ML Course – Notation

Prof. Dr. Isabel Valera

---

1 Probabilistic Theory

X	**random variable**	e.g., probability of a random variable X taking value 1 is $P_X(X = 1) = p$
$\mathcal{X}$	**feature space**	values that X can assume, e.g., $\mathcal{X} = \mathbb{R}$
x	**value**	value of a random variable X e.g., $X = x$
$\mathbf{x}$	**vector**	e.g. D-dimensional vector $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$, note: all vectors are assumed to be column vectors
$\mathbf{x}^T$	**transpose**	transpose of vector (or matrix), e.g. row vector
$\mathbf{A}$	**matrix**	
$\mathbf{A}^{-1}$	**inverse**	inverse of a matrix
Ω	**sample space**	set of all possible outcomes $\Omega = \{\omega_1, \ldots, \omega_n\}$, e.g. tossing a coin $\Omega = \{H, T\}$
P, Pr	**probability distribution, probability measure**	probability function that maps an event into the probability of the event, e.g., Bernoulli $P : 2^\Omega \to [0, 1]$
$p_X, p(X)$	**probability density function**	e.g., probability that continuous variable X will lie in an interval (a, b) is $p(X \in (a, b)) = \int_a^b p(x)\mathrm{d}x$
$F(x)$	**(cumulative) distribution function**	e.g. $F(x) = \Pr\Big(X \leq x\Big) = \int_{-\infty}^{x} p(t)dt$
$\mathbb{E}_X[f(X, Y)]$	**expectation**	expectation of a function $f(X, Y)$ with respect to a random variable X, note: omitted, if no ambiguity as to which variable is being averaged over, e.g., $\mathbb{E}[f(X)] = \mathbb{E}_X[f(X)]$

2 Supervised Learning notation

X	**r.v. input features**	
Y	**r.v. target/output/outcome variable**	
$D = (x_i, y_i)_{i=1}^n$	**observed dataset**	dataset containing features/outcome pairs with n observations.
$D = (X_i, Y_i)_{i=1}^n$	**dataset as r.v**	treats features and outcome as random variables.
$\hat{y}(x)$	**(any) classification function**	$f : \mathcal{X} \mapsto \{1, \dots, K\}$, with K being the number of classes.
$\hat{y}^*(x)$	**Bayes classifier**	Provides the optimal classification rule for any x.
$f(x)$	**regression function**	$f : \mathcal{X} \mapsto \mathbb{R}$
$\phi(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$	**basis function**	used to apply a non-linear transformation to the feature vector $\mathbf{x}$.
$\Phi(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}^m$	**vector of basis function**	used to apply m non-linear transformations to the observed feature vectors $\mathbf{x} \in \mathbb{R}^d$.
$\mathbf{X} \in \mathbb{R}^{n \times d}$	**observed feature matrix**	matrix containing observed feature vectors
$\mathbf{\Phi}$	**transformed feature matrix**	matrix containing transformed feature vectors using a set of basis functions Φ.
$\mathbf{Y}$	**vector containing observed values for the target vector**	