Machine Learning Notation

Shan-Hung Wu

1 Numbers & Arrays

a	A scalar (integer or real)
A	A scalar constant
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\mathsf{\mathbf{A}}$	A tensor
$\boldsymbol{I}_n$	The $n \times n$ identity matrix
$\boldsymbol{D}$	A diagonal matrix
$\mathrm{diag}(\boldsymbol{a})$	A square, diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

2 Sets & Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0,1\}$	The set containing 0 and 1
$\{0,1,\cdots,n\}$	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
$\mathbb{A}\backslash\mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathcal{G}$	A graph whose each vertex $\mathbf{x}^{(i)}$ denotes a random variable and edge denotes conditional dependency (directed) or correlation (undirected)
$\mathrm{Pa}(\mathbf{x}^{(i)})$	The parents of a vertex $\mathbf{x}^{(i)}$ in $\mathcal{G}$

3 Indexing

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

4 Functions

$f : \mathbb{A} \to \mathbb{B}$	A function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of functions f and g
$f(\boldsymbol{x};\boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$ (with $\boldsymbol{\theta}$ omitted sometimes)
$\ln x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, i..e, $(1+\exp(-x))^{-1}$
$\zeta(x)$	Softplus, $\ln(1+\exp(x))$
$\\|\boldsymbol{x}\\|_p$	L^p norm of $\boldsymbol{x}$
$\\|\boldsymbol{x}\\|$	L^2 norm of $\boldsymbol{x}$
x^+	Positive part of x, i.e., $\max(0,x)$
$1(x;\mathrm{cond})$	The indicator function of x: 1 if the condition is true, 0 otherwise
$g[f;x]$	A functional that maps f to $f(x)$

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

5 Calculus

$f'(a)$ or $\frac{df}{dx}(a)$	Derivative of $f : \mathbb{R} \to \mathbb{R}$ at input point a
$\frac{\partial f}{\partial x_i}(\boldsymbol{a})$	Partial derivative of $f : \mathbb{R}^n \to \mathbb{R}$ with respect to x_i at input $\boldsymbol{a}$
$\nabla f(\boldsymbol{a}) \in \mathbb{R}^n$	Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at input $\boldsymbol{a}$
$\nabla f(\boldsymbol{A}) \in \mathbb{R}^{m\times n}$	Matrix derivatives of $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ at input $\boldsymbol{A}$
$\nabla f(\mathsf{\mathbf{A}})$	Tensor derivatives of f at input $\mathsf{\mathbf{A}}$
$\boldsymbol{J}(\boldsymbol{f})(\boldsymbol{a}) \in \mathbb{R}^{m\times n}$	The Jacobian matrix of $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ at input $\boldsymbol{a}$
$\nabla^2 f(\boldsymbol{a})$ or $\boldsymbol{H}(f)(\boldsymbol{a}) \in \mathbb{R}^{n\times n}$	The Hessian matrix of $f : \mathbb{R}^n \to \mathbb{R}$ at input point $\boldsymbol{a}$
$\int f(\boldsymbol{x})d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_{\mathbb{S}} f(\boldsymbol{x})d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$

6 Linear Algebra

$\boldsymbol{A}^\top$	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^\dagger$	Moore-Penrose pseudo-inverse of $\boldsymbol{A}$
$\boldsymbol{A}\odot\boldsymbol{B}$	Element-wise (Hadamard) product of $\boldsymbol{A}$ and $\boldsymbol{B}$
$\det(\boldsymbol{A})$	Determinant of $\boldsymbol{A}$
$\mathrm{tr}(\boldsymbol{A})$	Trace of $\boldsymbol{A}$
$\boldsymbol{e}^{(i)}$	The i-th standard basis vector (a one-hot vector)

7 Probability & Info. Theory

$\mathrm{a}\perp\mathrm{b}$	Random variables a and b are independent
$\mathrm{a}\perp\mathrm{b}\,\vert\,\mathrm{c}$	They are conditionally independent given c
$\Pr(a\,\vert\,b)$ or $\Pr(\mathrm{a}\,\vert\,\mathrm{b})$	Shorthand for the probability $\Pr(\mathrm{a}=a\,\vert\,\mathrm{b}=b)$
$P_{\mathrm{a}}(a)$	A probability mass function of the discrete random variable a
$p_{\mathrm{a}}(a)$	A probability density function of the continuous random variable a
$\mathrm{P}(\mathrm{a}=a)$	Either $P_{\mathrm{a}}(a)$ or $p_{\mathrm{a}}(a)$
$\mathrm{P}(\theta)$	A probability distribution parametrized by θ
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathrm{x}\sim\mathrm{P}(\theta)$	Random variable x has distribution P
$\mathrm{E}_{\mathrm{x}\sim\mathrm{P}}[\mathrm{f}(\mathrm{x})]$	Expectation of f(x) with respect to P
$\mathrm{Var}[\mathrm{f}(\mathrm{x})]$	Variance of f(x)
$\mathrm{Cov}[\mathrm{f}(\mathrm{x}),\mathrm{g}(\mathrm{x})]$	Covariance of f(x) and g(x)
$\mathrm{H}(\mathrm{x})$	Shannon entropy of the random variable x
$\mathrm{D}_{\mathrm{KL}}(\mathrm{P}\Vert\mathrm{Q})$	Kullback-Leibler (KL) divergence from distribution Q to P

8 Machine Learning

$\mathbb{X}$	The set of training examples
N	Size of $\mathbb{X}$
$(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})$	The i-th example pair in $\mathbb{X}$ (supervised learning)
$\boldsymbol{x}^{(i)}$	The i-th example in $\mathbb{X}$ (unsupervised learning)
D	Dimension of a data point $\boldsymbol{x}^{(i)}$
K	Dimension of a label $\boldsymbol{y}^{(i)}$
$\boldsymbol{X}\in\mathbb{R}^{N\times D}$	Design matrix, where $\boldsymbol{X}_{i,:}$ denotes $\boldsymbol{x}^{(i)}$
$P(\mathbf{x},\mathbf{y})$	A data generating distribution
$\mathbb{F}$	Hypothesis space of functions to be learnt, i.e., a model
$C[f]$	A cost functional of $f\in\mathbb{F}$
$C(\theta)$	A cost function of θ parametrizing $f\in\mathbb{F}$
$(\boldsymbol{x}',\boldsymbol{y}')$	A testing pair
$\hat{\boldsymbol{y}}$	Label predicted by a function f, i.e., $\hat{\boldsymbol{y}}=f(\boldsymbol{x}')$ (supervised learning)

9 Typesetting

Section*	Section that can be skipped for the first time reading
Section**	Section for reference only (will not be taught)
[Proof]	Prove it yourself
[Homework]	You have homework