (with more variables):

$$\begin{aligned}
 P(X_1, X_2, X_3) &= P(X_1 | X_2, X_3)P(X_2, X_3) \\
 &= P(X_1 | X_2, X_3)P(X_2 | X_3)P(X_3) \\
 P(X_1, \dots, X_N) &= \prod_{n=1}^N P(X_n | X_1, \dots, X_{n-1})
 \end{aligned}$$

The chain rule decomposition may be done in any order. For example:

$$\begin{aligned}
 P(X_1, X_2, X_3) &= P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2) \\
 &= P(X_1)P(X_3 | X_1)P(X_2 | X_1, X_3) \\
 &= P(X_2)P(X_1 | X_2)P(X_3 | X_1, X_2) \\
 &= P(X_2)P(X_3 | X_2)P(X_1 | X_2, X_3) \\
 &= P(X_3)P(X_1 | X_3)P(X_2 | X_1, X_3) \\
 &= P(X_3)P(X_2 | X_3)P(X_1 | X_2, X_3)
 \end{aligned}$$

Bayes' Law:

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

Normalization:

$$\begin{aligned}
 P(Y | X) &= \frac{P(X, Y)}{P(X)} = \frac{P(X, Y)}{\sum_y P(X, Y = y)} \\
 P(Y | X) &\propto P(X, Y) \\
 P(Y | X) &= \alpha P(X, Y) \quad \text{Note this difference between } \propto \text{ and } \alpha \\
 \alpha &= \frac{1}{P(X)} = \frac{1}{\sum_y P(X, Y = y)}
 \end{aligned}$$

All of these basic probability rules hold when conditioning on a set of random variables or outcomes. To make this work, the conditioned variables need to be included in each term in the rule. For example, take Bayes' law from above, but now conditioned upon variables A and B :

$$P(Y | X, A, B) = \frac{P(X | Y, A, B)P(Y | A, B)}{P(X | A, B)}$$

1.3.4 Law of Total Probability (marginalization)

The law of total probability allows us to “sum out” variables from a joint distribution (sometimes called **marginalization**). This is useful when we are given the joint probability distribution and want to find the probability distribution over just a subset of the variables. Marginalization has the following forms:

To sum out a single variable:

$$P(X) = \sum_y P(X, Y = y)$$

To sum out multiple variables:

$$P(X) = \sum_z \sum_y P(X, Y = y, Z = z)$$

This also works for conditional distributions when summing out a variable that is not conditioned upon, i.e. a variable to the left of the $|$:

$$P(A \mid C, D = d) = \sum_b P(A, B = b \mid C, D = d)$$

This does NOT work when summing over a variable that is conditioned upon, i.e. a variable to the right of the $|$:

$$P(A, B = b \mid C) \neq \sum_d P(A, B = b \mid C, D = d)$$

1.3.5 Independence

If two variables X and Y are **independent** (denoted $X \perp\!\!\!\perp Y$), by definition the following are true:

- $P(X, Y) = P(X)P(Y)$
- $P(X) = P(X \mid Y)$
- $P(Y) = P(Y \mid X)$

If two variables X and Y are **conditionally independent given Z** (denoted $X \perp\!\!\!\perp Y \mid Z$), by definition the following are true:

- $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$
- $P(X \mid Y, Z) = P(X \mid Z)$
- $P(Y \mid X, Z) = P(Y \mid Z)$

1.3.6 Gaussian Distribution

Denoted as $X \sim \mathcal{N}(\mu, \sigma)$, the probability density function is given by:

$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

When X is a vector of K random variables, the associated multivariate Gaussian distribution has pdf:

$$p(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{K/2} \det(\Sigma)^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

1.3.7 Law of Total Expectation

For two given random variables X, Y :

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]]$$

If Y is discrete, this becomes:

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X | Y = y]P(Y = y)$$

1.3.8 Variance

It's a measure of spread for a distribution of a random variable that determines the degree to which the values of a random variable differ from the expected value.

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

$$Var(X) = \sigma^2 \text{ where } \sigma \text{ is the standard deviation of } X$$

1.3.9 Independent and identically distributed (i.i.d.)

A collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent i.e. each variable has the same chance of occurring as the others, and none of them have an influence on one another. This is a popular and important concept in statistics. A lot of the models and algorithms assume this property about their data.