

All Cheat Sheets

Machine Learning, Deep Learning, Artificial Intelligence



BY
STANFORD UNIVERSITY
AND
MASSACHUSETTS INSTITUTE OF
TECHNOLOGY



Stanford
University



**Massachusetts
Institute of
Technology**

COMPILED BY - NIKHIL YADAV

Probability—the Science of Uncertainty and Data

by Fabián Kozyński

PROBABILITY

Probability models and axioms

Definition (Sample space) A sample space Ω is the set of all possible outcomes. The set's elements must be mutually exclusive, collectively exhaustive and at the right granularity.

Definition (Event) An event is a subset of the sample space. Probability is assigned to events.

Definition (Probability axioms) A probability law \mathbb{P} assigns probabilities to events and satisfies the following axioms:

Nonnegativity $\mathbb{P}(A) \geq 0$ for all events A .

Normalization $\mathbb{P}(\Omega) = 1$.

(Countable) additivity For every sequence of events A_1, A_2, \dots

such that $A_i \cap A_j = \emptyset$: $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$.

Corollaries (Consequences of the axioms)

- $\mathbb{P}(\emptyset) = 0$.
- For any finite collection of disjoint events A_1, \dots, A_n ,
$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i)$$
.
- $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.
- $\mathbb{P}(A) \leq 1$.
- If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

Example (Discrete uniform law) Assume Ω is finite and consists of n equally likely elements. Also, assume that $A \subset \Omega$ with k elements. Then $\mathbb{P}(A) = \frac{k}{n}$.

Conditioning and Bayes' rule

Definition (Conditional probability) Given that event B has occurred and that $\mathbb{P}(B) > 0$, the probability that A occurs is

$$\mathbb{P}(A|B) \triangleq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Remark (Conditional probabilities properties) They are the same as ordinary probabilities. Assuming $\mathbb{P}(B) > 0$:

- $\mathbb{P}(A|B) \geq 0$.
- $\mathbb{P}(\Omega|B) = 1$
- $\mathbb{P}(B|B) = 1$.
- If $A \cap C = \emptyset$, $\mathbb{P}(A \cup C|B) = \mathbb{P}(A|B) + \mathbb{P}(C|B)$.

Proposition (Multiplication rule)

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}).$$

Theorem (Total probability theorem) Given a partition $\{A_1, A_2, \dots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and for every event B , we have

$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i) \mathbb{P}(B|A_i).$$

Theorem (Bayes' rule) Given a partition $\{A_1, A_2, \dots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and if $\mathbb{P}(A_i) > 0$ for all i , then for every event B , the conditional probabilities $\mathbb{P}(A_i|B)$ can be obtained from the conditional probabilities $\mathbb{P}(B|A_i)$ and the initial probabilities $\mathbb{P}(A_i)$ as follows:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i) \mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j) \mathbb{P}(B|A_j)}.$$

Independence

Definition (Independence of events) Two events are independent if occurrence of one provides no information about the other. We say that A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B).$$

Equivalently, as long as $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

$$\mathbb{P}(B|A) = \mathbb{P}(B) \quad \mathbb{P}(A|B) = \mathbb{P}(A).$$

Remarks

- The definition of independence is symmetric with respect to A and B .
- The product definition applies even if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

Corollary If A and B are independent, then A and B^c are independent. Similarly for A^c and B , or for A^c and B^c .

Definition (Conditional independence) We say that A and B are independent conditioned on C , where $\mathbb{P}(C) > 0$, if

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C) \mathbb{P}(B|C).$$

Definition (Independence of a collection of events) We say that events A_1, A_2, \dots, A_n are independent if for every collection of distinct indices i_1, i_2, \dots, i_k , we have

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Counting

This section deals with finite sets with uniform probability law. In this case, to calculate $\mathbb{P}(A)$, we need to count the number of elements in A and in Ω .

Remark (Basic counting principle) For a selection that can be done in r stages, with n_i choices at each stage i , the number of possible selections is $n_1 \cdot n_2 \cdots n_r$.

Definition (Permutations) The number of permutations (orderings) of n different elements is

$$n! = 1 \cdot 2 \cdot 3 \cdots n.$$

Definition (Combinations) Given a set of n elements, the number of subsets with exactly k elements is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Definition (Partitions) We are given an n -element set and nonnegative integers n_1, n_2, \dots, n_r , whose sum is equal to n . The number of partitions of the set into r disjoint subsets, with the i^{th} subset containing exactly n_i elements, is equal to

$$\binom{n}{n_1, \dots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

Remark This is the same as counting how to assign n distinct elements to r people, giving each person i exactly n_i elements.

Discrete random variables

Probability mass function and expectation

Definition (Random variable) A random variable X is a function of the sample space Ω into the real numbers (or \mathbb{R}^n). Its range can be discrete or continuous.

Definition (Probability mass function (PMF)) The probability law of a discrete random variable X is called its PMF. It is defined as

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}).$$

Properties

$$p_X(x) \geq 0, \forall x.$$

$$\sum_x p_X(x) = 1.$$

Example (Bernoulli random variable) A Bernoulli random variable X with parameter $0 \leq p \leq 1$ ($X \sim \text{Ber}(p)$) takes the following values:

$$X = \begin{cases} 1 & \text{w.p. } p, \\ 0 & \text{w.p. } 1 - p. \end{cases}$$

An indicator random variable of an event ($I_A = 1$ if A occurs) is an example of a Bernoulli random variable.

Example (Discrete uniform random variable) A Discrete uniform random variable X between a and b with $a \leq b$ ($X \sim \text{Uni}[a, b]$) takes any of the values in $\{a, a+1, \dots, b\}$ with probability $\frac{1}{b-a+1}$.

Example (Binomial random variable) A Binomial random variable X with parameters n (natural number) and $0 \leq p \leq 1$ ($X \sim \text{Bin}(n, p)$) takes values in the set $\{0, 1, \dots, n\}$ with probabilities $p_X(i) = \binom{n}{i} p^i (1-p)^{n-i}$.

It represents the number of successes in n independent trials where each trial has a probability of success p . Therefore, it can also be seen as the sum of n independent Bernoulli random variables, each with parameter p .

Example (Geometric random variable) A Geometric random variable X with parameter $0 \leq p \leq 1$ ($X \sim \text{Geo}(p)$) takes values in the set $\{1, 2, \dots\}$ with probabilities $p_X(i) = (1-p)^{i-1} p$. It represents the number of independent trials until (and including) the first success, when the probability of success in each trial is p .

Definition (Expectation/mean of a random variable) The expectation of a discrete random variable is defined as

$$\mathbb{E}[X] \triangleq \sum_x x p_X(x).$$

assuming $\sum_x |x| p_X(x) < \infty$.

Properties (Properties of expectation)

- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.
- If $X = c$ then $\mathbb{E}[X] = c$.

Example Expected value of know r.v.

- If $X \sim \text{Ber}(p)$ then $\mathbb{E}[X] = p$.
- If $X = I_A$ then $\mathbb{E}[X] = \mathbb{P}(A)$.
- If $X \sim \text{Uni}[a, b]$ then $\mathbb{E}[X] = \frac{a+b}{2}$.
- If $X \sim \text{Bin}(n, p)$ then $\mathbb{E}[X] = np$.
- If $X \sim \text{Geo}(p)$ then $\mathbb{E}[X] = \frac{1}{p}$.

Theorem (Expected value rule) Given a random variable X and a function $g: \mathbb{R} \rightarrow \mathbb{R}$, we construct the random variable $Y = g(X)$. Then

$$\sum_y y p_Y(y) = \mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_x g(x) p_X(x).$$

Remark (PMF of $Y = g(X)$) The PMF of $Y = g(X)$ is $p_Y(y) = \sum_{x: g(x)=y} p_X(x)$.

Remark In general $g(\mathbb{E}[X]) \neq \mathbb{E}[g(X)]$. They are equal if $g(x) = ax + b$.

Variance, conditioning on an event, multiple r.v.

Definition (Variance of a random variable) Given a random variable X with $\mu = \mathbb{E}[X]$, its variance is a measure of the spread of the random variable and is defined as

$$\text{Var}(X) \triangleq \mathbb{E}[(X - \mu)^2] = \sum_x (x - \mu)^2 p_X(x).$$

Definition (Standard deviation)

$$\sigma_X = \sqrt{\text{Var}(X)}.$$

Properties (Properties of the variance)

- $\text{Var}(aX) = a^2 \text{Var}(X)$, for all $a \in \mathbb{R}$.
- $\text{Var}(X + b) = \text{Var}(X)$, for all $b \in \mathbb{R}$.
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$.
- $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Example (Variance of known r.v.)

- If $X \sim \text{Ber}(p)$, then $\text{Var}(X) = p(1 - p)$.
- If $X \sim \text{Uni}[a, b]$, then $\text{Var}(X) = \frac{(b-a)(b-a+2)}{12}$.
- If $X \sim \text{Bin}(n, p)$, then $\text{Var}(X) = np(1 - p)$.
- If $X \sim \text{Geo}(p)$, then $\text{Var}(X) = \frac{1-p}{p^2}$.

Proposition (Conditional PMF and expectation, given an event) Given the event A , with $\mathbb{P}(A) > 0$, we have the following

- $p_{X|A}(x) = \mathbb{P}(X = x|A)$.
- If A is a subset of the range of X , then:
$$p_{X|A}(x) \triangleq p_{X|\{X \in A\}}(x) = \begin{cases} \frac{1}{\mathbb{P}(A)} p_X(x), & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$
- $\sum_x p_{X|A}(x) = 1$.
- $\mathbb{E}[X|A] = \sum_x x p_{X|A}(x)$.
- $\mathbb{E}[g(X)|A] = \sum_x g(x) p_{X|A}(x)$.

Proposition (Total expectation rule) Given a partition of disjoint events A_1, \dots, A_n such that $\sum_i \mathbb{P}(A_i) = 1$, and $\mathbb{P}(A_i) > 0$,

$$\mathbb{E}[X] = \mathbb{P}(A_1) \mathbb{E}[X|A_1] + \dots + \mathbb{P}(A_n) \mathbb{E}[X|A_n].$$

Definition (Memorylessness of the geometric random variable)

When we condition a geometric random variable X on the event $X > n$ we have memorylessness, meaning that the “remaining time” $X - n$, given that $X > n$, is also geometric with the same parameter. Formally,

$$p_{X-n|X>n}(i) = p_X(i).$$

Definition (Joint PMF) The joint PMF of random variables X_1, X_2, \dots, X_n is $p_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$.

Properties (Properties of joint PMF)

- $\sum_{x_1} \dots \sum_{x_n} p_{X_1, \dots, X_n}(x_1, \dots, x_n) = 1$.
- $p_{X_1}(x_1) = \sum_{x_2} \dots \sum_{x_n} p_{X_1, \dots, X_n}(x_1, x_2, \dots, x_n)$.
- $p_{X_2, \dots, X_n}(x_2, \dots, x_n) = \sum_{x_1} p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$.

Definition (Functions of multiple r.v.) If $Z = g(X_1, \dots, X_n)$, where $g: \mathbb{R}^n \rightarrow \mathbb{R}$, then $p_Z(z) = \mathbb{P}(g(X_1, \dots, X_n) = z)$.

Proposition (Expected value rule for multiple r.v.) Given $g: \mathbb{R}^n \rightarrow \mathbb{R}$,

$$\mathbb{E}[g(X_1, \dots, X_n)] = \sum_{x_1, \dots, x_n} g(x_1, \dots, x_n) p_{X_1, \dots, X_n}(x_1, \dots, x_n).$$

Properties (Linearity of expectations)

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
- $\mathbb{E}[X_1 + \dots + X_n] = \mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]$.

Conditioning on a random variable, independence

Definition (Conditional PMF given another random variable)

Given discrete random variables X, Y and y such that $p_Y(y) > 0$ we define

$$p_{X|Y}(x|y) \triangleq \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Proposition (Multiplication rule) Given jointly discrete random variables X, Y , and whenever the conditional probabilities are defined,

$$p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y|x) = p_Y(y) p_{X|Y}(x|y).$$

Definition (Conditional expectation) Given discrete random variables X, Y and y such that $p_Y(y) > 0$ we define

$$\mathbb{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y).$$

Additionally we have

$$\mathbb{E}[g(X)|Y = y] = \sum_x g(x) p_{X|Y}(x|y).$$

Theorem (Total probability and expectation theorems)

If $p_Y(y) > 0$, then

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x|y),$$

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

Definition (Independence of a random variable and an event) A discrete random variable X and an event A are independent if $\mathbb{P}(X = x \text{ and } A) = p_X(x) \mathbb{P}(A)$, for all x .

Definition (Independence of two random variables) Two discrete random variables X and Y are independent if

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \text{ for all } x, y.$$

Remark (Independence of a collection of random variables) A collection X_1, X_2, \dots, X_n of random variables are independent if

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \dots p_{X_n}(x_n), \quad \forall x_1, \dots, x_n.$$

Remark (Independence and expectation) In general, $\mathbb{E}[g(X, Y)] \neq g(\mathbb{E}[X], \mathbb{E}[Y])$. An exception is for linear functions: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Proposition (Expectation of product of independent r.v.) If X and Y are discrete independent random variables,

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)] \mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables) If X and Y are discrete independent random variables,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Continuous random variables

PDF, Expectation, Variance, CDF

Definition (Probability density function (PDF)) A probability density function of a r.v. X is a non-negative real valued function f_X that satisfies the following

- $\int_{-\infty}^{\infty} f_X(x) dx = 1$.
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$ for some random variable X .

Definition (Continuous random variable) A random variable X is continuous if its probability law can be described by a PDF f_X .

Remark Continuous random variables satisfy:

- For small $\delta > 0$, $\mathbb{P}(a \leq X \leq a + \delta) \approx f_X(a) \delta$.
- $\mathbb{P}(X = a) = 0, \forall a \in \mathbb{R}$.

Definition (Expectation of a continuous random variable) The expectation of a continuous random variable is

$$\mathbb{E}[X] \triangleq \int_{-\infty}^{\infty} x f_X(x) dx.$$

assuming $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.

Properties (Properties of expectation)

- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.
- $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$.
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.

Definition (Variance of a continuous random variable) Given a continuous random variable X with $\mu = \mathbb{E}[X]$, its variance is

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx.$$

It has the same properties as the variance of a discrete random variable.

Example (Uniform continuous random variable) A Uniform continuous random variable X between a and b , with $a < b$, ($X \sim \text{Uni}(a, b)$) has PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

We have $\mathbb{E}[X] = \frac{a+b}{2}$ and $\text{Var}(X) = \frac{(b-a)^2}{12}$.