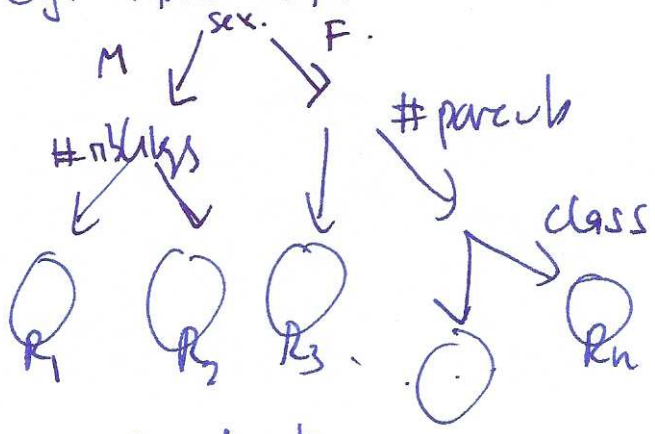
Basic decision trees.

e.g. Titanic data.



← problems: which tree? (exponentially many in data).
how do you tell if this is any good/or if there is a better tree?

in each chunk
vote for classification data.
take mean for continuous data.

tree splits data into chunks $R_1, \ldots, R_J$.

measure of error is $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$.

$\hat{y}_{R_j}$: prediction — vote or mean

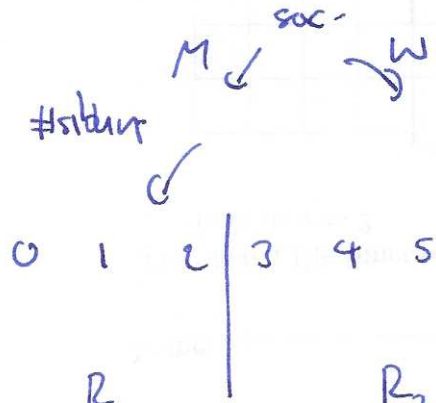
top down/greedy split -

— sometimes no choice in split: e.g. sex → M / F.

— if there is a choice: can choose at random.

better: choose best linear split which minimises RSS.

e.g.



compute RSS for this split

$$\sum_{x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{x_i \in R_2} (y_i - \hat{y}_{R_2})^2.$$

↑ choose split which minimises this

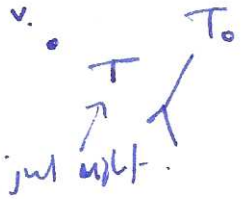
Pruning if there are n dim get many regions – extra regions may not be that useful.

pick a maximal tree T_0 and a parameter $\alpha \geq 0$.

- for each value of $\alpha \geq 0$ there is a subtree $T \leq T_0$ which minimizes: $\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$

if $\alpha = 0$ get $T = T_0$.

as $\alpha \to \infty$. end up with just single vertex.



Fact. if $\alpha_1 < \alpha_2$ then $T_{\alpha_1} \geq T_{\alpha_2}$

---

other classification methods. basic: majority vote.

error: $\sum_i (y_i - \hat{y})^2 \leftarrow$ # errors.

— spose you have more than just $0, 1$ classification. i.e. output is $\{1, 2, 3, \ldots K\}$.

common measures: • Gini coefficient: $G_m = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk})$

$\hat{p}_{mk}$ = proportion of observations in the mth region from the k-th class.

region ↑ ↑ class

G small if all $\hat{p}_{mk}$'s close to 0 or 1.

• entropy $D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \geq 0$.

entropy D close to 0 if all $\hat{p}_{mk}$'s close to 0 or 1.

advantage of better measures.

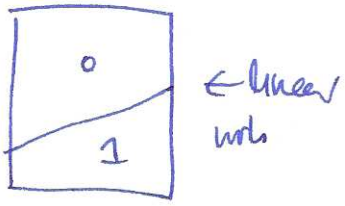
0	R_1	R_2
	10	10
	20	20

better split

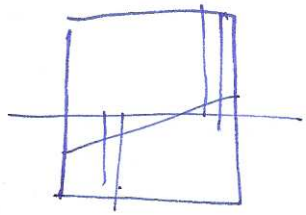
	R_1	R_2
0	20	
1		40.

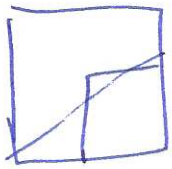
advantages: a simple tree is easy to explain!

tree vs linear regression:

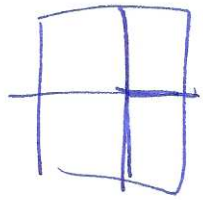


← linear works





linear doesn't work



split both ways!

<u>disadvantages</u>: trees can have have high variance

↝ e.g. small change in data might give very different tree.