Statistical Foundations of Learning
List of notations and key concepts[1]

Debarghya Ghoshdastidar
TUM Informatik
Summer 2020

Updated on: May 29, 2020(till lecture 11)

[1]This list will be regularly updated. Let us know if an improtant notation / term is missing from the list.

Chapter 1

List of notations and key concepts

This list ignores few notations that are used locally, in a specific section or proof.

Notation	First in Chapter	Description
$\mathbb{R}$, $\mathbb{R}^p$	1	Real line, or p-dimensional space
$\mathbb{N}$	1	Set of natural numbers $\{0, 1, 2, \ldots\}$
$\mathbb{E}[\cdot]$, $\mathbb{E}_S[\cdot]$	1	Expectation; Subscript identifies the random variable over which expectation is considered
$\mathbb{P}(\cdot)$, $\mathbb{P}_S(\cdot)$	1	Probability; Subscript identifies the random variable
$\mathcal{X}$	1	Feature space (possible values of the instances); We mostly have $\mathcal{X} \subseteq \mathbb{R}$ or $\mathbb{R}^p$ or finite set; In Chapter 9, it is set of vertices of a graph
$\mathcal{Y}$	1	Label space; For binary classification $\mathcal{Y} = \{\pm 1\}$ or $\{0, 1\}$; Multi-class classification / clustering $\mathcal{Y} = \{1, 2, \ldots, k\}$; Regression $\mathcal{Y} = \mathbb{R}$
$\mathcal{X} \times \mathcal{Y}$	1	Product of two spaces; In this lecture, space of data-label pairs
(x, y)	1	Data-label pair; Same notation also used for random instances
$\mathcal{Y}^{\mathcal{X}}$	1	Space of all functions $f : \mathcal{X} \to \mathcal{Y}$; If $\mathcal{Y} = \{\pm 1\}$, then space of all binary classification rules

$\mathcal{D}$	1	Joint distribution of data-label pairs over $\mathcal{X} \times \mathcal{Y}$
$\mathcal{D}_{\mathcal{X}}$	1	Marginal distribution of $\mathcal{D}$ of features in $\mathcal{X}$
$\mathbb{P}_{\mathcal{Y}\vert\mathcal{X}}(\cdot\vert x), \eta(x)$	1	Conditional distribution of label given feature x; For binary classification, we define $\eta(x) = \mathbb{P}_{\mathcal{Y}\vert\mathcal{X}}(y = 1\vert x)$
m	1	Number of training samples
$\mathcal{D}^m$	1	Joint distribution of m i.i.d. random variates, each distributed according to $\mathcal{D}$
$S = \{(x_i, y_i)\}_{i=1}^m$	1	Training sample; $S \in (\mathcal{X} \times \mathcal{Y})^m$ and often, we have $S \sim \mathcal{D}^m$
h, h_t $\widehat{h}, h^*$	1	Prediction rules; Subscript usually denotes a parameter Typically $\widehat{h}$ is output of an algorithm, and h^* is true / Bayes predictor
$\mathcal{H}, \mathcal{H}_{ds-1}$	1	Hypothesis class; Some subset of prediction rules in $\mathcal{Y}^{\mathcal{X}}$; Subscript is used to specify certain hypothesis class
$\mathcal{A}$, $\mathcal{A}_S$, $\mathcal{A}_{method}$	1	Learner / learning algorithm; Takes training set S as input, and returns a predictor $\widehat{h}$ Sometimes, we use $\mathcal{A}_S$ to denote the predictor $\mathcal{A}(S)$, whereas $\mathcal{A}_{method}$ is used to specify the learning approach (for instance, ERM)
$\ell, \ell^{\text{0-1}}$	1	Loss function; The superscript is used to specify the type of loss function
$L_{\mathcal{D}}(\cdot), L_{\mathcal{D}^{\text{0-1}}}(\cdot)$	1	Risk / Generalisation error; Expected error of a predictor with respect to distribution $\mathcal{D}$. Superscript used to specify loss function
$L_S(\cdot)$	1	Empirical risk / Training error; Error of a predictor with respect to sample S
$L^*_{\mathcal{D}}$	1	Bayes risk (minimum possible risk) for distribution $\mathcal{D}$
$L_{\mathcal{D}}(\mathcal{H})$	1	Minimum possible risk for $\mathcal{D}$ using predictors in $\mathcal{H}$
ERM	1	Empirical risk minimisation
$\mathcal{H}_{\vert C}$	1.2	Restriction of a hypothesis class $\mathcal{H}$ to a set $C \subset \mathcal{X}$
$\tau_{\mathcal{H}}(\cdot)$	1.2	Growth function for class $\mathcal{H}$, which is a function of sample size m
$\text{VCdim}(\mathcal{H})$ d	1.2	VC-dimension of $\mathcal{H}$; Mostly, we use d as notation for a finite VC-dimension
PAC	2	Probably Approximately Correct

ϵ	2 also 1.2	Excess risk; In PAC, ϵ is the allowable excess risk of learned predictor over the minimum possible risk (0 or $L_{\mathcal{D}}(\mathcal{H})$) Notation is used in a similar spirit in the uniform convergence results, but differs from the excess risk by factor 2 or 4 in some parts
δ	2 also 1.2	In PAC, allowable probability for excess risk bound not to be satisfied; Notation is used in a similar spirit in Chapter 1.2, but differs by constants
$m_{\mathcal{H}}(\cdot,\cdot)$	2	Sample complexity of $\mathcal{H}$; $m_{\mathcal{H}}(\epsilon,\delta)$ is minimum training sample size needed to (agnostic) PAC learn any distribution using $\mathcal{H}$ with specified error limits ϵ,δ
X_j	2 (NFL proof)	Particular sequence of m unlabelled examples
$S_{i,j}$	2 (NFL proof)	Labelled examples corresponding to X_j and labelled using function h_i
w, b	1	Parameter for linear prediction rule $h(x) = \text{sign}(\langle w, x\rangle + b)$. If $b = 0$, then it is called homogeneous linear classifier
w^*, b^*	2.2	Under realisable assumption, parameters for true linear classifier
T, t	2.2	For iterative algorithms (like perceptron), T is used for total number of iterations and t is the iteration counter
(γ, ρ)	2.2	Margin of linear separable data. (γ,ρ) satisfies $\|x\| \leq \rho$ and $(\langle w^*, x\rangle + b^*) \geq \gamma$ for all $(x, y) \sim \mathcal{D}$
γ-weak	3	Note: This use of γ has no connection with above γ-weak learner satisfies that generalisation error is smaller than $\frac{1}{2} - \gamma$
$\mathbf{D}$	3	Probability weight vector over m training examples
$L_{\mathbf{D}}(\cdot)$	3	Weighted empirical risk with weight vector $\mathbf{D}$
h_{ada}	3	Weighted majority predictor learned by AdaBoost
$\mathcal{B}, d_{\mathcal{B}}$	3	Base hypothesis class over which we apply majority voting. $d_{\mathcal{B}} = \text{VCdim}(\mathcal{B})$
$M_{\mathcal{B},T}$	3	Class of majority among T votes with hypotheses from base class $\mathcal{B}$
SRM	4.1	Structural risk minimisation
t_h	4.1	Degree of polynomial p that defines predictor $h(\cdot) = \text{sign}(p(\cdot))$
$L_V(\cdot)$	4.2	Validation error, computed on hold out set V

m_s, m_v	4.2	In case of validation with hold out set, m_s = number of examples used for training, and m_v = size of hold out set
S_i	4.2	i-th examples (x_i, y_i) in training sample S (for k-fold cross validation, it denotes i-th group)
S_{-i}	4.2	all examples in S other than S_i
S^i	4.3	Training sample S with i-th example replaced by an independent example
$L_{k-cv}(\mathcal{A}_S)$	4.2	k-fold cross validation error for learner $\mathcal{A}$ and total labelled examples S
$L_{loo}(\mathcal{A}_S)$	4.2	leave one out error for learner $\mathcal{A}$ and total labelled examples S
$\mathrm{Unif}(m)$	4.3	Uniformly sampled random variable from $\{1, \ldots, m\}$
β, β_m^{rep}	4.3	Stability rate function; $\beta_m = \beta(m)$; Superscript denotes type of stability
$\mathcal{A}_{S^i}$, $\mathcal{A}_{S_{-i}}$	4.3	Predictors learned by learner $\mathcal{A}$ with modified training data S^i or S_{-i}