Machine Learning Notation Template

Andreas Holzinger, a.holzinger@hci-kdd.org
Vienna University of Technology, Austria

Vienna, April 19, 2016

Abstract

The mathematical content of the course 185.A83 "Machine Learning for Health Informatics" is kept to a minimum, however, a nonzero level is necessary. It is hard to keep a consistent notation throughout the course, therefore in this short document the used mathematical notation is summarized.

1 Variables, Symbols, and Operations

An excellent compilation can be found in Duda, Hart and Stock (2000): Pattern Classification.

Introduction

It is very difficult to come up with a single, consistent notation to cover the wide variety of data, models and algorithms that we discuss. Furthermore, conventions differ between machine learning and statistics, and between different books and papers. Nevertheless, we have tried to be as consistent as possible. Below we summarize most of the notation used in this book, although individual sections may introduce new notation. Note also that the same symbol may have different meanings depending on the context, although we try to avoid this where possible.

General math notation

Symbol	Meaning

$\lfloor x \rfloor$	Floor of x, i.e., round down to nearest integer
$\lceil x \rceil$	Ceiling of x, i.e., round up to nearest integer
$\vec{x} \otimes \vec{y}$	Convolution of $\vec{x}$ and $\vec{y}$
$\vec{x} \odot \vec{y}$	Hadamard (elementwise) product of $\vec{x}$ and $\vec{y}$
$a \wedge b$	logical AND
$a \vee b$	logical OR
$\neg a$	logical NOT
$\mathbb{I}(x)$	Indicator function, $\mathbb{I}(x) = 1$ if x is true, else $\mathbb{I}(x) = 0$
∞	Infinity
$\rightarrow$	Tends towards, e.g., $n \rightarrow \infty$
$\propto$	Proportional to, so $y = ax$ can be written as $y \propto x$
$\lvert x \rvert$	Absolute value
$\lvert \mathcal{S} \rvert$	Size (cardinality) of a set
$n!$	Factorial function
∇	Vector of first derivatives
∇^2	Hessian matrix of second derivatives
$\triangleq$	Defined as
$O(\cdot)$	Big-O: roughly means order of magnitude
$\mathbb{R}$	The real numbers
$1:n$	Range (Matlab convention): $1:n = 1, 2, ..., n$
$\approx$	Approximately equal to
$\arg\max_x f(x)$	Argmax: the value x that maximizes f
$B(a, b)$	Beta function, $B(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
$B(\vec{\alpha})$	Multivariate beta function, $\dfrac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$
$\binom{n}{k}$	n choose k , equal to $n!/(k!(nk)!)$
$\delta(x)$	Dirac delta function,$\delta(x) = \infty$ if $x = 0$, else $\delta(x) = 0$
$\exp(x)$	Exponential function e^x
$\Gamma(x)$	Gamma function, $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u$
$\Psi(x)$	Digamma function,$Psi(x) = \dfrac{d}{dx} \log \Gamma(x)$
$\mathcal{X}$	A set from which values are drawn (e.g.,$\mathcal{X} = \mathbb{R}^D$)

Linear algebra notation

We use boldface lower-case to denote vectors, such as $\vec{x}$, and boldface upper-case to denote matrices, such as $\vec{X}$. We denote entries in a matrix by non-bold upper case letters, such as X_{ij}.

Vectors are assumed to be column vectors, unless noted otherwise. We use $(x_1, \cdots, x_D)$ to denote a column vector created by stacking D scalars. If we write $\vec{X} = (\vec{x}_1, \cdots, \vec{x}_n)$, where the left hand side is a matrix, we mean to stack the $\vec{x}_i$ along the columns, creating a matrix.

Symbol	**Meaning**
$\vec{X} \succ 0$	$\vec{X}$ is a positive definite matrix
$tr(\vec{X})$	Trace of a matrix
$det(\vec{X})$	Determinant of matrix $\vec{X}$
$\|\vec{X}\|$	Determinant of matrix $\vec{X}$
$\vec{X}^{-1}$	Inverse of a matrix
$\vec{X}^{\dagger}$	Pseudo-inverse of a matrix
$\vec{X}^T$	Transpose of a matrix
$\vec{x}^T$	Transpose of a vector
$\mathsf{diag}(x)$	Diagonal matrix made from vector $\vec{x}$
$\mathsf{diag}(X)$	Diagonal vector extracted from matrix $\vec{X}$
$\vec{I}$ or $\vec{I}_d$	Identity matrix of size $d \times d$ (ones on diagonal, zeros of)
$\vec{1}$ or $\vec{1}_d$	Vector of ones (of length d)
$\vec{0}$ or $\vec{0}_d$	Vector of zeros (of length d)
$\|\|\vec{x}\|\| = \|\|\vec{x}\|\|_2$	Euclidean or ℓ_2 norm $\sqrt{\sum_{j=1}^{d} x_j^2}$
$\|\|\vec{x}\|\|_1$	ℓ_1 norm $\sum_{j=1}^{d} \|x_j\|$
$\vec{X}_{:,j}$	j'th column of matrix
$\vec{X}_{i,:}$	transpose of i'th row of matrix (a column vector)
$\vec{X}_{i,j}$	Element (i, j) of matrix $\vec{X}$
$\vec{x} \otimes \vec{y}$	Tensor product of $\vec{x}$ and $\vec{y}$

Probability notation

We denote random and fixed scalars by lower case, random and fixed vectors by bold lower case, and random and fixed matrices by bold upper case. Occasionally we use non-bold upper case to denote scalar random variables. Also, we use $p()$ for both discrete and continuous random variables

Symbol	**Meaning**
X, Y	Random variable
$P()$	Probability of a random event
$F()$	Cumulative distribution function(CDF), also called distribution function
$p(x)$	Probability mass function(PMF)
$f(x)$	probability density function(PDF)
$F(x, y)$	Joint CDF
$p(x, y)$	Joint PMF
$f(x, y)$	Joint PDF
$p(X\|Y)$	Conditional PMF, also called conditional probability
$f_{X\|Y}(x\|y)$	Conditional PDF
$X \perp Y$	X is independent of Y
$X \not\perp Y$	X is not independent of Y
$X \perp Y\|Z$	X is conditionally independent of Y given Z
$X \not\perp Y\|Z$	X is not conditionally independent of Y given Z
$X \sim p$	X is distributed according to distribution p
$\vec{\alpha}$	Parameters of a Beta or Dirichlet distribution
$\mathtt{cov}[X]$	Covariance of X
$\mathbb{E}[X]$	Expected value of X
$\mathbb{E}_q[X]$	Expected value of X wrt distribution q
$\mathbb{H}(X)$ or $\mathbb{H}(p)$	Entropy of distribution $p(X)$
$\mathbb{I}(X; Y)$	Mutual information between X and Y
$\mathbb{KL}(p\|\|q)$	KL divergence from distribution p to q
$\ell(\vec{\theta})$	Log-likelihood function
$L(\theta, a)$	Loss function for taking action a when true state of nature is θ
λ	Precision (inverse variance) $\lambda = 1/\sigma^2$
Λ	Precision matrix $\Lambda = \Sigma^{-1}$
$\mathrm{mode}[\vec{X}]$	Most probable value of $\vec{X}$

μ	Mean of a scalar distribution
$\vec{\mu}$	Mean of a multivariate distribution
Φ	cdf of standard normal
ϕ	pdf of standard normal
$\vec{\pi}$	multinomial parameter vector, Stationary distribution of Markov chain
ρ	Correlation coefficient
$\text{sigm}(x)$	Sigmoid (logistic) function, $\frac{1}{1+e^{-x}}$
σ^2	Variance
Σ	Covariance matrix
$\text{var}[x]$	Variance of x
ν	Degrees of freedom parameter
Z	Normalization constant of a probability distribution

Machine learning/statistics notation

In general, we use upper case letters to denote constants, such as C, K, M, N, T, etc. We use lower case letters as dummy indexes of the appropriate range, such as $c = 1 : C$ to index classes, $i = 1 : M$ to index data cases, $j = 1 : N$ to index input features, $k = 1 : K$ to index states or clusters, $t = 1 : T$ to index time, etc.

We use x to represent an observed data vector. In a supervised problem, we use y or $\vec{y}$ to represent the desired output label. We use $\vec{z}$ to represent a hidden variable. Sometimes we also use q to represent a hidden discrete variable.

Symbol	**Meaning**
C	Number of classes
D	Dimensionality of data vector (number of features)
N	Number of data cases
N_c	Number of examples of class c,$N_c = \sum_{i=1}^{N} \mathbb{I}(y_i = c)$
R	Number of outputs (response variables)
$\mathcal{D}$	Training data $\mathcal{D} = \{(\vec{x}_i, y_i) \mid i = 1 : N\}$
$\mathcal{D}_{test}$	Test data
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
K	Number of states or dimensions of a variable (often latent)

$k(x, y)$	Kernel function
$\vec{K}$	Kernel matrix
$\mathcal{H}$	Hypothesis space
L	Loss function
$J(\vec{\theta})$	Cost function
$f(\vec{x})$	Decision function
$P(y\|\vec{x})$	TODO
λ	Strength of ℓ_2 or $\ell_1 regularizer$
$\phi(x)$	Basis function expansion of feature vector $\vec{x}$
Φ	Basis function expansion of design matrix $\vec{X}$
$q()$	Approximate or proposal distribution
$Q(\vec{\theta}, \vec{\theta}_{old})$	Auxiliary function in EM
T	Length of a sequence
$T(\mathcal{D})$	Test statistic for data
$\vec{T}$	Transition matrix of Markov chain
$\vec{\theta}$	Parameter vector
$\vec{\theta}^{(s)}$	s'th sample of parameter vector
$\hat{\vec{\theta}}$	Estimate (usually MLE or MAP) of $\vec{\theta}$
$\hat{\vec{\theta}}_{MLE}$	Maximum likelihood estimate of $\vec{\theta}$
$\hat{\vec{\theta}}_{MAP}$	MAP estimate of $\vec{\theta}$
$\bar{\vec{\theta}}$	Estimate (usually posterior mean) of $\vec{\theta}$
$\vec{w}$	Vector of regression weights (called $\vec{\beta}$ in statistics)
b	intercept (called ε in statistics)
$\vec{W}$	Matrix of regression weights
x_{ij}	Component (i.e., feature) j of data case i ,for $i = 1 : N, j = 1 : D$
$\vec{x}_i$	Training case, $i = 1 : N$
$\vec{X}$	Design matrix of size $N \times D$
$\bar{\vec{x}}$	Empirical mean $\bar{\vec{x}} = \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i$
$\tilde{\vec{x}}$	Future test case
$\vec{x}_*$	Feature test case
$\vec{y}$	Vector of all training labels $\vec{y} = (y_1, ..., y_N)$
z_{ij}	Latent component j for case i

2 Future Research

There are many future challenges in machine learning generally and in the application of machine learning to health informatics specifically. The ultimate goal is to design and develop algorithms which can learn from data, hence can improve with experience over time. Interactive Machine Learning (iML) is a relatively new approach and can be of particular important to solve problems in health informatics, where we are lacking big data sets, deal with complex data and/or rare events, where traditional learning algorithms suffer due to insufficient training samples. Here the doctor-in-the-loop can help, where human expertise and long-term experience can assist in solving hard problems.

- TODO
- TODO
- TODO

3 References