Math Formulas for Machine Learning: All taken from Introduction to Artificial Neural Networks with Applications in Python by Dr Sebastian Raschka

Method · April 2020

DOI: 10.13140/RG.2.2.25753.16485

1 author:

Burhan Rashid Hussein

Universiti Brunei Darussalam

All taken from Introduction to Artificial Neural Networks with Applications in Python Sebastian Raschka

A.1 Sets and Intervals

$\mathbb{Z}$	set of integers, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{N}$	set of natural numbers, $\{0, 1, 2, 3, \ldots\}$
$\mathbb{N}^+$	set of natural numbers excluding zero, $\{1, 2, 3, \ldots\}$
$\mathbb{R}$	set of real numbers
$\in$	*element of* symbol; for example, $x \in A$ translates to "x is an element of set A'
$\notin$	*not an element of* symbol
$\varnothing$	null set, empty set
$A \cup B$	union of two sets, A and B
$A \cap B$	intersection of two sets, A and B
$A \subseteq B$	A is a subset of B or included in B
$A \Delta B$	symmetric difference between two sets A and B
$\|A\|$	cardinality of a set A (number of elements in a set A)
(a, b)	open interval from a to b, excluding a and b
$[a, b]$	closed interval from a to b, including a and b
$[a, b)$	half-open interval from a to b, including a but not b
$(a, b]$	half-open interval from a to b, including b but not a

A.2 Sequences

$\sum_{i=1}^{n} x_i$	summation of an indexed variable x_i, defined as $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$
$\prod_{i=1}^{n} x_i$	product over an indexed variable x_i, defined as $\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \ldots \cdot x_n$

A.3 Functions

$f : A \to B$	function f with domain A and codomain B
$(g \circ f)(x)$	composition of two functions g and f alternative form: $g[f(x)]$
$f^{-1}(x)$	inverse of a function f, such that $f(y) = x$ if f^{-1} stands for y
$\|x\|$	absolute value of x; for example, $\|-2\| = 2$
$\log_b$	base-b logarithm
$\log$	natural logarithm (base-e logarithm)
$n!$	*n-factorial*, where $0! = 1$ and $n! = n(n-1)(n-2)\cdots 2 \cdot 1$ for $n > 0$
$\binom{n}{k}$	binomial coefficient ("n choose k"); $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ for $0 \leq k \leq n$
$\arg\max f(x)$	the x value that makes $f(x)$ as large as possible
$\arg\min f(x)$	the x value that makes $f(x)$ as small as possible

Source: https://sebastianraschka.com/resources/math-for-ml.htmlhttps://sebastianraschka.com/resources/math-for-ml.html

Compiled by Burhan Rashid Hussein

A.4 Linear Algebra

x scalar (lower-case italics notation)

$\mathbf{x}$ column vector (lower-case bold notation) or $n \times 1$-matrix

$\mathbf{a} \cdot \mathbf{b}$ dot product of two vectors, $\mathbf{a}$ and $\mathbf{b}$;
if $\mathbf{a}$ and $\mathbf{b}$ are $n \times 1$-matrices, also written as $\mathbf{a}^T\mathbf{b}$;
$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b} = \sum_i a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$

$\mathbf{X}$ $m \times n$-matrix (upper-case bold notation)

X 3D-tensor (upper-case italics notation)

$\mathbb{R}^n$ real coordinate space, written as a column vector with length n

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$\mathbf{x}^T$ transpose of a $n \times 1$-matrix

$$\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T$$

$\|\mathbf{x}\|_p$ L^p norm, vector p-norm,
$\|\mathbf{x}\|_p = \left(|x_1^p| + |x_2^p| + \cdots + |x_n^p|\right)^{1/p}$

$\|\mathbf{x}\|_\infty$ L^∞ norm, max norm; largest absolute value of a vector
$\|\mathbf{x}\|_\infty = \max_i |x_i|$

$\|\mathbf{x}\|$ vector norm, L^2-norm, $\|\mathbf{x}\| = \|\mathbf{x}\|_2$
$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

$\mathbf{A}_{i,:}$ ith row of matrix $\mathbf{A}$

$\mathbf{A}_{:,j}$ jth column of matrix $\mathbf{A}$

$\mathbf{A}^T$ transpose of a matrix, matrix element $\mathbf{A}_{i,j}$ becomes $\mathbf{A}^T_{j,i}$

for example, $\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$

I_n $n \times n$ identity matrix

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$\mathbf{A}^{-1}$ inverse of a matrix $\mathbf{A}$, such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \boldsymbol{I}$

$\operatorname{tr} \mathbf{A}$ trace of a matrix $\mathbf{A}$ (sum of the diagonal elements)
$\operatorname{tr} \mathbf{A} = \sum_{i=1}^{n} \mathbf{A}_{i,i}$

$\det \mathbf{A}$ determinant of a matrix $\mathbf{A}$

$\operatorname{diag}(a_1, a_2, ..., a_n)$ diagonal matrix, matrix whose
diagonal have the values $a_1, a_2, ..., a_n$ and all other elements are zero

$\mathbf{A} \odot \mathbf{B}$ Hadamard product, element-wise matrix multiplication

A.5 Calculus

$\lim_{x \to a} f(x)$ limit of $f(x)$ as x approaches a

$\lim_{x \to a-} f(x)$ limit of $f(x)$ as x approaches a from the left

$\lim_{x \to a+} f(x)$ limit of $f(x)$ as x approaches a from the right

$\frac{df}{dx}$ derivative of f

$\frac{d^n f}{dx^n}$ n-th derivative of f

$\frac{\partial f}{\partial x}$ partial derivative of $f(x, y, ...)$
with respect to variable x, where x is a scalar

∇f gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x_1, x_2, ..., x_n) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Δf Laplacian of a function $f : \mathbb{R}^n \to \mathbb{R}$
$\Delta f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}$

Hf Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$

$$Hf = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

$\frac{\partial f_j}{\partial x_i}$ partial derivative of component function f_j and the variable x_j, where $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, such that

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \frac{\partial \mathbf{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \frac{\partial f_2}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix}$$

$D\mathbf{f}$ Jacobian matrix of $\mathbf{f}$.

$$D\mathbf{f} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

$\int f(x)dx$ indefinite integral of f (derivative of F) with $f : \mathbb{R} \to \mathbb{R}$

$\int_a^b f(x)dx$ definite integral of f (derivative of F) with $f : \mathbb{R} \to \mathbb{R}$

A.6 Probability and Statistics

$P(A \cap B)$	probability that event A and B occur
$P(A \cup B)$	probability that event A or B occurs
$P(A \mid B)$	conditional probability of A given B
$E(X), \mu_X$	expected value (mean) of a random variable X $E(X) = \sum_{i=1}^{\infty} p_i x_i$ for a discrete random variable X with values $x_1, x_2, \ldots$ and probabilities $p_1, p_2, \ldots$. $E(X) = \int_{-\infty}^{\infty} x f(x) dx$ for a continuos random variable and probability density function $f(x)$.
$\bar{X}$	sample average of numerical data $X_1, ..., X_n$ $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
$\text{var}(X), \sigma_x^2$	variance of a random variable X $\text{var}(X) = E[(X - \mu_X)^2] = E(X^2) - E(X)^2$
s_X^2	sample variance of numerical data $X_1, ..., X_n$ $s_X^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$

std(X), σ_x	standard deviation of a random variable, square root of the variance
s_X	sample standard deviation, the square root of the sample variance s_X^2
$\text{cov}(X, Y)$	covariance of two random variables X and Y $\text{cov}(XY) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$
s_{XY}	sample covariance of numerical data $X_1, ..., X_n$, and $Y_1, ..., Y_n$ $s_{XY} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$
$\text{corr}(X, Y)$	correlation coefficient of two random variables X and Y, $\text{corr}(X, Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
$H(X)$	entropy of a random variable X discrete: $H(X) = -\sum_x P(X = x) \log_b P(X = x)$ continuous: $H(X) = -\int_{-\infty}^{\infty} f(x) \log_b f(x) dx$
PMF	probability mass function of a discrete random variable, $f(x) = P(X = x)$
CDF	cumulative distribution function of a continuous random variable, $F(x) = P(X \leq x)$
PDF	probability density function of a continuous random variable, $P(X \in [a, b]) = \int_a^b f(x) dx$
$X \sim D$	random variable X has a distribution D
$\hat{\theta}$	estimator of a parameter θ
$N(x, \mu, \sigma^2)$	normal (Gaussian) distribution over x with mean μ and variance σ^2

Source: https://sebastianraschka.com/resources/math-for-ml.htmlhttps://sebastianraschka.com/resources/math-for-ml.html

Compiled by Burhan Rashid Hussein

A.7 Numbers

e	Euler's number, mathematical constant approximated by 2.71828
π	"pi", mathematical constant approximated by 3.14159
∞	infinity symbol
1.234×10^5	scientific notation for $123,400$
or $1.234E05$	
$<$	*less than* sign, for example, $x < 10$ means that x is smaller than 10
$\ll$	*much less than* sign
$>$	*greater than* sign, for example, $x > 10$ means that x is larger than 10
$\gg$	*much greater than* sign
$\ll$	*much less than* sign

A.8 Approximation

$\approx$	approximate equality, for instance, $e \approx 2.71828$ is the approximation of Euler's number
$f(x) \sim g(x)$	symbol to assert that the ratio of two functions approaches 1 $\lim\limits_{x \to 0} \frac{f(x)}{g(x)} = 1$, if x is small $\lim\limits_{x \to \infty} \frac{f(x)}{g(x)} = 1$, if x is large
$f(x) \propto g(x)$	the two functions $f(x)$ and $g(x)$ are proportional to each other
$T(n) \in O(n^2)$	*big-O* notation, an algorithm is asymptotically bounded by n^2; an algorithm has an order of n^2 time complexity

A.9 Logic

$\Rightarrow$	*implication* operator for example, $A \Rightarrow B$ translates to "if A implies B" or "if A then B" (or "B only if A")
$\Leftrightarrow$	*equality* operator (*if and only if (iff)*) for example, $A \Leftrightarrow B$ translates to "A if and only if B" or "if A then B and if B then A"
$\wedge$	logical conjunction, *and* for example, $A \wedge B$ means "A and B"
$\vee$	logical (inclusive) disjunction, *or* for example, $A \vee B$ means "A or B"
$\neg$	negation, *not* for example, $\neg A$ means "not A" or "if A is true then $\neg A$ is false" and vice versa
$\forall$	universal quantifier, means *for all* for example, "$\forall x \in \mathbb{R}, x > 1$" translates to "for all real numbers x, x is greater than one"
$\exists$	existential quantifier, means *there exists* for example, "$\exists x \in A, f(x)$" translates to "there is an element in set A for which the predicate $f(x)$ holds true"

Source: https://sebastianraschka.com/resources/math-for-ml.htmlhttps://sebastianraschka.com/resources/math-for-ml.html

Compiled by Burhan Rashid Hussein