A Decision Tree for Interval-valued Data with Modal Dependent Variable

Djamal Seck[1], Lynne Billard[2], Edwin Diday[3] and Filipe Afonso[4]

[1] Departement de Mathematiques et Informatique, Université Cheikh Anta Diop de Dakar, Senegal *djamal.seck@ucad.edu.sn*
[2] Department of Statistics, University of Georgia, Athens GA 30605 USA *lynne@stat.uga.edu*
[3] CEREMADE, University of Paris Dauphine 75775 Paris Cedex 16 France *edwin.diday@ceremade.dauphine.fr*
[4] Syrokko, Aéropôle de Roissy, Bat. Aéronef, 5 rue de Copenhague, 95731 Roissy Charles de Gaulle Cedex France, *afonso@syrokko.com*

COMPSTAT - August 2010

The Future

Schweizer **(1985):** "Distributions are the numbers of the future"

Classical Data Value X:
- A single point in p-dimensional space
E.g., $X = 17$, $X = 2.1$, $X =$ blue

Types of Data

Classical Data Value X:
- A single point in p-dimensional space

E.g., $X = 17$, $X = 2.1$, $X =$ blue

Symbolic Data Value Y:
- Hypercube or Cartesian product of distributions in p-dimensional space

I.e. $Y =$ list, interval, modal in structure

Types of Data

Classical Data Value X:
- A single point in p-dimensional space

E.g., $X = 17$, $X = 2.1$, $X =$ blue

Symbolic Data Value Y:
- Hypercube or Cartesian product of distributions in p-dimensional space

I.e. $Y =$ list, interval, modal in structure

Modal data:
Histogram,
empirical distribution function,
probability distribution,
model, ...

Classical Data Value X:
- A single point in p-dimensional space

E.g., $X = 17$, $X = 2.1$, $X =$ blue

Symbolic Data Value Y:
- Hypercube or Cartesian product of distributions
in p-dimensional space

I.e. $Y =$ list, interval, modal in structure

Modal data:
Histogram,
empirical distribution function,
probability distribution,
model, ...

Weights:
Relative frequencies
capacities,
credibilities,
necessities,
possibilities, ...

How do symbolic data arise?

1. Aggregated data by classes or groups.
   - Research interest : classes or groups
2. Natural symbolic data.
   - Pulse rate : $64 \pm 2 = [62, 66]$.
   - Daily temperature : $[55, 67]$.
3. Published data : census data.
4. Symbolic data : range, list, and distribution, etc.

Literature Review

1. Clustering for classical data - CART, Breiman et al. (1984)

2. Clustering for symbolic data.
   - Agglomerative algorithm and dissimilarity measures for non-modal categorical and interval-valued data: Gowda and Diday (1991)
   - Pyramid clustering: Brito (1991, 1994), Brito and Diday (1990)
   - Spatial pyramids: Raoul Mohamed (2009)
   - Divisive monothetic algorithm for intervals: Chavent (1998,2000)
   - Divisive algorithms for histograms: Kim (2009)
   - Decision trees for non-modal dependent variables: Périnel (1996, 1999), Limam (2005), Winsberg et al. (2006),...
   - ......

Literature Review

1. Clustering for classical data - CART, Breiman et al. **(1984)**
2. Clustering for symbolic data.
   - Agglomerative algorithm and dissimilarity measures for non-modal categorical and interval-valued data: Gowda and Diday **(1991)**
   - Pyramid clustering: Brito **(1991, 1994)**, Brito and Diday **(1990)**
   - Spatial pyramids: Raoul Mohamed **(2009)**
   - Divisive monothetic algorithm for intervals: Chavent **(1998,2000)**
   - Divisive algorithms for histograms: Kim **(2009)**
   - Decision trees for non-modal dependent variables: Périnel **(1996, 1999)**, Limam **(2005)**, Winsberg et al. **(2006)**,...
   - ......
   - Decision tree for interval data and modal dependent variable (STREE): Seck **(2010)** (a CART methodology for symbolic data)

We have observations $\Omega = \{\omega_1, \ldots, \omega_n\}$, where ω_i has realization
$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip}), \; i = 1, \ldots, n.$

Modal multinominal (Modal categorical):
$Y_{ij} = \{m_{ijk}, p_{ijk}; \; k = 1, \ldots, s_i\}, \quad \sum_{k=1}^{s_i} p_{ijk} = 1,$
with $m_{ijk} \in \mathcal{O}_j = \{m_{j1}, \ldots, m_{js}\}, j = 1, \ldots, p \; i = 1, \ldots, n.$
(Take $s_i = s$, wlg.)

Multi-valued (non-modal):
$Y_{ij} = \{m_{ijk}, \; k = 1, \ldots, s_i\}$, i.e., $p_{ijk} = 1/s$ or 0,
with $m_{ijk} \in \mathcal{O}_j, j = 1, \ldots, p, \; i = 1, \ldots, n.$

Intervals:
$\mathbf{Y}_i = ([a_{i1}, \; b_{i1}], \ldots, [a_{ip}, \; b_{ip}]),$
with $a_{ij}, b_{ij} \in \mathcal{R}_j, \; j = 1, \ldots, p, \; i = 1, \ldots, n.$

Nominal (classical categorical):
Special case of modal multinominal with $s_i = 1, \quad p_1 = 1$; write
$Y_{ij} \equiv m_{ij1} = \delta_{ij}, \, \delta_{ij} \in \mathcal{O}_j.$

Classical continuous variable:
Special case of interval with $a_{ij} = [a_{ij}, \; a_{ij}]$ for $a_{ij} \in \mathcal{R}_j$.

STREE Algorithm

Have at r^{th} stage the partition $P_r = (C_1, \ldots, C_r)$
Discrimination criterion: $D(N)$ - explains partition of node N as in CART analysis
Homogeneity criterion: $H(N)$ - inertia associated with explanatory variables as in pure hierarchy tree analysis
We take the mixture, for $\alpha > 0,\ \beta > 0$,

$$I = \alpha D(N) + \beta H(N) \quad \text{with} \quad \alpha + \beta = 1.$$

The $D(N)$ is taken as the Gini measure (as in CART)

$$D(N) = \sum_{i \neq f} p_i p_f = 1 - \sum_{i=1,\ldots,r} p_i^2$$

with $p_i = n_i/n$, $n_i = \text{card}(N \bigcap C_i)$, $n = card(N)$;
the $H(N)$ is

$$H(N) = \sum_{\omega_{i_1} \in \Omega} \sum_{\omega_{i_2} \in \Omega} \frac{p_{i_1} p_{i_2}}{2\mu} d^2(\omega_{i_1}, \omega_{i_2})$$

where $d(\omega_{i_1}, \omega_{i_2})$ is a distance measure between ω_{i_1} and ω_{i_2}, p_i is the weight associated with ω_i and $\mu = \sum_{i=1}^{N} p_i$.

STREE Algorithm

Have at r^{th} stage the partition $P_r = (C_1, \ldots, C_r)$
Discrimination criterion: $D(N)$ - explains partition of node N as in CART analysis
Homogeneity criterion: $H(N)$ - inertia associated with explanatory variables as in pure hierarchy tree analysis
We take the mixture, for $\alpha > 0,\ \beta > 0$,

$$I = \alpha D(N) + \beta H(N) \quad \text{with} \quad \alpha + \beta = 1.$$

The $D(N)$ is taken as the Gini measure (as in CART)

$$D(N) = \sum_{i \neq f} p_i p_f = 1 - \sum_{i=1,\ldots,r} p_i^2$$

with $p_i = n_i/n$, $n_i = \text{card}(N \bigcap C_i)$, $n = card(N)$;
the $H(N)$ is

$$H(N) = \sum_{\omega_{i_1} \in \Omega} \sum_{\omega_{i_2} \in \Omega} \frac{p_{i_1} p_{i_2}}{2\mu} d^2(\omega_{i_1}, \omega_{i_2})$$

where $d(\omega_{i_1}, \omega_{i_2})$ is a distance measure between ω_{i_1} and ω_{i_2}, p_i is the weight associated with ω_i and $\mu = \sum_{i=1}^{N} p_i$.

Select the partition $C = \{C_1,\ C_2\}$ for which the reduction in I is greatest; i.e.,
maximize $\Delta I = I(C) - I(C_1,\ C_2)$.

The homogeneity criterion $H(N)$

$$H(N) = \sum_{\omega_{i_1} \in \Omega} \sum_{\omega_{i_2} \in \Omega} \frac{p_{i_1} p_{i_2}}{2\mu} d^2(\omega_{i_1}, \omega_{i_2})$$

where $d(\omega_{i_1}, \omega_{i_2})$ is a distance measure between ω_{i_1} and ω_{i_2}, p_i is the weight associated with ω_i and $\mu = \sum_{i=1}^{N} p_i$.

Decision Tree - Distance Measures

The homogeneity criterion $H(N)$

$$H(N) = \sum_{\omega_{i_1} \in \Omega} \sum_{\omega_{i_2} \in \Omega} \frac{p_{i_1} p_{i_2}}{2\mu} d^2(\omega_{i_1}, \omega_{i_2})$$

where $d(\omega_{i_1}, \omega_{i_2})$ is a distance measure between ω_{i_1} and ω_{i_2}, p_i is the weight associated with ω_i and $\mu = \sum_{i=1}^{N} p_i$. The STREE algorithm uses

Modal categorical variables - L_1 distance:

$$d_j(\omega_{i_1},\ \omega_{i_2}) = \sum_{k \in \mathcal{O}} |p_{i_1 jk} - p_{i_2 jk}|;$$

or, L_2 distance:

$$d_j(\omega_{i_1},\ \omega_{i_2}) = \sum_{k \in \mathcal{O}} (p_{i_1 jk} - p_{i_2 jk})^2$$

Interval variables - Hausdorff distance:

$$d_j(\omega_{i_1},\ \omega_{i_2}) = \max(|a_{i_1 j} - a_{i_2 j}|,\ |b_{i_1 j} - b_{i_2 j}|)$$

Classical categorical variables - $(0,\ 1)$ distance:

$$d_j(\omega_{i_1},\ \omega_{i_2}) = \begin{cases} 0, & \text{if } m_{i_1 j} = m_{i_2 j} \\ 1, & \text{if } m_{i_1 j} \neq m_{i_2 j} \end{cases}$$

Classical continuous variables - Euclidean distance:

$$d_j(\omega_{i_1},\ \omega_{i_2}) = (a_{i_1 j} - a_{i_2 j})^2$$

Decision Tree - Distance Measures

The homogeneity criterion $H(N)$

$$H(N) = \sum_{\omega_{i_1} \in \Omega} \sum_{\omega_{i_2} \in \Omega} \frac{p_{i_1} p_{i_2}}{2\mu} d^2(\omega_{i_1}, \omega_{i_2})$$

where $d(\omega_{i_1}, \omega_{i_2})$ is a distance measure between ω_{i_1} and ω_{i_2}, p_i is the weight associated with ω_i and $\mu = \sum_{i=1}^{N} p_i$. The STREE algorithm uses

Modal categorical variables - L_1 distance:

$d_j(\omega_{i_1},\ \omega_{i_2}) = \sum_{k \in \mathcal{O}} |p_{i_1 jk} - p_{i_2 jk}|$;

or, L_2 distance:

$d_j(\omega_{i_1},\ \omega_{i_2}) = \sum_{k \in \mathcal{O}} (p_{i_1 jk} - p_{i_2 jk})^2$

Interval variables - Hausdorff distance:

$d_j(\omega_{i_1},\ \omega_{i_2}) = \max(|a_{i_1 j} - a_{i_2 j}|,\ |b_{i_1 j} - b_{i_2 j}|)$

Classical categorical variables - $(0,\ 1)$ distance:

$$d_j(\omega_{i_1},\ \omega_{i_2}) = \begin{cases} 0, & \text{if } m_{i_1 j} = m_{i_2 j} \\ 1, & \text{if } m_{i_1 j} \neq m_{i_2 j} \end{cases}$$

Classical continuous variables - Euclidean distance:

$d_j(\omega_{i_1},\ \omega_{i_2}) = (a_{i_1 j} - a_{i_2 j})^2$

Hence,

$d(\omega_{i_1},\ \omega_{i_2}) = \sum_{j=1}^{p} d_j(\omega_{i_1},\ \omega_{i_2})$.

Decision Tree - Cut Points

Cut points: Take Modal categorical case – Recall

$Y_{ij} = \{m_{ijk}, p_{ijk};\ k = 1, \ldots, s_i\}, \quad \sum_{k=1}^{s_i} p_{ijk} = 1, \qquad$ (Take $s_i = s$, wlg.)

with $m_{ijk} \in \mathcal{O}_j = \{m_{j1}, \ldots, m_{js}\}$, $j = 1, \ldots, p$ $i = 1, \ldots, n$.

First: For each k in turn, order p_{ijk} from smallest to largest.

There are $L_k \leq n$ distinct values of $p_{jkr}, r = 1, \ldots, L_k$.

Then, cut point for this modality (m_{jk}) is the probability

$$c_{jkr} = (p_{jkr} + p_{jk,r+1})/2, \quad r = 1, \ldots, L_k - 1, \quad k = 1, \ldots, s.$$

There are $\sum_{k=1}^{s}(L_k - 1)$possible partitions for each j.

Decision Tree - Cut Points

Cut points: Take Modal categorical case – Recall

$Y_{ij} = \{m_{ijk}, p_{ijk};\ k = 1, \ldots, s_i\}$, $\sum_{k=1}^{s_i} p_{ijk} = 1$, (Take $s_i = s$, wlg.)

with $m_{ijk} \in \mathcal{O}_j = \{m_{j1}, \ldots, m_{js}\}$, $j = 1, \ldots, p$ $i = 1, \ldots, n$.

First: For each k in turn, order p_{ijk} from smallest to largest.

There are $L_k \leq n$ distinct values of $p_{jkr}, r = 1, \ldots, L_k$.

Then, cut point for this modality (m_{jk}) is the probability

$$c_{jkr} = (p_{jkr} + p_{jk,r+1})/2, \quad r = 1, \ldots, L_k - 1, \quad k = 1, \ldots, s.$$

There are $\sum_{k=1}^{s}(L_k - 1)$possible partitions for each j.

Similarly, take pairs $(m_{ijk_1},\ m_{ijk_2})$ with probability $(p_{ijk_1} + p_{ijk_2}) = p_{ijk_1k_2}$.

Repeat previous process using now these probabilities $p_{ijk_1k_2}$, for the $L_{k_1k_2}$ distinct probabilities among the $s(s+1)/2$ possible pairs.

Decision Tree - Cut Points

Cut points: Take Modal categorical case – Recall
$Y_{ij} = \{m_{ijk}, p_{ijk};\ k = 1, \ldots, s_i\}, \quad \sum_{k=1}^{s_i} p_{ijk} = 1, \qquad$ (Take $s_i = s$, wlg.)
with $m_{ijk} \in \mathcal{O}_j = \{m_{j1}, \ldots, m_{js}\}$, $j = 1, \ldots, p$ $i = 1, \ldots, n$.

First: For each k in turn, order p_{ijk} from smallest to largest.

There are $L_k \leq n$ distinct values of $p_{jkr}, r = 1, \ldots, L_k$.

Then, cut point for this modality (m_{jk}) is the probability
$$c_{jkr} = (p_{jkr} + p_{jk,r+1})/2, \quad r = 1, \ldots, L_k - 1, \quad k = 1, \ldots, s.$$

There are $\sum_{k=1}^{s}(L_k - 1)$possible partitions for each j.

Similarly, take pairs $(m_{ijk_1},\ m_{ijk_2})$ with probability $(p_{ijk_1} + p_{ijk_2}) = p_{ijk_1k_2}$.

Repeat previous process using now these probabilities $p_{ijk_1k_2}$, for the $L_{k_1k_2}$ distinct probabilities among the $s(s+1)/2$ possible pairs.

Likewise, take sets of three, four, $\ldots$, $(s-1)$ of the s values of $m_{ijk}, k = 1, \ldots, s$ in $\mathcal{O}_j$.

The total number of possible cuts points is L. It can be shown that
$$maxL = (n-1)\sum_{q=1}^{s-1}\binom{s}{q} = (n-1)2(2^{s-1} - 1).$$

Cut points: Take Intervls case – Recall

$\mathbf{Y}_i = ([a_{i1},\ b_{i1}], \ldots, [a_{ip},\ b_{ip}])$,
with $a_{ij}, b_{ij} \in \mathcal{R}_j,\ j = 1, \ldots, p,\ i = 1, \ldots, n$.

First: For each j, let $\mathcal{D}_j = \{d_{jr}, r = 1, \ldots, L\}$ be the set of n a_{ij} and n b_{ij} values, ordered from smallest to largest. Thus, e.g.,

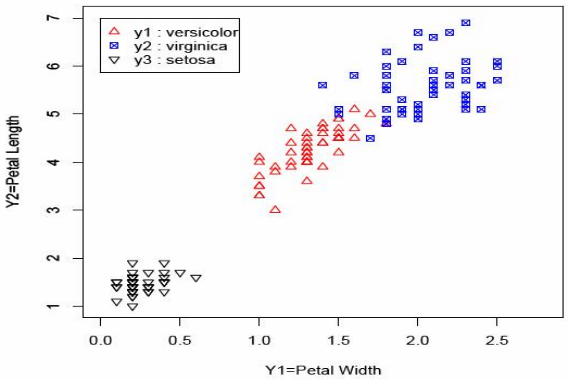
$$d_{j1} = min_{i \in \Omega}(a_{ij}), \quad d_{jL} = min_{i \in \Omega}(b_{ij}), \quad j = 1, \ldots, p.$$

There are $L \leq 2n$ distinct values of $d_{jr}, r = 1, \ldots, L$.

The cut points are
$c_{jr} = (d_{jr} + d_{j,r+1})/2, \quad r = 1, \ldots, L$

Fisher (1936): IRIS dataset -
150 observations, 50 for each species *versicolor*, *virginica*, *setosa*

Y_1 = Sepal Length, Y_2 = Sepal Width, Y_3 = Petal Length, Y_4 = Petal Width



Fisher **(1936):** IRIS dataset -
150 observations, 50 for each species *setosa*, *versicolor*, *virginica*

Y_1 = Sepal Length, Y_2 = Sepal Width, Y_3 = Petal Length, Y_4 = Petal Width

Clustered into 30 sets of observations, by k-means clustering method

Fisher **(1936):** IRIS dataset -
150 observations, 50 for each species *setosa*, *versicolor*, *virginica*

Y_1 = Sepal Length, Y_2 = Sepal Width, Y_3 = Petal Length, Y_4 = Petal Width

Clustered into 30 sets of observations, by k-means clustering method

Table 1: Fisher's Iris Data as Intervals

Concept	Species[a]	Sepal length	Sepal width	Petal length	Petal width
ω_1	{1, 1.0}	[4.8, 5.4]	[3.3, 3.8]	[1.5, 1.9]	[0.2, 0.6]
...					
ω_4	{1,1.0}	[4.5, 4.5]	[2.3, 2.3]	[1.3, 1.3]	[0.3, 0.3]
...					
ω_{12}	{2,.9; 3,.1}	[4.9, 5.7]	[2.5, 3.0]	[4.1, 4.5]	[1.2, 1.7]
...					
ω_{30}	{2,1.0}	[6.2, 6.3]	[2.2, 2.3]	[4.4, 4.5]	[1.3, 1.5]

[a]Species identified by 1,2,3 for *setosa*, *versicolor*, *virginica*, respectively.

Fisher **(1936):** IRIS dataset -
150 observations, 50 for each species *setosa*, *versicolor*, *virginica*

Y_1 = Sepal Length, Y_2 = Sepal Width, Y_3 = Petal Length, Y_4 = Petal Width

Clustered into 30 sets of observations, by k-means clustering method

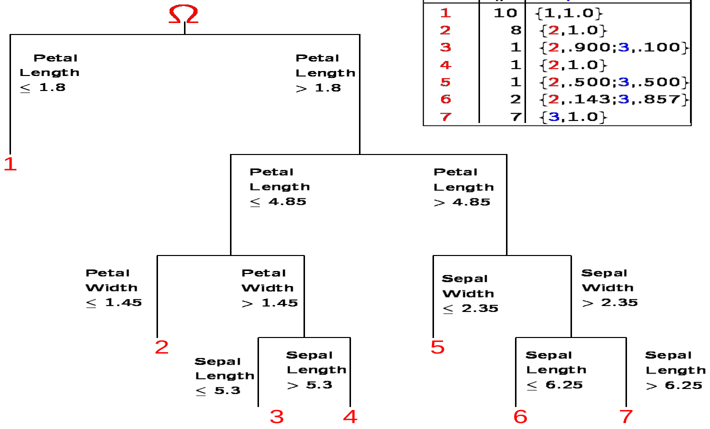
Table 1: Fisher's Iris Data as Intervals

Concept	Species[a]	Sepal	length	Sepal	width	Petal	length	Petal	width
ω_1	{1, 1.0}	[4.8,	5.4]	[3.3,	3.8]	[1.5,	1.9]	[0.2,	0.6]
...									
ω_4	{1,1.0}	[4.5,	4.5]	[2.3,	2.3]	[1.3,	1.3]	[0.3,	0.3]
...									
ω_{12}	{2,.9; 3,.1}	[4.9,	5.7]	[2.5,	3.0]	[4.1,	4.5]	[1.2,	1.7]
...									
ω_{30}	{2,1.0}	[6.2,	6.3]	[2.2,	2.3]	[4.4,	4.5]	[1.3,	1.5]

[a]Species identified by 1,2,3 for *setosa*, *versicolor*, *virginica*, respectively.

Species – modal categorical data: $Y_u = \{y_k, p_k;\ k = 1, \ldots, s_u\},\ u = 1, \ldots, m$
$Y_1,\ Y_2,\ Y_3, Y_4$ – interval data: $Y_{uj} = [a_{uj}, b_{uj}],\ j = 1, \ldots, p,\ u = 1, \ldots, m$

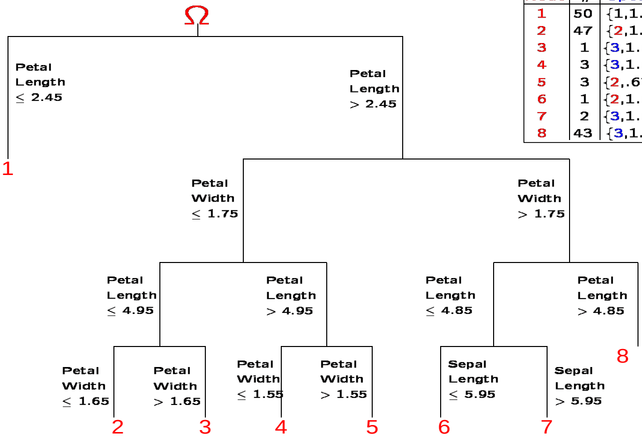
Pure Decision Tree on 30 IRIS intervals: $\alpha = 0$



Node	#	Species
1	10	{1,1.0}
2	8	{2,1.0}
3	1	{2,.900;3,.100}
4	1	{2,1.0}
5	1	{2,.500;3,.500}
6	2	{2,.143;3,.857}
7	7	{3,1.0}

species: *setosa*, *versicolor*, *virginica*

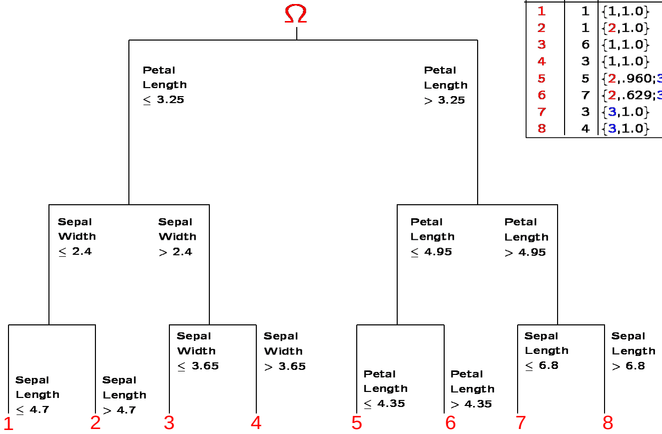
Pure CART Tree on original 150 IRIS observations: $\alpha = 0$



Node	#	Species
1	50	{1,1.0}
2	47	{2,1.0}
3	1	{3,1.0}
4	3	{3,1.0}
5	3	{2,.67;3,.33}
6	1	{2,1.0}
7	2	{3,1.0}
8	43	{3,1.0}

species: *setosa*, *versicolor*, *virginica*

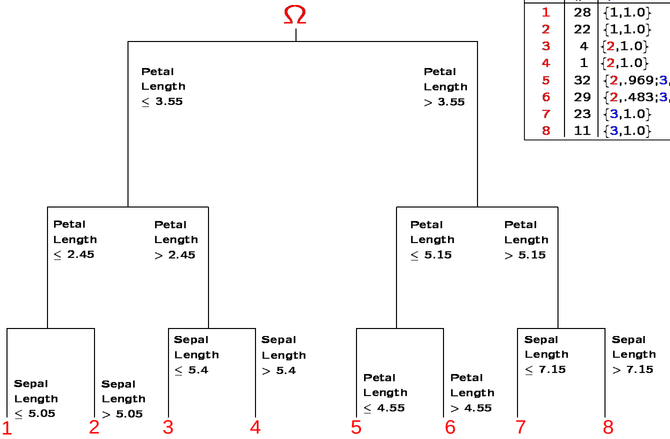
Pure DIV Tree on 30 IRIS intervals: $\alpha = 1$



Node	#	Species
1	1	{1,1.0}
2	1	{2,1.0}
3	6	{1,1.0}
4	3	{1,1.0}
5	5	{2,.960;3,.040}
6	7	{2,.629;3,.371}
7	3	{3,1.0}
8	4	{3,1.0}

species: *setosa*, *versicolor*, *virginica*

Pure DIV Tree on original 150 IRIS observations: $\alpha = 1$

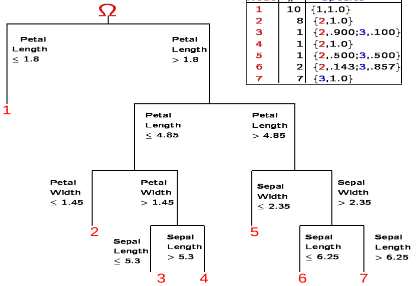


Node	#	Species
1	28	{1,1.0}
2	22	{1,1.0}
3	4	{2,1.0}
4	1	{2,1.0}
5	32	{2,.969;3,.031}
6	29	{2,.483;3,.517}
7	23	{3,1.0}
8	11	{3,1.0}

species: *setosa*, *versicolor*, *virginica*

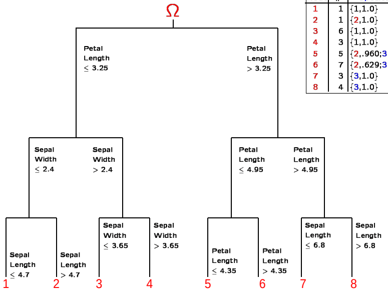
Decision Tree

Trees on 30 IRIS intervals:



Node	#	Species
1	10	{1,1.0}
2	8	{2,1.0}
3	1	{2,.900;3,.100}
4	1	{2,1.0}
5	1	{2,.500;3,.500}
6	2	{2,.143;3,.857}
7	7	{3,1.0}

Pure Decision Tree: $\alpha = 0$

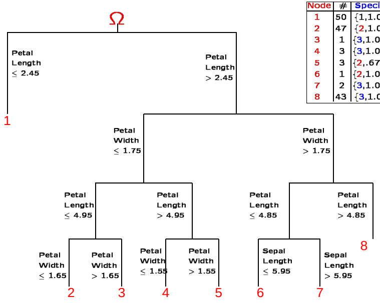


Node	#	Species
1	1	{1,1.0}
2	1	{2,1.0}
3	6	{1,1.0}
4	3	{1,1.0}
5	5	{2,.960;3,.040}
6	7	{2,.629;3,.371}
7	3	{3,1.0}
8	4	{3,1.0}

Pure DIV Tree: $\alpha = 1$

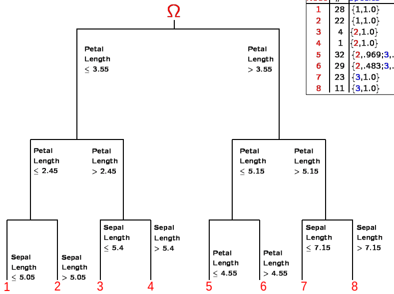
Decision Tree

Trees on original 150 IRIS observations:



Node	#	Species
1	50	{1,1.0}
2	47	{2,1.0}
3	1	{3,1.0}
4	3	{3,1.0}
5	3	{2,.67;3,.33}
6	1	{2,1.0}
7	2	{3,1.0}
8	43	{3,1.0}

Pure Decision Tree: $\alpha = 0$



Node	#	Species
1	28	{1,1.0}
2	22	{1,1.0}
3	4	{2,1.0}
4	1	{2,1.0}
5	32	{2,.969;3,.031}
6	29	{2,.483;3,.517}
7	23	{3,1.0}
8	11	{3,1.0}

Pure DIV Tree: $\alpha = 1$

Comparison of STREE and CART Algorithms
Randomly divided 150 observations into

Training subset (size n_1), and

Test subset (size n_2), with

$n_1 + n_2 = n = 150$

Several sets of (n_1, n_2)

1. For CART:

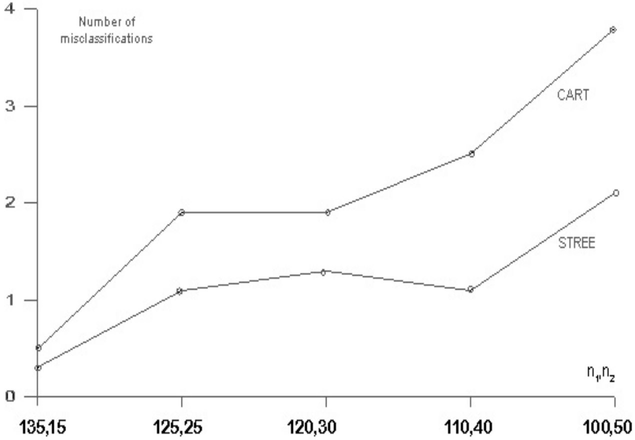
   Run CART algorithm on the n_1 observations in Training subset

1. For STREE:

   First find 30 clusters from the n_1 observations in Training subset
   Run decision tree analysis ($\alpha = 0$) on the 30 clusters

2. Test tree on the n_1 observations in Test subset
3. Obtain number of misclassifications
4. Repeat 10 times for each (n_1, n_2)
5. Calculate average number of misclassifications for each (n_1, n_2) and for each Algorithm

Comparison of STREE and CART Average Misclassifications for Test subsets (n_2)



n_1 = Size of Training subset, n_2 = Size of Test subset

~ Merci Bien ~

$\sim$ Merci Bien $\sim$

$\sim$ Thank You $\sim$

Partial support from National Science Foundation gratefully acknowledged