

Chapter ML:I

I. Introduction

- ❑ Examples of Learning Tasks
- ❑ Specification of Learning Tasks
- ❑ Elements of Machine Learning
- ❑ Notation Overview
- ❑ Classification Approaches Overview

Notation Overview

Data, Sets, and Distributions

Symbol	Semantics
x, x_i, x_1, \dots, x_p	Feature
$\mathbf{x} = (x_1, \dots, x_p)^T \in \mathbf{R}^p$	Feature vector
$\mathbf{x} = (1, x_1, \dots, x_p)^T \in \mathbf{R}^{p+1}$, i.e., $x_0 = 1$	Extended feature vector
\mathbf{X}	Feature space, Cartesian product of the domains of the p dimensions of a feature vector \mathbf{x} .
$X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$	Multiset of feature vectors
X	Random variable (randomness regarding feature x of an object o)
\mathbf{X}	Multivariate random variable, random vector (randomness regarding feature vector \mathbf{x} of an object o)

Notation Overview

Indexing

Running	Sequence	Semantics of maximum
φ_s	$\in \{\varphi_1, \dots, \varphi_d\}$	Number of layers in a multilayer perceptron
φ_i	$\in \{\varphi_1, \dots, \varphi_k\}$	Number of classes Number of folds during cross validation
φ_l	$\in \{\varphi_1, \dots, \varphi_m\}$	Number of elements in a domain of a feature Number of hyperparameter values during model selection
φ_i	$\in \{\varphi_1, \dots, \varphi_n\}$	Number of elements in a data set D
φ_j	$\in \{\varphi_1, \dots, \varphi_p\}$	Dimension of a feature space or a feature vector

Notation Overview

Functions

Function definition	Function name	Occurrence
$I_{\neq}(a, b) = \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases}$	Indicator function	Part II: Machine Learning Basics Part III: Linear Models
$f(x) = \dots$	function	Part :

Notation Overview

Algorithms

Signature	Algorithm name	Occurrence
$LMS(D, \eta)$	Least Mean Squares	Part I: Introduction Examples of Learning Tasks
$ALG(\dots)$	algorithm	Part : ...

Classification Approaches Overview

				Search in hypothesis space											
Taxonomy		Model function	Classification rule	Optimization principle	Optimization objective (loss/cost function [+ regularization])		Optimization approach (algorithm)								
Classification approaches	Discriminative approaches	Linear decision boundary (in inner product space)	Linear decision boundary in input space	Perceptron: $y(\mathbf{x}) = \text{heaviside}(\mathbf{w}^T \mathbf{x})$	$\mathbf{w}^T \mathbf{x} \begin{cases} \geq 0 \\ < 0 \end{cases}$ $\mathbf{w}^T = (w_0, \dots, w_p)$ $x_0 = 1$	Exploit misclassified examples individually: Hebbian learning		\sim	No misclassified example		\sim	Perceptron training algorithm			
				Linear function: $y(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$		+ Regularization	Squared loss (residual sum of squares, RSS)		+ L_1 or L_2 norm on $\mathbf{w} _{1, \dots, p}$	Gradient descent: – batch – incremental – stochastic Newton-Raphson, BFGS					
				Logistic function: $y(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$			Logistic loss (derived via ML)								
				SVM w/o kernel (aka linear kernel)			Regularized hinge loss								
				$y(\mathbf{x}) = \text{sign}(\mathbf{w}^T \phi(\mathbf{x}))$			+ Regularization					Squared loss	L_1 or L_2 norm on $\mathbf{w} _{1, \dots, \mathbf{w} }$	Gradient descent: – batch – incremental – stochastic Newton-Raphson, BFGS	
		$y(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \phi(\mathbf{x})}}$	Log. regression (nonlinear in predictors)	Logistic loss											
		SVM with nonlinear kernel	Empirical risk minimization		Regularized hinge loss			Quadratic prog., sub-grad. descent							
		Multilayer percep.: $\mathbf{y}(\mathbf{x}) = \sigma(W^0(\sigma^1(W^h \mathbf{x})))$	Regression		Squared loss (residual sum of squares, RSS)				Backpropagation algorithm						
		Unrestricted decision boundary	Polythetic	Monothetic feature analysis	Test if \mathbf{x} is a model for α (= fulfills α). α is a formula in DNF.	Maximize version space				No misclassified example		Candidate elimination algorithm			
						+ Regularization	Decision tree: (greedy) feature-wise splitting of example set			0/1 Loss (= number of misclassified examples)	+ Tree height, external path length		Algorithms: ID3, C4.5, C5.0, CART (exhaustive) search in space of domain splittings		
Maximum a-posteriori hypothesis							Goodness of fit, e.g. according to chi-squared, Kolmogorov-Smirnov								
Generative approaches	Statistical approaches														
												Bayes rule for combined conditional events	$\text{argmax}_{c \in C} \{ \text{Naive Bayes probabilities} \}$	Maximum a-posteriori hypothesis	Goodness of fit, e.g. according to chi-squared, Kolmogorov-Smirnov