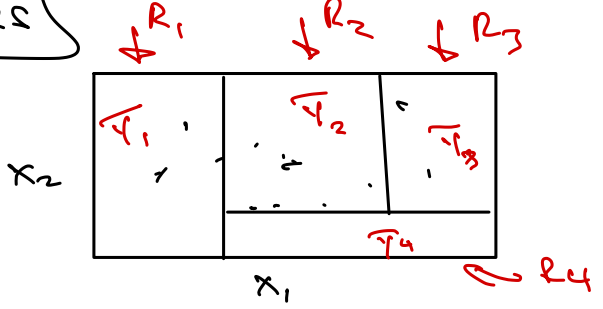
10.3 Classification trees

Regression tree:



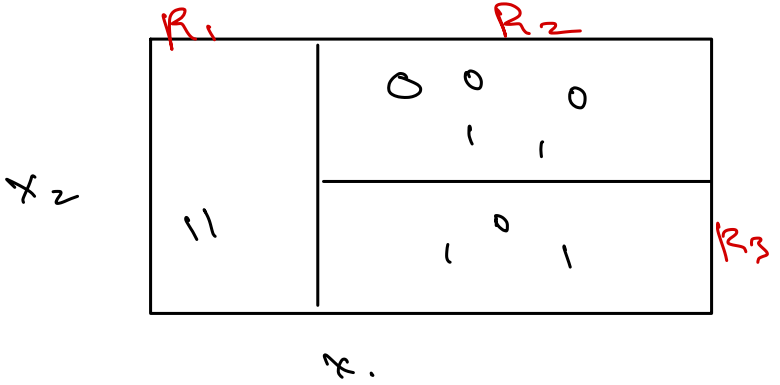
A classification tree refers to the case that y is categorical. Still split feature space into regions & use a single predictor within those regions. The difference is to use the most common class. If region R_m has N_m samples in it, let

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} \mathbb{1}[y_i = k]$$

$k = 1, \ldots, K$ = empirical prob of class k in region m.

We classify new obs in region m into class k when

$$k(m) = \underset{k}{\operatorname{argmax}}\ \hat{P}_{mk}$$



$\hat{P}_{10} = 0$
$\hat{P}_{11} = 1$
→ $R_1 : \hat{y} = 1$

$\hat{P}_{20} = 3/5$
$\hat{P}_{21} = 2/5$
→ $R_2 : \hat{y} = 0$

$\hat{P}_{30} = \frac{1}{3}$
$\hat{P}_{31} = \frac{2}{3}$
→ $R_3 : \hat{y} = 1$

$\hat{y}$ in region

Need a different notion of goodness-of-fit. Two most common:

Gini index

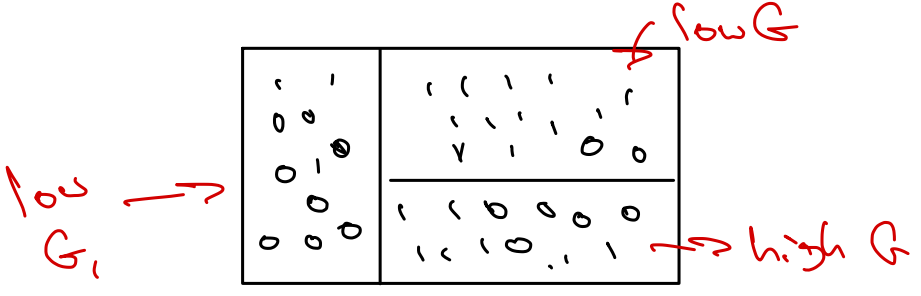
$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) = 1 - \sum_{k=1}^{K} p_{mk}^2$$

mth node (region R_m)

G_m is negatively-oriented

$\Rightarrow$ smaller values are better

$$G = \sum_{m=1}^{M} G_m = \text{total Gini index}$$



Cross-entropy or deviance index

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

m → region m

→ wants $\hat{p}$ close to 0 or 1

Prefer trees / cuts that minimize D_m

$$D = \sum_{m=1}^{M} D_m$$

Can do cost-complexity fit to control size of tree, or bagging. Predicted values use majority vote.

For B trees, use the most commonly predicted class

$B = 100$

$\hat{Y}_1 = 0, \ldots, \hat{Y}_{80} = 0$, $\hat{Y}_{81} = 1, \ldots, \hat{Y}_{100} = 1$ $\Rightarrow \hat{Y} = 0.$

80 $\hat{Y} = 0$ 20 $\hat{Y} = 1$