DISCRIMINANT ANALYSIS

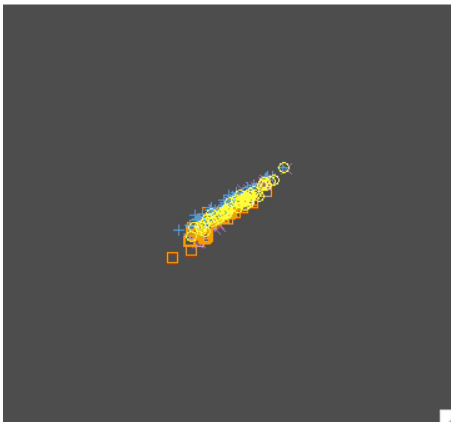
Statistic 407, ISU

WHAT IS?

- Supervised classification, alternatively called discriminant analysis, includes multivariate techniques finding a rule for separating observations/cases into known classes, and using this rule to classify new observations.
- The process starts with a training sample, that is the full data set with known classes. Typically the variables that will be used to generate the classification rule are easy/cheap to measure, but the class is more difficult to measure. It is important to be able to classify new observations using variables that are easy to measure.

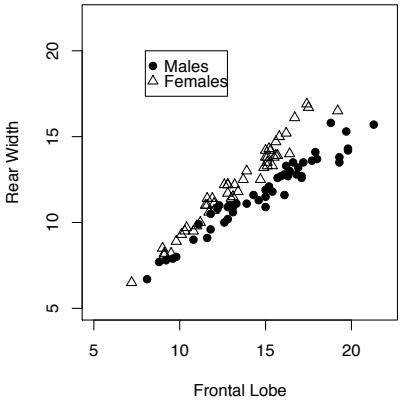
VISUAL METHODS FOR DISCRIMINATION

Use color/glyph(symbol) to code the class/group information in the plots. Then use the full range of plotting methods described in the section of graphics. Look for separations of the points into the color/glyph grouping. Determine what variables are potentially good separators.



EXAMPLE: AUSTRALIAN CRABS

This data is from a study of australian crabs. There are 5 physical measurements recorded on 2 species (blue and orange) and both sexes of each species, giving 4 groups. This is a scatterplot of the blue species with the two sexes identified.



Where would you draw the boundary for this data?

LINEAR DISCRIMINANT ANALYSIS

- LDA is based on the assumption that the data comes from a multivariate normal distribution with equal variance-covariance matrices. Comparing the density functions reduces the rule to:

2 groups, p variables

Allocate a new observation, $\mathbf{x}_0$ to group 1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{pooled}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}_{pooled}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq 0$$

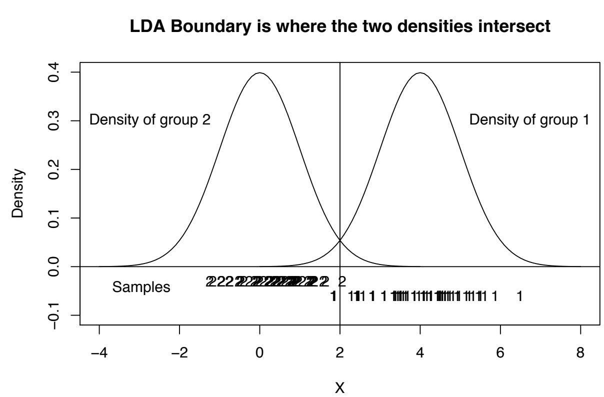
else allocate to group 2.

Average of the means of the groups

Projection, linear coefficients space / reduces the dimension to 1D

Rule: if the linear discriminant of the new observation is closest to the mean for group 1 assign it to group 1.

LDA RULE FOR P=1, G=2



The LDA rule results from assuming that data for each class comes from a MVN with different means but the same variance-covariance matrix.

The boundary between the two groups is half-way between the two means.

EXAMPLE

$\bar{\mathbf{x}}_{Male} = (14.8 \quad 11.7)', \quad \bar{\mathbf{x}}_{Fem} = (13.3 \quad 12.1)' \qquad n_{Male} = 50, \quad n_{Fem} = 50$

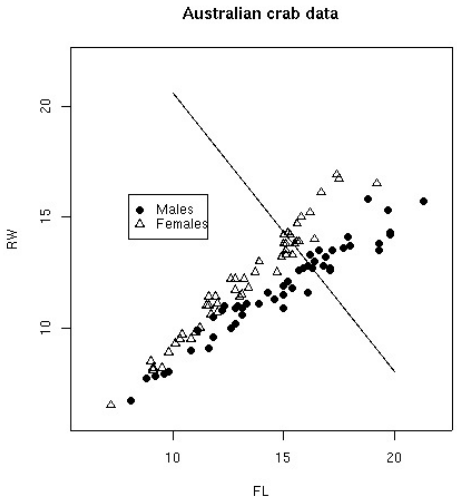
$$\mathbf{S}_{Male} = \begin{bmatrix} 10.3 & 6.5 \\ 6.5 & 4.5 \end{bmatrix} \qquad \mathbf{S}_{Fem} = \begin{bmatrix} 6.9 & 6.3 \\ 6.3 & 5.9 \end{bmatrix}$$

$$\begin{aligned} \mathbf{S}_{pooled} &= \frac{(n_1-1)\mathbf{S}_1}{(n_1-1)+(n_2-1)} + \frac{(n_2-1)\mathbf{S}_2}{(n_1-1)+(n_2-1)} \\ &= \begin{bmatrix} 8.6 & 6.4 \\ 6.4 & 5.2 \end{bmatrix} \end{aligned} \qquad \mathbf{S}_{pooled}^{-1} = \begin{bmatrix} 1.47 & -1.81 \\ -1.81 & 2.42 \end{bmatrix}$$

$$\begin{aligned} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1} &= \begin{bmatrix} 1.5 & -0.4 \end{bmatrix} \begin{bmatrix} 1.47 & -1.81 \\ -1.81 & 2.42 \end{bmatrix} \\ &= \begin{bmatrix} 3.01 \\ -3.86 \end{bmatrix} \end{aligned}$$

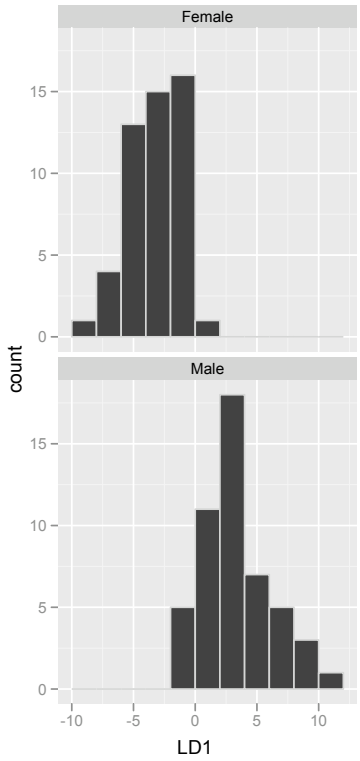
This forms the coordinates of a vector giving the direction of maximum separation.

EXAMPLE



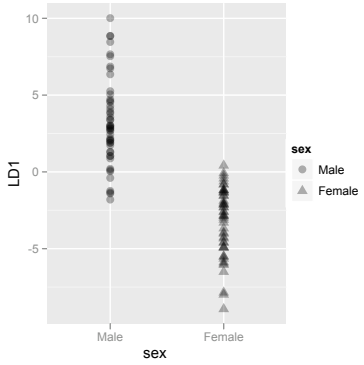
- Direction of maximum separation:

$$\mathbf{x}_2 = \frac{-3.86}{3.01}\mathbf{x}_1 = -1.26\mathbf{x}_1$$



EXAMPLE

- Data projected into the discriminant space.
- Boundary between groups is at 0.



EXAMPLE

The resulting rule is:

Classify the new observation, $\mathbf{X}_0$ as Male if

$$[3.01 \quad -3.86]\mathbf{X}_0 + 2.93 \geq 0$$

else allocate as Female.

Suppose a new crab has values FL=8.1, RW=6.7, what is it's predicted class?

INCORPORATING PRIORS

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_{pooled}^{-1} \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' S_{pooled}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left(\frac{p_2}{p_1}\right)$$

- Where p_1 p_2 are the prior probabilities for group 1 and group 2.
- It shifts the boundary away from the group with the highest prior.

MISCLASSIFICATION TABLE

Predict the class of the training sample. Tabulate against the true class.

		Predicted membership	
		Group 1	Group 2
Actual membership	Group 1	n_{1C}	$n_{1M} = n_1 - n_{1C}$
	Group 2	$n_{2M} = n_2 - n_{2C}$	n_{2C}

The apparent error rate is

$$\frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

The method is biased because it is the best model for this sample, but may under-estimate error with future data.

EXAMPLE

	Male	Female
Male	45	5
Female	1	49

APR=6/100=0.06

DISCRIMINANT FUNCTIONS

The LDA rule can be divided into parts:

$$c_j = \bar{\mathbf{x}}_j' \mathbf{S}_{pooled}^{-1} \mathbf{x}_0 - \frac{1}{2} \bar{\mathbf{x}}_j' \mathbf{S}_{pooled}^{-1} \bar{\mathbf{x}}_j + \ln(p_j) \qquad j = 1, 2; i \neq j$$

And the rule it to allocate the new observation to the group with the largest value of the discriminant function.

CLOSEST MEAN?

The LDA rule corresponds to allocating a new observation to the group that has the smallest squared Mahalanobis distance between the new observation and the group mean.

$$d_j = \frac{1}{2}(\mathbf{x}_0 - \bar{\mathbf{x}}_j)'\mathbf{S}_{pooled}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_j) - \ln(p_j) \qquad j = 1, 2; i \neq j$$

MORE THAN 2 GROUPS

There are now g groups, and the rule is the same, allocate to the group with the largest value of the discriminant function

$$c_j = \bar{\mathbf{x}}_j'\mathbf{S}_{pooled}^{-1}\mathbf{x}_0 - \frac{1}{2}\bar{\mathbf{x}}_j'\mathbf{S}_{pooled}^{-1}\bar{\mathbf{x}}_j + \ln(p_j) \qquad j = 1, ..., g; i \neq j$$

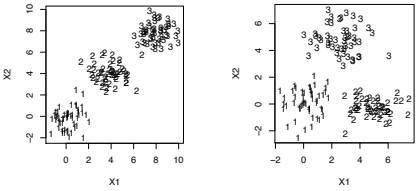
CANONICAL COORDINATES

The low-dimensional space which best separates the groups is given by the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ where

$$\mathbf{B} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})', \quad \mathbf{W} = \sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i$$

g is the number of groups, and $\bar{\mathbf{x}}$ is the overall mean.
At most (g-1) dimensions are needed.

Eg, g=3,
1 or 2 dim
needed



QUADRATIC DISCRIMINANT ANALYSIS

Suppose that the variance-covariances are not the same for each group, then the rule becomes:

Allocate a new observation, $\mathbf{x}_0$ to group 1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - \frac{1}{2}(\ln\left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\right) + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1}\bar{\mathbf{x}}_2)) \geq \ln\left(\frac{p_2}{p_1}\right)$$

else allocate to group 2.

DISCRIMINANT FUNCTIONS

Allocate the new observation to the group with the largest value of the discriminant function:

$$c_j = -\frac{1}{2}\mathbf{x}_0'\mathbf{S}_j^{-1}\mathbf{x}_0 + \bar{\mathbf{x}}_j\mathbf{S}_j^{-1}\mathbf{x}_0 - \frac{1}{2}\ln(|\mathbf{S}_j|) + \bar{\mathbf{x}}_j'\mathbf{S}_j^{-1}\bar{\mathbf{x}}_j + \ln(p_j) \quad j = 1, 2; i \neq j$$

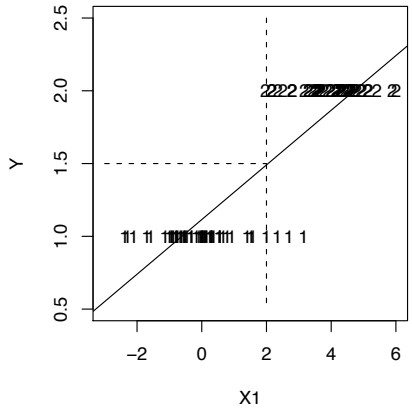
RELATIONSHIP BETWEEN LDA AND REGRESSION

- A matrix of variables is used to predict a categorical response:

$$\mathbf{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}_{n\times p} \qquad \mathbf{Y} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 2 \\ \vdots \\ 2 \end{bmatrix}$$

LINEAR REGRESSION

$$\hat{\mathbf{Y}} = b_0 + b_1\mathbf{X}_1 + \ldots + b_p\mathbf{X}_p$$



Problems: Predictions outside range of 1-2

CLASSIFICATION TREES

The tree algorithm generates classification rules by sequentially doing binary splits on the data. Splits are made on individual variables. On each variable the values are sorted, and splits between each pair of values are examined for quality of the split using a criterion function. Of the cases to the left of the split, the criterion compares the purity, the proportion which are in each class, and similarly for cases to the right of the split. A common criterion is entropy, which for two classes, would be computed as:

$$-\hat{p}_0 \log \hat{p}_0 - \hat{p}_1 \log \hat{p}_1$$

where $\hat{p}_0 = \frac{N_0}{N}, \hat{p}_1 = \frac{N_1}{N} = 1 - \hat{p}_0$ are the relative proportions of cases in classes 0,1.

This is lowest if either N_0 or $\dot{N}_1$ is 0. A good split has pure groups to each side (bucket), all class 0 on the left and all class 1 to the right. To measure the quality of a split we need to measure the impurity in each bucket:

$$\hat{p}^L(-\hat{p}_0^L \log \hat{p}_0^L - \hat{p}_1^L \log \hat{p}_1^L) + \hat{p}^R(-\hat{p}_0^R \log \hat{p}_0^R - \hat{p}_1^R \log \hat{p}_1^R)$$

where p^L, p^R is the proportion of cases in the left, right buckets, respectively. This is a weighted average of the impurity, as measured by entropy, in each bucket.

ALGORITHM

1. For each variable, and for each possible split calculate the the impurity measure.
2. Pick the split with the smallest impurity, subset the data into two using this split. Each split is called a node on the resulting tree.
3. On each subset, repeat step 1-2.
4. Splitting a node is controlled by number of cases in the subset at that node, and also the amount of impurity the node. Stop splitting when either of these gets below a tolerance.

EXAMPLE: OLIVE OILS 3 REGIONS, ALL VARIABLES

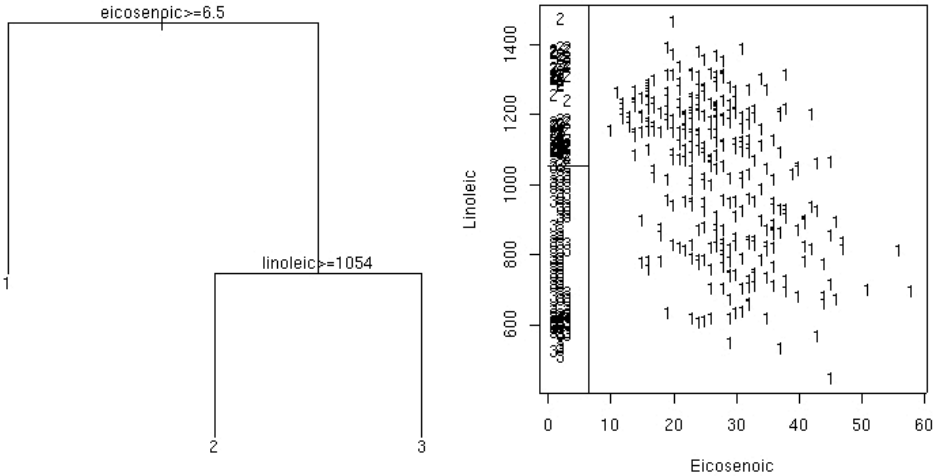
```
> library(rpart)
> olive.rp<-rpart(d.olive[,1]~.,d.olive[,-c(1,2)],
+   method="class", parms=list(split='information'))
> olive.rp
n= 572

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 572 249 1 (0.5646853 0.1713287 0.2639860)
  2) eicosenoic>=6.5 323   0 1 (1.0000000 0.0000000 0.0000000) *
  3) eicosenoic< 6.5 249  98 3 (0.0000000 0.3935743 0.6064257)
    6) linoleic>=1053.5 98   0 2 (0.0000000 1.0000000 0.0000000) *
    7) linoleic< 1053.5 151   0 3 (0.0000000 0.0000000 1.0000000) *
```

node) is the arbitrary numbering of nodes from top to bottom of the tree
split is the rule for the split from that node
n is the number of cases at this node
loss is the number of cases misclassified at this node
yval is the predicted value for all cases at this node
(yprob) are the proportions in each class

EXAMPLE: OLIVE OILS 3 REGIONS, ALL VARIABLES



The first split is on eicosenoic acid and the next split is on linoleic acid.

It only uses these two variables!

And there is no error!

A CLOSER LOOK.....

Consider the data $x = (1, 2, 3, 4, 5, 6, 7, 8)$ and

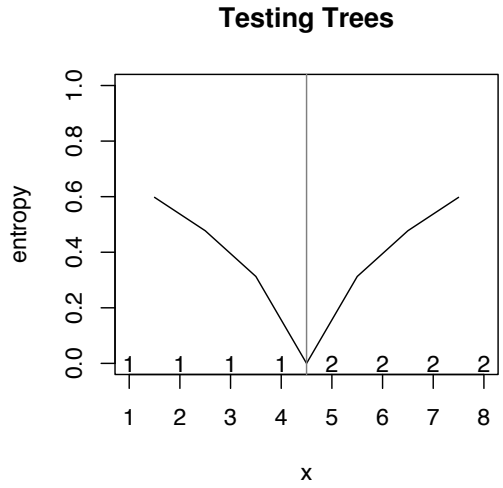
$class = (1, 1, 1, 1, 2, 2, 2, 2)$

then all possible splits would be

Left	Right
(1,0)	(3,4)
(2,0)	(2,4)
(3,0)	(1,4)
(4,0)	(0,4)
(4,1)	(0,3)
(4,2)	(0,2)
(4,3)	(0,1)

Calculate the impurity (on slide 2) for each possible split

lowest value is between points 4 and 5. That's the split to use.

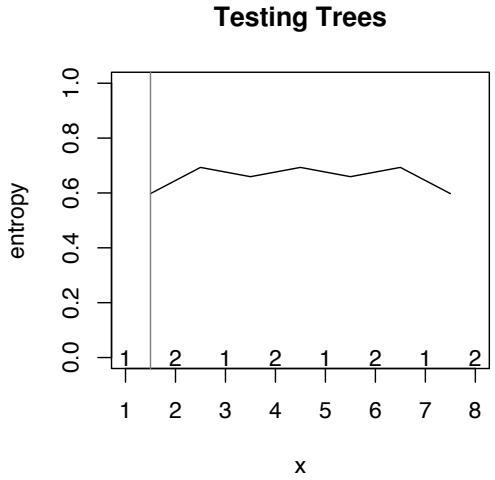


How does it work for a nonsensical class structure?
Consider data:

$x = (1, 2, 3, 4, 5, 6, 7, 8)$

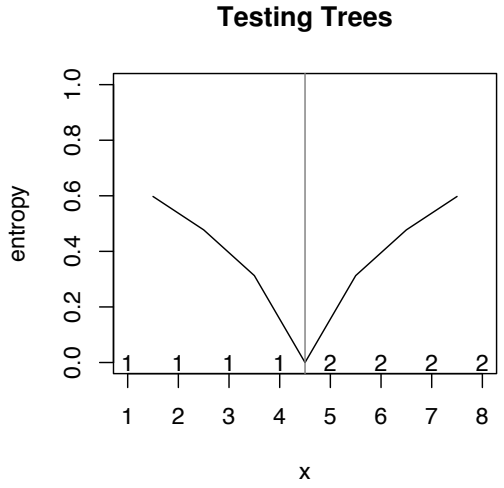
$class = (1, 2, 1, 2, 1, 2, 1, 2)$

The split chosen will most likely be the first one, between points 1 and 2.



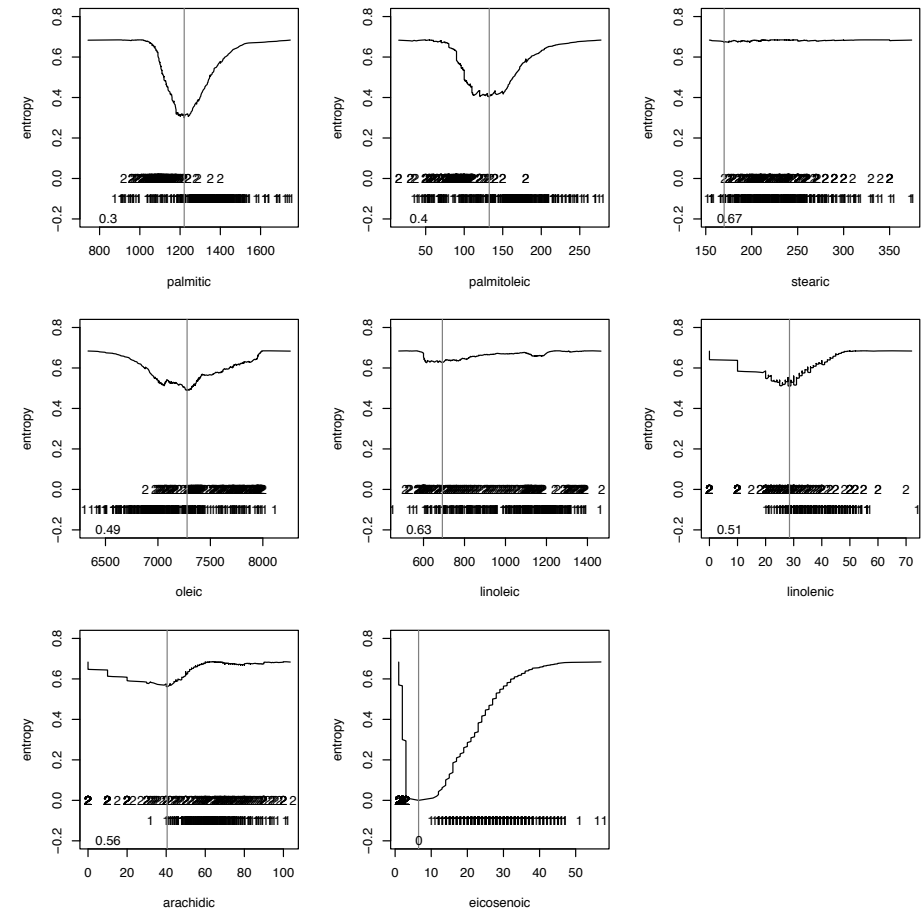
Left	Right
(1,0)	(3,4)
(2,0)	(2,4)
(3,0)	(1,4)
(4,0)	(0,4)
(4,1)	(0,3)
(4,2)	(0,2)
(4,3)	(0,1)

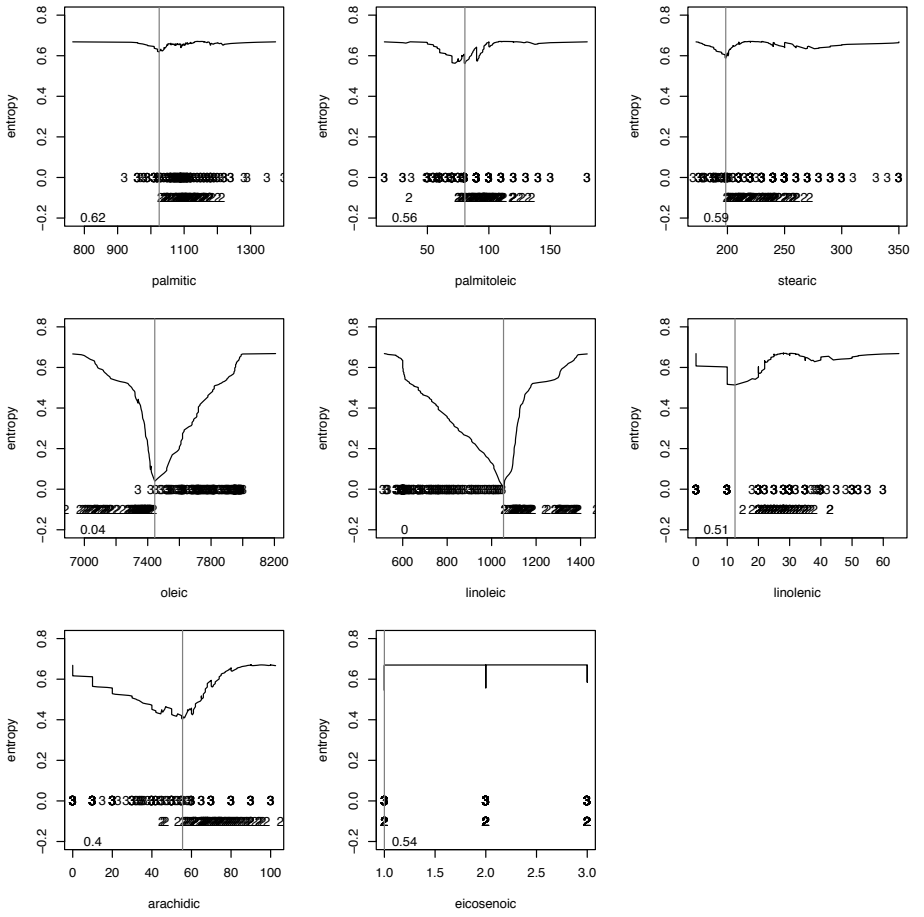
$$\hat{p}^L(-\hat{p}_0^L \log \hat{p}_0^L - \hat{p}_1^L \log \hat{p}_1^L) + \hat{p}^R(-\hat{p}_0^R \log \hat{p}_0^R - \hat{p}_1^R \log \hat{p}_1^R)$$



HOW DOES IT WORK ON THE OLIVE OILS DATA?

- In practice the impurity functions can be quite noisy.
- The next two sets of plots show the impurity measure calculated to separate the (1) southern oils from the other two regions, and (2) northern from Sardinian oils.
- Eicosenoic acid is the variable with the lowest impurity overall, 0. It would be chosen as the most important variable at the top of the tree.
- Linoleic acid is the variable with the lowest impurity, 0, when region 1 is removed. It would be chosen as the second split variable.


entropy
palmitic
0.3
palmitoleic
0.4
stearic
0.67
oleic
0.49
linoleic
0.63
linolenic
0.51
arachidic
0.56
eicosenoic
0


entropy
palmitic
0.62
palmitoleic
0.56
stearic
0.59
oleic
0.04
linoleic
0
linolenic
0.51
arachidic
0.4
eicosenoic
0.54

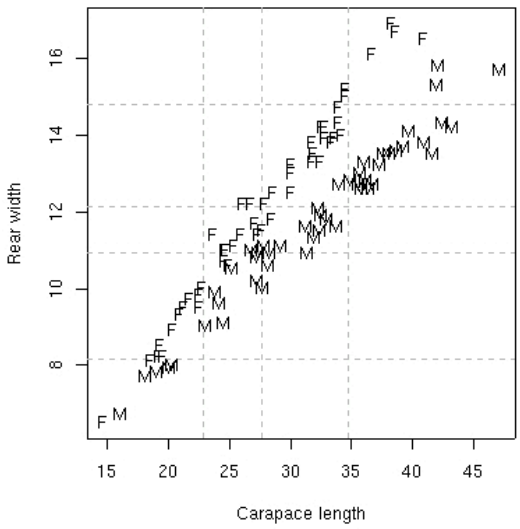
STRENGTHS AND WEAKNESSES

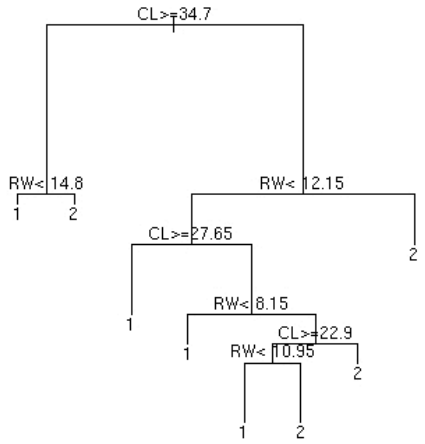
- The solutions are usually _____, and easy to implement. There are few probabilistic assumptions underlying trees, which complicate the solution. For example, because LDA assumed that the variance-covariance of the groups are equal it doesn't see the ``perfect'' split of northern and sardinian oils in linoleic acid.
- The fitting _____ in the sense that the first best fit will be used at each split, but it may be a better final result might be obtained by a less optimal previous step.

STRENGTHS AND WEAKNESSES

- The additive model approach, _____, is too limited for problems where separations between groups is due to combinations of variables. But because it works variable-by-variable it can _____, using complete data on each variable. Trees can also accommodate complex data, where some variables are continuous and some are categorical.
- Because it is an algorithmic method it can be easy to _____ (_____) the data. The tree will then not have inferential power: it will have worse error on new data. Split the current data into training and test sets, use the training subset to build the tree, and the test set to estimate the error.

TREES DON'T DO SO WELL IN THE PRESENCE OF COVARIANCE BETWEEN VARIABLES





OTHER COMMON CLASSIFICATION METHODS

- _____ - fit many trees to samples of the data, and subsets of the variables, and combine the predictions.
- _____ - a mixture of logistic regression models.
- _____ - find "gaps" between groups and fit a hyperplane to the points bordering the gaps.

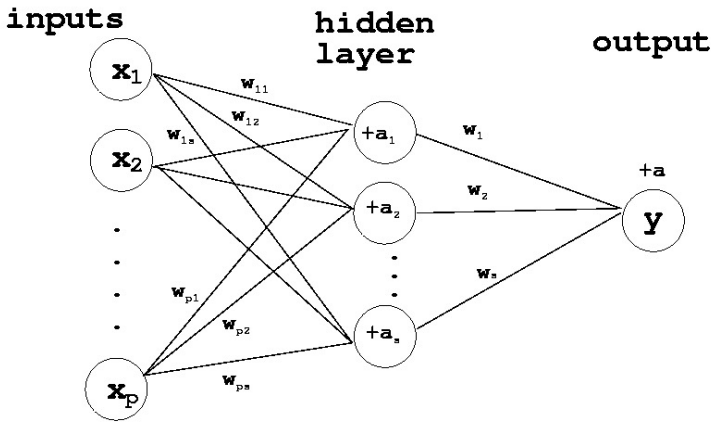
NEURAL NETWORK

Feed-forward neural networks (FFNN) were developed from this concept, that combining small components is a way to build a model from predictors to response. They actually generalize _____. A simple network model is represented by:

$$\hat{y} = f(\mathbf{x}) = \phi(\alpha + \sum_{h=1}^{s} w_h \phi(\alpha_h + \sum_{i=1}^{p} w_{ih} x_i))$$

where x is the vector of explanatory variable values, y is the target value, p is the number of variables, s is the number of nodes in the single hidden layer and phi is a fixed function, usually a linear or logistic function. This model has a single hidden layer, and univariate output values.

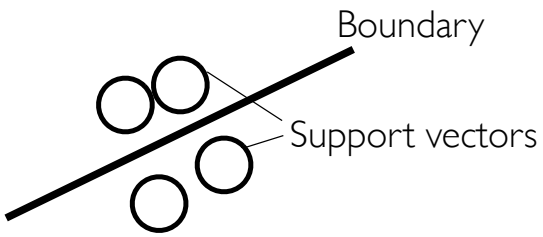
$$\hat{y} = f(\mathbf{x}) = \phi(\alpha + \sum_{h=1}^{s} w_h \phi(\alpha_h + \sum_{i=1}^{p} w_{ih} x_i))$$



The network is fit by minimizing a squared error

$$\sum_{i=1}^{n} (y_i - f(\mathbf{x}))^2$$

SUPPORT VECTOR MACHINES



- The algorithm finds a hyperplane that maximizes the _____ (gap) between the two classes.

$\mathbf{w}.\mathbf{x} + b = 0$

- The points on the edge of the margin are called _____, and are used to define the

$\mathbf{w} = \sum_{i=1}^{N_S} \alpha_i y_i \mathbf{x}_i$

N_S is the number of support vectors