Outline

Classification and Regression Trees

CART is arguably the most widely used tree algorithm in the statistics community, but most tree algorithms are quite similar.
We suppose in the following that the p predictor variables are real-valued but extensions to categorical variables are straightforward.
A tree is then equivalent to a partition of $\mathbb{R}^p$ into R disjoint sets $\mathcal{P} = \{\mathcal{R}_1, \ldots, \mathcal{R}_R\}$, where each $\mathcal{R}_j \subset \mathbb{R}^p$ and has constant fitted values in each region $\mathcal{R}_j$, $j = 1, \ldots, R$.

Example I : NHS Direct self-help guide

Colds and flu

Colds and flu

This advice is suitable for children and adults.

Are you developing a rash that does not fade when you press a glass tumbler or finger against it?

Yes → **Dial 999**

No

Are you suffering from a stiff neck, headache and do you find light hurts your eyes and / or do you feel very sleepy and confused?

Yes → **Dial 999**

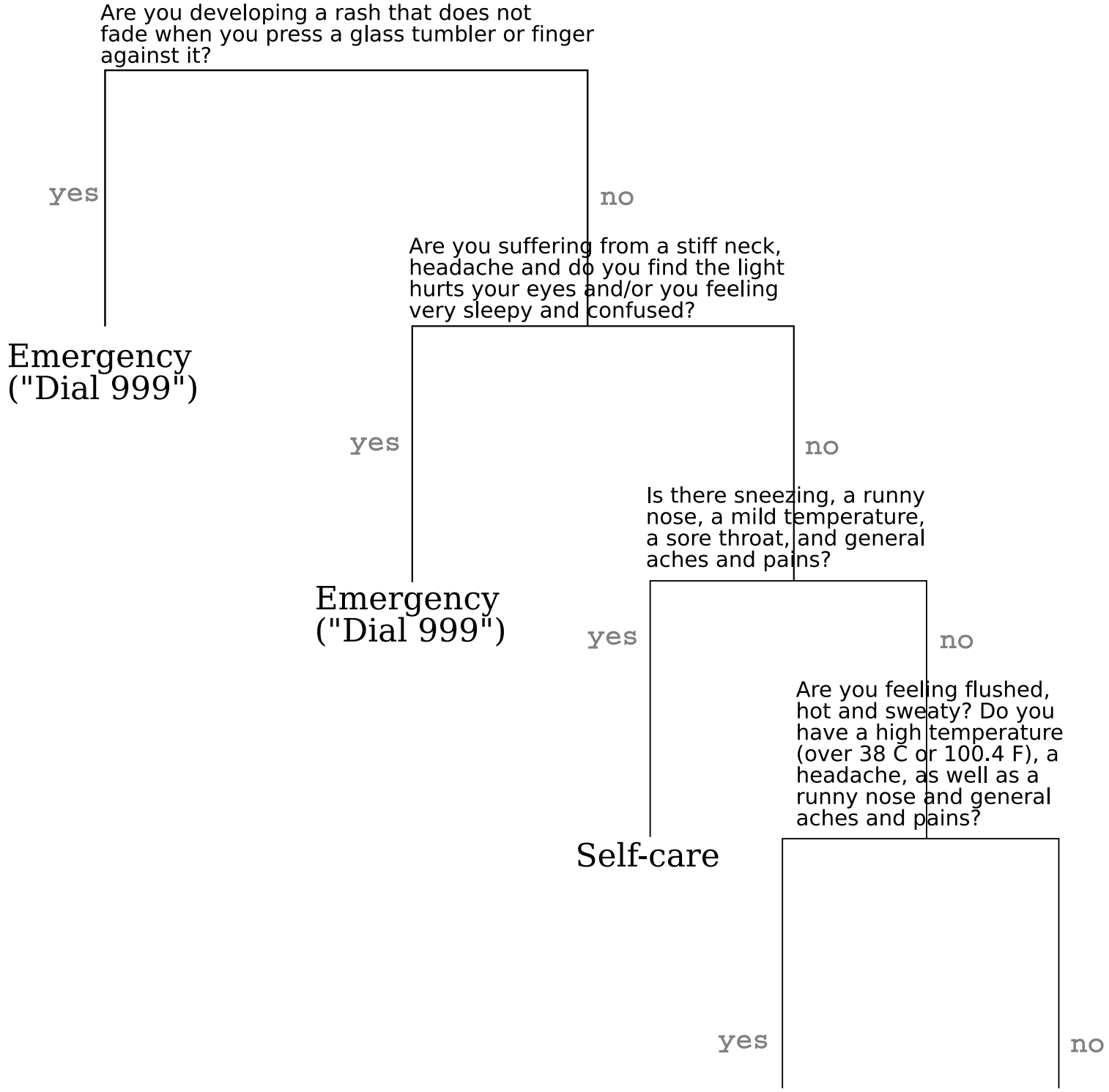
No

Is there sneezing, a runny nose, a mild temperature, a sore throat, and general aches and pains?

Yes → **Self-care**

It could be a common cold which antibiotics cannot treat effectively. Unless the person is very old, frail or has some other serious condition, you **do not need to see your doctor**. Take paracetamol (or, for children use paediatric paracetamol oral suspension, available from pharmacists), warm

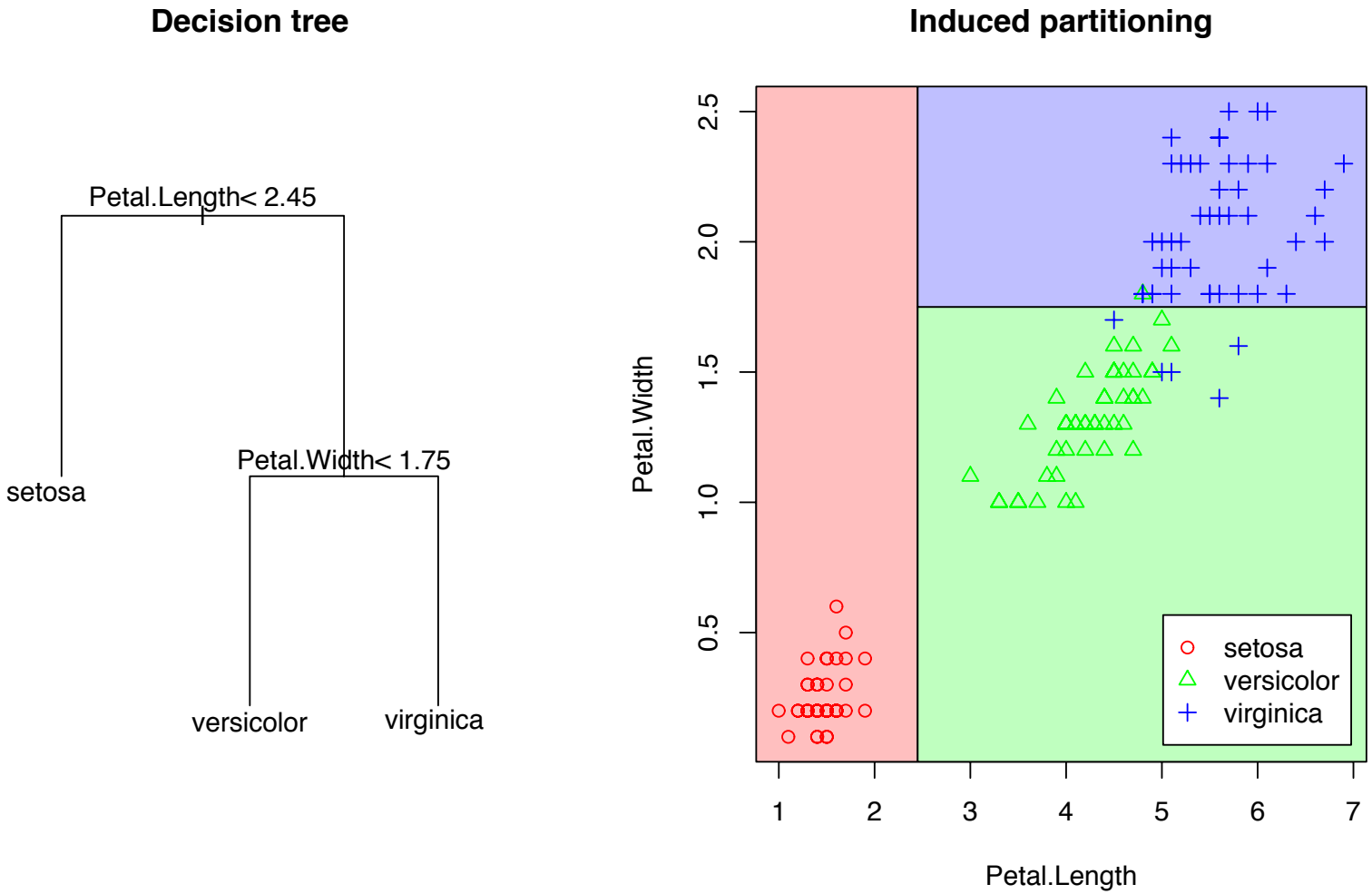
No


Are you developing a rash that does not fade when you press a glass tumbler or finger against it?
yes
no
Emergency ("Dial 999")
Are you suffering from a stiff neck, headache and do you find the light hurts your eyes and/or you feeling very sleepy and confused?
yes
no
Emergency ("Dial 999")
Is there sneezing, a runny nose, a mild temperature, a sore throat, and general aches and pains?
yes
no
Self-care
Are you feeling flushed, hot and sweaty? Do you have a high temperature (over 38 C or 100.4 F), a headache, as well as a runny nose and general aches and pains?
yes
no

Example II: Iris Data

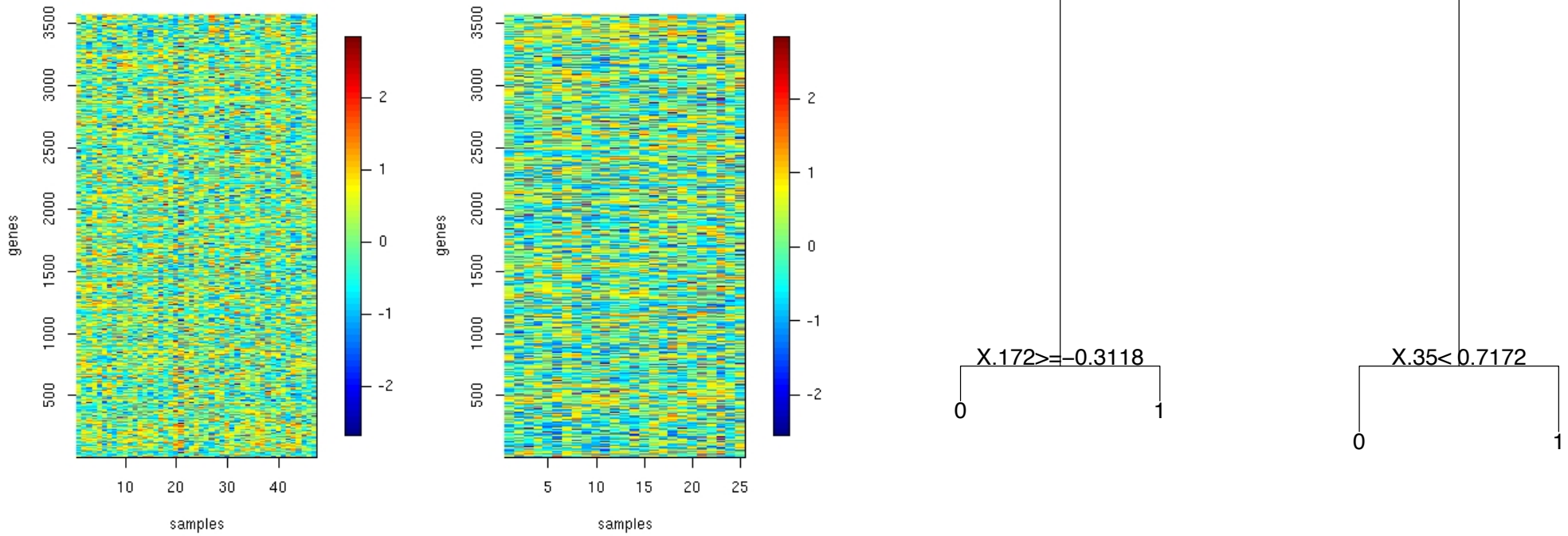
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.4	3.2	1.3	0.2	setosa
5.9	3.0	5.1	1.8	virginica
6.3	3.3	6.0	2.5	virginica
5.3	3.7	1.5	0.2	setosa
5.5	2.5	4.0	1.3	versicolor
6.1	2.9	4.7	1.4	versicolor
6.1	3.0	4.9	1.8	virginica
5.7	2.8	4.5	1.3	versicolor
5.4	3.0	4.5	1.5	versicolor
4.8	3.4	1.6	0.2	setosa
4.6	3.1	1.5	0.2	setosa
4.9	3.1	1.5	0.2	setosa
6.4	2.9	4.3	1.3	versicolor
.....				

Previously seen Iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.



The decision tree derived from the Iris data (left) and the partitioning of feature space.

Example III: Leukemia Prediction



Leukemia Dataset: Expression values of 3541 genes for 47 patients with Leukemia ALL subtype (left) and 25 patients with AML (middle). Decision tree (right).

Some terminology

We will use the following terminology.

- **Parent** of a node c is the set of nodes (here maximally of size 1) which have an arrow pointing towards c.
- **Children** of a node c are those nodes which have node c as a parent.
- **Root node** is the top node of the tree; the only node without parents.
- **Leaf nodes** are nodes which do not have children.
- **Stumps** are trees with just the root node and two leaf nodes.
- A K**-ary tree** is a tree where each node (except for leaf nodes) has K children. Usually working with binary trees ($K = 2$).
- The **depth** of a tree is the maximal length of a path from the root node to a leaf node.

Regression

For regression, CART provides a piecewise constant prediction on each region $\mathcal{R}_j$,

$$\hat{Y}_{tree}(x) = \sum_{r=1}^{R} \beta_r \mathbf{1}_{[x \in \mathcal{R}_r]},$$

where β_j is the constant fitted value in $\mathcal{R}_j$.
The function $\hat{Y}(\cdot)$ is hence determined if the (a) partition and (b) the fitted values β are chosen. These choices are made such as to minimize the expected squared error loss for future observations (x, Y),

$$E(Y - \hat{Y}(x))^2.$$

Classification

For classification with two classes, the response is in $Y \in \{0, 1\}$. CART chooses again a piece-wise constant function

$$\hat{Y}_{tree}(x) = \sum_{r=1}^{R} \beta_r \mathbf{1}_{[x \in \mathcal{R}_r]}.$$

This time, $\beta_r \in [0, 1]$. The default classification is

$$\eta_{tree}(x) = \begin{cases} 0 & \hat{Y}_{tree}(x) \leq 1/2 \\ 1 & \hat{Y}_{tree}(x) > 1/2 \end{cases}.$$

A good choice of $\hat{Y}_{tree}$ is one that leads to a small misclassification error

$$P(\eta_{tree}(x) \neq Y)$$

or, in general, to a small loss

$$E(L(Y, \eta_{tree}(X))),$$

for a loss function $\{0, 1\} \times \{0, 1\} \mapsto \mathbb{R}^+$.

Parameter Estimation

Recall model

$$\hat{Y}_{tree}(x) = \sum_{r=1}^{R} \beta_r \mathbf{1}_{[x \in \mathcal{R}_r]},$$

Parameter estimation $\hat{\beta}_1, \ldots, \hat{\beta}_R$ is easy if the partition $\mathcal{P} = \{\mathcal{R}_1, \ldots, \mathcal{R}_R\}$ were given.
We use

$$\begin{aligned} \hat{\beta}_r &= \sum_{i=1}^{n} Y_i \mathbf{1}_{[x_i \in \mathcal{R}_r]} / \sum_{i=1}^{n} \mathbf{1}_{[x_i \in \mathcal{R}_r]} \\ &= \text{mean}\{Y_i : X_i \in \mathcal{R}_r\}. \end{aligned}$$

for regression and binary classification (where $\hat{\beta}_r$ is just the proportion of samples from class 1 in region $\mathcal{R}_r$).

Partition Estimation

Ideally, would like to find partition that allows (with the previous parameter estimates) to achieve lowest mean-squared error loss (prediction) or misclassification rate (classification).
Number of potential partitions is too large to search exhaustively for problems of even just small to moderate size (in terms of number p of variables and number n of samples).
Need 'greedy' search for a good partition. First search for a good split for the root node and then work successively downwards in the tree.

Splitpoint estimation for regression trees

Given are data-points $(X_1, Y_1), \ldots, (X_n, Y_n)$, where each $X_i = (X_i^{(1)}, \ldots, X_i^{(p)})$ is a p-dimensional vector.
For continuous predictor variables, the search for a partition works as follows.

1. Start with $\mathcal{R}_1 = \mathbb{R}^p$.
2. Given a partition $\mathcal{R}_1, \ldots, \mathcal{R}_r$, split each region $\mathcal{R}_j$ into two parts $\mathcal{R}_{j_1}, \mathcal{R}_{j_2}$, where

$$\begin{aligned} \mathcal{R}_{j_1} &= \{x \in \mathbb{R}^p : x \in \mathcal{R}_j \text{ and } X^{(k)} \leq c\}. \\ \mathcal{R}_{j_2} &= \{x \in \mathbb{R}^p : x \in \mathcal{R}_j \text{ and } X^{(k)} > c\}, \end{aligned}$$

   where splitpoint c and splitting variable k are found as

$$\text{argmin}_{c,k} \min_{\beta_1, \beta_2} \Big(\sum_{i: X_i \in \mathcal{R}_{j_1}} (Y_i - \beta_1)^2 + \sum_{i: X_i \in \mathcal{R}_{j_2}} (Y_i - \beta_2)^2 \Big).$$

   Let $\mathcal{R}_{1_1}, \mathcal{R}_{1_2}, \ldots, \mathcal{R}_{r_1}, \mathcal{R}_{r_2}$ be the new partition.
3. Repeat step 2) d times to get tree of depth d.

Boston Housing Data

The original data are 506 observations on 14 variables, medv being the response variable:

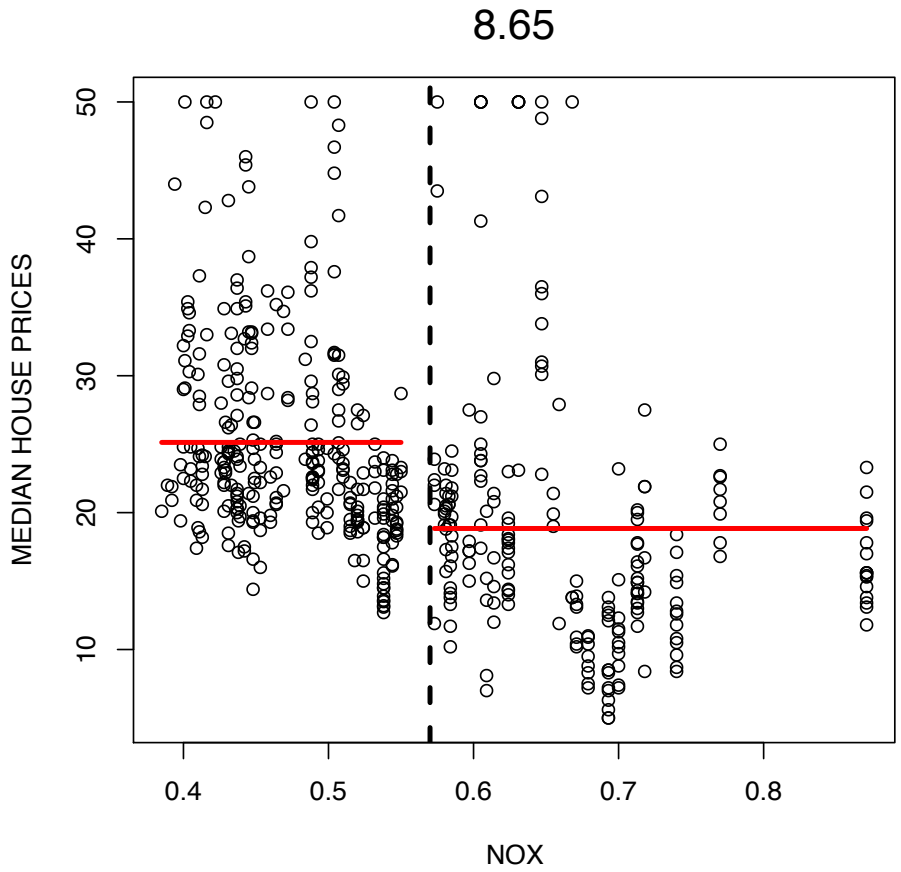
```
crim    per capita crime rate by town
zn      proportion of residential land zoned for lots over 25,000
indus   proportion of non-retail business acres per town
chas    Charles River dummy variable (= 1 if tract bounds river; (
nox     nitric oxides concentration (parts per 10 million)
rm      average number of rooms per dwelling
age     proportion of owner-occupied units built prior to 1940
dis     weighted distances to five Boston employment centres
rad     index of accessibility to radial highways
tax     full-value property-tax rate per USD 10,000
ptratio pupil-teacher ratio by town
b       1000(B - 0.63)^2 where B is the proportion of blacks by to
lstat   percentage of lower status of the population
medv    median value of owner-occupied homes in USD 1000's
```

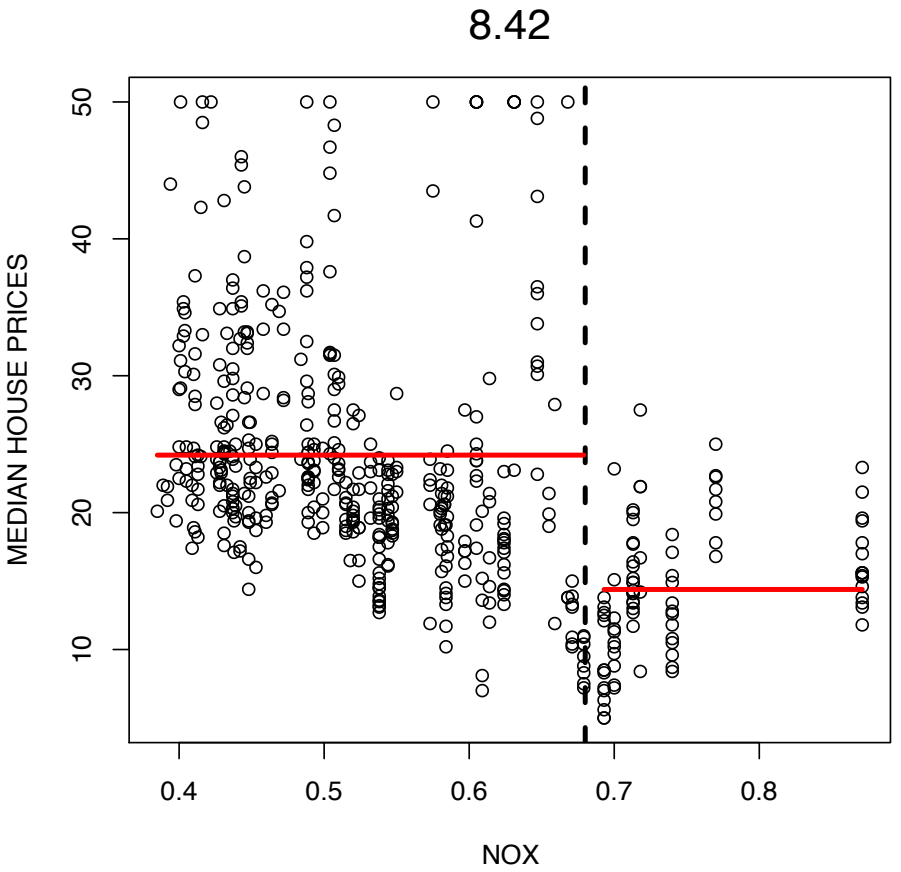
```
> library(MASS)
> data(Boston)
> str(Boston)
'data.frame': 506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0
 $ rm     : num  6.58 6.42 7.18 7.00 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 .
```

...predict median house price given 13 predictor variables.

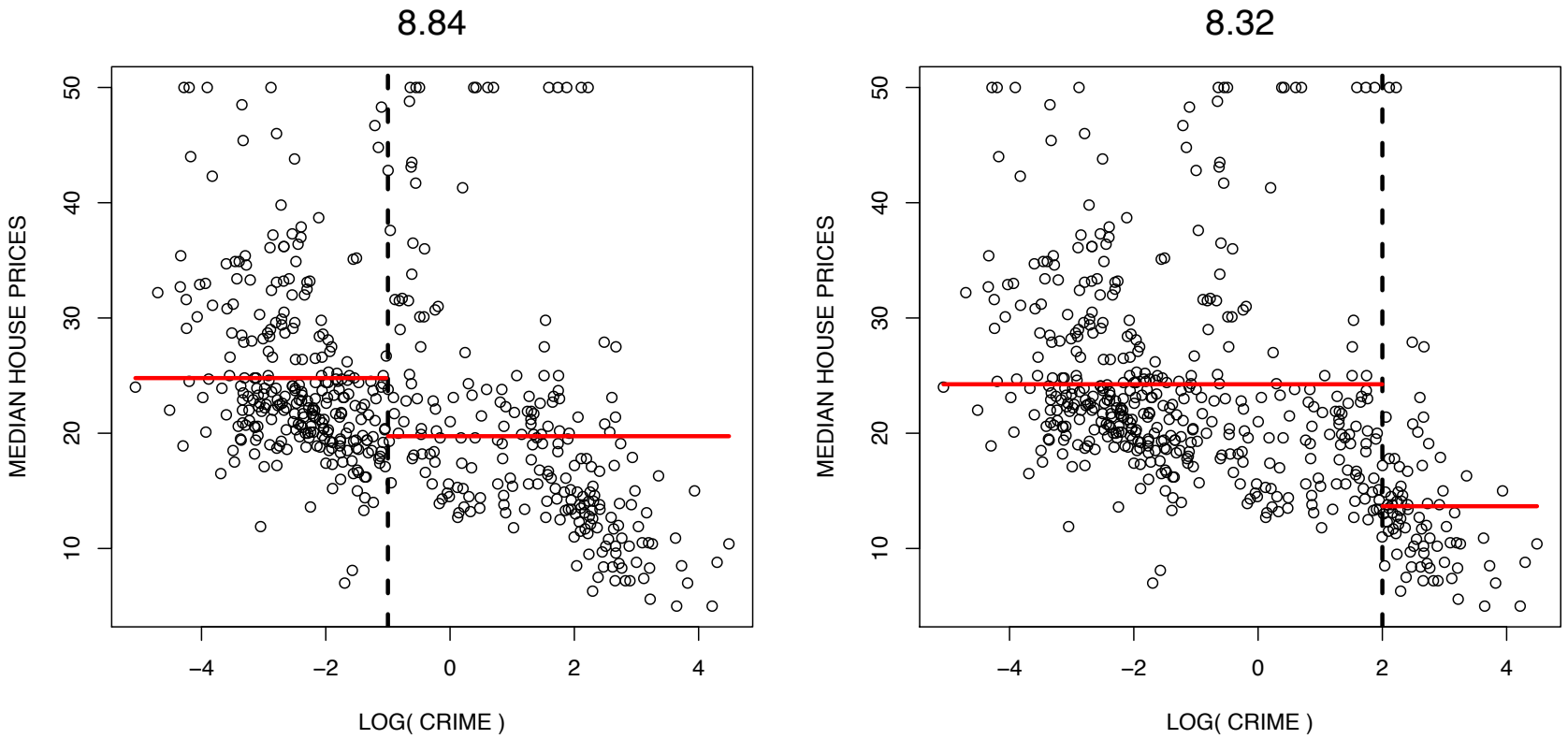
crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat
medv

Try splitting on variable ‘NOX’ at two different splitpoints and look at the residual sum of squares.



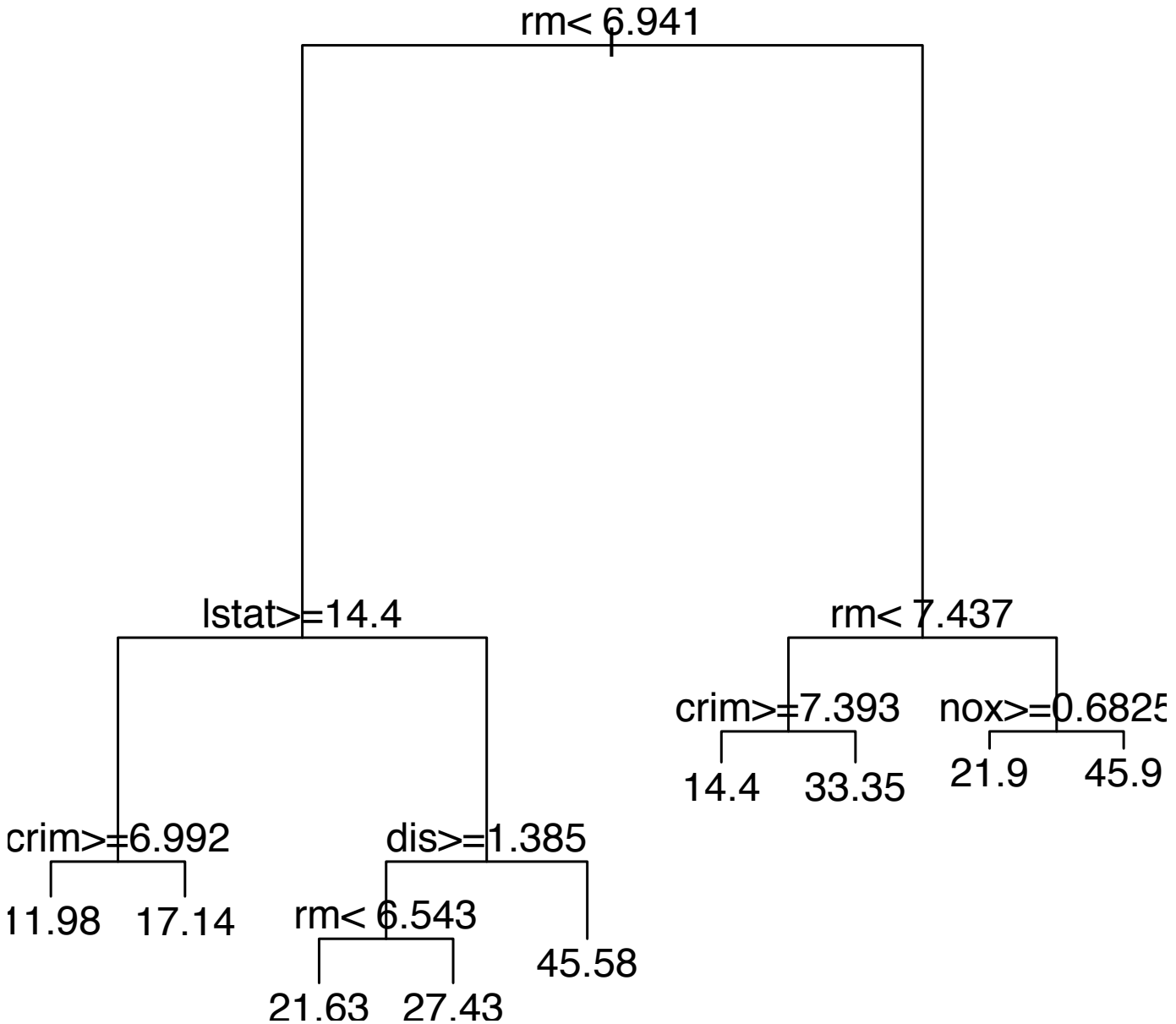


Try splitting now on variable ‘CRIME at two different splitpoints instead.



...last split is most favourable among the four considered splits as it leads to largest reduction in MSE. Choose this split.

Overall, the best first split is on variable `rm`, average number of rooms per dwelling. Final tree contains predictions in leaf nodes.



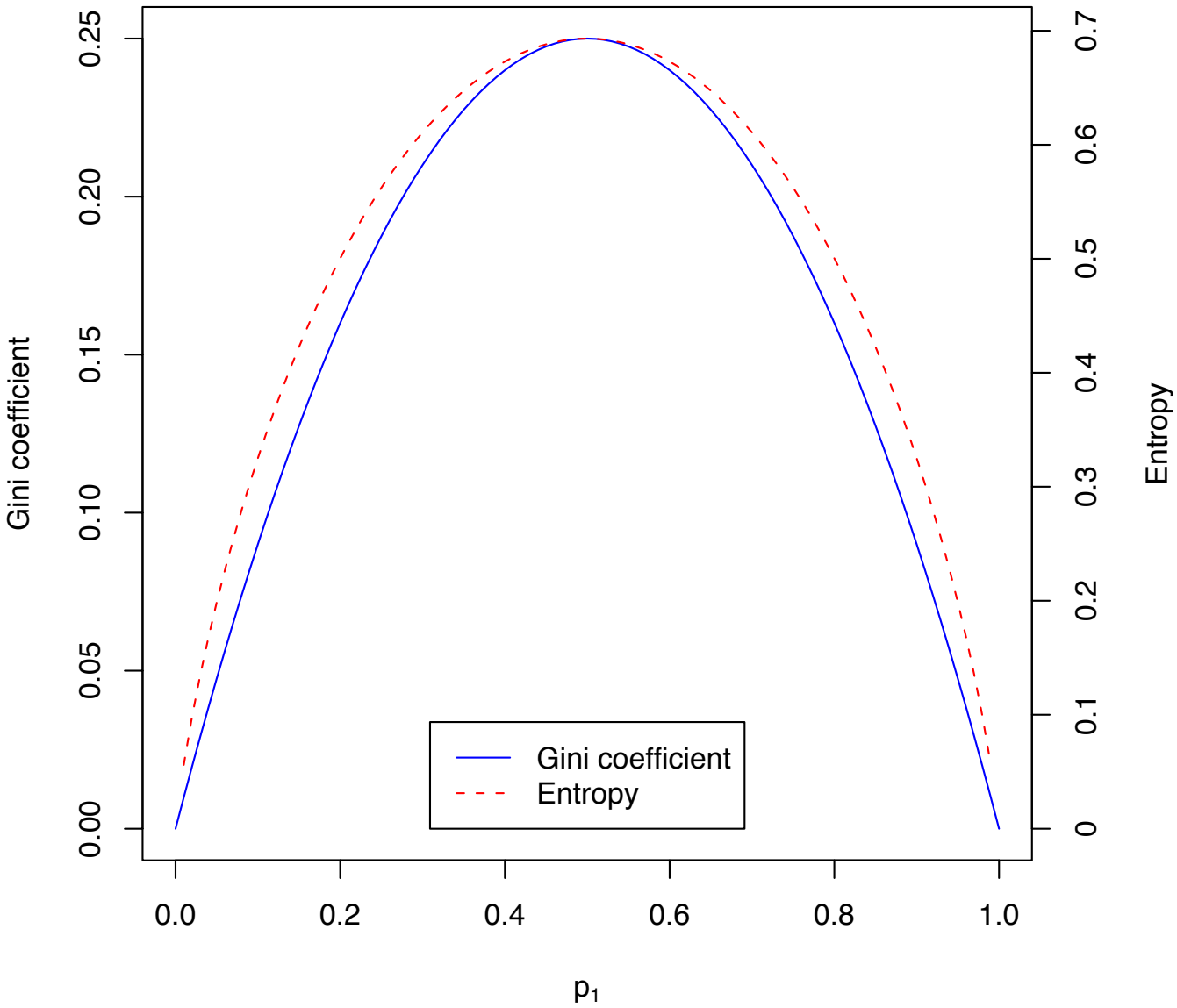
Classification

Remember that for binary classification,

$$\hat{p}_r = \hat{\beta}_r = \sum_{i=1}^{n} Y_i \mathbf{1}_{[x_i \in \mathcal{R}_r]} / \sum_{i=1}^{n} \mathbf{1}_{[x_i \in \mathcal{R}_r]}$$

is just the estimated probability for class $Y = 1$ in each partition $\mathcal{R}_r$.
The tree growth algorithm is identical to the regression case, except that one is using a different measure of node impurity. For regression, the residual sum of squares $\sum_{i:\mathcal{R}_r} (Y_i - \hat{\beta}_r)^2$ was used for each region $\mathcal{R}_r$.
For classification, try instead to minimize a measure of node impurity:

- Misclassification error: $1 - \max\{\hat{p}_r, 1 - \hat{p}_r\}$.
- Gini Index: $2\hat{p}_r(1 - \hat{p}_r)$.
- Cross-entropy: $-\hat{p}_r \log \hat{p}_r - (1 - \hat{p}_r) \log(1 - \hat{p}_r)$.



Misclassification error?

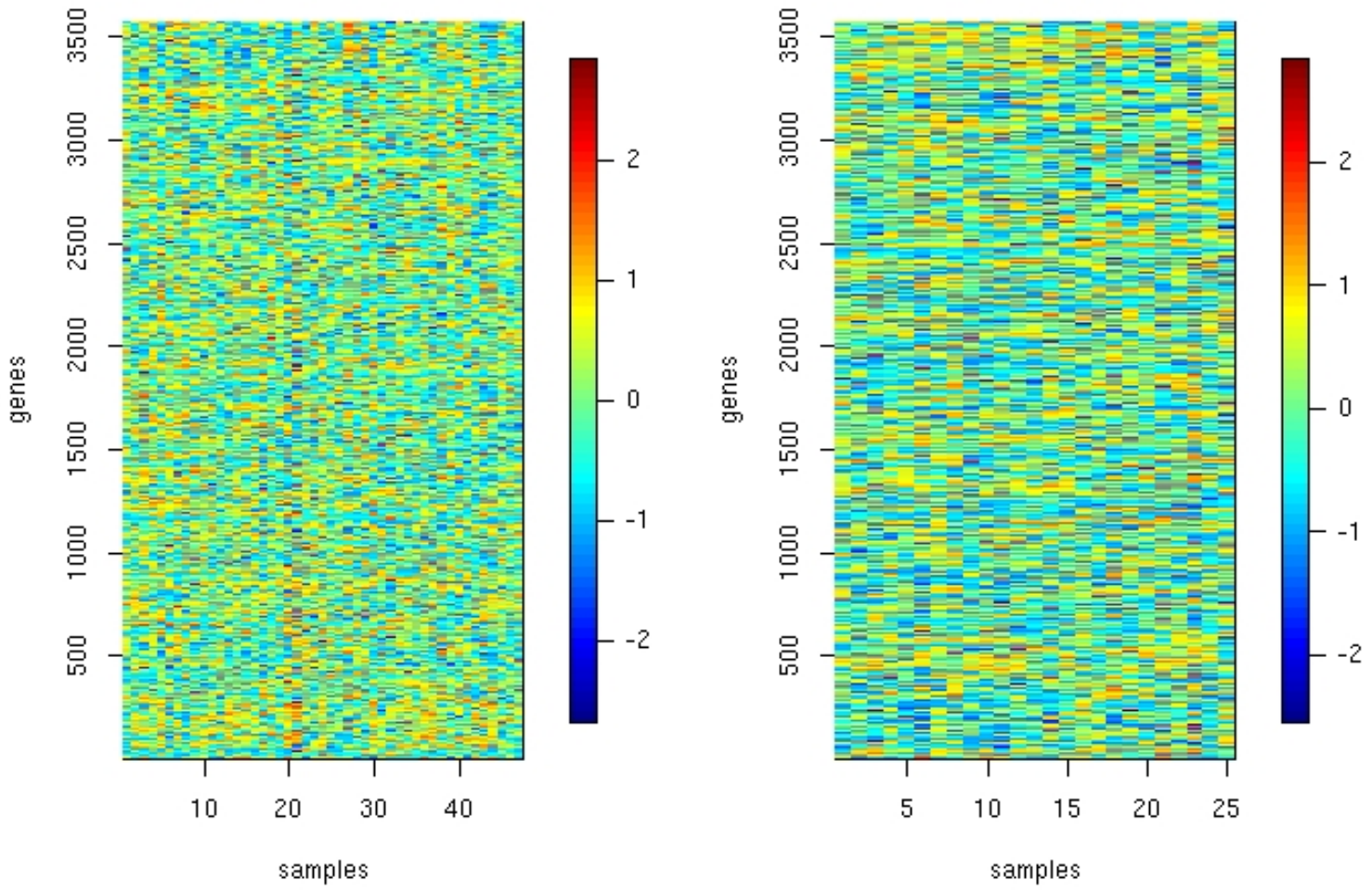
All three criteria of misclassification error are similar, but Gini Index and Cross-entropy usually preferred. Gini Index and Cross-entropy are differentiable; misclassification error not.
Gini Index and Cross-entropy also favour purer nodes. Consider example where 400 observations of class 0 and 400 observations of class 1 are present. Possible splits into
A: (300,100) and (100,300) vs.
B: (200,400) and (200,0).

A: (300,100) and (100,300) vs.
B: (200,400) and (200,0).

Split A	$\hat{\beta}$	Misclassification error	Gini Index
Node 1	1/4	1/4	$2\frac{1}{4}(1-\frac{1}{4}) = 3/8$
Node 2	3/4	1/4	$2\frac{3}{4}(1-\frac{3}{4}) = 3/8$
Total		$\frac{400}{800} \cdot 1/4 + \frac{400}{800} \cdot 1/4 = 1/4$	$\frac{400}{800} \cdot 3/8 + \frac{400}{800} \cdot 3/8 = 3/8$

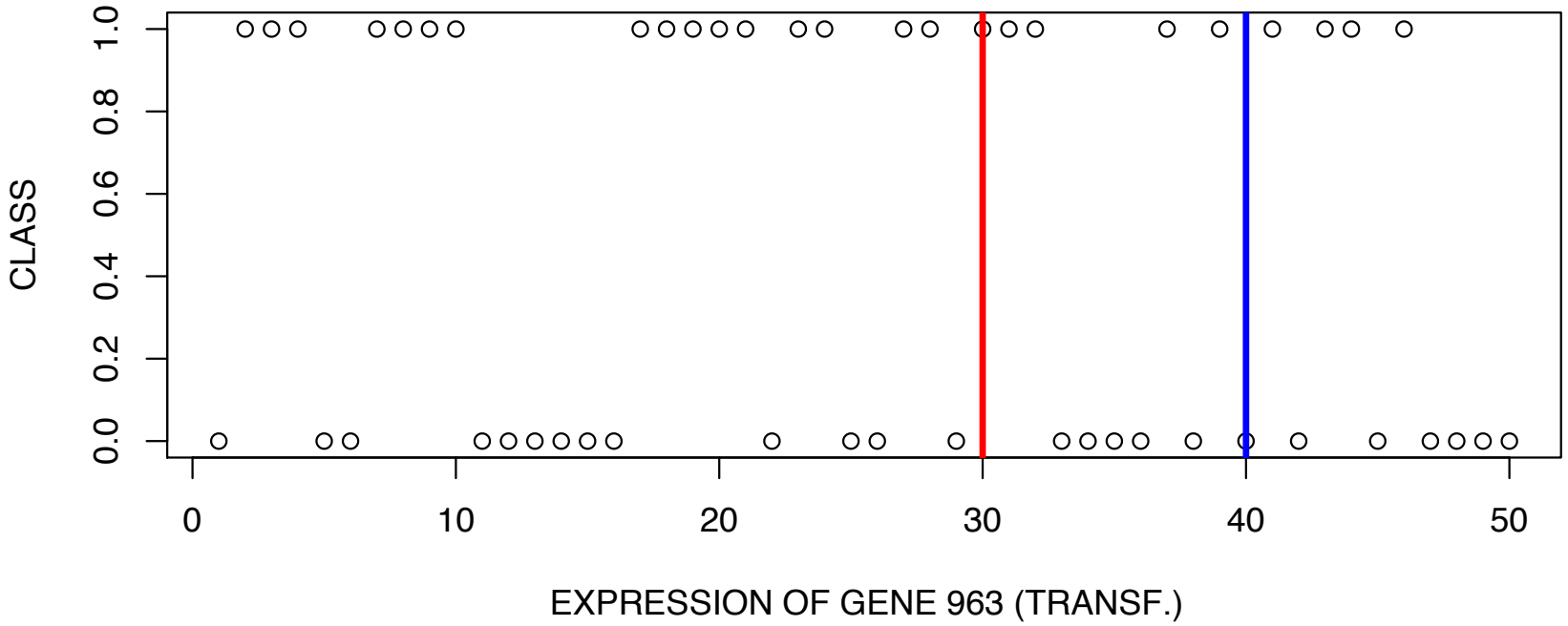
Split B	$\hat{\beta}$	Misclassification error	Gini Index
Node 1	2/3	1/3	$2\frac{2}{3}(1-\frac{2}{3}) = 4/9$
Node 2	0	0	0
Total		$\frac{600}{800} \cdot 1/3 + \frac{200}{800} \cdot 0 = 1/4$	$\frac{600}{800} \cdot 4/9 + \frac{200}{800} \cdot 0 = 1/3$

Example: Leukemia Prediction



Leukemia Dataset: Expression values of 3541 genes for 47 patients with Leukemia ALL subtype (left) and 25 patients with AML (right).

Compare two potential splits (red and blue) on gene 963.



	$X \geq 30$	$X < 30$	
Y=0	12	13	25
Y=1	9	16	25
$\hat{\beta}$	0.42	0.55	

	$X \geq 40$	$X < 40$	
Y=0	7	18	25
Y=1	4	21	25
$\hat{\beta}$	0.36	0.53	

	$X \geq 30$	$X < 30$
Y=0	12	13
Y=1	9	16
$\hat{\beta}$	0.42	0.55

Misclassification error:

$$\frac{12+9}{50} \cdot 0.42 + \frac{13+16}{50} \cdot (1 - 0.55) = 0.4374$$

Gini Index:

$$2\frac{12+9}{50} \cdot 0.42(1-0.42) + 2\frac{13+16}{50} \cdot 0.55(1-0.55) = 0.4917$$

	$X \geq 40$	$X < 40$
Y=0	7	18
Y=1	4	21
$\hat{\beta}$	0.36	0.53

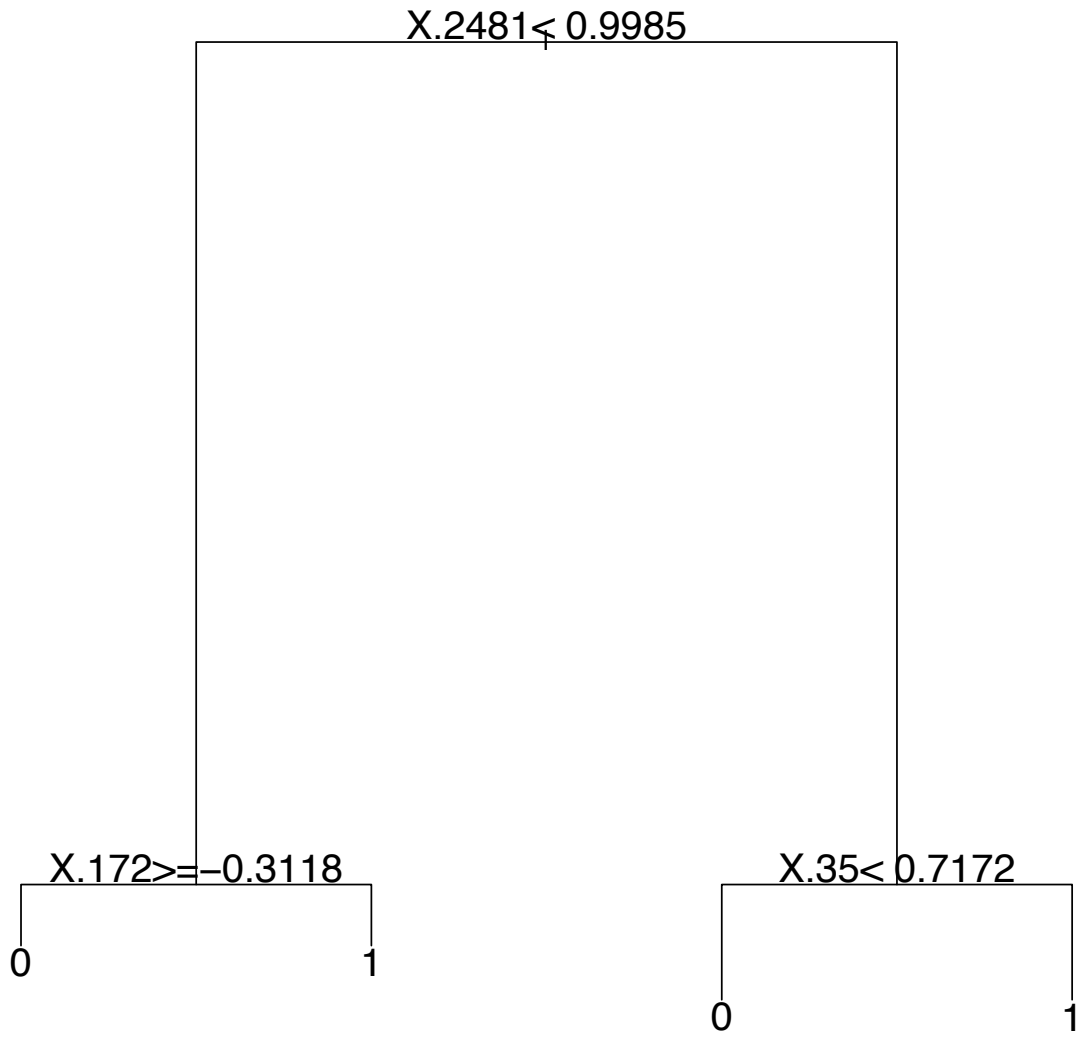
Misclassification error:

$$\frac{7+4}{50} \cdot 0.36 + \frac{18+21}{50}(1 - 0.53) = 0.445$$

Gini Index:

$$2\frac{7+4}{50} \cdot 0.36(1-0.36) + 2\frac{18+21}{50}0.53(1-0.53) = 0.4899$$

Final tree is of depth 2. This tree is very interpretable as it selects 3 out of 4088 genes and bases prediction only on these (as opposed to LDA, QDA, LR and k-NN which perform no variable selection).



Extension to multi-class problems

Let $Y \in \{1, \ldots, K\}$. The empirical probability of class m in a region $\mathcal{R}_r$ is

$$\hat{p}_{m,r} = \frac{1}{N_r} \sum_{i: X_i \in \mathcal{R}_r} \mathbf{1}\{Y_i = m\},$$

where $N_r = \sum_{i: X_i \in \mathcal{R}_r} 1$ are the number of samples in region $\mathcal{R}_r$. Let m_r^* be the class with the highest probability

$$m_r^* = \text{argmax}_m \, \hat{p}_{m,r}.$$

The measures of node impurity generalize then as

- Misclassification error: $1 - \hat{p}_{m_r^*,r}$.
- Gini Index: $\sum_{m \neq m'} \hat{p}_{m,r} \hat{p}_{m',r} = \sum_{m=1}^{k} \hat{p}_{m,r}(1 - \hat{p}_{m,r})$.
- Cross-entropy: $-\sum_{m=1}^{k} \hat{p}_{m,r} \log \hat{p}_{m,r}$.