

# 机器学习的数学笔记

X

# 目录

# Notation

This section provides a concise reference describing notation used throughout this document. If you are unfamiliar with any of the corresponding mathematical concepts, ? describe most of these ideas in chapters 2–4.

## Numbers and Arrays

$a$	A scalar (integer or real)
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\mathbf{A}$	A tensor
$\mathbf{I}_n$	Identity matrix with $n$ rows and $n$ columns
$\mathbf{I}$	Identity matrix with dimensionality implied by context
$\mathbf{e}^{(i)}$	Standard basis vector $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position $i$
$\text{diag}(\mathbf{a})$	A square, diagonal matrix with diagonal entries given by $\mathbf{a}$
$a$	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

## Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and $n$
$[a, b]$	The real interval including $a$ and $b$
$(a, b]$	The real interval excluding $a$ but including $b$
$\mathbb{A} \setminus \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathcal{G}$	A graph
$Pag(x_i)$	The parents of $x_i$ in $\mathcal{G}$

## Indexing

$a_i$	Element $i$ of vector $\mathbf{a}$ , with indexing starting at 1
$a_{-i}$	All elements of vector $\mathbf{a}$ except for element $i$
$A_{i,j}$	Element $i, j$ of matrix $\mathbf{A}$
$\mathbf{A}_{i,:}$	Row $i$ of matrix $\mathbf{A}$
$\mathbf{A}_{:,i}$	Column $i$ of matrix $\mathbf{A}$
$A_{i,j,k}$	Element $(i, j, k)$ of a 3-D tensor $\mathbf{A}$
$\mathbf{A}_{::,i}$	2-D slice of a 3-D tensor
$\mathbf{a}_i$	Element $i$ of the random vector $\mathbf{a}$

## Linear Algebra Operations

$\mathbf{A}^\top$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^+$	Moore-Penrose pseudoinverse of $\mathbf{A}$
$\mathbf{A} \odot \mathbf{B}$	Element-wise (Hadamard) product of $\mathbf{A}$ and $\mathbf{B}$
$\det(\mathbf{A})$	Determinant of $\mathbf{A}$

## Calculus

$\frac{dy}{dx}$	Derivative of $y$ with respect to $x$
$\frac{\partial y}{\partial x}$	Partial derivative of $y$ with respect to $x$
$\nabla_x y$	Gradient of $y$ with respect to $\mathbf{x}$
$\nabla_{\mathbf{X}} y$	Matrix derivatives of $y$ with respect to $\mathbf{X}$
$\nabla_{\mathbf{X}} y$	Tensor containing derivatives of $y$ with respect to $\mathbf{X}$
$\frac{\partial f}{\partial \mathbf{x}}$	Jacobian matrix $\mathbf{J} \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$
$\nabla_x^2 f(\mathbf{x})$ or $\mathbf{H}(f)(\mathbf{x})$	The Hessian matrix of $f$ at input point $\mathbf{x}$
$\int f(\mathbf{x}) d\mathbf{x}$	Definite integral over the entire domain of $\mathbf{x}$
$\int_{\mathbb{S}} f(\mathbf{x}) d\mathbf{x}$	Definite integral with respect to $\mathbf{x}$ over the set $\mathbb{S}$

## Probability and Information Theory

$a \perp b$	The random variables $a$ and $b$ are independent
$a \perp b \mid c$	They are conditionally independent given $c$
$P(a)$	A probability distribution over a discrete variable
$p(a)$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$a \sim P$	Random variable $a$ has distribution $P$
$\mathbb{E}_{x \sim P}[f(x)]$ or $\mathbb{E}f(x)$	Expectation of $f(x)$ with respect to $P(x)$
$\text{Var}(f(x))$	Variance of $f(x)$ under $P(x)$
$\text{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $P(x)$
$H(x)$	Shannon entropy of the random variable $x$
$D_{\text{KL}}(P \parallel Q)$	Kullback-Leibler divergence of $P$ and $Q$
$\mathcal{N}(x; \mu, \Sigma)$	Gaussian distribution over $x$ with mean $\mu$ and covariance $\Sigma$

## Functions

$f : \mathbb{A} \rightarrow \mathbb{B}$	The function $f$ with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions $f$ and $g$
$f(\mathbf{x}; \boldsymbol{\theta})$	A function of $\mathbf{x}$ parametrized by $\boldsymbol{\theta}$ . (Sometimes we write $f(\mathbf{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of $x$
$\sigma(x)$	Logistic sigmoid, $\frac{1}{1 + \exp(-x)}$
$\zeta(x)$	Softplus, $\log(1 + \exp(x))$
$\ \mathbf{x}\ _p$	$L^p$ norm of $\mathbf{x}$
$\ \mathbf{x}\ $	$L^2$ norm of $\mathbf{x}$
$x^+$	Positive part of $x$ , i.e., $\max(0, x)$
$\mathbf{1}_{\text{condition}}$	is 1 if the condition is true, 0 otherwise

Sometimes we use a function  $f$  whose argument is a scalar but apply it to a vector, matrix, or tensor:  $f(\mathbf{x})$ ,  $f(\mathbf{X})$ , or  $f(\mathbf{X})$ . This denotes the application of  $f$  to the array element-wise. For example, if  $\mathbf{C} = \sigma(\mathbf{X})$ , then  $C_{i,j,k} = \sigma(X_{i,j,k})$  for all valid values of  $i$ ,  $j$  and  $k$ .



## Datasets and Distributions

$p_{\text{data}}$	The data generating distribution
$\hat{p}_{\text{data}}$	The empirical distribution defined by the training set
$\mathbb{X}$	A set of training examples
$\mathbf{x}^{(i)}$	The $i$ -th example (input) from a dataset
$y^{(i)}$ or $\mathbf{y}^{(i)}$	The target associated with $\mathbf{x}^{(i)}$ for supervised learning
$\mathbf{X}$	The $m \times n$ matrix with input example $\mathbf{x}^{(i)}$ in row $\mathbf{X}_{i,:}$

# Chapter 1

## 逻辑回归

This document is an example of how to use the accompanying files as well as some commentary on them. The files are `math_commands.tex` and `notation.tex`. The file `math_commands.tex` includes several useful  $\text{\LaTeX}$  macros and `notation.tex` defines a notation page that could be used at the front of any publication.

We developed these files while writing ?. We release these files for anyone to use freely, in order to help establish some standard notation in the deep learning community.

你好

假设

### 1.1 Examples

We include this section as an example of some  $\text{\LaTeX}$  commands and the macros we created for the book.

Citations that support a sentence without actually being used in the sentence should appear at the end of the sentence using `citep`:

One of the simplest and most common kinds of parameter norm penalty is the squared  $L^2$  parameter norm penalty commonly known as **weight decay**. In other academic communities,  $L^2$  regularization is also known as **ridge regression** or **Tikhonov regularization**.

To refer to a figure, use either `figref` or `Figref` depending on whether you want to capitalize the resulting word in the sentence.

See figure ?? for an example of a how to include graphics in your document. Figure ?? shows how to include graphics in your document.

Similarly, you can refer to different sections of the book using `partref`, `Partref`, `secref`, `Secref`, etc.

You are currently reading section ??.

## Acknowledgments

We thank Catherine Olsson and Úlfar Erlingsson for proofreading and review of this manuscript.