SpringerBriefs in Computer Science

More information about this series at http://www.springer.com/series/10028

Rodrigo C. Barros · André C.P.L.F. de Carvalho
Alex A. Freitas

Automatic Design of Decision-Tree Induction Algorithms

Springer

Rodrigo C. Barros
Faculdade de Informática
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, RS
Brazil

André C.P.L.F. de Carvalho
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos, SP
Brazil

Alex A. Freitas
School of Computing
University of Kent
Canterbury, Kent
UK

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-3-319-14230-2 ISBN 978-3-319-14231-9 (eBook)
DOI 10.1007/978-3-319-14231-9

Library of Congress Control Number: 2014960035

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to my family:
Alessandra, my wife;
Marta and Luís Fernando, my parents;
Roberta, my sister; Gael, my godson;
Lygia, my grandmother.

Rodrigo C. Barros

To Valeria, my wife, and to Beatriz,
Gabriela and Mariana, my daughters.

André C.P.L.F. de Carvalho

To Jie, my wife.

Alex A. Freitas

Contents

Notations

T	A decision tree
$\mathbf{X}$	A set of instances
N_x	The number of instances in $\mathbf{X}$, i.e., $\lvert\mathbf{X}\rvert$
$\mathbf{x}^j$	An instance—n-dimensional attribute vector $[x_1^j, x_2^j, \ldots, x_n^j]$—from $\mathbf{X}$, $j = 1, 2, \ldots, N_x$
$\mathbf{X_t}$	A set of instances that reach node t
A	The set of n predictive (independent) attributes $\{a_1, a_2, \ldots, a_n\}$
y	The target (class) attribute
Y	The set of k class labels $\{y_1, \ldots, y_k\}$ (or k distinct values if y is continuous)
$y(x)$	Returns the class label (or target value) of instance $\mathbf{x} \in \mathbf{X}$
$a_i(x)$	Returns the value of attribute a_i from instance $\mathbf{x} \in \mathbf{X}$
$dom(a_i)$	The set of values attribute a_i can take
$\lvert a_i\rvert$	The number of partitions resulting from splitting attribute a_i
$\mathbf{X_{a_i=v_j}}$	The set of instances in which attribute a_i takes a value contemplated by partition v_j. Edge v_j can refer to a nominal value, to a set of nominal values, or even to a numeric interval
$N_{v_j,\bullet}$	The number of instances in which attribute a_i takes a value contemplated by partition v_j, i.e., $\lvert\mathbf{X_{a_i=v_j}}\rvert$
$\mathbf{X_{y=y_l}}$	The set of instances in which the class attribute takes the label (value) y_l
$N_{\bullet,y_l}$	The number of instances in which the class attribute takes the label (value) y_l, i.e., $\lvert\mathbf{X_{y=y_l}}\rvert$
$N_{v_j \cap y_l}$	The number of instances in which attribute a_i takes a value contemplated by partition v_j and in which the target attribute takes the label (value) y_l
v_X	The target (class) vector $[N_{\bullet,y_1}, \ldots, N_{\bullet,y_k}]$ associated to $\mathbf{X}$
p_y	The target (class) probability vector $[p_{\bullet,y_1}, \ldots, p_{\bullet,y_k}]$
$p_{\bullet,y_l}$	The estimated probability of a given instance belonging to class y_l, i.e., $\frac{N_{\bullet,y_l}}{N_x}$

$p_{v_j,\bullet}$	The estimated probability of a given instance being contemplated by partition v_j, i.e., $\frac{N_{v_j,\bullet}}{N_x}$
$p_{v_j \cap y_l}$	The estimated joint probability of a given instance being contemplated by partition v_j and also belonging to class y_l, i.e., $\frac{N_{v_j \cap y_l}}{N_x}$
$p_{y_l \mid v_j}$	The conditional probability of a given instance belonging to class y_l given that it is contemplated by partition v_j, i.e., $\frac{N_{v_j \cap y_l}}{N_{v_j,\bullet}}$
$p_{v_j \mid y_l}$	The conditional probability of a given instance being contemplated by partition v_j given that it belongs to class y_l, i.e., $\frac{N_{v_j \cap y_l}}{N_{\bullet,y_l}}$
ζ_T	The set of nonterminal nodes in decision tree T
λ_T	The set of terminal nodes in decision tree T
$\aleph_T$	The set of nodes in decision tree T, i.e., $\aleph_T = \zeta_T \cup \lambda_T$
$T^{(t)}$	A (sub)tree rooted in node t
$E^{(t)}$	The number of instances in t that do not belong to the majority class of that node