APRENENTATGE AUTOMÀTIC (APA)
Grau en Enginyeria Informàtica - UPC
Notació 2018-2019 (DRAFT)

Lluís A. Belanche, belanche@cs.upc.edu

Symbol	Description
a, b, c, α, β, $\gamma, \ldots$	scalar values
$\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}', \boldsymbol{y}, \boldsymbol{z}$	(column) vectors
$\boldsymbol{A}$, $\boldsymbol{B}, \boldsymbol{K}_{n\times n}, \ldots$	matrices
$\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$	sets of real, integral, and natural numbers
$\boldsymbol{x}^\top \boldsymbol{x'}, \langle \boldsymbol{x}, \boldsymbol{x'} \rangle$	standard dot product in $\mathbb{R}$ and general inner product in an abstract space of vectors
$\boldsymbol{I}_d$	identity matrix in $\mathbb{R}^d$
$\boldsymbol{J}_d$	'all ones' matrix in $\mathbb{R}^d$
$x_{i,j}$, a_{ij}	vector components and matrix elements
(a_{ij})	matrix with elements a_{ij}
$\lvert\boldsymbol{A}\rvert, \text{rank}(\boldsymbol{A}), \text{Tr}(\boldsymbol{A})$	determinant, rank, and trace of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$, $\boldsymbol{A}^\top$	inverse and transpose of matrix $\boldsymbol{A}$
$\frac{\partial \boldsymbol{y}_j}{\partial x_i}$, $\partial_{x_i} \boldsymbol{y}_j$	derivative of $\boldsymbol{y}_j$ with respect to $\boldsymbol{x}_i$
$\nabla_{\boldsymbol{x}} f$	gradient of function f with respect to $\boldsymbol{x}$
$f^*, \hat{f}$	the best (theoretical) model, an estimated model
$\mathcal{X}, \mathcal{Y}$	input and output data spaces; Normally, $\mathcal{X} = \mathbb{R}^d$
$D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$	a dataset of size n, where $\boldsymbol{x}_i \in \mathcal{X}, \boldsymbol{y}_i \in \mathcal{Y}, i = 1, \ldots, n$; if $\mathcal{Y} = \emptyset$, it is understood that there is no output data, in which case $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$
d, m	dimensions of $\mathcal{X}, Y$
$\boldsymbol{X}_{n\times d}$	a data matrix, built from D, with vectors $\boldsymbol{x}_i^\top$ as rows; that is, $x_{ij} = x_{i,j}$
$\boldsymbol{y} = (y_1, \ldots, y_n)^\top$	a data vector, built from D, with values y_i as elements
$\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^\top$	an estimated vector (predicted or fitted values), built from a model $\hat{f}$, with values $\hat{y}_i$ as elements

Notes:

1. Often, we augment the input vectors $\boldsymbol{x}_i$ as $\boldsymbol{x}_i = (x_{i,0} = 1, \boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,d})^\top$; this translates into a data matrix $\boldsymbol{X}_{n\times(d+1)}$ with a column vector of ones in the lefmost column

2. The data matrix $\boldsymbol{X}$ can also be seen as formed by d vectors of length n; the j-th vector (the j-th column of $\boldsymbol{X}$) is an i.i.d. random sample of the random variable X_j of size n, $j = 1, \ldots, d$. In this interpretation, the data matrix can be seen as an i.i.d. random sample of the random vector $(X_1, \ldots, X_d)^\top$. This will also be valid for Y seen as a random variable, of which $\boldsymbol{y}$ is an i.i.d. random sample of size n.

Symbol	Description
$\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{v}_i, \ldots$	(column) vectors of parameters

Symbol	Description
L	arbitrary loss function (penalty for the discrepancy between a real and a predicted value)
λ, τ	regularization parameters
$\|\boldsymbol{x}\|_p$	p-norm of $\boldsymbol{x}$, $p \in [1, \infty]$
ln	natural logarithm
e	base for the natural logarithm (often called Euler's number)
$:=$	"by definition"
$\mathbb{E}[\cdot]$	expected value
$\mathbb{1}_{\{z\}}$	logical evaluator function, $\mathbb{1}_{\{z\}} := \begin{cases} 1 & \text{if } z \\ 0 & \text{if } \neg z \end{cases}$
ϕ_i	i-th base function
$\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_h)^\top$	vector of base functions
$\boldsymbol{\Phi}$	design matrix, obtained from $\boldsymbol{\phi}$ and X
k	kernel function
$\boldsymbol{K}$	kernel matrix, obtained from k and X
c	number of classes; in a classification setting, $\mathcal{Y} = \{1, 2, \ldots, c\}, c \geq 2$. When $c = 2$, it is customary to use the sets $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$
g	activation function

Notes:

1. As mentioned above, Y may be seen as a random variable; in a classification setting, we will write $P(Y = k)$ to refer to the (prior) probability of the class being the k-th class
2. In general, the variable i runs through input vectors (as in $\boldsymbol{x}_i$), j runs through input dimensions, and k runs through output dimensions. Note that $(\boldsymbol{x}_i)_j = x_{i,j}$ and x_{ij} refer to the same quantity, but are conceptually different mathematical objects.

1. In general, a hat on top of something (as in $\hat{\alpha}$) means that this is an estimation (so $\hat{\alpha}$ is an estimation of α)
2. We follow the usual (but not universal) conventions in statistical notation:
   (a) Upcase letters denote random variables and lowcase letters denote their values (so a is a value of the random variable A)
   (b) Upcase letters ($P, Q, \ldots$) denote probability mass functions, and lowcase letters ($p, q, \ldots$) denote probability density functions
   (c) The notation is greatly simplified by omitting both the distribution and the random variable; for example, $P_A(A = a)$ will be simply written $P(a)$.

Abbreviations

Abbreviation	Meaning
pmf	"probability mass function" (funció de massa de probabilitat)
pdf	"probability density function" (funció de densitat de probabilitat)
rv	"random variable" (variable aleatòria)
ml	"maximum likelihood" (màxima versemblança)
mle	"maximum likelihood estimator" (estimador de màxima versemblança)
pd	"positive definite" (definida positiva)
psd	"positive semi-definite" (semi-definida positiva)
bf	"basis function" (funció de base)
iid	"independent and identically distributed" (independent i idènticament distribuït)