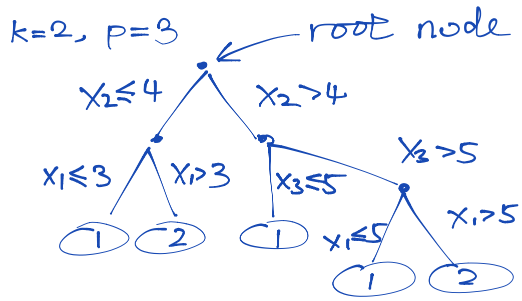
Classification Tree

Want to model $\theta_j(\underset{\sim}{x}) = P(G=j \mid \underset{\sim}{X} = \underset{\sim}{x})$ for $j=1,\cdots,k$



Question: How to construct (grow) tree?

Recursive partitioning

At a given node, define

$$B = \{(g_i, \underset{\sim}{x}_i) : i \in I(B)\}$$

$g_i \in 1,\cdots,k$

For $\underset{\sim}{x}_i \in B$, we have

$$\hat{\theta}_j(B) = \frac{1}{n(B)} \underbrace{\sum_{i\in I(B)} I(g_i = j)}_{= n_j(B)} = \frac{n_j(B)}{n(B)}$$

$n(B)$ = # of observations in B

= proportions of group j in node B.

Now define possible new nodes: B_1, B_2 with $B = \underbrace{B_1 \cup B_2}_{\text{disjoint}}$

$$\hat{\theta}_j(B_1) = \frac{1}{n(B_1)} \sum_{i\in I(B_1)} I(g_i=j) = \frac{n_j(B_1)}{n(B_1)} \quad (j=1,\cdots,k)$$

$$\hat{\theta}_j(B_2) = \frac{1}{n(B_2)} \sum_{i\in I(B_2)} I(g_i=j) = \frac{n_j(B_2)}{n(B_2)}$$

Define $D(B_1, B_2 : B) = \sum_{j=1}^{k} \left\{ n_j(B_1) \ln\left(\frac{n_j(B_1)}{n(B_1)}\right) + n_j(B_2) \ln\left(\frac{n_j(B_2)}{n(B_2)}\right) - n_j(B) \ln\left(\frac{n_j(B)}{n(B)}\right) \right\}$

B_1, B_2: variables; B: fixed

- Find $\underbrace{B_1 + B_2}_{\text{many choices}}$ s.t. $B_1 \cup B_2 = B$ to maximize D.

- restrict maximization to simple one variable splits

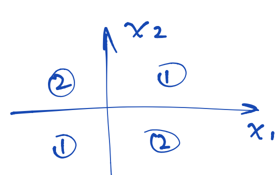
$$B_1 = \{(g_i, \underset{\sim}{x}_i) \in B,\ x_{il} \le d\} \leftarrow \text{threshold}$$
$$B_2 = \{(g_i, \underset{\sim}{x}_i) \in B,\ x_{il} > d\}$$

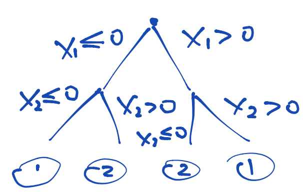
$l = 1,\cdots,p$

- also require other constraints e.g. $n(B_1), n(B_2) \ge 5$.

Does procedure always work?

Example: $k=2$, $p=2$





How does tree algorithm work here?

- $P(G=j \mid X_1) = \frac{1}{2}$ (for $j=1,2$)
  $P(G=j \mid X_2) = \frac{1}{2}$

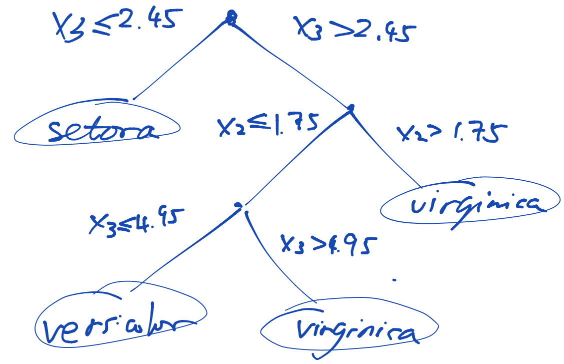
- recursive partitioning algorithm has trouble getting started

– better performance if allow splitting along linear combinations (i.e. projections) of variables.

Example Iris data

3 species $\begin{cases} \text{setosa} \\ \text{virginica} \\ \text{versicolor} \end{cases}$

4 vars $\begin{cases} x_1 = \text{sepal length} \\ x_2 = \text{width} \\ x_3 = \text{petal length} \\ x_4 = \text{width} \end{cases}$



– compare favourably to LDA

error rate = 4/150 (tree)

error rate = 3/150 (LDA)

Regression models for multivariate data

– repeated measures

Multivariate Analysis of Variance (MANOVA)

Problem k treatments (or groups)

n_i subjects in treatment i

Multivariate response $\underset{\sim}{x}_{ij}$ $\begin{cases} i=1,\cdots,k \\ j=1\cdots,n_i \end{cases}$

p vectors

Model: $\underset{\sim}{X}_{ij} = \underset{\sim}{\mu}_i + \underset{\sim}{\varepsilon}_{ij}$ $i=1,\cdots,k$, $j=1,\cdots,n_i$

where $\{\underset{\sim}{\varepsilon}_{ij}\}$ are independent $N_p(\underset{\sim}{0}, C)$ random vectors i.e. $\underset{\sim}{X}_{ij} \sim N_p(\underset{\sim}{\mu}_i, C)$

One question of interest: Is there a difference between k treatments? (For example, is $\underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \cdots = \underset{\sim}{\mu}_k$?)

Univariate ($p=1$) case: Decompose total sum of squares

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

$\bar{x}$: overall sample mean

$\bar{x}_i$: sample mean for treatments

$$SS_{Total} = SS_{\substack{between\\group}} + SS_{\substack{within\\group}}$$

To test $H_0: \mu_1 = \cdots = \mu_k$, we compare $SS_{between}$ to SS_{within}

Test statistic $F = \dfrac{SS_{between}(k-1)}{SS_{within}\ (n-k)} \sim F_{k-1,n-k}$ under H_0

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})^T = \underbrace{\sum_{i=1}^{k} n_i(\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})(\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})^T}_{SS_{between}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}}_i)^T}_{SS_{within}}$$

Question:

How to compare $SS_{between}$ to SS_{within}