Symbols

In general, scalars are printed as normal letters, e. g. $\mathtt{p}, \mathtt{q}$; vectors are bolded, e. g. $\boldsymbol{\theta}$; and matrices are denoted by bold upper-case letters, e. g. $\boldsymbol{\Phi}$.

Latin letters

$\mathtt{a}$	action (RL)
$\boldsymbol{A}$	adjacency matrix
$\mathcal{A}$	action space (RL)
$\mathtt{C}$	inverse regularization parameter (SVMs), cluster
$\mathcal{C}$	set of classes
$\mathcal{C}_\mathtt{k}$	$\mathtt{k}$-th class
D	number of dimensions
$\boldsymbol{D}$	degree matrix
$\mathcal{D}$	data set, $\{\boldsymbol{x}^{(\mathtt{i})}, \mathtt{y}^{(\mathtt{i})}\}_{\mathtt{i}=1}^{\mathtt{n}}$
$\mathcal{D}_\mathtt{j}$	$\mathtt{j}$-th partition of data set $\mathcal{D}$
$\mathcal{E}$	set of edges of a graph
$\mathcal{F}$	feature space
$\mathcal{G}$	graph, goal state (RL)
$\mathtt{h}$ / $\mathtt{h}_{\boldsymbol{\theta}}$	hypothesis / model
$\mathtt{H}$	kernel function for kernel density estimation
$\boldsymbol{H}$	Hessian matrix (matrix of second derivatives)
$\mathcal{H}$	hypothesis space
$\mathtt{i}$	index over the data examples
$\boldsymbol{I}$	identity matrix
$\mathtt{j}$	index over the features
$\mathcal{J}$	loss function / error function

$\mathtt{k}$	index over classes, number of clusters / neighbors / base classifiers
$\boldsymbol{K}$	kernel matrix
$\mathcal{K}$	kernel function
ℓ	loss
$\boldsymbol{L}$	graph Laplacian matrix
$\mathcal{L}$	likelihood, lagrangian, loss
$\mathcal{LL}$	log-likelihood
$\mathtt{m}$	number of features ($\mathtt{j} = 1, \ldots, \mathtt{m}$)
$\mathtt{M}$	number of mixture components
$\mathcal{M}$	margin
$\mathtt{n}$	number of examples ($\mathtt{i} = 1, \ldots, \mathtt{n}$)
$\mathtt{N}$	node in a tree
$\mathtt{p}$	degree of a polynomial
$\mathtt{Q}$	true q value (RL)
$\widehat{\mathtt{Q}}$	approximate q value (RL)
$\mathtt{r}$	cost ratio, $\mathtt{c}_{\mathtt{fp}}/\mathtt{c}_{\mathtt{fn}}$, reward (RL)
$\mathcal{R}$	region
$\mathtt{s}$	state (RL)
$\mathtt{s}_0$	initial state (RL)
$\mathtt{s}'$	next state (RL)
$\mathcal{S}$	state space (RL)
$\mathtt{T}$	sequence length
ν	bin volume
$\boldsymbol{v}$	vector
$\mathtt{V}^{\pi}$	value function when using policy π (RL)

$\mathcal{V}$	vocabulary, set of vertices (nodes) of a graph
w	window size
$\boldsymbol{x}$	data instance
$\boldsymbol{x}^{(i)}$	i-th data instance
x_j	j-th feature of $\boldsymbol{x}$
$x_j^{(i)}$	j-th feature of the i-th data instance
$\boldsymbol{x}_\perp$	orthogonal projection of $\boldsymbol{x}$
$\widehat{\boldsymbol{x}}$	data instance with attached 1 entry, $[1\boldsymbol{x}]$
$\boldsymbol{X}$	design matrix / feature matrix / regressor matrix
$\mathcal{X}$	data input space
y	label
$\boldsymbol{y}$	label vector
z	center of radial basis function, activation

Greek letters

α	learning rate, responsibility (mixture models), model weight
β	precision $\equiv \sigma^{-1}$
γ	variance threshold (PCA), discount factor (RL)
δ	error gradient (neural networks), state transition function (RL)
ε	noise, error rate
Θ, θ	trainable parameters of the model
κ	number of classes ($k = 1, \ldots, \kappa$)
λ	regularization parameter, eigenvalue, lagrange multiplier
μ	mean of a distribution
ξ	slack variable (SVMs)

π	prior (mixture models), policy (RL)
π^*	optimal policy (RL)
σ	sigmoid, standard deviation
σ^2	variance
$\boldsymbol{\Sigma}$	covariance matrix
τ	trajectory (RL)
$\varphi(\boldsymbol{x})$	basis function of $\boldsymbol{x}$
Φ	design matrix with basis functions

Mathematical symbols ($\bullet$ is a placeholder)

$\mathbb{1}\{\dots\}$	indicator function
$\arg\max_x f(x)$	value of x which maximizes $f(x)$
$\arg\min_x f(x)$	value of x which minimizes $f(x)$
$\mathcal{Bern}(x\|\mu)$	bernoulli distribution, $\mu^x(1-\mu)^{1-x}$
$\mathcal{Bin}(m\|n,\mu)$	binomial distribution, $\binom{n}{m}\mu^m(1-m)^{n-m}$
$c(\mathcal{C}_i\|\mathcal{C}_j)$	cost for predicting class $\mathcal{C}_i$ instead of $\mathcal{C}_j$
$\mathcal{C}_{MAP}$	maximum aposteriori class
$d(\dots)$	distance metric
$\mathrm{dom}(f)$	domain of function f
$E(\mathcal{D})$	entropy of a data set, $-\sum_{c\in\mathcal{C}} p_c \log_2 p_c$
$\exp\{\dots\}$	exponential function
$\mathbb{E}\{x\}$	expectation of a random variable x
$\mathbb{E}_{x\sim p(x)}\{f(x)\}$	expectation of $f(x)$ where x is drawn from distribution $p(x)$
$g(x)$	activation function, sigmoid
$KL(p\|q)$	Kullback-Leibler divergence between p and q

$\max_{\mathtt{x}} \mathtt{f}(\mathtt{x})$	maximum value of $\mathtt{f}(\mathtt{x})$
$\min_{\mathtt{x}} \mathtt{f}(\mathtt{x})$	minimum value of $\mathtt{f}(\mathtt{x})$
$\mathcal{N}(\mathtt{x}\|\mu, \sigma)$	normal (Gaussian) distribution
$\mathcal{N}(\boldsymbol{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate normal (Gaussian) distribution
$\widehat{\mathtt{p}}(\mathtt{x})$	approximate probability of $\mathtt{x}$
$\mathtt{p}(\mathtt{x})$	probability of $\mathtt{x}$, marginal probability
$\mathtt{p}(\mathtt{x}, \mathtt{y})$	probability of $\mathtt{x}$ and $\mathtt{y}$, joint probability
$\mathtt{p}(\mathtt{x}\|\mathtt{y})$	probability of $\mathtt{x}$ given $\mathtt{y}$
$\mathtt{R}(\alpha_{\mathtt{i}}\|\boldsymbol{x})$	risk of a decision $\alpha_{\mathtt{i}}$ given $\boldsymbol{x}$
$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	set of positive real numbers (including 0)
$\mathtt{ReLU}(\mathtt{x})$	ReLU activation function
$\tanh(\mathtt{x})$	tangent hyperbolic activation function
$\mathrm{var}\{\ \mathrm{x}\ \}$	variance of a random variable $\mathtt{x}$
$\mathtt{x} \sim \mathtt{p}(\mathtt{x})$	$\mathtt{x}$ is distributed according to probability distribution $\mathtt{p}(\mathtt{x})$
$\bullet^{\intercal}$	transpose of $\bullet$
$\bullet^{-1}$	inverse of $\bullet$ (of a matrix for example)
$\bullet^{\#}$	pseudo-inverse of $\bullet$
$\boldsymbol{a}^{\intercal}\boldsymbol{b}$	dot product of $\boldsymbol{a}$ and $\boldsymbol{b}$
$\langle \bullet, \bullet \rangle$	dot product
$\boldsymbol{a}\boldsymbol{b}^{\intercal}$	outer product
$\lvert\ \bullet\ \rvert$	number of elements in the set $\bullet$
$\lVert\ \bullet\ \rVert$	vector norm
$\measuredangle$	angle
$\binom{\mathtt{n}}{\mathtt{k}}$	binomial coefficient, $\mathtt{n}$ choose $\mathtt{k}$

$\nabla_{\boldsymbol{\theta}}$	gradient with respect to $\boldsymbol{\theta}$ / nabla operator
∂	partial derivative
$\sum_{i=1}^{n} \cdots$	sum
$\prod_{i=1}^{n} \cdots$	product
$\oplus$	positive class
$\ominus$	negative class
$\overline{x}$	sample set mean of x
$\star$	convolution operator

Abbreviations

AUC	area-under-the-curve
CNN	convolutional neural network
fp	false positives
fn	false negatives
GRU	gated recurrent unit
i. i. d.	independent and identically distributed
Lasso	least absolute shrinkage and selection operator
LSTM	long short-term memory
MAE	mean absolute error
MDP	Markov decision process
MLE	maximum Likelihood Estimation
MLP	multi-layer perceptron
MNIST	
OCR	optical character recognition
PSD	positive semi-definite

ROC	receiver operating characteristic
RMSE	root mean square error
RNN	recurrent neural network
s. t.	subject to
SVM	support vector machine
$\mathfrak{tp}$	true positives
$\mathfrak{tn}$	true negatives
X-Val	cross validation