10-315 Notes
Math Background

Carnegie Mellon University
Machine Learning Department

Contents

1 Linear algebra

Most of the linear algebra listed here should be prerequisite material for you. The exceptions might be vector and matrix norms and any notational changes.

1.1 Notation

Matrix notation: $A \in \mathbb{R}^{N \times M}$:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix}$$

Summation notation:

Matrix multiplication with $A \in \mathbb{R}^{M \times K}, B \in \mathbb{R}^{K \times N}, C \in \mathbb{R}^{M \times N}$. If $C = AB$, then $C_{i,j} = \sum_{k=1}^{K} a_{i,k} b_{k,j}$. The number of columns in A must match the number of rows in B. C will have the same number of rows as A and the same number of columns as B.

Vector notation: $\mathbf{v} \in \mathbb{R}^N$, in this course, we'll assume all vectors are column vectors unless specified otherwise:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}$$

$\mathbf{u} \in \mathbb{R}^{M \times 1}$	Column vector of length M
$\mathbf{v} \in \mathbb{R}^{1 \times M}$	Row vector of length M
$\mathbf{z} \in \mathbb{R}^{M}$	Ambiguous. In this course, assume column vector unless stated otherwise.

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_M \end{bmatrix}$$

1.2 Linear systems of equations

Consider matrix $A \in \mathbb{R}^{N \times M}$ and vectors $\mathbf{v} \in \mathbb{R}^M$ and $\mathbf{u} \in \mathbb{R}^N$. The following is a linear system of equations with N equations and M unknowns, $v_1, \ldots, v_M$:

$$
\begin{aligned}
\mathbf{u} &= A\mathbf{v} \\
u_1 &= a_{1,1}v_1 + a_{1,2}v_2 + \cdots + a_{1,M}v_M \\
u_2 &= a_{2,1}v_1 + a_{2,2}v_2 + \cdots + a_{2,M}v_M \\
&\vdots \\
u_1 &= a_{N,1}v_1 + a_{N,2}v_2 + \cdots + a_{N,M}v_M
\end{aligned}
$$

underdetermined: If there are fewer equations than variables, the system is underdetermined and cannot have exactly 1 solution, it must have either infinitely many or no solutions.

overdetermined: A system with more equations than variables. An overdetermined system may have 1 solution, 0 solutions, or infinitely many solutions.

inconsistent: when the system of equations does not have a solution.

consistent: when the system of equations has at least one solution.

1.3 Vectors

dot product: $\mathbf{a}^T\mathbf{b} = a_1b_1 + a_2b_2 + ... + a_Mb_M$ for two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^M$. It is the sum of the products of corresponding entries of the two vectors.

inner product (more general than dot product): is a way to multiply vectors, resulting in a scalar. Let $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be vectors and let α be a scalar. Then, the inner product satisfies the following properties: $\backslash\ \langle \mathbf{u} + \mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{u}, \mathbf{w}\rangle + \langle \mathbf{v}, \mathbf{w}\rangle \backslash\ \langle \alpha\mathbf{v}, \mathbf{w}\rangle = \alpha\langle \mathbf{v}, \mathbf{w}\rangle \backslash\ \langle \mathbf{v}, \mathbf{w}\rangle = \langle \mathbf{w}, \mathbf{v}\rangle \backslash\ \langle \mathbf{v}, \mathbf{v}\rangle \geq 0$ and $\langle \mathbf{v}, \mathbf{v}\rangle = 0$ if and only if $\mathbf{v} = 0$

outer product: Outer product of vectors $\mathbf{u}$ and $\mathbf{v}$ is $Y = \mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^T$, and $Y_{i,j} = u_iv_j$.

magnitude: Magnitude (length) of vector $\mathbf{u}$ is $|\mathbf{u}| = \|\mathbf{u}\|_2 = \left(\sum_i u_i^2\right)^{\frac{1}{2}}$.

L2 norm: Also known as Euclidean norm $\|\mathbf{v}\|_2 = \left(\sum_i v_i^2\right)^{\frac{1}{2}} = \left(\mathbf{v}^T\mathbf{v}\right)^{\frac{1}{2}}$

L1 norm: The sum of absolute values of the entries of the vector $\|\mathbf{v}\|_1 = \sum_i |v_i|$

L0 "norm": Number of non-zero entries in a vector (not technically a norm) $\|\mathbf{v}\|_0 = \sum_i |v_i|^0$, where 0^0 is defined as being equal to zero.

p-norm: $\|\mathbf{v}\|_p = \left(\sum_i |v_i|^p\right)^{\frac{1}{p}}$ (Only a norm for $p \geq 1$).

span: Set of all linear combinations of a set of vectors. For example, given a set of vectors $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}, \mathbf{v_i} \in \mathbb{R}^M$, $span(\mathcal{S}) = \{\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 \mid \alpha_i \in \mathbb{R}\}$. Span is an example of a vector space.

vector space: Span of a set of vectors is an example of a vector space. Vector space is a more general term for the result of combining a set of vectors with addition and scalar multiplication. For example, if we changed span to include multiplication by complex scalars, that would be a different vector space.

linearly dependent: A vector, $\mathbf{u}$, is linearly dependent with a set of vectors, $\mathcal{S} = \{\mathbf{v}_k\}_{k=1}^K$, if it is possible to represent $\mathbf{u}$ as a linear combination of vectors in $\mathcal{S}$. For example, if $\mathbf{u} = 3\mathbf{v}_1 - 3.5\mathbf{v}_3$, then $\mathbf{u}$ is linearly dependent with $\mathcal{S}$.

A set of vectors in linearly dependent if any of the vectors in the set can be represented by a linear combination of the remaining vectors in the set.

linearly independent: A vector, $\mathbf{u}$ is linearly independent from a set of vectors, $\mathcal{S} = \{\mathbf{v}_k\}_{k=1}^K$, if it is not possible to represent $\mathbf{u}$ as a linear combination of vectors in $\mathcal{S}$.

A set of vectors is linearly independent if no single vector in the set can be represented by a linear combination of the remaining vectors in the set.

1.4 Matrices

identity matrix: A matrix with all ones on the diagonal and zeros elsewhere. Represented as I, or more specifically I_N, where the N indicates that it is an $N \times N$ identity matrix. $I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

matrix inverse: The inverse of matrix $A \in \mathbb{R}^{N \times N}$ is denoted A^{-1}. A^{-1} is also an $N \times N$ matrix. The inverse of a square, $N \times N$ matrix will exist if the matrix is full rank, i.e., the column rank and row rank is N. If the inverse of a square matrix exists then $A^{-1}A = AA^{-1} = I$, where I is the $N \times N$ identity matrix.

column rank of a matrix: the maximal number of linearly independent vectors among the column vectors in a given matrix. This is also the dimensionality of the vector space spanned by the column vectors of the matrix.

row rank of a matrix: the maximal number of linearly independent vectors among the row vectors in a given matrix. This is also the dimensionality of the vector space spanned by the row vectors of the matrix.

rank: the row rank and column rank of a matrix are always equal, so there are often just referred to as matrix "rank".

full rank: A matrix is full rank is the rank is equal to the minimum of its number of rows and number of columns. If a matrix is square and full rank then the inverse of that matrix exists.

singular matrix: A square matrix is singular if it is not full rank and thus it's inverse doesn't exist.

Frobenius norm of a matrix: Basically the L2 norm if we were to flatten the matrix into a vector. For matrix $A \in \mathbb{R}^{N \times M}$,

$$\|A\|_F = \left(\sum_{i=1}^{N} \sum_{j=1}^{M} {a_{i,j}}^2 \right)^{\frac{1}{2}}$$

$$\|A\|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} {a_{i,j}}^2$$

2 Multivariate calculus

While you may not have explicitly learned multivariate calculus, if very much just builds on top of normal (scalar) calculus. In multivariate calculus, we are just shifting to working with functions that have multiple inputs variables and potentially multiple outputs.

2.1 Partial derivatives

A **partial derivative** is when we take the derivative of a function f with respect to one of its many input variables. Notation-wise, you'll see it written as $\frac{\partial}{\partial z} f(x, z)$ or $\frac{\partial f}{\partial z}$. (It could also be written as $f_z(x, z)$, but we won't use that in this course.)

When we take the partial derivative with respect to one variable, we just hold all other variables constant.

For example:

$$f(x, z) = 2x^3 z^5 \tag{1}$$
$$\frac{\partial f}{\partial x} = 6x^2 z^5 \tag{2}$$
$$\frac{\partial f}{\partial z} = 10x^3 z^4 \tag{3}$$
$$\tag{4}$$

You can think of linear algebra as having many individual variables. Take, for example, the L2 norm squared of $\mathbf{x} \in \mathbb{R}^3$:

$$f(\mathbf{x}) = \|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \sum_i {x_i}^2 = {x_1}^2 + {x_2}^2 + {x_3}^2 \tag{5}$$
$$f(x_1, x_2, x_3) = \|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \sum_i {x_i}^2 = {x_1}^2 + {x_2}^2 + {x_3}^2 \tag{6}$$
$$\frac{\partial f}{\partial x_1} = 2x_1 \tag{7}$$
$$\frac{\partial f}{\partial x_2} = 2x_2 \tag{8}$$
$$\frac{\partial f}{\partial x_3} = 2x_3 \tag{9}$$
$$\tag{10}$$

2.2 Gradients

Given a scalar function with vector input, $f : \mathbb{R}^M \to \mathbb{R}$, $f(\mathbf{x}) = f(x_1, ..., x_M)$, the **gradient** is a column vector where the i-th entry is the partial derivative of the function with respect to the i-th input entry in the input vector.

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_M} \end{bmatrix}$$

We call this the gradient of f with respect to $\mathbf{x}$. The $\mathbf{x}$ in the subscript is redundant if $\mathbf{x}$ is the only argument to f and would typically be dropped, $\nabla f(\mathbf{x})$.

Using the same example from above, the L2 norm squared of $\mathbf{x} \in \mathbb{R}^3$:

$$f(\mathbf{x}) = \|\mathbf{x}\|^2 = \begin{bmatrix} {x_1}^2 \\ {x_2}^2 \\ {x_3}^2 \end{bmatrix} \tag{11}$$

$$\frac{\partial f}{\partial x_1} = 2x_1 \tag{12}$$

$$\frac{\partial f}{\partial x_2} = 2x_2 \tag{13}$$

$$\frac{\partial f}{\partial x_3} = 2x_3 \tag{14}$$

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{bmatrix} = 2\mathbf{x} \tag{15}$$

3 Optimization notation

We can formalize an optimization problem with the following form:

$$y^* = \min_{x \in \mathcal{X}} f(x)$$

where

- $f : \mathbb{R}^k \to \mathbb{R}$ is known as the objective function that we are trying to minimize
- $\mathcal{X} \subset \mathbb{R}^k$ is the set of feasible inputs that we are trying to minimize f over
- y^* is the smallest value of $f(x)$ for all possible values $x \in \mathcal{X}$ (which is why we have a min in the formulation)

For all possible values x in the set $\mathcal{X}$ and return **the** x **corresponding to** the output of $f(x)$ that has the smallest value (i.e. return the argument, not the value of the function):

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} f(x)$$

For example, suppose that we wanted to minimize the objective function $f(x) = 3(x-5)^2 - 200$:

$$y^* = \min_{x \in \mathbb{R}} 3(x-5)^2 - 200 \tag{16}$$

$$= -200 \tag{17}$$

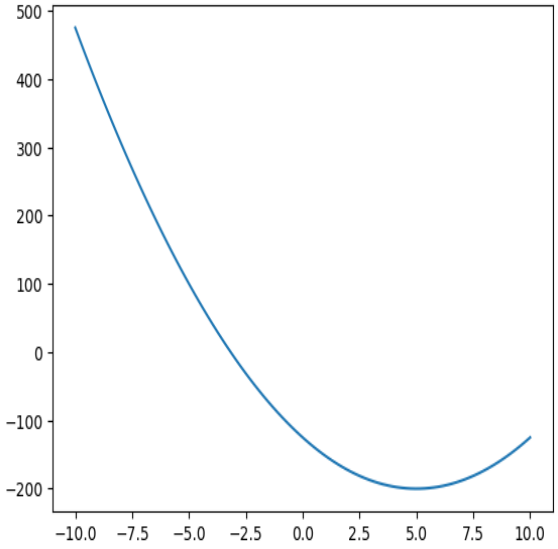
$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}} 3(x-5)^2 - 200 \tag{18}$$

$$= 5 \tag{19}$$

$$\tag{20}$$

Plot for the above example:

```
# Plot for above example
def f(x):
    return 3*(x-5)**2 - 200
    x_grid = np.linspace(-10, 10, 100)
    y_grid = f(x_grid)
    plt.plot(x_grid, y_grid, '-');
```



4 Probability

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

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.

4.2 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.

4.2.1 Probability (of an event)

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

$P(X = x)$: 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.

$P(X = heads)$: Likely an abuse of notation for a discrete random variable that maps the outcome "heads" one-to-one to a value. For example, if $X(heads) = 1$ and nothing else maps to 1, this shorthand notation could be more precisely written as $P(X = 1)$. Writing just $P(heads)$ 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.

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

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

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

4.2.2 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 $P(X = x) \forall x \in range(X)$, 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 $p(x \mid y = 1)$, which could be written more precisely as $p_{X,Y}(x \mid 1)$.

$p(x; \theta)$ or $p(x \mid \theta)$ or $p(x)$ or $p_\theta(x)$: (marginal) distribution (pmf/pdf) defined by parameter(s) θ. The $\mid$ 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 $\mid$ 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.

$p(x \mid y)$ or $p_{X|Y}(x \mid y)$: 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.

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

$x \sim DistributionName(\theta)$: 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.

4.2.3 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 $\mathcal{D} = \{z^{(1)}, z^{(2)}, ..., z^{(n)}\}$, where $n = |\mathcal{D}|$, the number of elements in the set. This will become helpful when we get into a set of training examples, $\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(n)}, y^{(n)})\}$. 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.

4.3 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.

4.3.1 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:

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

4.3.2 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:

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

4.3.3 Basic Rules

Definition of Conditional Probability:

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

Chain Rule:

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

Chain Rule (with more variables):

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

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

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

Bayes' Law:

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

Normalization:

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

$$P(Y \mid X) \propto P(X,Y)$$

$$P(Y \mid X) = \alpha P(X,Y) \qquad \textit{Note this difference between } \propto \textit{ and } \alpha$$

$$\alpha = \frac{1}{P(X)} = \frac{1}{\sum_y P(X, Y = y)}$$

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:

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

4.3.4 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 |:

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

4.3.5 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:

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

4.3.6 Gaussian Distribution

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

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

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

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

4.3.7 Law of Total Expectation

For two given random variables X, Y:

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

If Y is discrete, this becomes:

$$\mathbb{E}[X] = \sum_{y} \mathbb{E}[X \mid Y = y] P(Y = y)$$

4.3.8 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.

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

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

4.3.9 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.