Introduction to machine learning Non Parametric classification

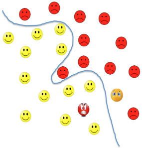
Classification Trees

Mathilde Mougeot

ENSIIE

2019

Classification Trees



Classification tree.

Application :
Cardiac Heart Disease (CHD variable $\{0, 1\}$)

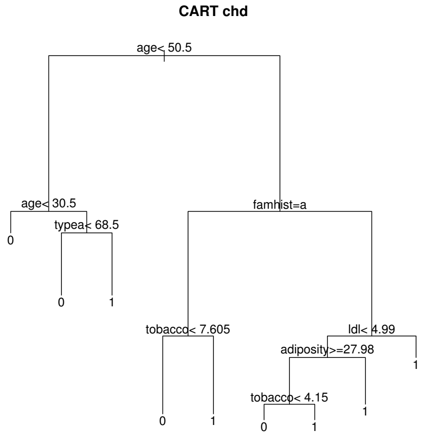
The data :

nř	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1
6	132	6.20	6.47	36.21	Present	62	30.77	14.14	45	0
7	142	4.05	3.38	16.20	Absent	59	20.81	2.62	38	0
8	114	4.08	4.59	14.60	Present	62	23.11	6.72	58	1
9	114	0.00	3.83	19.40	Present	49	24.86	2.49	29	0
10	132	0.00	5.80	30.96	Present	69	30.11	0.00	53	1

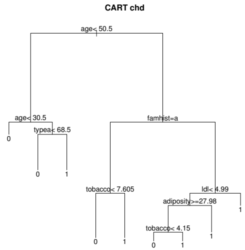
Classification tree.

Application : Cardiac Heart Disease (CHD variable $\{0, 1\}$)

Decision tree :

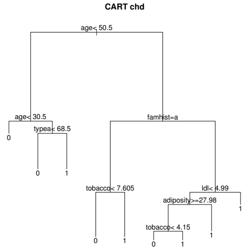


Decision tree
Some vocabulary



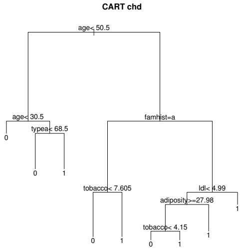
- Root : first node of the tree

Decision tree
Some vocabulary



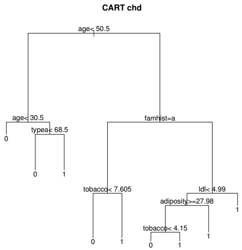
- Root : first node of the tree
- leaf : terminal node

Decision tree
Some vocabulary



- Root : first node of the tree
- leaf : terminal node
- Rule between the root and a leaf

Decision tree
Some vocabulary



- Root : first node of the tree
- leaf : terminal node
- Rule between the root and a leaf
- Regions : spaces

Decision Trees

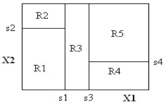
Leo Breiman, Friedman, Olshen 1984

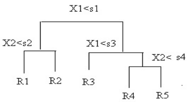
Decision tree is a method which splits the input space in a set of rectangularly domains, in which a **constant model** is adjusted.

$\rightarrow$ The global classification function is given by

$$f(x) = \sum_{m=1}^{M} c_m 1_{x \in \mathcal{R}_m}$$

c_m is a modality for each region $\mathcal{R}_m$





CART

Classification And Regression Tree, Arbres de Décision

- Y : target variable
  - Qualitative : Classification Tres
  - Quantitative : Regression Tree
  - The same approach let to study regression or classification problems

CART

Classification And Regression Tree, Arbres de Décision

- Y : target variable
  - Qualitative : Classification Tres
  - Quantitative : Regression Tree
  - The same approach let to study regression or classification problems
- X^j : covariables, $1 \leq j \leq p$ qualitatives ou quantitatives

CART

Classification And Regression Tree, Arbres de Décision

- Y : target variable
  - Qualitative : Classification Tres
  - Quantitative : Regression Tree
  - The same approach let to study regression or classification problems

- X^j : covariables, $1 \leq j \leq p$ qualitatives ou quantitatives

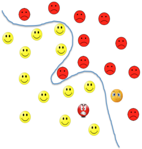
CART belongs to the Non Parametric method family.

No assumption is made on the data distribution

$\rightarrow$ The method builds a binary tree

CART-

Classification tree



Classification Trees

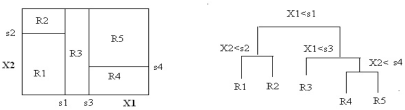
Classification tree is a method which splits the input space in a set of rectangularly domains, in which a **constant model** is adjusted.

$\rightarrow$ The global classification function is given by

$$f(x) = \sum_{m=1}^{M} c_m 1_{x \in \mathcal{R}_m}$$

c_m is a constant modality for each region $\mathcal{R}_m$

Question : How to compute c_m, $1 \leq m \leq M$, for the M regions ?



Classification Trees

Classification trees are used for a qualitative target variable and are associated in this case with a classification criteria.

Classification Trees

Classification trees are used for a qualitative target variable
and are associated in this case with a classification criteria.

- Using the Training set.
  For node m corresponding to region $\mathcal{R}_m$ with N_m observations

Classification Trees

Classification trees are used for a qualitative target variable
and are associated in this case with a classification criteria.

- Using the Training set.
  For node m corresponding to region $\mathcal{R}_m$ with N_m observations
  - In node m, the frequency for modality k is estimated :
    for $k \in \{1, ..., K\}$

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in \mathcal{R}_m} I(y_i = k; (x_i, y_i) \in \text{TrainDataSet})$$

  - In region (node) m, an **new** observation is affected to class k_0 if

$$k_0 = \arg\max_{k \in 1..K} \{\hat{p}_{mk}\}$$

which represents the most represented class in node m.

Classification Trees

- The classification function is given by

$$f(x) = \sum_{m=1}^{M} c_m 1_{x \in \mathcal{R}_m}$$

- Notations : c_m corresponds to the main modality on the training set for the region $\mathcal{R}_m$
- The estimated classification function is given by

$$\hat{f}(x) = \sum_{m=1}^{M} \hat{c}_m 1_{x \in \mathcal{R}_m}$$

- For an observation $x \in \mathcal{R}_m$, $\hat{y} = \hat{c}_m$. $\hat{c}_m$ estimated on Train DataSet.

Classification Tree. Construction of the Tree

In the classification setting, ther are several ways to measure the quality of a split.
Node Impurity measures (Left ou Right) :

- Missclassification :
  $\mathcal{D}_{\mathcal{R}_m} = \frac{1}{N_m} \sum_{i \in \mathcal{R}_m} I(y \neq k(m)) = 1 - \hat{p}_m$

Classification Tree. Construction of the Tree

In the classification setting, ther are several ways to measure the quality of a split.
Node Impurity measures (Left ou Right) :

- Missclassification :
  $\mathcal{D}_{\mathcal{R}_m} = \frac{1}{N_m} \sum_{i \in \mathcal{R}_m} I(y \neq k(m)) = 1 - \hat{p}_m$
- Gini index (most used) :
  $\mathcal{D}_{\mathcal{R}_m} = \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
  $= 2p(1-p)$

Classification Tree. Construction of the Tree

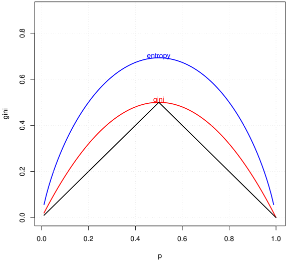
In the classification setting, ther are several ways to measure the quality of a split.
Node Impurity measures (Left ou Right) :

- Missclassification :
  $\mathcal{D}_{\mathcal{R}_m} = \frac{1}{N_m} \sum_{i \in \mathcal{R}_m} I(y \neq k(m)) = 1 - \hat{p}_m$
- Gini index (most used) :
  $\mathcal{D}_{\mathcal{R}_m} = \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
  $= 2p(1-p)$
- Entropy :
  $\mathcal{D}_{\mathcal{R}_m} = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}$
  $= -p \log_2 p + (1-p) \log_2 (1-p)$

with $p_m = \frac{1}{N_m} \sum_{i \in \mathcal{R}_m} I(y = k(m))$.
* For 2 classes and for each region

Decision trees

Comparaisons of node impurity



A function of impurity is by default chosen (R : gini)

Classification trees. Node creation

For a impurity measure

- k : node number

Classification trees. Node creation

For a impurity measure

- k : node number
- $\mathcal{R}_1$ et $\mathcal{R}_2$ regions of the two leaves

Classification trees. Node creation

For a impurity measure

- k : node number
- $\mathcal{R}_1$ et $\mathcal{R}_2$ regions of the two leaves
- The algorithm computes the optimal partition for which the value of $\mathcal{D}_{\mathcal{R}_1} + \mathcal{D}_{\mathcal{R}_2}$ is minimal

Classification trees. Node creation

For a impurity measure

- k : node number
- $\mathcal{R}_1$ et $\mathcal{R}_2$ regions of the two leaves
- The algorithm computes the optimal partition for which the value of $\mathcal{D}_{\mathcal{R}_1} + \mathcal{D}_{\mathcal{R}_2}$ is minimal
- i.e. at each step k , the split of "one upper region" in "two lower regions" (corresponding to the recursive construction of the tree) maximizes the difference of node impurity measure (deviance) :

$$\begin{array}{ll} \Delta\mathcal{D}_{\mathcal{R}\to\mathcal{R}_1+\mathcal{R}_2} & = \mathcal{D}_{\mathcal{R}} - \left(\frac{N_{\mathcal{R}_1}}{N_{\mathcal{R}}}\mathcal{D}_{\mathcal{R}_1} + \frac{N_{\mathcal{R}_2}}{N_{\mathcal{R}}}\mathcal{D}_{\mathcal{R}_2}\right) \\ \{X^j, 1 \leq j \leq p\} & \end{array}$$

Classification tree. Node creation. Illustration

Impurity measure : Gini index

		$Y=0$	$Y=1$	
Left	$X < S$	n_{11}	n_{12}	n_{1+}
Right	$X \geq S$	n_{21}	n_{22}	n_{2+}
Top	global	n_{+1}	n_{+2}	n_{++}

Classification tree. Node creation. Illustration

Impurity measure : Gini index

		$Y=0$	$Y=1$	
Left	$X < S$	n_{11}	n_{12}	n_{1+}
Right	$X \geq S$	n_{21}	n_{22}	n_{2+}
Top	global	n_{+1}	n_{+2}	n_{++}

Gini indices :

Top G	$2 * \frac{n_{+1}}{n_{++}}(1 - \frac{n_{+1}}{n_{++}})$
Left G_L	$2 * \frac{n_{11}}{n_{1+}} * (1 - \frac{n_{11}}{n_{1+}})$
Right G_R	$2 * \frac{n_{21}}{n_{2+}} * (1 - \frac{n_{21}}{n_{2+}})$

Classification tree. Node creation. Illustration

Impurity measure. Entropy

		$Y=0$	$Y=1$	
Left	$X < S$	n_{11}	n_{12}	n_{1+}
Right	$X \geq S$	n_{21}	n_{22}	n_{2+}
Top	(global)	n_{+1}	n_{+2}	n_{++}

Classification tree. Node creation. Illustration

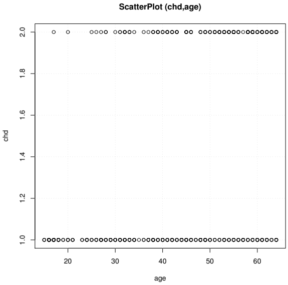
Impurity measure. Entropy

		$Y=0$	$Y=1$	
Left	$X < S$	n_{11}	n_{12}	n_{1+}
Right	$X \geq S$	n_{21}	n_{22}	n_{2+}
Top	(global)	n_{+1}	n_{+2}	n_{++}

Entropy	
top H	$\frac{n_{+1}}{n_{++}} log \frac{n_{+1}}{n_{++}} + \frac{n_{+2}}{n_{++}} log \frac{n_{+2}}{n_{++}}$
Left H_L	$\frac{n_{11}}{n_{1+}} log \frac{n_{11}}{n_{1+}} + \frac{n_{12}}{n_{1+}} log \frac{n_{12}}{n_{1+}}$
Right H_R	$\frac{n_{21}}{n_{2+}} log \frac{n_{21}}{n_{2+}} + \frac{n_{22}}{n_{2+}} log \frac{n_{22}}{n_{2+}}$

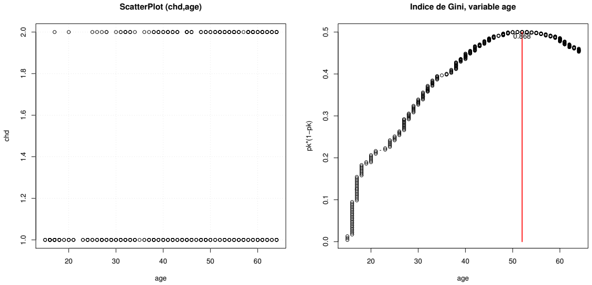
Decision tree. Cardiac Heart Disease application (chd)

The target variable is chd (binary variable).
$\rightarrow$ **Computation of the decision threshold for the quantitative co-variable age and for the Gini index**



Decision tree. Cardiac Heart Disease application (chd)

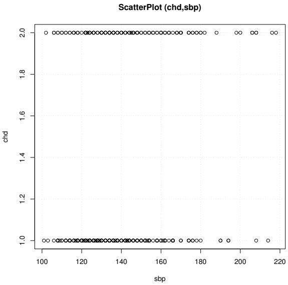
The target variable is chd (binary variable).
$\rightarrow$ **Computation of the decision threshold for the quantitative co-variable age and for the Gini index**



$\mathcal{D}$(age)=0.868

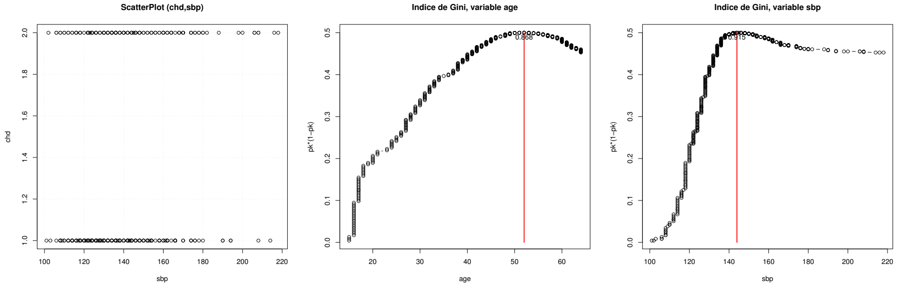
Decision tree. Cardiac Heart Disease application (chd)

Decision threshold computation and variable selection between two covariables (age, spb) based on the Gini Index



Decision tree. Cardiac Heart Disease application (chd)

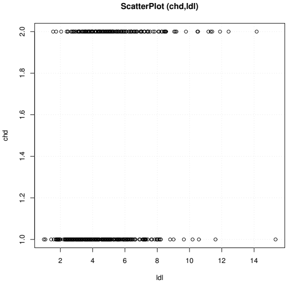
Decision threshold computation and variable selection between two covariables (age, spb) based on the Gini Index



The age variable is selected : $\mathcal{D}$(age)=0.868 < $\mathcal{D}$(spb)=0.915

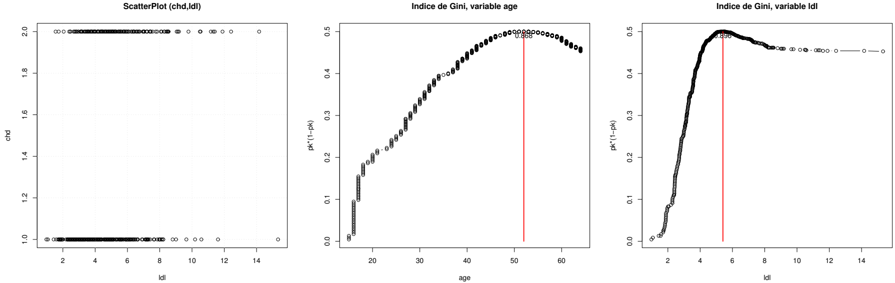
Arbre de décision, Application chd

Seuil de Décision, et choix de variable (age, ldl) à l'aide de l'indice de Gini



Arbre de décision, Application chd

Seuil de Décision, et choix de variable (age, ldl) à l'aide de l'indice de Gini



The age variable is selected : $\mathcal{D}(\text{age})=0.868 < \mathcal{D}(\text{ldl})=0.896$

Classification Trees

Stopping the recursive split process

A node is terminal if :

- the region is homogeneous (only one label)
- The is no authorized partitions regarding the algorithmic rule of decreasing the variance criteria ($\Delta\mathcal{D}$).
- The number of observations in the region NCut (or in the sub regions Nsize) is lower than a given threshold then No authorized split. (algorithm parameters).

Tuning parameters : (NCut, Nsize, $\Delta\mathcal{D}$)

Classification trees. Estimation. Prediction

From Estimation to Prediction :

- Y quantitative : average of observations of the training data set
- Y qualitative. Each leave is affected to one given class C_k of Y regarding a conditional approach regarding the training data set
  - **The more frequently class represented in the node (training data set)**
  - or

Classification trees. Estimation. Prediction

From Estimation to Prediction :

- Y quantitative : average of observations of the training data set
- Y qualitative. Each leave is affected to one given class C_k of Y regarding a conditional approach regarding the training data set
  - **The more frequently class represented in the node (training data set)**
  - or
  - the "more probable" class if some apriori exist (training data set)

Classification trees. Estimation. Prediction

From Estimation to Prediction :

- Y quantitative : average of observations of the training data set
- Y qualitative. Each leave is affected to one given class C_k of Y regarding a conditional approach regarding the training data set
  - **The more frequently class represented in the node (training data set)**
  - or
  - the "more probable" class if some apriori exist (training data set)
  - The "cheapest" class if there exists some coast indications.

Classification trees. Estimation. Prediction

From Estimation to Prediction :

- Y quantitative : average of observations of the training data set
- Y qualitative. Each leave is affected to one given class C_k of Y regarding a conditional approach regarding the training data set
  - **The more frequently class represented in the node (training data set)**
  - or
  - the "more probable" class if some apriori exist (training data set)
  - The "cheapest" class if there exists some coast indications.

Classification trees. Model selection. Pruning

In order to avoid (minimize) overfitting, the length of the tree is penalized.

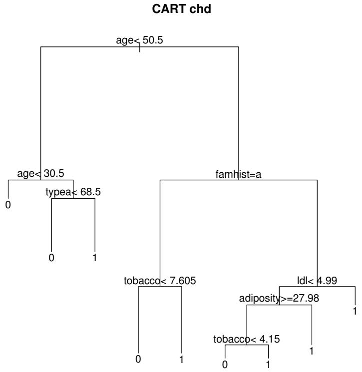
One the maximal tree is built (one tree), the pruning algorithm proposes several trees by pruning. A comparison between all these the trees helps to select the tree which minimize the following complexity criteria.
The tree with the lowest error is finally selected.
Complexity criteria :

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

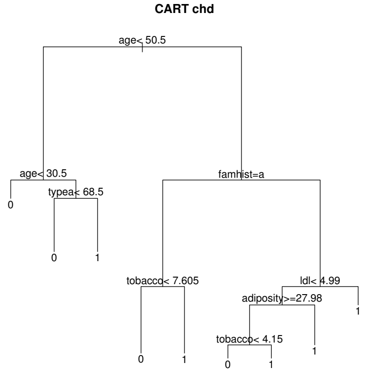
avec

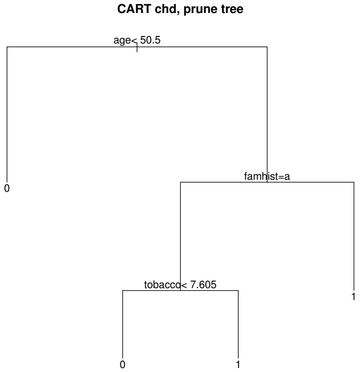
- $|T|$: terminal node number
- $N_m = \sharp\{x_i \in \mathcal{R}_m\}$
- $\hat{c}_m = \frac{1}{N_m} \sum_{x_1 \in \mathcal{R}_m} y_i$
- $\hat{Q}_m(T) = \frac{1}{N_m} \sum_{x_i \in \mathcal{R}_m} 1_{y_i \neq k_m}$
- α is selected by cross-validation

Classification trees. Pruning. Illustration



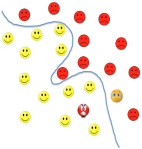
Classification trees. Pruning. Illustration





Decision trees.

Ensemble methods



Classification Trees

Model aggregation

- Bagging. Random strategies on the set of observations.
  Bagging for Bootstrap Aggeging (Breiman, 1996)

Classification Trees

Model aggregation

- Bagging. Random strategies on the set of observations.
  Bagging for Bootstrap Aggeging (Breiman, 1996)

- Random Forest. Random strategies on the observations and on the variables (Breiman, 2001)

Classification Trees

Model aggregation

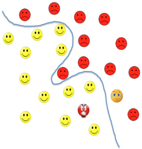
- Bagging. Random strategies on the set of observations.
  Bagging for Bootstrap Aggeging (Breiman, 1996)

- Random Forest. Random strategies on the observations and on the variables (Breiman, 2001)

→ This approach provides non linear classifiers

Decision trees

Bagging



Bagging. Breiman 1996

- Y target variable, qualitative or quantitative

Bagging. Breiman 1996

- Y target variable, qualitative or quantitative
- $X = (X^1, \ldots, X^p)$ Multi dimensional co variables ($\mathbb{R}^p$)

Bagging. Breiman 1996

- Y target variable, qualitative or quantitative
- $X = (X^1, \dots, X^p)$ Multi dimensional co variables ($\mathbb{R}^p$)
- $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ un n sample of $\mathcal{F}$ distribution law.

Bagging. Breiman 1996

- Y target variable, qualitative or quantitative
- $X = (X^1, \ldots, X^p)$ Multi dimensional co variables ($\mathbb{R}^p$)
- $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un n sample of $\mathcal{F}$ distribution law.
- $\Phi(x)$ a given model function of the covariable $x = (X^1, \ldots, X^p)$

Bagging provides a family of random models :

- B independant samples are generated with replacement : $\{\mathcal{S}_b\}_{b=1,B}$
  - Y quatitative : $\hat{\Phi}_B(.) = \arg\max_j \text{card}\{b | \hat{\Phi}_{\mathcal{S}_b}\}$ (vote)
  - Y quantitative : $\hat{\Phi}_B(.) = \frac{1}{B}\sum_{b=1}^{B} \hat{\Phi}_{\mathcal{S}_b}(.)$ average

Bagging. Breiman 1996

- Y target variable, qualitative or quantitative
- $X = (X^1, \ldots, X^p)$ Multi dimensional co variables ($\mathbb{R}^p$)
- $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un n sample of $\mathcal{F}$ distribution law.
- $\Phi(x)$ a given model function of the covariable $x = (X^1, \ldots, X^p)$

Bagging provides a family of random models :

- B independant samples are generated with replacement : $\{\mathcal{S}_b\}_{b=1,B}$
  - Y quatitative : $\hat{\Phi}_B(.) = \arg\max_j \text{card}\{b|\hat{\Phi}_{\mathcal{S}_b}\}$ (vote)
  - Y quantitative : $\hat{\Phi}_B(.) = \frac{1}{B}\sum_{b=1}^{B} \hat{\Phi}_{\mathcal{S}_b}(.)$ average
- $\Phi(.) = \mathbb{E}_{\mathcal{F}}(\hat{\Phi}_{\mathcal{S}})$ estimator with no biais

Bagging. Breiman 1996

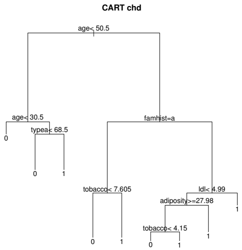
- Y target variable, qualitative or quantitative
- $X = (X^1, \ldots, X^p)$ Multi dimensional co variables ($\mathbb{R}^p$)
- $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un n sample of $\mathcal{F}$ distribution law.
- $\Phi(x)$ a given model function of the covariable $x = (X^1, \ldots, X^p)$

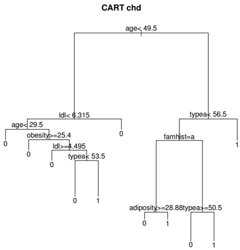
Bagging provides a family of random models :

- B independant samples are generated with replacement : $\{\mathcal{S}_b\}_{b=1,B}$
  - Y quatitative : $\hat{\Phi}_B(.) = \arg\max_j \text{card}\{b|\hat{\Phi}_{\mathcal{S}_b}\}$ (vote)
  - Y quantitative : $\hat{\Phi}_B(.) = \frac{1}{B}\sum_{b=1}^{B} \hat{\Phi}_{\mathcal{S}_b}(.)$ average
- $\Phi(.) = \mathbb{E}_{\mathcal{F}}(\hat{\Phi}_{\mathcal{S}})$ estimator with no biais

Averaging "independent" predictions in order to reduce the variance B using independant samples (B bootstrap replications). $\text{Var}(\bar{Z}) = \frac{\text{Var}(Z)}{\#\{B\}}$

Bagging. Illustration





Random observations 1^{st} sampling

Random observations 2^{nd} sampling

...

...

...

$\rightarrow$ A tree is built for each Bootstrap sample

Bagging. Benefits and drawbacks

(+) Each global tree is characterize by a low biais (well approximation)

(+) The variance is reduced by the aggregation of the different models.

Biais-Variance Trade-Off

Bagging. Benefits and drawbacks

(+) Each global tree is characterize by a low biais (well approximation)

(+) The variance is reduced by the aggregation of the different models.

Biais-Variance Trade-Off

(-) All the models need to be stored. How to choose $|B|$?

Bagging. Benefits and drawbacks

(+) Each global tree is characterize by a low biais (well approximation)

(+) The variance is reduced by the aggregation of the different models.

Biais-Variance Trade-Off

(-) All the models need to be stored. How to choose $|B|$?

(-) Computation time

Bagging. Benefits and drawbacks

(+) Each global tree is characterize by a low biais (well approximation)

(+) The variance is reduced by the aggregation of the different models.

Biais-Variance Trade-Off

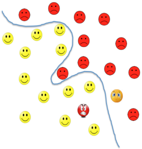
(-) All the models need to be stored. How to choose $|B|$?

(-) Computation time

(-) Black box models

Decision Trees

Random Forest



Random Forest. Breiman 2001.

- They are a modification of bagging of cart models

Random Forest. Breiman 2001.

- They are a modification of bagging of cart models
- Randomization of the set of variables at each node.
  At each node, the final variable is selected in a sub-set of variables chosen at random. Tuning parameters (classification : p/3 ; regression $\sqrt{p}$.

Random Forest. Breiman 2001.

- They are a modification of bagging of cart models
- Randomization of the set of variables at each node.
  At each node, the final variable is selected in a sub-set of variables chosen at random. Tuning parameters (classification : p/3 ; regression $\sqrt{p}$.
- Random choice of the variables

Random Forest. Breiman 2001.

- They are a modification of bagging of cart models
- Randomization of the set of variables at each node.
  At each node, the final variable is selected in a sub-set of variables chosen at random. Tuning parameters (classification : p/3 ; regression $\sqrt{p}$.
- Random choice of the variables

Illustration. SPAM classifier

Problem : to be able to automatically classify a regular email from a spam email

SPAM

From : Felix Damians, From .Abidjan Cote D.Ivoire

Hello Dear, Pardon me for not having the pleasure of knowing your mindset before making you this offer and it is utterly confidential and genuine by virtue of its nature. I want someone like you to help me out after i had pray ,then believes that you are a good person and that i can stay with you for the rest of my life , am 24 years old, My late father is a wealthy and successful business man before he died , My mum died when i was a baby, am the only child in my family. Honestly speaking , i am ready to give you 15percent of this total money for your assistance and with extra 5percent for your expenses on phone call, please reply me now if really serious to help me out so that i can tell you more about my intention and forward to you some of the legal papers after knowing you more better. Yours Felix Damians.

Illustration. SPAM classifier

A Text Mining is first performed to compute features

SPAM

From : Felix Damians, From .Abidjan Cote D.Ivoire

Hello Dear, Pardon me for not having the pleasure of knowing your mindset before making you this offer and it is utterly confidential and genuine by virtue of its nature. I want someone like you to **help** me out after i had pray ,then believes that you are a good person and that i can stay with you for the rest of my life , am 24 years old, My late father is a wealthy and successful business man before he died , My mum died when i was a baby, am the only child in my family. And he told me that he used my name to deposit (us dollars 12.5million) in the bank and he seriously advise me to transfer this total **money** to oversea account for my investment, where i will start my new life and finish my education , Because of this reason, i am soliciting your assistance for the claim and transfer to your bank account for the business. Honestly speaking , i am ready to give you 15percent of this total money for your assistance and with extra 5percent for your expenses on phone call, please reply me now if really serious to **help** me out so that i can tell you more about my intention and forward to you some of the legal papers after knowing you more better. All about the money are legal and i have all the legal documents and papers of the money money with me which the bank issued to my late father the day of the deposit, Because of the war in our country now the bank manager here has promised to me that they will use their branch corrospodant bank in Europ or Asia.to transfer the money into any account i provided to them that is why i contacted you to **help** me out in receiveing of the moneymoney into an account over there in your country before i join you over to has rest of mind. Thanks and remain bless with your family as i wait for your urgent reply soonest, Yours Felix Damians.

Illustration. SPAM classifier

SPAM data base

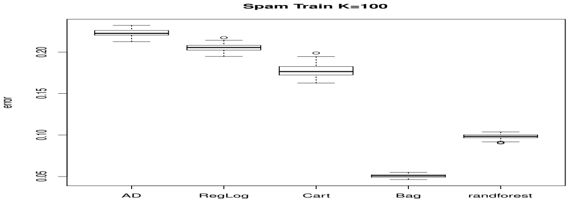
- 48 continuous real $[0, 100]$ attributes of type word-freq-WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.
- 6 continuous real $[0, 100]$ attributes of type char-freq-CHAR = percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail
- 1 continuous real $[1, \ldots]$ attribute of type capital-run-length-average = average length of uninterrupted sequences of capital letters
- 1 continuous integer $[1, \ldots]$ attribute of type capital-run-length-longest = length of longest uninterrupted sequence of capital letters
- 1 continuous integer $[1, \ldots]$ attribute of type capital-run-length-total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail
- 1 nominal 0,1 class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

Historical data base

- $p = 57$ features computed from the initial texts ($p = 56$)
- $n = 4601$ Emails with
- $Y \in \{0, 1\}$ a binary indicator
- $n >> p$

Classifier comparison

Train data



Test data