View Package Documentation

対応 NYSOL バージョン: Ver. 1.2, 2.0

revise history:
October 6, 2014 : first release

2014 年 10 月 6 日

目次

第 1 章

はじめに

1.1 概要

本「View（眺）」パッケージは、複数のデータ視覚化用コマンドから構成されている。視覚化とは、データをグラフやチャートとして描画することをいい、データを概観したり、資料に掲載したりする際には重要な役割を果たす。

本パッケージには、円グラフを描画する `mpie.rb` コマンド、Sankey ダイアグラム（流量グラフ）を描画する `msankey.rb` コマンド、グラフデータを汎用的な形式に変換する `mgv.rb` コマンドが含まれている。

1.2 インストール

本パッケージは全て nysol パッケージに含まれている。 nysol パッケージをインストールすれば必要なソフトウェアは全てインストールされる。詳しくは nysol パッケージのインストールの説明 (http://www.nysol.jp/install) を参照のこと。

第 2 章

視覚化コマンド

2.1　msankey.rb sankey ダイアグラムの描画

Sankey ダイアグラムとは，閉路のない有向グラフ (DAG:Directed Acyclic Graph) を視覚化する手法の一つで，枝の重みとして定義される流量が接点間でどのような割合で流れていくかを直感的に理解することができ，送電ネットワークの視覚化などに利用される．

内部では、D3 ライブラリ (Data-Driven Documents) で作成された視覚化アプリケーション sankey diagram (http://bost.ocks.org/mike/sankey/) を利用している。

入力データとしては，表 2.1 に示されるような枝を節点ペアとその値を一行で示した CSV データである。

出力は sankey ダイアグラムを組み込んだ単体の html ファイルで、インターネットへの接続がなくてもブラウザがあれば描画できる (2.1)。グラフの向きは左から右の方向で、コマンドパラメータ `f=`で指定した 1 番目の項目が左、2 番目の項目が右に対応する。色のついたバーが節点に対応し、接点間を結ぶ帯が枝に対応する。節点の出力位置の決定には interval relaxztion 法が用いられている。詳細はオリジナルの URL(上述) を参照のこと。経験的には 5 節点 × 水平位置 10 箇所 = 50 節点ともなると、描画に非常に時間を要する。

なお、本コマンドを利用するためには、nysol/mcmd ライブラリの他に json ライブラリが必要となる。

表 2.1　入力データ (閉路のない有向グラフ)

node1	node2	val
a	b	1
a	c	2
a	d	1
b	c	3
b	d	3
c	f	1
c	e	4
d	e	1
e	f	3

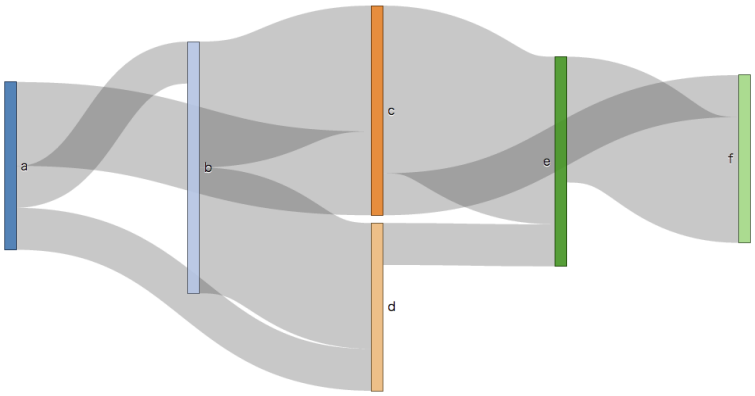


図 2.1　sankey ダイアグラム

2.1.1 書式

```
msankey [i=] f= v= [o=] [t=] [T=] [--help]
```

```
 i=       : 枝データファイル
 f=       : 枝データ上の 2 つの節点項目名
 v=       : 枝の重み項目名
 o=       : 出力ファイル (HTML ファイル)
 t=       : タイトル文字列
 -T       : ワークディレクトリ (default:/tmp)
 --help   : ヘルプの表示
```

2.1.2 利用例

例 1: **基本例**

前節の解説で用いてる例。

```
$ more dat1.csv
node1,node2,val
a,b,1
a,c,2
a,d,1
b,c,3
b,d,3
c,f,1
c,e,4
d,e,1
e,f,3
$ msankey.rb i=dat1.csv f=node1,node2 v=val o=output.html
$ head output.html
<!DOCTYPE html>
<html class="ocks-org do-not-copy">
<meta charset="utf-8">
<!--
<title>Sankey Diagram</title>
-->
<title></title>
<style>
```

2.2　mpie.rb 円グラフの描画

円グラフを描画するコマンドである。x 軸・y 軸に展開する属性項目を指定することで、1 次元もしくは 2 次元の円グラフ行列を描画することができる。グラフは単独の HTML ファイルとして出力されるので、一般的なブラウザで表示が可能である。

入力データには、表 2.2 のような CSV を用いる。円グラフを構成する扇となる項目を構成要素項目といい、`f=`パラメータで指定する。x 軸・y 軸に展開する属性項目は `k=`パラメータで指定する。`k=`パラメータで 1 項目を指定すると 1 次元の（x 軸に展開された）円グラフ行列が、2 項目を指定すると 2 次元の（x 軸・y 軸に展開された）円グラフ行列が描画される。`k=`パラメータを省略した場合は、1 個の円グラフが描画される。

なお円グラフの描画には、内部的に JavaScript ライブラリ D3.js (Data-Driven Documents) を使用している。D3.js の詳細は公式ページ（http://d3js.org/）を参照のこと。

また本コマンドを利用するためには、nysol/mcmd ライブラリが必要となる。

表 2.2　都道府県と年代別の個体数

Pref	Age	Population
奈良	10	310504
奈良	20	552339
奈良	30	259034
奈良	40	450818
奈良	50	1231572
奈良	60	1215966
奈良	70	641667
北海道	10	310504
北海道	20	252339
北海道	30	859034
北海道	40	150818
北海道	50	9231572
北海道	60	4215966
北海道	70	341667

2.2.1 書式

```
mpie.rb [i=] f= v= [o=] [k=] [title=] [pr=] [cc=] [--help]
```

i=	入力データファイル名（CSV 形式）
f=	構成要素項目名を指定する。 データに null が含まれる場合は無視する。
v=	構成比項目（円グラフの円弧の長さを決定する項目）を指定する。 データに null が含まれる場合は 0 として扱う。 先頭の 0 は無視する。数字以外の場合はエラーとなる。
o=	出力ファイル名（HTML ファイル）
k=	x 軸・y 軸に展開する属性項目名を 2 つ以内で指定する。 省略した場合は円グラフを 1 つ作成する。 項目を 1 つ指定した場合は 1 次元の円グラフ行列を、 項目を 2 つ指定した場合は 2 次元の円グラフ行列を作成する。
title=	グラフのタイトル文字列を指定する。
pr=	円グラフの半径を指定する（デフォルトは 160）。
cc=	1 行に表示する円グラフの最大数を指定する（デフォルトは 5）。 1 次元の円グラフ行列のときのみ指定できる。
--help	ヘルプの表示

なお mpie.rb コマンドには、f=パラメータや k=パラメータで指定した項目を自動的に並べ替える機能はない。グラフに表示したい順に、あらかじめ並べ替えておく必要がある。

2.2.2 利用例

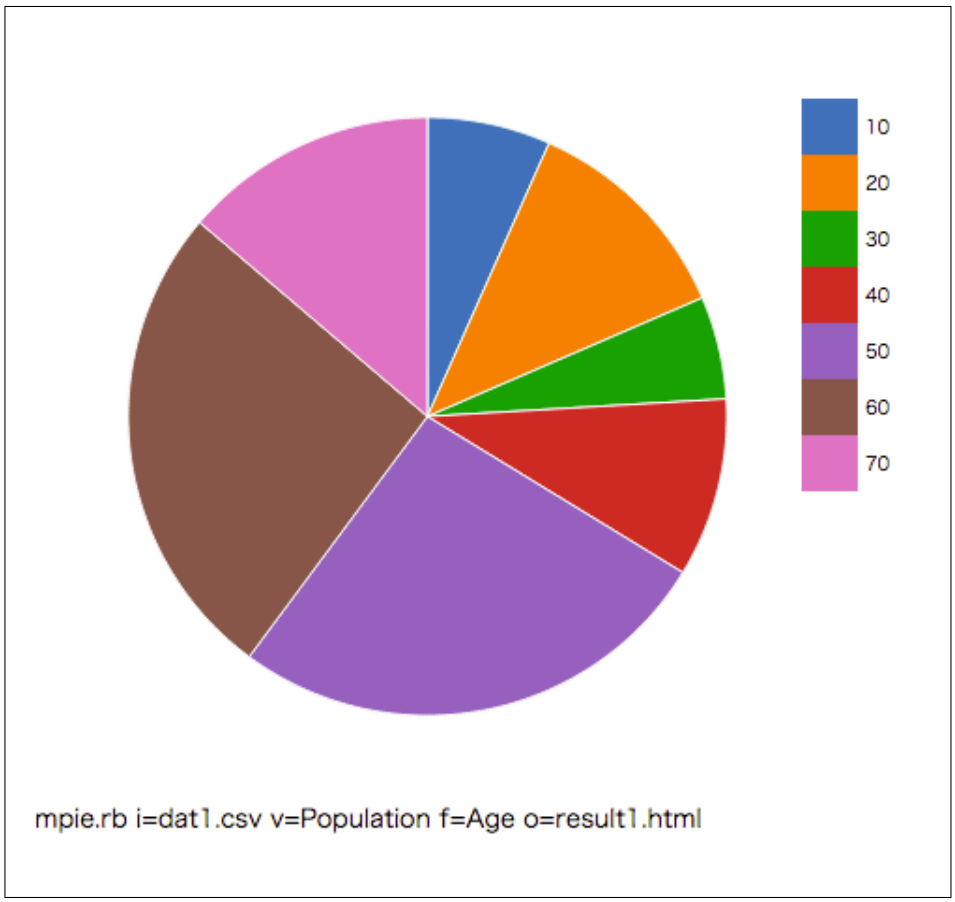
例 1: 円グラフを 1 つ描画する

dat1.csv ファイルの Age を構成要素項目に、Population を構成比項目として円グラフを 1 つ描画する。

```
$ more dat1.csv
Age,Population
10,310504
20,552339
30,259034.5555
40,0450818
50,1231572
60,1215966
70,641667
$ mpie.rb i=dat1.csv v=Population f=Age o=result1.html
#END# mpie.rb i=dat1.csv v=Population f=Age o=result1.html;
```

以下の円グラフが描画される。

ブラウザで表示した円グラフにマウスカーソルを置くと、構成要素項目とその構成比がポップアップで表示される。グラフはマウスによるドラッグ操作で移動することができ、またマウスのスクロール操作によって拡大縮小もできる。

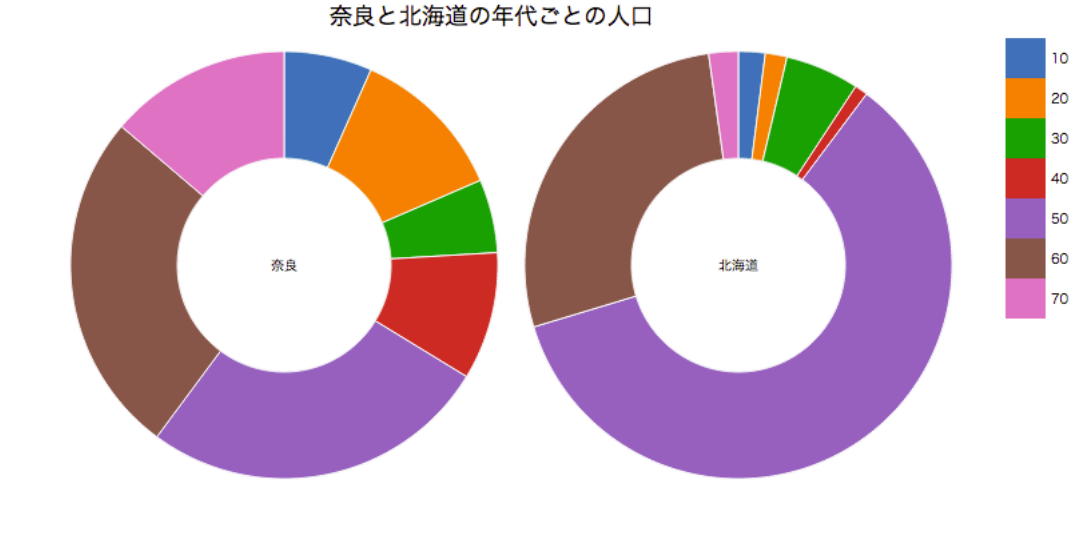


例 2: 1 **次元の円グラフ行列を描画する**

dat2.csv ファイルの Age を構成要素項目に、Population を構成比項目として円グラフを描画する。k=パラメータに Pref 項目を指定しているので、Pref 項目の値を x 軸（横方向）に展開した 1 次元の円グラフ行列が描画される。title=パラメータでグラフのタイトルも指定している。

```
$ more dat2.csv
Pref,Age,Population
奈良,10,310504
奈良,20,552339
奈良,30,259034
奈良,40,450818
奈良,50,1231572
奈良,60,1215966
奈良,70,641667
北海道,10,310504
北海道,20,252339
北海道,30,859034
北海道,40,150818
北海道,50,9231572
北海道,60,4215966
北海道,70,341667
$ mpie.rb i=dat2.csv k=Pref v=Population f=Age o=result2.html
 title=奈良と北海道の年代ごとの人口
#END# mpie.rb i=dat2.csv k=Pref v=Population f=Age o=result2.html
title=奈良と北海道の年代ごとの人口;
```

以下の円グラフ行列が描画される。

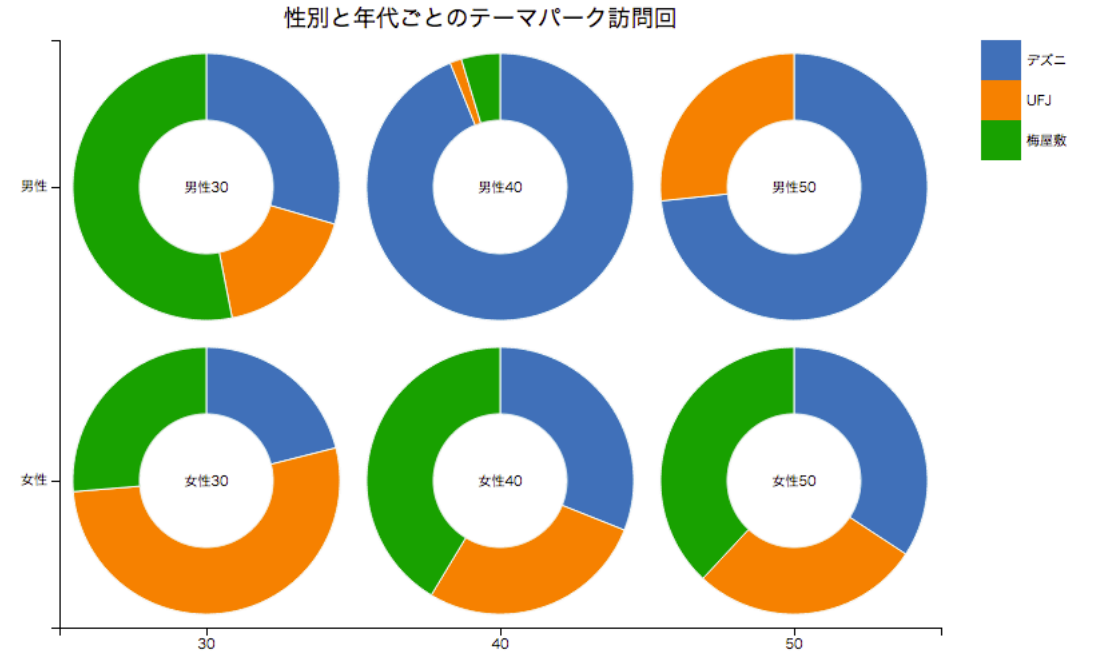


例 3: 2 **次元の円グラフ行列を描画する**

dat3.csv ファイルの**テーマパーク名**を構成要素項目、Number を構成比項目とし、pr=パラメータに半径 100 を指定して円グラフを描画する。k=パラメータに Gender と Age 項目を指定しているので、Gender 項目の値を x 軸（横方向）に、Age 項目の値を y 軸（縦方向）に展開した 2 次元の円グラフ行列が描画される。

```
$ more dat3.csv
Gender,Age,テーマパーク名,Number
男性,30,デズニ,100
男性,30,UFJ,59
男性,30,梅屋敷,180
男性,40,デズニ,200
男性,40,UFJ,3
男性,40,梅屋敷,10
男性,50,デズニ,110
男性,50,UFJ,40
女性,30,梅屋敷,100
女性,30,デズニ,80
女性,30,UFJ,200
女性,40,デズニ,90
女性,40,UFJ,80
女性,40,梅屋敷,120
女性,50,デズニ,99
女性,50,UFJ,80
女性,50,梅屋敷,110
$ mpie.rb i=dat3.csv k=Gender,Age v=Number f=テーマパーク名 o=result3.html
 title=性別と年代ごとのテーマパーク訪問回 pr=100
#END# mpie.rb i=dat3.csv k=Gender,Age v=Number f=テーマパーク名 o=result3.html
title=性別と年代ごとのテーマパーク訪問回 pr=100;
```

以下の円グラフ行列が描画される。




```
mpie.rb i=dat3.csv k=Gender,Age v=Number f=テーマパーク名 o=result3.html title=性別と年代ごとのテーマパーク訪問回 pr=100
```

2.3 mbar.rb 棒グラフの描画

棒グラフを描画するコマンドである。x 軸・y 軸に展開する属性項目を指定することで、1 次元もしくは 2 次元の棒グラフ行列を描画することができる。グラフは単独の HTML ファイルとして出力されるので、一般的なブラウザで表示が可能である。

入力データには、表 2.3 のような CSV を用いる。棒グラフを構成する項目を構成要素項目といい、f=パラメータで指定する。x 軸・y 軸に展開する属性項目は k=パラメータで指定する。k=パラメータで 1 項目を指定すると 1 次元の（x 軸に展開された）棒グラフ行列が、2 項目を指定すると 2 次元の（x 軸・y 軸に展開された）棒グラフ行列が描画される。k=パラメータを省略した場合は、1 個の棒グラフが描画される。

なお棒グラフの描画には、内部的に JavaScript ライブラリ D3.js (Data-Driven Documents) を使用している。D3.js の詳細は公式ページ（http://d3js.org/）を参照のこと。

また本コマンドを利用するためには、nysol/mcmd ライブラリが必要となる。

表 2.3 都道府県と年代別の個体数

Pref	Age	Population
奈良	10	310504
奈良	20	552339
奈良	30	259034
奈良	40	450818
奈良	50	1231572
奈良	60	1215966
奈良	70	641667
北海道	10	310504
北海道	20	252339
北海道	30	859034
北海道	40	150818
北海道	50	9231572
北海道	60	4215966
北海道	70	341667

2.3.1 書式

```
mbar.rb [i=] f= v= [o=] [k=] [title=] [width=] [height=] [cc=] [--help]
```

i=	入力データファイル名（CSV 形式）
f=	構成要素項目名を指定する。 データに null が含まれる場合は無視する。
v=	構成量項目（棒グラフの高さを決定する項目）を指定する。 データに null が含まれる場合は 0 として扱う。 マイナス、小数点に対応している。 先頭の 0 は無視する。数字以外の場合はエラーとなる。
o=	出力ファイル名（HTML ファイル）
k=	x 軸・y 軸に展開する属性項目名を 2 つ以内で指定する。 省略した場合は棒グラフを 1 つ作成する。 項目を 1 つ指定した場合は 1 次元の棒グラフ行列を、 項目を 2 つ指定した場合は 2 次元の棒グラフ行列を作成する。
title=	グラフのタイトル文字列を指定する。
width=	棒グラフ用描画枠の横幅を指定する（デフォルトは 250、1 つの棒グラフは 600）。
height=	棒グラフ用描画枠の縦幅を指定する（デフォルトは 250、1 つの棒グラフは 400）。
cc=	1 行に表示する棒グラフの最大数を指定する（デフォルトは 5）。 1 次元の棒グラフ行列のときのみ指定できる。
--help	ヘルプの表示

なお mbar.rb コマンドには、f=パラメータや k=パラメータで指定した項目を自動的に並べ替える機能はない。グラフに表示したい順に、あらかじめ並べ替えておく必要がある。

2.3.2 利用例

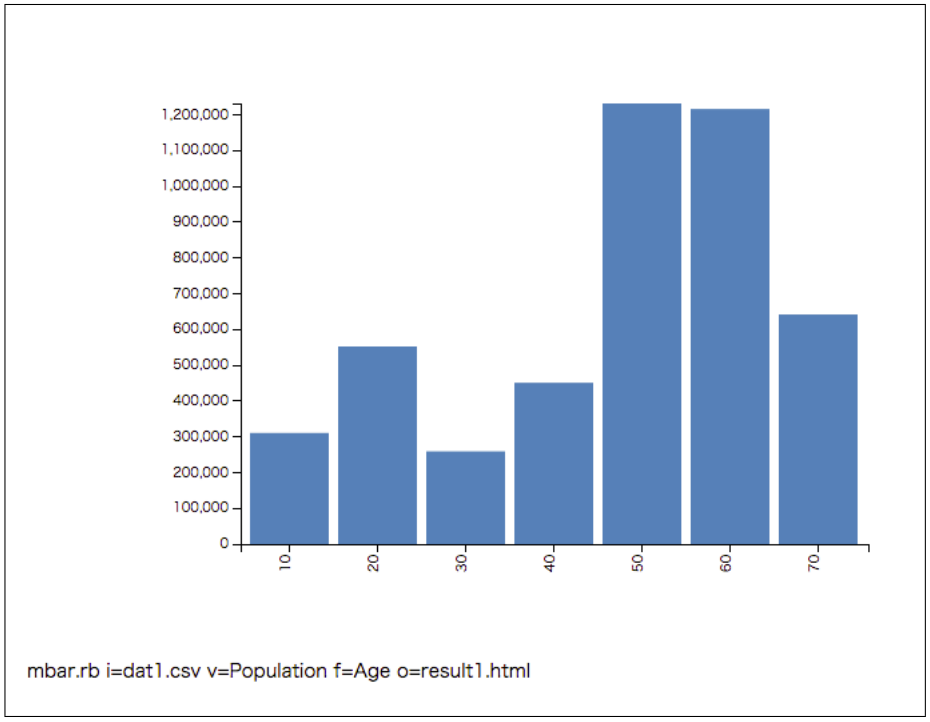
例 1: 棒グラフを 1 つ描画する

dat1.csv ファイルの Age を構成要素項目に、Population を構成量項目として棒グラフを 1 つ描画する。

```
$ more dat1.csv
Age,Population
10,310504
20,552339
30,259034.5555
40,0450818
50,1231572
60,1215966
70,641667
$ mbar.rb i=dat1.csv v=Population f=Age o=result1.html
#END# mbar.rb i=dat1.csv v=Population f=Age o=result1.html;
```

以下の棒グラフが描画される。

ブラウザで表示した棒グラフにマウスカーソルを置くと、構成要素項目とその構成量がポップアップで表示される。グラフはマウスによるドラッグ操作で移動することができ、またマウスのスクロール操作によって拡大縮小もできる。

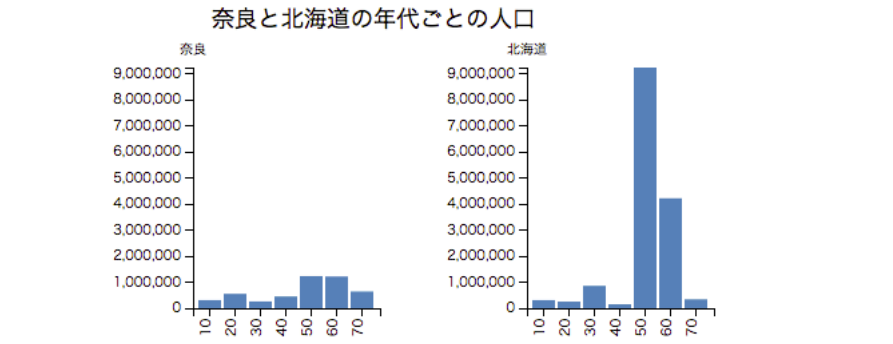


例 2: 1 **次元の棒グラフ行列を描画する**

dat2.csv ファイルの Age を構成要素項目に、Population を構成量項目として棒グラフを描画する。k=パラメータに Pref 項目を指定しているので、Pref 項目の値を x 軸（横方向）に展開した 1 次元の棒グラフ行列が描画される。title=パラメータでグラフのタイトルも指定している。

```
$ more dat2.csv
Pref,Age,Population
奈良,10,310504
奈良,20,552339
奈良,30,259034
奈良,40,450818
奈良,50,1231572
奈良,60,1215966
奈良,70,641667
北海道,10,310504
北海道,20,252339
北海道,30,859034
北海道,40,150818
北海道,50,9231572
北海道,60,4215966
北海道,70,341667
$ mbar.rb i=dat2.csv k=Pref v=Population f=Age o=result2.html
 title=奈良と北海道の年代ごとの人口
#END# mbar.rb i=dat2.csv k=Pref v=Population f=Age o=result2.html
title=奈良と北海道の年代ごとの人口;
```

以下の棒グラフ行列が描画される。



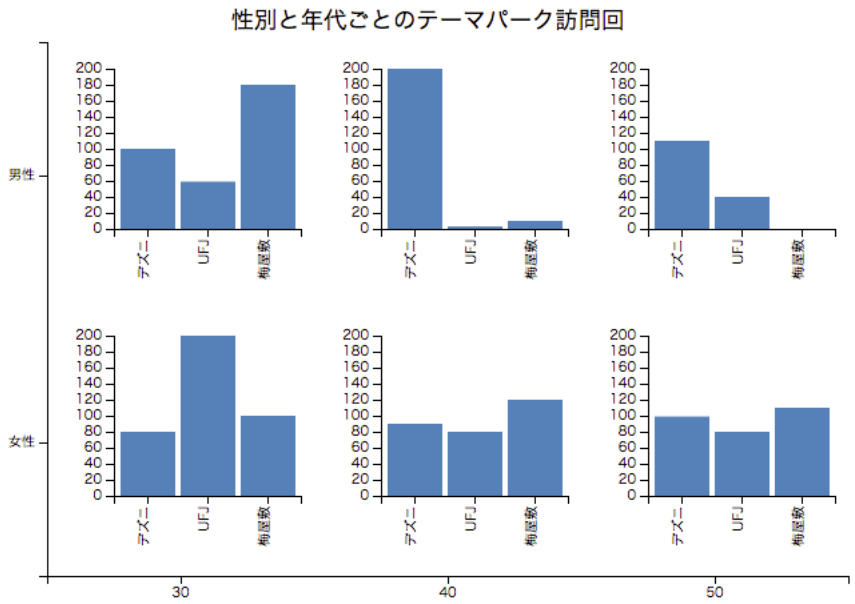
mbar.rb i=dat2.csv k=Pref v=Population f=Age o=result2.html title=奈良と北海道の年代ごとの人口

例 3: 2 **次元の棒グラフ行列を描画する**

dat3.csv ファイルの**テーマパーク名**を構成要素項目、Number を構成量項目とし、width=に幅 200、height=に高さ 150 を指定して、棒グラフを描画する。k=パラメータに Gender と Age 項目を指定しているので、Gender 項目の値を x 軸（横方向）に、Age 項目の値を y 軸（縦方向）に展開した 2 次元の棒グラフ行列が描画される。

```
$ more dat3.csv
Gender,Age,テーマパーク名,Number
男性,30,デズニ,100
男性,30,UFJ,59
男性,30,梅屋敷,180
男性,40,デズニ,200
男性,40,UFJ,3
男性,40,梅屋敷,10
男性,50,デズニ,110
男性,50,UFJ,40
女性,30,梅屋敷,100
女性,30,デズニ,80
女性,30,UFJ,200
女性,40,デズニ,90
女性,40,UFJ,80
女性,40,梅屋敷,120
女性,50,デズニ,99
女性,50,UFJ,80
女性,50,梅屋敷,110
$ mbar.rb i=dat3.csv k=Gender,Age v=Number f=テーマパーク名 o=result3.html
 title=性別と年代ごとのテーマパーク訪問回 width=200 height=150
#END# ./bin/mbar.rb i=dat3.csv k=Gender,Age v=Number f=テーマパーク名 o=result3.html
title=性別と年代ごとのテーマパーク訪問回 width=200 height=150;
```

以下の棒グラフ行列が描画される。



mbar.rb i=dat3.csv k=Gender,Age v=Number f=テーマパーク名 o=result3.html title=性別と年代ごとのテーマパーク訪問回 width=200 height=200

2.4　mgv.rb Graphviz 用グラフデータ (.dot) の作成

CSV 形式のグラフデータを、Graphviz が読み込める.dot 形式に変換する。

Take パッケージに含まれる mpolishing コマンド等では、グラフの入出力に CSV 形式を用いている。グラフを資料に掲載したり、グラフの規模や密度を目視するには、グラフを視覚化（画像として描画）する必要がある。

.dot 形式に変換することで、Graphviz（http://www.graphviz.org）のほか Gephi（http://www.gephi.org）などグラフ視覚化ソフトウェアに読み込ませることが可能となる。

ただし、Graphviz は比較的小規模なグラフの描画を目的としているため、頂点数が数百〜数千となると描画時間の面で実用的でなくなる。大規模グラフの操作・描画にあたっては Gephi の使用を推奨する。

2.4.1　書式

```
mgv.rb [ni=] [nf=] [nv=] [nr=] [-nl] ei= ef= [ev=] [er=] [-el] [-d] [o=] [--help]
```

ni=	: 頂点集合ファイル名
nf=	: 頂点 ID 項目名
nv=	: 頂点属性 (頂点の大きさ) 項目名
nr=	: グラフ描画時のノードの拡大率。1 から 10 までの実数値を指定できる。デフォルト値は 3
-nl	: nv=で指定した値をノードの名称に加える
ei=	: 枝集合ファイル名
ef=	: 開始頂点 ID 項目名, 終了頂点 ID 項目名
ev=	: 枝属性 (枝の太さ) 項目名
er=	: グラフ描画時のエッジの拡大率。1 から 20 までの実数値を指定できる。デフォルト値は 10
-el	: ev=で指定した値をエッジの横に表示する
-d	: 有向グラフとみなすとき指定する
o=	: 出力ファイル名 (.dot ファイル)
--help	: ヘルプの表示

入力するグラフデータは、ei=パラメータで指定する枝集合の CSV のみでかまわない。頂点にも属性（大きさ）を与えたい場合は、ni=パラメータを用いて頂点集合の CSV を指定することができる。

CSV 形式のグラフデータ例

1 行が 1 本の枝を表し、枝は開始頂点と終了頂点の 2 項目で表されている。

```
node1,node2
A,B
B,C
C,A
C,D
E,D
```

.dot 形式のグラフデータ例

頂点に関する情報と、枝に関する情報からなる。

```
digraph G {
  edge [dir=none]

    n0 [label="A" height=0.5 width=0.75]
n1 [label="B" height=0.5 width=0.75]
n2 [label="C" height=0.5 width=0.75]
n3 [label="D" height=0.5 width=0.75]
n4 [label="E" height=0.5 width=0.75]
```

```
    n0 -> n1 [style="setlinewidth(1.0)"]
n1 -> n2 [style="setlinewidth(1.0)"]
n2 -> n0 [style="setlinewidth(1.0)"]
n2 -> n3 [style="setlinewidth(1.0)"]
n4 -> n3 [style="setlinewidth(1.0)"]

}
```

Graphviz による描画例

Graphviz の GUI から対話的に読み込むことができるほか、Graphviz と一緒にインストールされる `dot` コマンドを使用して画像ファイル（.png ファイル）に直接変換することもできる。視覚化したグラフを図 2.2 に示す。

```
$ dot -Tpng rsl1.dot > rsl1.png
$ open rsl1.png
```

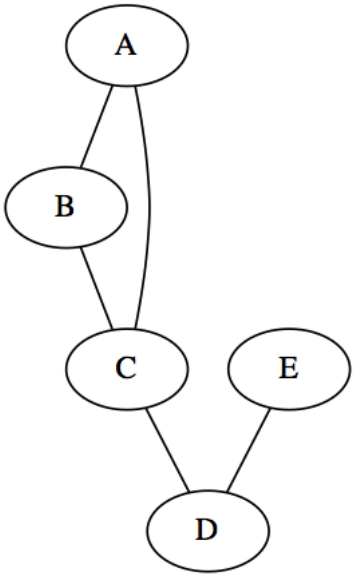


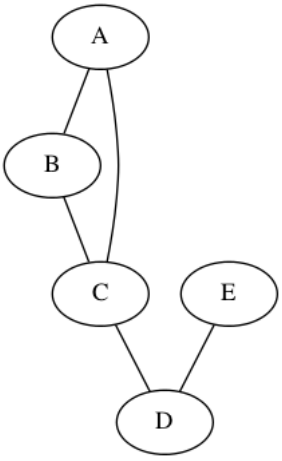
図 2.2　Graphviz による描画例

2.4.2 利用例

例 1: **基本例**

開始頂点と終了頂点からなる枝集合ファイルのみを与える。

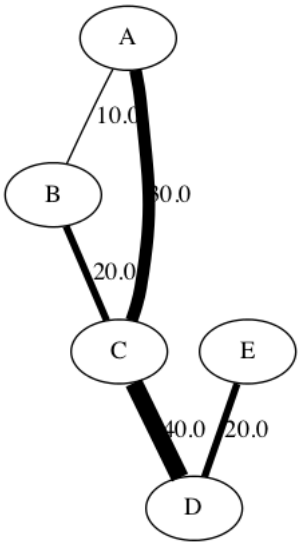
```
$ more edge1.csv
node1,node2
A,B
B,C
C,A
C,D
E,D
$ mgv.rb ei=edge1.csv ef=node1,node2 o=rsl1.dot
```



例 2: **枝に属性 (太さ) を指定する例**

ev=パラメータで val 項目を属性 (太さ) として指定している。同時に-el オプションを付けることで、属性値もグラフに描画される。

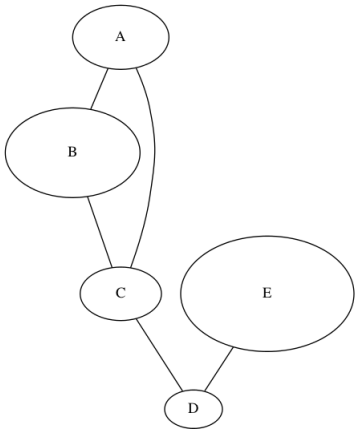
```
$ more edge2.csv
node1,node2,val
A,B,10
B,C,20
C,A,30
C,D,40
E,D,20
$ mgv.rb ei=edge2.csv ef=node1,node2 ev=val -el o=rsl2.dot
```



例 3: 頂点に属性 (大きさ) を指定する例

ni=パラメータで頂点集合ファイルを指定する。nv=パラメータで、val 項目を属性 (大きさ) として指定している。

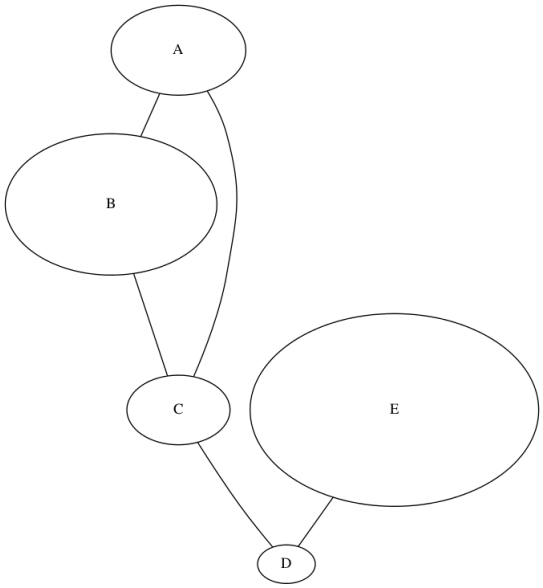
```
$ more node1.csv
node,val
A,10
B,15
C,8
D,5
E,20
$ more edge1.csv
node1,node2
A,B
B,C
C,A
C,D
E,D
$ mgv.rb ei=edge1.csv ef=node1,node2 ni=node1.csv nf=node nv=val o=rsl3.dot
```



例 4: 頂点に属性 (大きさ) と拡大率を指定する例

nr=パラメータで、ノードの拡大率を指定している。

```
$ more node1.csv
node,val
A,10
B,15
C,8
D,5
E,20
$ more edge1.csv
node1,node2
A,B
B,C
C,A
C,D
E,D
$ mgv.rb ei=edge1.csv ef=node1,node2 ni=node1.csv nf=node nv=val nr=5 o=rsl4.dot
```


A
B
C
E
D

2.5 mdtree.rb PMML による決定木モデルの描画

本コマンドは、PMML(Predictive Model Markup Language) で記述された決定木モデルを D3 ライブラリにより HTML 文書として視覚化する。mbonsai コマンド用の出力結果を視覚化する目的で作成したコマンドであるが、他のソフトウェアで生成される決定木の PMML データであっても視覚化できるであろう。PMML では数値とカテゴリの分岐ルールの記述方法は定義されているが、mbonsai で扱う系列パターンの有無の記述は定義されていない。そのため、mbonsai では、系列パーンによる分岐を記述するための拡張タグを定義しており、本コマンドもその拡張タグに対応して決定木を視覚化できる。

mbonsai で決定木を構築してから本コマンドで視覚化する一連の流れを以下に例示する。

表 2.4 入力データ dat1.csv。全データは例を参照のこと。

性別	来店距離	購入パターン	入院歴
男	1.2	ABCAAA	あり
男	10.5	BCDADD	あり
男	0.5	AAAA	なし
男	2.0	BBCC	なし
男	3.1	DEDDA	あり
女	0.7	CCCAA	なし
女	1.5	DDDEEE	あり
女	2.6	BACD	あり
女	3.5	ABBB	あり
女	4.0	DDDD	あり
女	2.1	DEDE	なし
:	:	:	:

表 2.4 に示されるデータを訓練データとして mbonsai コマンドで決定木を構築する。決定木は、O=に指定したディレクトリに PMML ファイル model.pmml として保存されている。

```
$ mbonsai c=入院歴 n=来店距離 p=購入パターン d=性別 i=dat1.csv O=outdat
#END# kgbonsai O=outdat c=入院歴 d=性別 i=dat1.csv n=来店距離 p=購入パターン; IN=81;
$ ls outdat
alpha_list.csv model.pmml      model.txt      model_info.csv param.csv      predict.csv
```

そして、model.pmml を視覚化するには以下のように本コマンドを実行すればよい。出力された model.html をブラウザで描画したものを図 2.3 に示す。

```
$ mdtree.rb i=outdat/model.pmml o=model.html
#END# mdtree.rb i=outdat/model.pmml o=model.html;
$ open model.html # mac の場合は html ファイルを open すればブラウザが起動され描画される
```

また、mbonsai で構築される決定木には最大木が保存されているので、枝刈り度を alpha=で指定することで枝刈りされた決定木を描画することもできる。alpha=は 0 以上の実数で、大きくすると枝が多く刈られる。alpha を指定しなかった場合、mbonsai で交差検証を指定しなければ、alpha=0.01 が指定されたことになり、交差検証を指定していれば、誤分類率最小のモデルが描画される。

図 2.4 に、alpha=0.1 で枝刈りした決定木を示す。

```
$ mdtree.rb alpha=0.1 i=outdat/model.pmml o=model2.html
#END# mdtree.rb alpha=0.1 i=outdat/model.pmml o=model.html;
$ open model2.html # mac の場合は html ファイルを open すればブラウザが起動され描画される
```

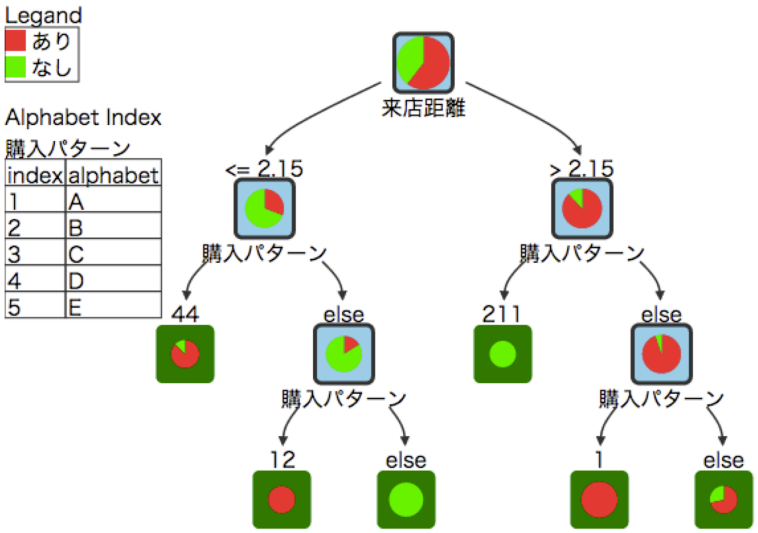


図 2.3　本コマンドによる決定木の描画。接点内の円グラフは、クラスの分布を示している (色は凡例に表示)。背景が水色の節点は中間節点を、緑色の節点は葉節点を表している。節点の直ぐ下には分岐に使う項目名が閉められており、節点の直ぐ上には分岐のルールが示されている。例えば、最上位の節点では、来店距離が 2.15 以下であれば左に、2.15 より長ければ右に分岐する。系列パターンの場合は、そのパターンを含めば左に、含まなければ右に分岐する。例えば、上から 2 段目の左の節点は、購入パターンに”44”を含んでいれば左に、含まなければ右に分岐することを意味している。また、系列パターンの文字は、図の左上に示された alphabet-index の対応表における index に示されている。

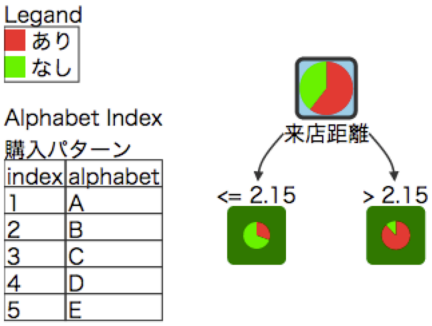


図 2.4　枝刈り度 alpha=0.1 で描画した決定木

2.5.1　R との連携

統計解析パッケージ R には多くの決定木構築パッケージが用意されている。以下では、`rpart` で構築した決定木を本コマンドで描画する方法を解説する。

以下では、アヤメデータセット (`iris`) と前立腺がんデータセット (`stagec`) の 2 つのデータセットから決定木を構築する R スクリプトである。そこでは、決定木を構築する `rpart` ライブラリとモデルを PMML 出力する `pmml` ライブラリを最初に読み込んでいる。データセットの内容および決定木モデルの構築方法についての詳細は省略する。最終的に、アヤメの決定木と前立腺がんの決定木が、それぞれ PMML ファイル `model_r1.pmml`、`model_r1.pmm2` として出力される。

```
library(pmml)
library(rpart)
iris.rp=rpart(Species~.,data=iris)
sink("model_r1.pmml")
pmml(iris.rp)
```

```
sink()

stagec$progstat <- factor(stagec$pgstat, levels = 0:1, labels = c("No", "Prog"))
cfit <- rpart(progstat ~ age + eet + g2 + grade + gleason + ploidy, data = stagec, method = 'class')
sink("model_r2.pmml")
pmml(cfit)
sink()
```

得られた 2 つの PMML ファイルについて、本コマンドで決定木を描画する手順は以下のとおりである。そして、それぞれの決定木は図 2.5、図 2.5 に示されるように描画される。

```
$ mdtree.rb i=model_r1.pmml o=$op/out_r1.html
#END# mdtree.rb i=model_r1.pmml o=model_r1.html;
$ mdtree.rb i=model_r2.pmml o=$op/out_r2.html
#END# mdtree.rb i=model_r2.pmml o=model_r2.html;
$ open model1_r1.html
$ open model1_r2.html
```

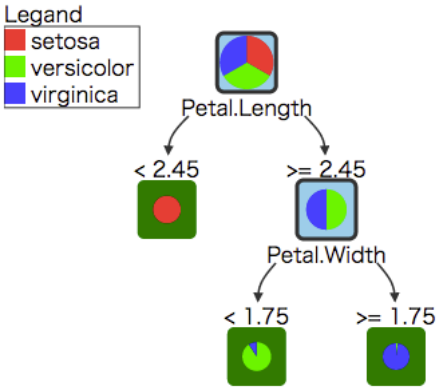


図 2.5 アヤメデータセットの決定木

2.5.2 書式

```
mdtree.rb i= o= [alpha=] [--help]
```

```
i=       : 決定木モデルの PMML ファイル
o=       : 出力ファイル (HTML ファイル)
alpha=   : 枝刈り度を指定する (0 以上の実数で、大きくすると枝が多く刈られる)。
         : 指定しなかった場合、mbonsai で交差検証を指定しなければ、
         : 0.01 が指定されたことになり、交差検証を指定していれば、誤分類率最小のモデルが描画される。
         : このパラメータは mbonsai で構築した決定木のみ有効。
--help   : ヘルプの表示
```

2.5.3 利用例

例 1: 基本例

前節の解説で用いてる例。

```
$ cat dat1.csv
性別,来店距離,購入パターン,入院歴
男,1.2,ABCAAA,あり
男,10.5,BCDADD,あり
```

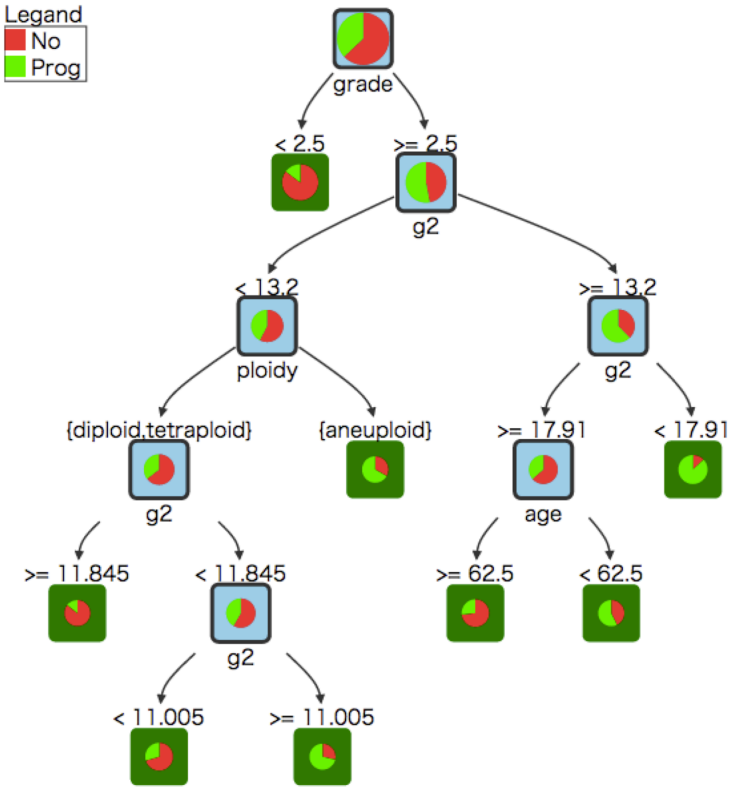


図 2.6　前立腺がんの決定木

```
男,0.5,AAAA,なし
男,2.0,BBCC,なし
男,3.1,DEDDA,あり
女,0.7,CCCAA,なし
女,1.5,DDDEEE,あり
女,2.6,BACD,あり
女,3.5,ABBB,あり
女,4.0,DDDD,あり
女,2.1,DEDE,なし
男,1.2,ABCAAA,あり
男,10.5,BCDADD,あり
男,0.5,AAAA,なし
男,2.0,BBCC,なし
男,3.1,DEDDA,あり
男,0.7,CCCAA,なし
男,1.5,DDDEEE,なし
男,2.6,BACD,あり
男,3.5,ABBB,あり
男,4.0,DDDD,あり
男,2.1,DEDE,なし
男,1.2,ABCAAA,あり
男,10.5,BCDADDA,あり
男,0.5,AAAAA,なし
男,2.0,BBCCA,なし
男,3.1,DEDDA,あり
男,0.7,CCCAA,なし
男,1.5,ADDDEEE,あり
男,2.6,BACD,あり
男,3.5,ABBB,あり
男,4.0,DDDD,あり
女,2.1,DEDE,なし
女,1.2,ABCAAA,あり
女,10.5,BCDADD,あり
```

```
女,0.5,AAAA,なし
女,2.0,BBCC,なし
女,3.1,DEDDA,あり
女,0.7,CCCAA,なし
女,1.5,DDDEEE,あり
女,2.6,BACD,あり
女,3.5,ABBB,あり
女,4.0,DDDD,あり
女,2.1,DEDE,なし
女,1.2,ABCAAA,あり
女,10.5,BCDADD,あり
女,0.5,AAAA,なし
女,2.0,BBCC,なし
女,3.1,DEDDA,あり
女,0.7,CCCAA,なし
女,1.5,DDDEEE,あり
女,2.6,BACD,あり
女,3.5,ABBB,あり
女,1.0,DDDD,あり
女,2.5,DEDE,なし
女,2.5,ABBB,あり
女,1.0,DDDD,あり
女,1.1,DEDE,なし
女,2.2,ABCAAA,あり
女,10.5,BCDADD,あり
女,1.5,AAAA,なし
女,2.6,BBCC,なし
女,3.3,DEDDA,あり
女,1.7,CCCAA,なし
女,1.5,DDDEEE,あり
女,2.6,BACD,あり
女,3.9,ABBB,あり
女,4.5,DDDD,あり
女,2.1,DEDE,なし
女,3.9,BABB,あり
男,4.5,BAA,なし
女,2.1,DEDE,なし
男,3.9,BABB,あり
女,3.9,BABB,あり
男,4.5,BAA,なし
女,2.1,DEDE,なし
男,3.9,BABB,あり
女,3.9,BABB,あり
男,4.5,BAA,なし
女,2.1,DEDE,なし
男,3.9,BABB,あり
$ mbonsai c=入院歴 n=来店距離 p=購入パターン d=性別 i=dat1.csv O=outdat
ABCDE = 12345  *improved(errev:0.037037 *improved(errMin:0,leaf:1)
#END# kgbonsai O=outdat c=入院歴 d=性別 i=dat1.csv n=来店距離 p=購入パターン
$ mdtree.rb i=outdat/model.pmml o=model1.html
#END# /usr/bin/mdtree.rb i=outdat/model.pmml o=model1.html
$ mdtree.rb alpha=0.1 i=outdat/model.pmml o=model2.html
#END# /usr/bin/mdtree.rb alpha=0.1 i=outdat/model.pmml o=model2.html
$ head model1.html
<html lang="ja">
<head>
  <meta charset="utf-8">
        <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <style type="text/css">
          p.title { border-bottom: 1px solid gray
                g > .type-node > rect { stroke-width: 3px
                g > .type-leaf > rect { fill: green
                .edge path {  fill: none
                svg >.legend > rect { stroke-width: 1px
```

```
$ head model2.html
<html lang="ja">
<head>
  <meta charset="utf-8">
        <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <style type="text/css">
          p.title { border-bottom: 1px solid gray
                g > .type-node > rect { stroke-width: 3px
                g > .type-leaf > rect { fill: green
                .edge path {  fill: none
                svg >.legend > rect { stroke-width: 1px
```