

Classification

Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino

Copyright - All rights reserved DATA MINING: CLASSIFICATION - 1 Elena Baralis Politecnico di Torino

Associative classification

- The classification model is defined by means of Association rules
 - The rule head is the class label
 - Rules are selected based on
 - Support, confidence and correlation thresholds
 - Database coverage: procedure to cover the training set by using the extracted rules, sorted according to proper criteria
- Strength
 - Easy to interpret model
 - It achieves higher classification accuracy than Decision Trees
 - It considers the correlation among more attributes at the same time
 - Efficient in classification
 - Unaffected by missing data
 - Good scalability in the training set size
- Weakness
 - Rule generation may be slow
 - Low scalability in the number of attributes

Database and data mining group, Politecnico di Torino

Copyright - All rights reserved DATA MINING: CLASSIFICATION - 2 Elena Baralis Politecnico di Torino

Bayesian classification

- Based on computing probabilities
- Strength
 - The model can be incrementally updated
 - Efficient in classification
 - Quite easy to interpret model
- Weakness
 - The generation of the reference model is computationally unfeasible
 - Simplified hypothesis: naïve hypothesis
 - Naïve hypothesis significantly reduces the accuracy

Database and data mining group, Politecnico di Torino

Copyright - All rights reserved DATA MINING: CLASSIFICATION - 3 Elena Baralis Politecnico di Torino

Bayesian classification

- Use of the Bayes theorem
 - $P(C|X)$ = probability that $X = \langle x_1, \dots, x_k \rangle$ belongs to class C
 - X is labeled with the class that maximizes $P(C|X)$
 - Bayes theorem

$$P(C|X) = P(X|C) \cdot P(C) / P(X)$$
 - $P(X)$ constant for all class labels C
 - $P(C)$ a-priori probability of C
 - $P(X|C)$ **cannot** be completely computed
- Naïve hypothesis
 - It assumes statistical independence among attributes x_1, \dots, x_k
 - It is not always verified
 - It may affect the quality of the model
- Bayesian networks
 - They allow specifying a subset of dependencies among attributes

Database and data mining group, Politecnico di Torino

Copyright - All rights reserved DATA MINING: CLASSIFICATION - 4 Elena Baralis Politecnico di Torino

Bayesian classification

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

From Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2002

Database and data mining group, Politecnico di Torino

Copyright - All rights reserved DATA MINING: CLASSIFICATION - 5 Elena Baralis Politecnico di Torino

Bayesian classification

outlook	
$P(\text{sunny} p) = 2/9$	$P(\text{sunny} n) = 3/5$
$P(\text{overcast} p) = 4/9$	$P(\text{overcast} n) = 0$
$P(\text{rain} p) = 3/9$	$P(\text{rain} n) = 2/5$
temperature	
$P(\text{hot} p) = 2/9$	$P(\text{hot} n) = 2/5$
$P(\text{mild} p) = 4/9$	$P(\text{mild} n) = 2/5$
$P(\text{cool} p) = 3/9$	$P(\text{cool} n) = 1/5$
humidity	
$P(\text{high} p) = 3/9$	$P(\text{high} n) = 4/5$
$P(\text{normal} p) = 6/9$	$P(\text{normal} n) = 2/5$
windy	
$P(\text{true} p) = 3/9$	$P(\text{true} n) = 3/5$
$P(\text{false} p) = 6/9$	$P(\text{false} n) = 2/5$

$P(p) = 9/14$
 $P(n) = 5/14$

Data object to classify
 $X = \langle \text{rain, hot, high, false} \rangle$

$P(X|p) \cdot P(p) = P(\text{rain}|p) \cdot P(\text{hot}|p) \cdot P(\text{high}|p) \cdot P(\text{false}|p) \cdot P(p)$
 $= 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$

$P(X|n) \cdot P(n) = P(\text{rain}|n) \cdot P(\text{hot}|n) \cdot P(\text{high}|n) \cdot P(\text{false}|n) \cdot P(n)$
 $= 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$

From Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2002

Database and data mining group, Politecnico di Torino

Copyright - All rights reserved DATA MINING: CLASSIFICATION - 6 Elena Baralis Politecnico di Torino

Neural Networks

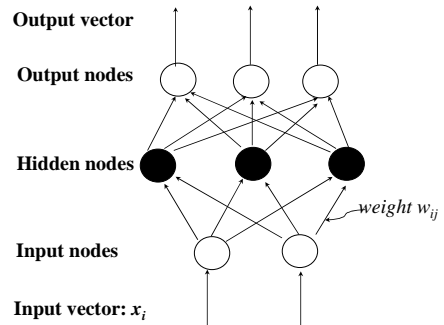
- Inspired to the structure of the human brain
 - Neurons as elaboration units
 - Synapses as connection network
- Strength
 - High accuracy
 - High tolerance to noise and outliers
 - Efficient in classification
 - It supports both discrete and continuous output
- Weakness
 - Long training time
 - No interpretable model
 - Hard introducing knowledge from the application domain

Copyright - All rights reserved

DATA MINING: CLASSIFICATION - 7

Elena Baralis
Politecnico di Torino

Structure of a Neural Network



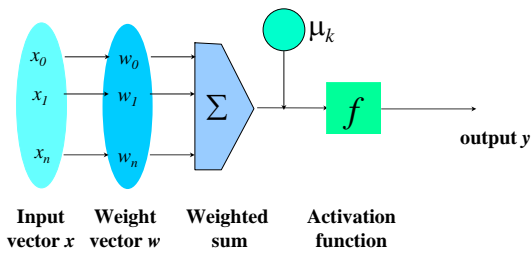
From Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2002

Copyright - All rights reserved

DATA MINING: CLASSIFICATION - 8

Elena Baralis
Politecnico di Torino

Structure of a neuron



From Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2002

Copyright - All rights reserved

DATA MINING: CLASSIFICATION - 9

Elena Baralis
Politecnico di Torino

Construction of the Neural Network

- Objective
 - Definition of an optimal set of weights and offset
- Basic algorithm
 - Initially assign random values to weights and offset
 - Process instances in the training set one at a time
 - For each neuron, compute the result when applying weights, offset and activation function for the instance
 - Forward propagation until the output is computed
 - Compare the computed output with the expected output, and error evaluation
 - Backpropagation of the error, by updating weights and offset
 - The process ends when
 - % of accuracy above a given threshold
 - % of error below a given threshold
 - The maximum number of epoches is reached

Copyright - All rights reserved

DATA MINING: CLASSIFICATION - 10

Elena Baralis
Politecnico di Torino