
Notations

x	variable
\mathbf{x}	vector
\mathbf{A}	matrix
\mathbf{I}	identity matrix
\mathcal{X}, \mathcal{Y}	input and output spaces
\mathcal{D}	probability distribution
D	data sample (data set)
\mathcal{N}	normal distribution
\mathcal{U}	uniform distribution
\mathcal{H}	hypothesis space
H	set of hypotheses
$h(\cdot)$	hypothesis (learner)
\mathcal{L}	learning algorithm
$p(\cdot)$	probability density function
$p(\cdot \cdot)$	conditional probability density function
$P(\cdot)$	probability mass function
$P(\cdot \cdot)$	conditional probability mass function
$\mathbb{E}_{\cdot \sim \mathcal{D}}[f(\cdot)]$	mathematical expectation of function $f(\cdot)$ to \cdot under distribution \mathcal{D} . \mathcal{D} and/or \cdot is ignored when the meaning is clear
$var_{\cdot \sim \mathcal{D}}[f(\cdot)]$	variance of function $f(\cdot)$ to \cdot under distribution \mathcal{D}
$\mathbb{I}(\cdot)$	indicator function which takes 1 if \cdot is true, and 0 otherwise
$\text{sign}(\cdot)$	sign function which takes -1, 1 and 0 when $\cdot < 0$, $\cdot > 0$ and $\cdot = 0$, respectively
$err(\cdot)$	error function
$\{\dots\}$	set
(\dots)	row vector
$(\dots)^\top$	column vector

vi

$|\cdot|$ size of data set
 $\|\cdot\|$ L_2 -norm

DRAFT

Contents

1	Introduction	1
1.1	Basic Concepts	1
1.2	Popular Learning Algorithms	3
1.2.1	Linear Discriminant Analysis	3
1.2.2	Decision Trees	4
1.2.3	Neural Networks	6
1.2.4	Naïve Bayes Classifier	8
1.2.5	k -Nearest Neighbor	9
1.2.6	Support Vector Machines and Kernel Methods	9
1.3	Evaluation and Comparison	12
1.4	Ensemble Methods	15
1.5	Applications of Ensemble Methods	17
1.6	Further Readings	20
2	Boosting	23
2.1	A General Boosting Procedure	23
2.2	The AdaBoost Algorithm	24
2.3	Illustrative Examples	28
2.4	Theoretical Issues	32
2.4.1	Initial Analysis	32
2.4.2	Margin Explanation	32
2.4.3	Statistical View	35
2.5	Multiclass Extention	39
2.6	Noise Tolerance	40
2.7	Further Readings	44
3	Bagging	47
3.1	Two Ensemble Paradigms	47
3.2	The Bagging Algorithm	48
3.3	Illustrative Examples	50
3.4	Theoretical Issues	53
3.5	Random Tree Ensembles	57
3.5.1	Random Forest	57
3.5.2	Spectrum of Randomization	59
3.5.3	Random Tree Ensembles for Density Estimation	61
3.5.4	Random Tree Ensembles for Anomaly Detection	64

3.6	Further Readings	66
4	Combination Methods	67
4.1	Benefits of Combination	67
4.2	Averaging	68
4.2.1	Simple Averaging	68
4.2.2	Weighted Averaging	70
4.3	Voting	71
4.3.1	Majority Voting	72
4.3.2	Plurality Voting	73
4.3.3	Weighted Voting	73
4.3.4	Soft Voting	75
4.3.5	Theoretical Issues	77
4.4	Combining by Learning	83
4.4.1	Stacking	83
4.4.2	Infinite Ensemble	85
4.5	Other Combination Methods	87
4.5.1	Algebraic Methods	87
4.5.2	Behavior Knowledge Space Method	88
4.5.3	Decision Template Method	88
4.6	Relevant Methods	89
4.6.1	Error-Correcting Output Codes	89
4.6.2	Dynamic Classifier Selection	92
4.6.3	Mixture of Experts	93
4.7	Further Readings	94
5	Diversity	97
5.1	Ensemble Diversity	97
5.2	Error Decomposition	98
5.2.1	Error-Ambiguity Decomposition	98
5.2.2	Bias-Variance-Covariance Decomposition	100
5.3	Diversity Measures	103
5.3.1	Pairwise Measures	103
5.3.2	Non-Pairwise Measures	104
5.3.3	Summary and Visualization	107
5.3.4	Limitation of Diversity Measures	108
5.4	Information Theoretic Diversity	109
5.4.1	Information Theory and Ensemble	109
5.4.2	Interaction Information Diversity	110
5.4.3	Multi-Information Diversity	111
5.4.4	Estimation Method	112
5.5	Diversity Generation	114
5.6	Further Readings	116

6	Ensemble Pruning	117
6.1	What is Ensemble Pruning	117
6.2	Many Could be Better Than All	118
6.3	Categorization of Pruning Methods	121
6.4	Ordering-Based Pruning	122
6.5	Clustering-Based Pruning	125
6.6	Optimization-Based Pruning	126
6.6.1	Heuristic Optimization Pruning	126
6.6.2	Mathematical Programming Pruning	127
6.6.3	Probabilistic Pruning	129
6.7	Further Readings	130
7	Clustering Ensemble	133
7.1	Clustering	133
7.1.1	Clustering Methods	133
7.1.2	Clustering Evaluation	135
7.1.3	Why Clustering Ensembles	137
7.2	Categorization of Clustering Ensemble Methods	139
7.3	Similarity-Based Methods	140
7.4	Graph-Based Methods	142
7.5	Relabeling-Based Methods	145
7.6	Transformation-Based Methods	150
7.7	Further Readings	152
8	Advanced Topics	155
8.1	Semi-Supervised Learning	155
8.1.1	Usefulness of Unlabeled Data	155
8.1.2	Semi-Supervised Learning with Ensembles	157
8.2	Active Learning	161
8.2.1	Usefulness of Human Intervention	161
8.2.2	Active Learning with Ensembles	162
8.3	Cost-Sensitive Learning	164
8.3.1	Learning with Unequal Costs	164
8.3.2	Ensemble Methods for Cost-Sensitive Learning	165
8.4	Class-Imbalance Learning	169
8.4.1	Learning with Class Imbalance	169
8.4.2	Performance Evaluation with Class Imbalance	170
8.4.3	Ensemble Methods for Class-Imbalance Learning	173
8.5	Improving Comprehensibility	177
8.5.1	Reduction of Ensemble to Single Model	177
8.5.2	Rule Extraction from Ensemble	178
8.5.3	Visualization of Ensembles	179
8.6	Future Directions of Ensembles	179
8.7	Further Readings	182