

$\alpha$  by cross-validation

	$\alpha = \alpha_1$	$\alpha = \alpha_2$	...	$\alpha = \alpha_B$
test $K=1$	$C_{\alpha_1}(J)$			
$K=2$				
...				
$K=K$	$C_{\alpha_1}(J)$			
	$CV(\alpha_1)$	$CV(\alpha_2)$	...	$CV(\alpha_B)$

$\underbrace{\hspace{10em}}_{\text{min } \alpha \text{ which minimize the CV}}$

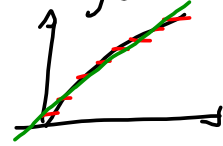
Advantages of trees

- simplicity and ease of communication;
- compact representation;
- speed of computation (easy to parallelize)
- easy to handle both continuous and categorical variables (mixture of them)
- easily implement different loss functions;
- easy to handle missing data;
- automatic variable selection



Disadvantages

- instability of the results  $\rightarrow$  bagging, boosting and random forest
- difficulties with online computation (difficult to integrate new information)
- very hard to approximate very steep functions
- there are no formal statistical procedures (tests of hypotheses, confidence intervals, ...)
- difficult to evaluate variable importance



# Classification trees

Start  $p=1, k=2$

Instead of estimating  $f(x)$ , we try to estimate  $\hat{\pi}_1(x) = \Pr[Y=1|X=x]$

$$\hat{\pi}_1(x) = \sum_{h=1}^J p_h \mathbb{1}(x \in R_h)$$

where  $p_h$  is the probability that  $Y=1$  given  $x \in R_h$ . We allocate each observation to 0 or 1 depending  $\hat{\pi}_1(x)$ , if  $\hat{\pi}_1(x) > t$ ,  $t$  being a threshold (usually  $1/2$ )

To estimate  $p_h$ , we can simply use the frequency of 1 in  $R_h$

$$\hat{p}_h = \frac{1}{n_j} \sum_{i: x_i \in R_h} \mathbb{1}(y_i = 1)$$

Given the binary nature of  $Y$ , we need to use a specific version of the deviance (binomial deviance, see notes of lecture 5) as the loss function

$$D = -2 \sum_{i=1}^n y_i \log \hat{\pi}_1 + (1 - y_i) \log (1 - \hat{\pi}_1)$$

which can be rewritten as

$$D = -2 \sum_{h=1}^J n_h \left( \hat{p}_h \log \hat{p}_h + (1 - \hat{p}_h) \log (1 - \hat{p}_h) \right) = \sum_{h=1}^J D_h$$

because the probability of  $Y=1$  is constantly equal to  $p_h$  in each region  $R_h$

Consider  $D$  in the form

$$D = 2n \sum_{h=1}^J \frac{n_h}{n} Q(\hat{p}_h)$$

weight proportional to the size of leaves  $\rightarrow$  entropy  $\rightarrow$  measure of impurity  
 how much the elements of a region are non-homogeneous  
 $\rightarrow Q=0$  if all observations are 0 or 1  
 $\rightarrow Q=1/2$  in the case of maximum entropy (50%/50%)

where

$$Q(p_h) = - \sum_k p_{jk} \log p_{jk} = - (p_{j1} \log p_{j1} + p_{j0} \log p_{j0}) \quad k=2$$

$$= - (p_{j1} \log p_{j1} + (1 - p_{j1}) \log (1 - p_{j1}))$$

We can then use different measures of impurity, e.g.

Gini index  $Q(p_h) = \sum_{k \in \{0,1\}} p_{jk} (1 - p_{jk})$

Missclassification error  $\sum_{h=1}^J \frac{1}{n_h} \sum_{i: x_i \in R_h} \mathbb{1}(y_i \neq k_h)$

