Introduction- to t'ecnicas classi fi caci'on

Maria-Amparo Vila

vila@decsai.ugr.es

Investigaci'on group Databases and Intelligent Systems Informaci'on https://idbis.ugr.es/

Department of Sciences of the computation and artificial intelligence

University of Granada

Introduction

B'asicos concept

De fi nition

Classi fi caci'on *is the process of learning a funci'on applying a set of attributes* $X_{one..}$ X_n *in another attribute* Y *. Yes:*

- *Yes Y It is discreet, boolean, nominal etc. we have Models classified caci'on themselves*
- *Yes Y is continuous we Regresi'on models*

The function is called learns tambi'en Model classi fi caci'on in general

Introduction

B'asicos concept

Seg'one we have the goal of learning:

explanatory models Tambi'en called descriptive. try

mostar c'omo depends Y of $X_{one..}$ $X_{N:}$ decisi'on spanning trees of classi fi ers Bayesian models regresi'on, model rules

predictive models . They do not seek both show the dependence

as given an item or_i with values x_{ij}, j = one.. N get the value Y_i of the target variable. Yes Y It is discreet, the class to which it belongs. M'etodos the nearby M'as neighbor, m'etodos based on neural networks. SVM

Introduction

The general process of classifying caci' on

one.- September 1 data values are considered in Y. set of training

items variables X_{one} X_{two}			· · · · · ·	X_N	Y
or_{one}	x_{eleven}	x_{12}	· · · · · ·	$x_{one\,N}$	Y_{one}
⋮	⋮	⋮	· · · · · ·	⋮	⋮
or_M	$x_{M\,one}$ $x_{M\,two}$		· · · · · ·	x_{MN} Y_M	

two.- Is constructed (learn) the classified models caci'on

Introduction

The general process of classifying caci' on

3.- Tested in another dataset joint test caculando the

values Y_{pred}

items variables X_{one}	· · · · · · · X_N YY	$pred$	
or_{one}	x_{eleven} · · · · · · · $x_{one\ N}$ Y_{one}		Y_{pred} one
⋮	⋮ · · · · · · · ⋮	⋮	⋮
or_n	$x_{n\ one}$ · · · · · · · x_{nN} Y_n	Y_{pred}	n

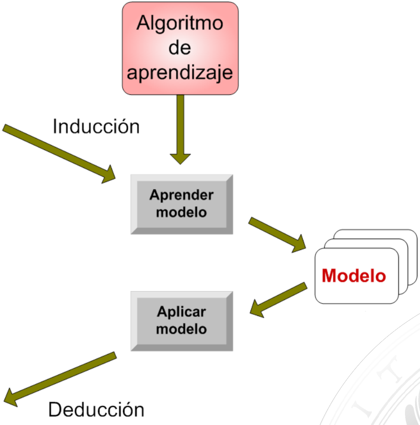
4.- is eval' ua in the model seg' one different criteria: precisi'on, error caci'on classified, scalability, interpretability, complexity, etc.

Introduction

The general process of classifying caci' on

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Conjunto de entrenamiento



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Conjunto de prueba

Caci'on classified by trees Decisi'on

b'asicas ideas

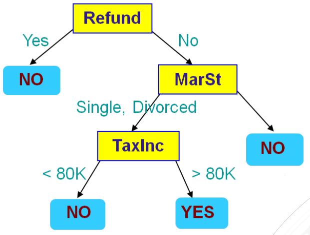
- *The spanning trees of decisi'on try to find a structure for explaining jer'arquica c'omo different parts of the input space corresponding to different values of the attribute object.*
- *The tree has three types of nodes:*
  - *Root node where begins*
  - *internal nodes each of which has an input shaft and two or more output partitions the subspace corresponding to this node*
  - *Leaf node has output shafts and est'a labeled with a target attribute valos*
- *At each node that is not part of sheet input space is divided into several subsets seg' one value of a given attribute, to reach leaf nodes.*
- *In principle the importance of each attribute in the process of classifying caci'on so it is possible that different results are obtained for the same problem is unknown*

Caci'on classified by trees Decisi'on

Example: tree from data

categórico categórico continuo clase

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



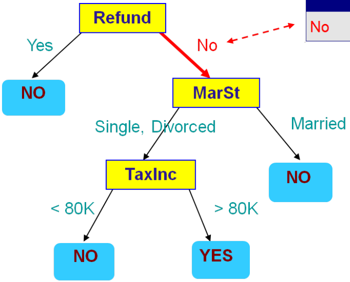
Conjunto de entrenamiento → **Modelo de clasificación: Árbol de decisión**

15

Caci'on classified by trees Decisi'on

Example: classifying caci'   on an item

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Modelo de clasificación:
Árbol de decisión

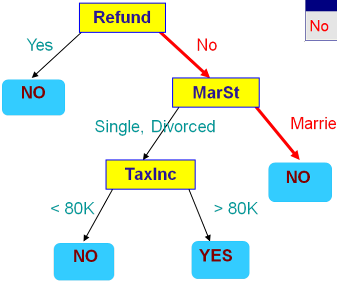
Caci'on classified by trees Decisi'on

Example: classifying caci' on an item

Caso de prueba

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	No

Clase 'No'



Modelo de clasificación: Árbol de decisión

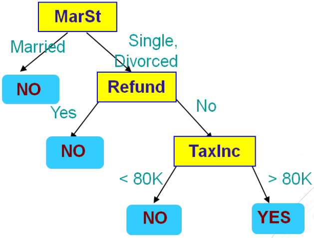
Caci'on classified by trees Decisi'on

Example: other tree

categórico categórico continuo clase

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Podemos construir distintos árboles: ¿cuál es mejor?



Conjunto de entrenamiento → **Modelo de clasificación: Árbol de decisión**

Construcci' Trees on Decisi' on

b'asicas ideas

- Find a spanning tree of decisi'on'optimo is not trivial. Is two *M* where *M* It is the n' umber of examples
- peque~ reasonably spanning tree is sought not explain properly training set
- It uses a Greedy strategy converting the problem NP problem.
- It is part of a root node and goes rami fi ed each node of the "best" possible way

Construcci' Trees on Decisi' on

b'asicas ideas

? Algorithm "divide and vencer'as"

one. We started with all training examples in ra'ız

spanning tree of decisi'on.

two. Examples are dividing into funci'on of the attribute

Select to rami fi ed the spanning tree on each node.

3. If a node contains examples of s'olo a class is transformed into leaf

Four. The attributes used to rami fi cation are chosen from funci'on

one heuristic.

5. The form of rami fi ed tambi'en

When building a tree decisi'on stops?

- When all remaining examples belong to the same class (is å~ nade a leaf spanning tree with the class label).
- When there are attributes for which rami fi car (is å~ nade one leaf labeled M'as frequent class in the node).
- When we are not classifying data.

¿Given a non-leaf node, as is partitioned

binary nodes It has no problem

nominal nodes Two options:

- Partitioning all values (partici'on nary)
- Group and converted into binary

ordinal nodes Two options:

- Repartition all values
- Group and converted into binary by fixing a cut point ($\leq v, > v$)

? Criteria for Node partici'on

¿Given a non-leaf node, as is partitioned

num'ericos nodes We have two options:

- Discretize the attribute and treated as ordinal
- Group and converted into binary by fixing a cut point ($\leq v, > v$)

Utilzan different algorithms partici'on different forms: single binary CART

ID3 s'olo discrete attributes and particici'on nary attributes for categ'oricos

C4.5 partici'on nary and binary to continuous, etc.

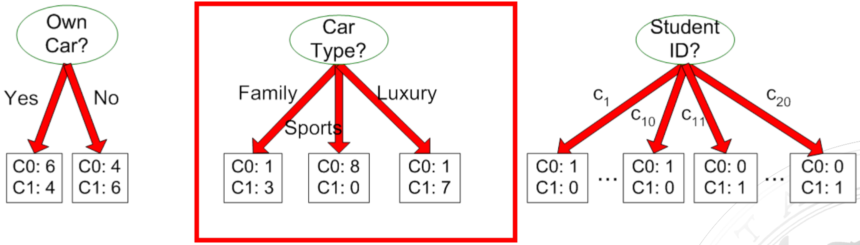
Given a non-leaf node, ¿Qu'e attribute is chosen to partition?

- We want a tree peque~   do not. The goal is to reach as leaf nodes before.
- Partitions need s'olo elements class
- Partitions are looking very homog'eneos nodes
- We will use measures based on the diversity of classes in each element of the partici'on. *measures impurity*

Construcci' Trees on Decisi' on

Selecci' criteria on attributes

Example



Construcci' Trees on Decisi' on

? Selecci'on measures

- Be *p (i / t) i* ∈ *{* one, two.. *c}* the fracci'on of items belonging to the class *i* that est'a in a given node *t,* obviously *c* It is the n'

  umber of classes
- *p (i / t)* It is an approximation of the probability of finding an item class *i* in that partici'on *t* It represents.
- According to the above idea, the more uniform are the values of *p (i / t)* It is less desirable *t* to be selected.
- The worst possible value for *p (i / t) i* ∈ *{* one, two.. *c}* is (one/ *nt,* ..., one/ *nt)* where *nt* It is the n' umber of element *t.* The best is (0 .., one, ..0)

Construcci' Trees on Decisi' on

Selecci' criteria on attributes

? Selecci'on measures based on Entropy

entropy *Entropy (t) = - Σ c i = one p (i / t) log two(p (i / t))*

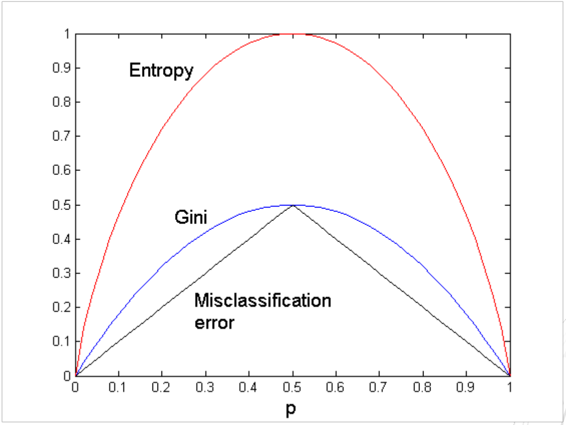
Gini index *Gini =* one *- Σ c i =* one *p (i / t)* two

Error classi fi caci'on *error =* one *- max i (p (i / t))*

Example

Distribuci'on node		Gini	entropy	Error
None	Class1 0 / Class2 1	0	0	0
Ntwo	Class1 Class2 / 0.2 0.8	0.278	0.650	0.167
None	Class1 Class2 / 0.5 0.5	0.5	one	0.5

? Selecci'on measures based on Entropy, for two classes



Construcci' Trees on Decisi' on

Selecci' criteria on attributes

? Informaci'on gain

To see c'omo operates a divisi'on compare the measure of a parent node to the child nodes. Sean p a parent node

v_j, j = one... k their children, $N(p)$ n' umber of elements in the node p Y

$N(v_j)$ n' umber of elements v_j. We define:

Gain $\Delta = I(p) - \sum_{j=one}^{k} \frac{N(v_j)}{N(p)} I(v_j)$ where $I(.)$ It is one of the measures before fi ned. Sis $I(.)$ is the Entropy is called *Informaci'on gain Δ_{info}*

Proporci'on Gain Gain $Gainratio = \frac{\Delta_{info}}{SplitInfo}$ where

$$SplitInfo = -\sum_{j=one}^{k} \frac{N(v_j)}{N(p)} \log_{two}\left(\frac{N(v_j)}{N(p)}\right)$$

Δ_{info} used ID3 and *GainRatio* in C4.5. CART, SLIQ..utilizan the Gini'ındice

Construcci' Trees on Decisi' on

Selecci' criteria on attributes

Comparaci'on rules divisi'on

Informaci'on gain Biased towards attributes with many different values.

Proporci'on gain Partitions tend to prefer slightly balanced (with a partici'on M'as much larger than the other)

Gini index It works worse when there are many kinds and tends to partitions favor tama~ and no similar purity.

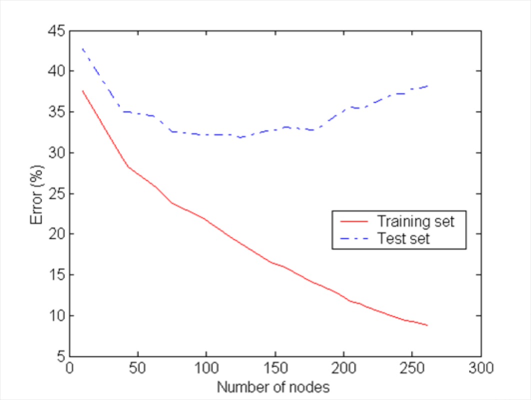
No rule of divisi'on is significantly better than dem'as

Construcci' Trees on Decisi'   on

Additional questions: overlearning

The greater complexity, models classified caci'on M'as fit the training set overlearning

.



Construcci' Trees on Decisi' on

Additional questions: overlearning

A soluci'on to overlearning are *T'ecnicas pruning* which they are developed to simplify the spanning tree. To prune a spanning tree of decisi'on

- one sub'arbol by a leaf node (corresponding to frequent M'as class in the sub'arbol) is replaced
- Or, a sub'arbol other sub'arbol contained in the first. There are techniques

previous Poda It is shrinking the spanning tree when it is generating

Poda post the spanning tree once generated is reduced For criteria see

bibliogr'a fi b'asica

Construcci' Trees on Decisi'   on

additional issues

- The classified caci'on / predicci'on by spanning trees of decisi'on is one of the M'as t'ecnicas studied within the classification caci'on.
- There are many variants and extensions of algorithms b'asicos funci'on of: partici'on mechanisms of space, use of t'ecnicas pruning, use of additional mechanisms as the rules of asociaci'on etc.
- The idea has spread to predict m'etodos for continuous attributes *Regresi'on trees*
- They have been extended to consider criteria partici'on fuzzy partitions domain labels establishing LING u'ısticas for attributes discretized *Fuzzy Decison Trees*

Construcci' Trees on Decisi' on

additional issues

Advantages of spanning trees of decisi'on

- F'acil interpretaci'on (when peque~ us).
- Quickly to classify new data.
- Precisi'on comparable to other t'ecnicas.

Some algorithms ef fi cient and scalable

- PUBLIC (Rastogi and Shim, VLDB'1998) integrates pruning process spanning tree construcci'on
- RainForest (Gehrke et al., VLDB'1998) separates what algorithm determines scalability
- BOAT (Gehrke et al., PODS'1999) s'olo needs to run 2 times the dataset

Caci'on classified by rules

b'asicas ideas

objective

Classify records using a colecci'on rule "if then"

The form of a rule is Condition - → Y Where:

- condition is a conjunci'on conditions on the value of various attributes, tambi'en antecedent
- Y is the value of third chapter is devoted class consequent

Examples of rules

- Blood Type = hot ∧ Lays eggs = Si - → Bird
- Income 6 30 ∧ Devoluci'on = yes - → no evader

Caci'on classified by rules

b'asicas ideas

Given a rule r we say that it covers an instance x the dataset if that body satisfies the history of the rule

Example

R1: (Viviparo = no) $\wedge$ (Puede volar = yes) $\rightarrow$ Pajaro

R2: (Viviparo = no) $\wedge$ (Acuatico = yes) $\rightarrow$ Pez

R3: ((Viviparo = yes) $\wedge$ (Sangre = caliente) $\rightarrow$ Mamífero

R4: ((Viviparo = no) $\wedge$ (Puede volar = no) $\rightarrow$ Reptil

R5: (Acuatico = a veces) $\rightarrow$ Anfibios

First name	Blood	Viviparous	Can fly	Water	Class
Hawk	Hot	do not	yes	do not	bird
Bear	Hot	yes	do not	do not	mammal
Platypus	Hot	do not	do not	Sometimes	mammal
Lemur	Hot	yes	do not	do not	mammal
Turtle	cold	do not	do not	Sometimes	an bio fi

Caci'on classified by rules

b'asicas ideas

Example

- Halcon is covered by R1
- Platypus and turtle are covered by R5 A rule is:

Coverage Proporci'on of records that satisfy their background

Precision Proportion of records and meeting records consistent

Example

Ruler	Cober.	Preci.
R1	1/5	1/5
R2	0	0
R3	2/5	2/5
R4	2/5	0
R5	2/5	1/5

Caci'on classified by rules

b'asicas ideas

A test set is classi fi ed register to register firing rules corresponding to the values of each of them, and scoring his class

Example

Hot Gorri'on no yes no bird					

Register shoot *Rule R1*

Caci'on classified by rules

b'asicas ideas

Regarding your aplicaci'on a set of rules can be:

Mutually exclusive Yes:

- Each rule can be applied independently
- Any registration est'a covered as much by a rule

Comprehensive Yes:

- There is a rule for any possible attribute values combinaci'on
- Any registration est'a covered by at least one rule

Caci'on classified by rules

b'asicas ideas

Example. set of mutually exclusive and exhaustive rules

r1: (Sangre= fria) $\rightarrow$ No mamífero

r2: (Sangre=caliente) $\wedge$ (Vivíparo = yes) $\rightarrow$ Mamífero

R3: ((Sangre = caliente) $\wedge$ (Viviparo = No) $\rightarrow$ No Mamífero

If a set of rules is mutually exclusive and exhaustive, each record to classify triggers a rule ys'olo one

Caci'on classified by rules

b'asicas ideas

¿Qu'e do when a set of rules does not have these properties?

- If not comprehensive malaria, it defines a default class and assigned to it the uncovered records
- If not exclusive, a record may be covered by several contradictory rules. solutions:
  - the rules are sorted by alg' criteria: coverage, precisi'on class that define etc., and priority rule applies M'as
  - all the rules for the registration skyrocket and assigned the class "voted M'as" quiz'as weighted by the weight of the rules

The mayor'ıa algorithms do not produce unique rules follow a criterion ordenaci'on

Caci'on classified by rules

Extracci'   on rules: general ideas

Given a set of TRAINING WHAT C'omo extract a set of rules?

From a spanning tree of decisi'on Simply describe the spanning tree by a

Set of rules.

- They are mutually exclusive
- They are exhaustive
- Informaci'on contain all the spanning tree

direct M'etodos Act'   ohn directly on the data, known they are those of M'as

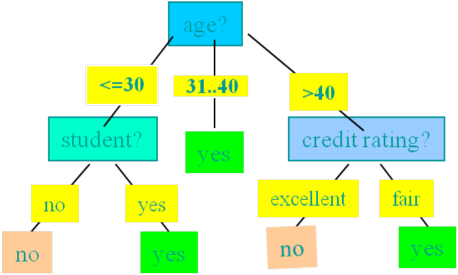
m'etodos *sequential coating* .

CN2, RIPPER and its variants etc.

Caci'on classified by rules

Extracci' Rules on using spanning trees of decisi' on

Example



IF (age<=30) AND (student=no) THEN buys_computer = no

IF (age<=30) AND (student=yes) THEN buys_computer = yes

IF (30<age<=40) THEN buys_computer = yes

IF (age>40) AND (credit_rating=excellent) THEN buys_computer = no

IF (age>40) AND (credit_rating=fair)THEN buys_computer = yes

Caci'on classified by rules

Extracci' direct on rules: b'asicas Ideas

The process is the b'asico sequential coating

one. Start with a set of rules vac'ıo

two. generate best rule covering a particular class

3. TO nadir learned rule to set

Four. Remove the examples set of training covered by Rule

5. Failure to comply with the stopping rule go 2 otherwise stop

Caci'on classified by rules

Extracci' direct on rules: b'asicas Ideas

To extract the best rule:

one. classes are ordered (no criteria high to low or contrary)

two. They are considered positive examples of that class and the negative rest

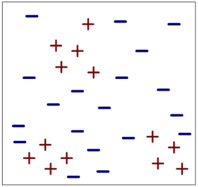
3. The best rule is covering many positive examples and few negative. evaluaci'on measures used for this purpose.

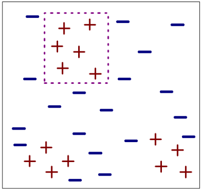
Four. A rule may need to be pruned if it causes problems of overlearning.

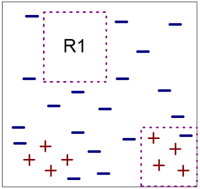
Caci'on classified by rules

Extracci'   direct on rules: b'asicas Ideas

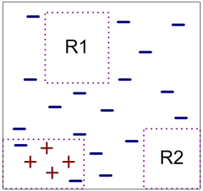
Example rules selecci'on


(i) Original Data


(ii) Step 1



(iii) Step 2



(iv) Step 3

Caci'on classified by rules

some algorithms

- FOIL (Quinlan, Machine Learning, 1990)
- CN2 (Clark and Boswell, EWSL'1991)
- RIPPER (Cohen, ICML'1995)
- PNrule (Joshi and Kumar Agarwal, SIGMOD'2001

Bayesian classi fi caci'on

b'asicas ideas

- There are problems in which the relaci'on between items and classes have a random component
- Even items (instances) with equal values in the attributes may belong to different classes (diagn'ostico of seg' diseases

  one s'ıntomas)
- b'asico principle If you can not ensure qu'e class belongs to an instance, assign the class that is more likely to belong

Bayesian classi fi caci'on

b'asicas ideas

- Two problems:

  TO Given a particular instance, it is estimated the ¿C'omo
  M'as likely class to which it belongs? *By Bayes theorem*

  B How are stored / calculate M'as classes
  likely seg' one possible combination of values
  attributes so efficient? *Hip'otesis simplified cadoras*

  Hip'otesis of independence for discrete attributes
  M'etodos leads to "Naive Bayes"
  Hip'otesis joint standard for continuous attributes
  distribuci'on Leads to An'alisis
  Discriminant and its variants

Bayesian classi fi caci'on

probabil'ıstico Model: Bayes theorem

- Given random events A and C we have:

$$P(C \mid A) = \frac{P(A, C)}{P(A)} \qquad P(C)\,P(A \mid C) = P(A, C) \Rightarrow$$

$$= \Rightarrow P(C \mid A) = \frac{P(A \mid C)\,P(C)}{P(A)}$$

- *Example of Use*
    - It is known that meningitis causes stiff neck in 50% of cases.
    - It is known that the probability of having meningitis is 1/50000 and that a patient has a stiff neck 1/20.
    - So:

$$P(Men \mid Rg) = \frac{P(Rg \mid Men)\,P(Men)}{P(Rg)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayesian classi fi caci'on

probabil'ıstico model

- Suppose that both attributes $X_{one..}$ X_N as the class Y They are random variables.

- Given an instance (item) with attribute values $x_{one..}$ x_N

  we want to predict the value of its class Y
- Speci fi cally want to find the value that maximizes expresi'on:

$$Prob\ (Y = y \mid X_1 = x_{one},\ X_2 = x_{two},\ ...,\ X_n = x_N)$$

to simplify $P\ (y \mid x_{one},\ x_{two..},\ x_N)$

- *issue* :

  Can we calculate $P\ (y \mid x_{one},\ x_{two..},\ x_N)$, known ($x_{one},\ ..\ x_N$) from the data?

- *Solution* The use of Bayes' theorem.

Bayesian classi fi caci'on

probabil'ıstico model algorithm b'asico

one. For all $Y \in do\ my\ )$ calculate:

$$P(y \mid x_{one}, x_{two..}, x_{n}) = \frac{P(x_{one}, x_{two..}, x_{N} \mid y)\, P(y)}{P(x_{one..}\ x_{N})}$$

two. To choose $\hat{Y}$ such that

$$P(\hat{and} \mid x_{one}, x_{two..}, x_{n}) = \max_{Y \in do\ my\ )} \frac{P(x_{one}, x_{two..}, x_{N} \mid y)\, P(y)}{P(x_{one..}\ x_{N})}$$

3. It really is equivalent to choose $\hat{Y}$ such that

$$P(\hat{and} \mid x_{one}, x_{two..}, x_{n}) = \max_{Y \in do\ my\ )} P(x_{one}, x_{two..}, x_{N} \mid y)\, P(y)$$

Four. *issue* . ¿C'omo estimate $P(x_{one}, x_{two..}, x_{N} \mid Y)$?. In principle it is joint distributions of N random variables, each conditional value of the domain class

Bayesian classi fi caci'on

Naive Bayes classi fi er

conditional independence between attributes is assumed

$X_{one..}$ X_N *so that:*

$$\forall Y \in Dom(Y)\ P(x_{one},\ x_{two..},\ x_{N}\,/\,y) = P(x_{1}\,|\,y)\ P(x_{2}\,|\ and)\ ..\ P(x_{N}\,/\,Y)$$

one. $\forall Y \in Do\ my)\ \forall j \in \{$one, .. $N\}$ It can be estimated $P(x_{j}\,/\,Y)$ Y $P(y)$ using the training set

two. Since a new item value ($v_{one..}$ $v_{N})$ It is categorized in class z such that:

$$P(z)\ P(v_{1}\,|\,z)\ ..\ P(v_{N}\,/\,z) = \max_{Y \in do\ my)}\ P(y)\ P(v_{1}\,|\,y)\ ..\ P(v_{N}\,/\,Y)$$

Bayesian classi fi caci'on

Naive Bayes classi fi er: simple example

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Chance of classes

$P(y) = M_{Y}/M$, P (Si) = 10.03, P (N) = 7/10

Probability of discrete attributes

$$P(x_j \mid y) = m_{x_j Y}/m_Y$$

P (Status = Married | no) = 4/7

Probability of continuous attributes

Hip'otesis normal

$$P(x_j \mid y) = (\text{one}/\sqrt{\text{two}\ \pi\sigma^{\text{two}}_{jy}})\ \exp\left(\frac{x_j - \mu_{jy})^{\text{two}}}{\text{two}\ \sigma^{\text{two}}_{jy}}\right)$$

P (Income = 120 | No) = 0.0072

Bayesian classi fi caci'on

Naive Bayes classi fi er: simple example

Be X = (Refund = NO, Married, Income = 120)

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

P(X|Class=No) = P(Refund=No|Class=No)
× P(Married| Class=No)
× P(Income=120K| Class=No)
= 4/7 × 4/7 × 0.0072 = 0.0024

P(X|Class=Yes) = P(Refund=No| Class=Yes)
× P(Married| Class=Yes)
× P(Income=120K| Class=Yes)
= 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

Bayesian classi fi caci'on

Naive Bayes classi fi er: estimaci' on the probabilities of discrete attributes

In general $\forall Y \in Do\ my\ )\ \forall j \in \{$one, .. $N\}\ P\ (x_{j\,|}\ Y)$ it is estimated:

$$P\ (x_{j\,|}\ y) = \frac{\gamma + m_{x_j Y}}{\gamma m_{X_j} + m_Y}$$

where m_{X_j} It is the n' umber of elements $dom\ (X_{j)})$ Y $p\ (y)$ it is estimated as:

$$p\ (y) = \frac{\gamma + m_Y}{\gamma m_Y + m}$$

where m_Y It is the n' umber of elements having $do\ my\ )$ Y m the number of elements of the dataset

- constant γ It is called *CORRECTION Laplace* .
- Usually it is taken equal to zero; but to treat cases of non-existent attribute values is taken equal to 1 or 1/2
- It is used when the training set is peque~ do not

Bayesian classi fi caci'on

Naive Bayes classi fi er: estimaci' on the probabilities of continuous attributes

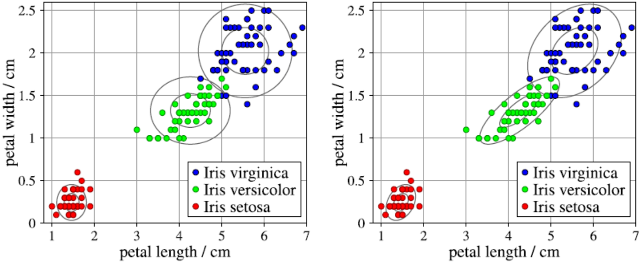
- A continuous attributes X_j It is considered distributed $N(\mu_{jY}, \sigma^2_{jy})$ all class value Y. And his conditional probability is given by $f(x_j \mid y) = N(\mu_{jY}, \sigma^2_{j})(x_j)$
- When it has a set of continuous attributes $\bar{X}$, It can hip'otesis avoid assuming conditional independence, $\forall Y, (\bar{X} \mid y)$ It is distributed seg' a A Multivariate Normal $N(\mu_{\bar{X} \mid and}, \Sigma_{\bar{X} \mid y})$, You can then calculate the joint conditional probability $F(\bar{x} \mid y)$.
- When s'olo have num'ericos attributes do not impose the m'etodos *hip'otesis independence* and we have *classi fi ers complete Bayes*

Bayesian classi fi caci'on

Naive Bayes classi fi er: estimaci' on the probabilities of continuous attributes

Example

Iris type	Iris setosa	Iris versicolor	Iris virginica
Prior probability	0.333	0.333	0.333
Petal length	1.46 ± 0.17	4.26 ± 0.46	5.55 ± 0.55
Petal width	0.24 ± 0.11	1.33 ± 0.20	2.03 ± 0.27



Bayesian classi fi caci'on

Naive Bayes classi fi er overview

- They are robust against noise isolated points and working on midrange
- Ignoring orders let you manage instances values that are at the stage of ESTIMATION odds.
- They are robust to irrelevant attributes because if an attribute *X* It has no influence on *YP (X / Y)* It tends to distribuci'on uniform.
- The conditional independence hip'otesis everyone attributes can be very strong.
  - Bayesian networks generalize the model and make this hip'otesis fl exible M'as

Bayesian classi fi caci'on

Discriminant an'alisis

- It is a particular case of *classi fi er full Bayes*
- Hip'otesis Simpli fi cadoras:
  - Num'ericos attributes
  - Distribuci'on regular mutivariante. With restrictions:
    - Covariance matrix equal for classes
    - very different half
  - Initially two classes s'olo
- The calculation is based on class'optima calculate a set of hyperplanes that divide the space for classes.

Classify instances based caci'on

- The classi fi ers studied so far, working in two stages:

  inductively Learning model classi fi caci'on

  inferential Apply the model to the test examples set

- They are " forward classi fi ers "(eager learners)

b'asica Idea

Why qu'e not store the entire training set and when you get a test example search, items that "M'as will appear" and assign class ?.

They are m'etodos "lazy" Lazy Learners

Classify instances based caci'on

Set of Stored Cases

Atr1	………	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

Unseen Case

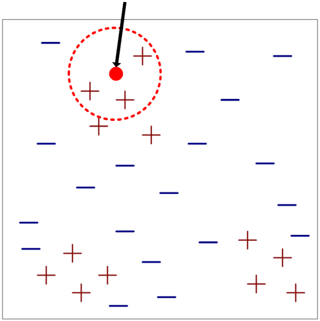
Atr1	………	AtrN

Classify instances based caci'on

- M'etodos obviously are predictive. They explain nothing.
- Examples:
  - The "memor'ısticos classi fi ers" (Rote Learners)
    - Stores the training set and only assigns a class to an example when there is a training item that is exactly like'el.
  - K-nearest neighbors M'as (K nearest neighbor) (K-NN)
    - Select the k-items that "M'as resemble" the example and assign the frequent M'as class or "important M'as"

K-nearest neighbors M'as (K-NN)

b'asicas ideas



requirements

Distance between records *d (.,.)*

The value k fi xed

Algorithm

1. Sea *x* calculate *d (x, e), ∀ and ∈ AND*

two.- Identify { *and i, i =* one... *k}*

k nearest neighbors M'as ax

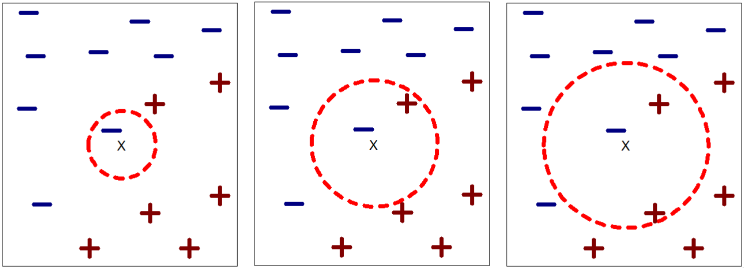
3.- classes { *and i, i =* one... *k},*

get the kind of *x*

Mayor'ıa or weighted by mayor'ıa

K-nearest neighbors M'as (K-NN)

b'asicas ideas



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

It is important to fi x the value of k. Since it can lead to overlearning or error.

K-nearest neighbors M'as (K-NN)

The function away .

- Est'a's method designed to num'ericos attributes
- You can use distances proposals on the subject grouping

- Should take into account problems of scale

The value of k .

- It is best to get it through validaci'on cross
  (*is subsequently ver'a)*

Class mechanism selecci'on .

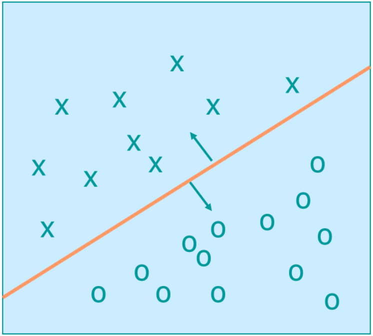
- Usually choose the majority class
- It can be weighted by the distance from the neighbor point or
  more functions so this fi sticadas

Other predictive models classi fi caci' on

Classi fi ers based on Neural Networks

They have been studied in other subjects

Support Vector Machines SVMs



Other predictive models classi fi caci' on

Support Vector Machines SVMs

Advantage . • Precisi'

on high

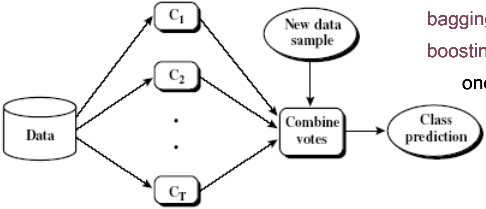
- Robustness against noise

drawbacks • Expensive to train. (Little scalable and ef fi cient)

- Di fi cult to interpret (Increase dimension space to separate classes hierplanos)

Other models classified caci'on

"Emsembles"



Combine several models to improve the classi fi er precisi'on deciding qu'e class is assigned:

bagging Votaci'on by majority
boosting Votaci'on weighted seg'
one classi fi er quality (*AdaBoost)*

Evaluaci'on models classified caci'on

b'asicas ideas

Process steps classi fi caci'on:

one.- September 1 data values are considered in Y. set of training

two.- Is constructed (learn) the classified models caci'on

3.- Tested in another dataset joint test caculando the values Y_{pred}

4.- is eval' ua in the model seg' one different criteria: precisi'on, Error classi fi caci'on, scalability, interpretability, complexity, etc.

Evaluaci'on models classified caci'on

b'asicas ideas

- *Evaluaci'on different aspects of modeling*

M'etricas They are measures of the "quality" of a process classi fi caci' on.

Models So they used to estimate reliably measures quality.

Comparison Are t'ecnicas that compare performance concerning two models classified caci'on

Evaluation of models classified caci' on

b'asicas ideas

- *Criteria for quality measures*

Precision ¿C'omo well classifies the model?

Ef fi ciency needed to build / use the classi fi er time

Sturdiness Against noise and nulls

scalability You admit large datasets?

interpretability Does it explain the model?

Complexity Tree with many nodes etc.

The first two may be m'etricas while ' last

s'olo can be measured in some cases.

Precisi'on measures ser'an key

Evaluaci'on models classified caci'on

Precisi' measures on

Sean: M one classi fi er, and $x_i = x_{i\,\text{one}}, .. x_{iN}$ an example of the independent variable of a test set $T = \{x_{\text{one}}\ Y_{\text{one}}, .. x_n\ Y_n\}$.

Be $\hat{Y}_i = M(x_i)$ the result of applying the process to M x_i.

We define:

Precisi'on M .

$$Acc = \text{one}/\, n \sum_{i=\text{one}}^{n} I(and_i = \hat{Y}_i)$$

where $I(e)$ It equals 1 if *and* It is true, and 0 if *and* it is false.

Error rate M .

$$Er = \text{one}/\, n \sum_{i=\text{one}}^{n} I(and_i \,6= \hat{Y}_i) = \text{one} - Acc$$

It seems that it is best to have a low error rate / high precisi'on but:

If it is excessively adjusted to set s'olo learn this training model. overlearning

The aparici'on of overlearning of due to various causes, some dependent models; but others can be smoothed.

Evaluation of models classified caci' on

Precisi' measures on

Some considerations on overlearning

- The greater the complexity of a model classi fi caci'on, M'as excessively adjusted to the training set (in spanning trees of decisi'on)

- Another cause of overlearning is the presence of noise points (on models that partition the space: an'alisis discriminant, SVM

  eg)

- By Finally a shortage of points in a class against other overlearning can give problems, since the precisi'on not take into account the tama~

  no classes

To avoid this' Last issue must take into account the weight classes precisi'on measures for this purpose

Arrays Confusi'on

Evaluation of models classified caci' on

Precisi' measures on: matrices confusi' on

Given a classi fi caci'on process if we define n_{ij} n' umber of elements assigned to the class i when est'an in class j we have:

Matrix confusi'on

predicted some	Y_{one}	Y_{two}	. . . .	Y_k	TOT_{pred}
$\hat{Y}_{one}$	n_{eleven}	n_{12}	. . . .	$n_{one\ k}$	m_{one}
⋮	⋮	⋮	. . . .	⋮	⋮
$\hat{Y}_k$	$n_{k\,one}$	$n_{k\,two}$	. . . .	n_{kk}	m_k
TOT_{CIER}	n_{one}	n_{two}	. . . .	n_k	n

Confusi'on matrices allow define measures associated with the classes and global measures weighted

Evaluation of models classified caci' on

Precisi' measures on: matrices confusi' on

They can define:

Precisi'on of a class $\forall i \in \{one..\ k\},\ acc_i = n_{ii} / m_i$

"Recall" of a class $\forall i \in \{one..\ k\},\ recc_i = n_{ii} / n_i$

F as a class $\forall i \in \{one..\ k\},\ F_i = two\ n_{ii} / (m_i + n_i)$

Precisi'on and recall the overall remain the same but you can define:

Global F $F = one / k \Sigma^{k}_{i = one} F_i$

Evaluation of models classified caci' on

Precisi' measures on: matrices confusi' on

Example

Iris data using "sepal length" and "sepal width". Naive Bayes classi fi er. 120 training examples and test data 30.

global data $acc = 0.733$ $er = 0.267$

Confusion Matrix

predicted certain	setosa	versicolor	Viginica	
setosa	10	0	0	10
versicolor	0	7	5	12
virginica	0	3	5	8
	10	10	10	30

Data associated with classes

	Accuracy	Recall	F-measure
setosa	one	one	one
versicolor	0.583	0.7	0.636
virginica	0.625	0.5	0.556

Overall = 0.731 F

Evaluation of models classified caci' on

binary problems: matrices confusi' on

Binary matrices confusi'on

- Classes are now P (positive), N (negative)
- The confusion matrix is

predicted some	positives	negatives	
positives	TP	FP	$P_{pred} = TP + FP$
negatives	FN	TN	$N_{pred} = FN + TN$
	$P_{ci} = TP + FN$	$N_{ci} = FP + TN$	n

- the measurements are calculated seg' one previous expressions.
- In the event that there is a great descompensaci'on between cases you can use the cost matrix

Evaluation of models classified caci' on

binary problems: ROC curves (Receiver Operating Characteristics)

Hypothesis

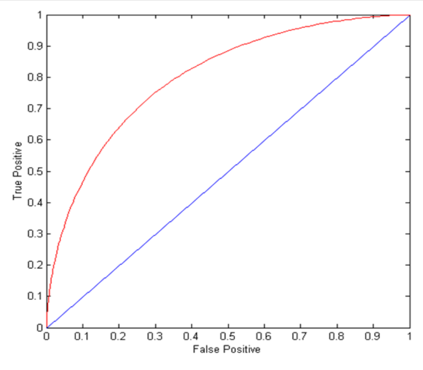
- A binary problem
- There is a measure $S(.)$ and a threshold ρ such that if $S(x_i) > \rho$ is categorized x_i as a positive case. For example, a classi fi er Bayes $S(x_i) = Prob(P \mid x_i)$ ys'olo we consider a positive example c'omo $S(x_i) > 0.8$

The ROC curve is constructed by varying the threshold ρ between its m'ınimo ym'aximo value and representing:

- On the Y axis the "positive rate certain" $TPR = \frac{TP}{P_{ci}} = \frac{TP}{(TP + FN)}$
- X in the "False Positive Rate" $FPR = \frac{FP}{N_{ci}} = \frac{FP}{(FP + TN)}$

Evaluation of models classified caci' on

binary problems: ROC curves



Yes ρ It is m'ınimo all positive (1.1) If ρ The same is m'aximo all negative n' umber of false to true value on the diagonal If the curve above est'a Good

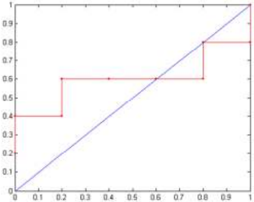
If the curve below est'a Bad

With values close to the curve (0.1) the best area under the curve $\approx$ 1 perfect

Evaluation of models classified caci' on

binary problems: ROC curves

Example constructi'on



Ejemplo	P(+|E)	Clase
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

Class	+	-	+	-	-	-	+	-	+	+	
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	two	two	one	0
FP	5	5	4	4	3	two	one	one	0	0	0
TN	0	0	one	one	two	3	4	4	5	5	5
FN	0	one	one	two	two	two	two	3	3	4	5
TPR	one	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	one	one	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

88

M'etodos for evaluaci' on models classified caci'on

issue In a real case, with a given dataset. ¿C'omo organize training and test sets?

- Precisi'on to evaluate a model classi fi caci'on, the whole test must be independent
- Divide the dataset into two. For example, podr'ıamos reserve 2/3 of the examples available for training and the remaining 1/3 utilizar'ıamos the test set.
- *issue* ¿C'omo split, qu'e data will training and what to test?
- A first idea *random Selecci'on* . Examples of the test set are drawn
    - By drawing overall uniform
    - Strati fi ed by lot seg' one classes

issue

Selecci'on random s'olo once made can be biased

- As soluci'on repeats *h* Sometimes the process and the average precisi'on ser'a $acc = one/ h \Sigma_{i = one}^{h} acc_i$
- Another alternative: *Validaci'on cross*

one. the dataset is divided into *h* equal parts

two. They are caught *h* - one parts and training of test.

3. Leaves of varying the test to repeat process *h* times. The precisi'on is average.

Variants validaci'on cross

Two- fl od cross validation . In this case h = 2.

Leave-one-out . If we have N examples in the data set, we divide N times, leaving N-1 training and test case. N execution process performed classified caci' on.

Validaci'on cross stratify each . Partitions are maintaining the initial proporci'on of elements in each class

M'etodos for evaluaci' on models classified caci'on

Bootrapping

- the training set is sampled with replacement. With what examples can be repeated
- If N is sufficiently large sample fi tama~ It contains no N about 63.2% of the examples.
- The data are chosen not part of the set test
- The process is repeated b times with an accuracy of ε_i, $i =$ one.. b.
- Precisi'on total can be calculated in various ways, the usual M'as is:

$$acc_{Boost} = 0632\ (1\ /\ b) \sum_{i=\text{one}}^{b} \varepsilon_i + 0.368\ acc_s$$

where acc_s precisi'on is obtained using the set of Total Training

Comparaci' on classi fi ers

b'asicas ideas

Classi fi ers to compare:

Comparaci'on gen'erica .

one. a set of problems is chosen

two. a quality measure is chosen. Habitually precision.

3. is eval'   ohn and performs alg'   a statistical test (Mean difference, ANOVA etc.) for comparing results

Comparaci' on classi fi ers

b'asicas ideas

Classi fi ers to compare:

Comparaci'on against particular problem .

- If the classi fi ers are binary ROC curves can be used to compare your actuaci'on on a particular problem.
- They can be used, validaci'on cross or bootstrapping to generate experiments and compare results without socks.

Not classi fi ers f'acil compare in general, M'as likely that some work better than others seg' one class problems