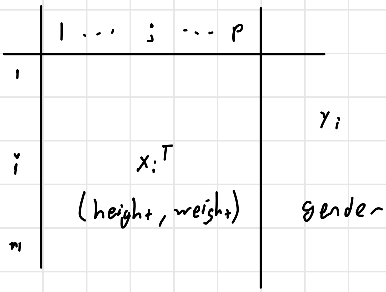
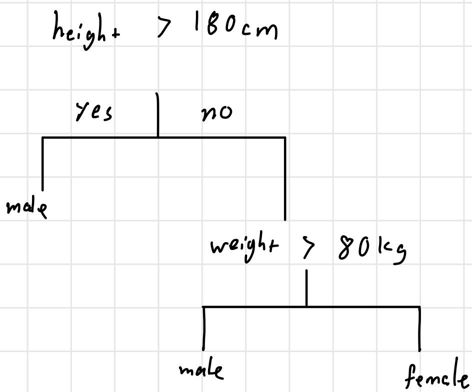
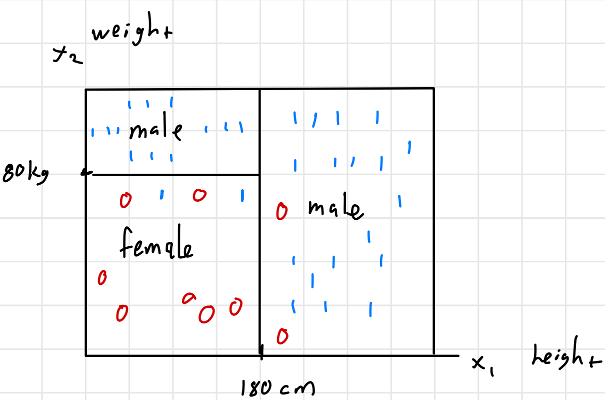
Lecture 7

- Part 2: Trees, Forest, Boosting

- Tree: CART classification and Regression Trees
  - classification / Decision







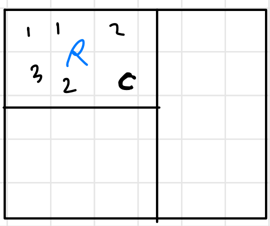
Recursive Partition

- Each iteration : $(m \times p \times n)$

  - Choose a region $m \in \{1, \dots, M\}$:
    choose a variable $j \in \{1, \dots, p\}$;
    choose a cut-off $t \in \{x_{ij},\ i = 1, \dots, n\}$

    max reduction of Loss function

- Principled way of choosing a loss is to use log-likelihood .
- But there are intuitive ideas for the loss function $\to$ purity



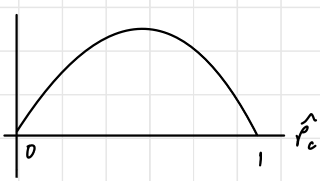
Within a region R

	1	2	...	c		C
counts	n_1	n_2	...	n_c	...	n_C
props	p_1	p_2	...	p_c	...	p_C
estimates	$\hat{p}_1$	$\hat{p}_2$	...	$\hat{p}_c$	...	$\hat{p}_C$

Note : $\hat{p}_c = \frac{n_c}{n}$

Let $c^* = \arg\max_c \hat{p}_c$

purity $= 1 - \hat{p}_{c^*}$ (smaller $\Rightarrow$ more pure)



Gini-index $= \sum_{c=1}^{C} \hat{p}_c (1 - \hat{p}_c)$

how close $\hat{p} = (\hat{p}_1, \dots, \hat{p}_c, \dots, \hat{p}_C)$ to one-hat (assume observations are in $\mathbb{R}$ for simplicity)

- $$\text{Log-likelihood}(p) = \sum_{i=1}^{n} \log p(y_i)$$
$$= \sum_{c=1}^{C} n_c \log(p_c)$$
$$= n \sum_{c=1}^{C} \hat{p}_c \log(p_c)$$
$$= -n\, H(\hat{p}, p) \quad , \quad H: \text{cross-entropy}$$

$$\max_{p} \text{ log-like}(p) \longrightarrow \hat{p}_{MLE} = \hat{p}$$

&

$$\text{log-like}(\hat{p}) = n \sum_{c=1}^{C} \hat{p}_c \log \hat{p}_c$$
$$= -n\, H(\hat{p}) \quad , \quad H: \text{entropy}$$

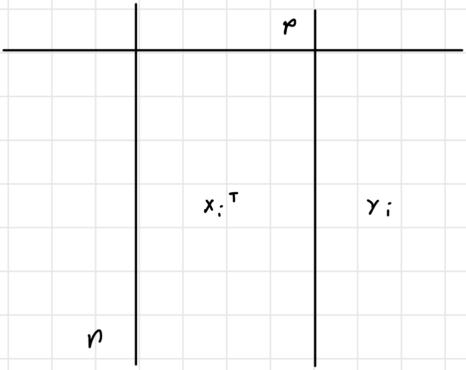
- $$\text{log-like}(\hat{p}) - \text{log-like}(p) = n \sum_{c=1}^{C} \hat{p}_i \log \frac{\hat{p}_c}{p_c}$$
$$= n\, D_{KL}(\hat{p} \,\|\, p) \geq 0$$

- So we can define purity $\equiv n H(\hat{p})$

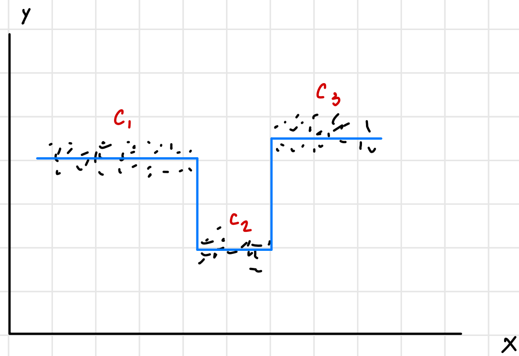
- Also the Gini-index needs n as a factor:

$$\underbrace{\text{Gini-index} = n \sum \hat{p}_c (1-\hat{p}_c)}_{\text{A surrogate for entropy}}$$

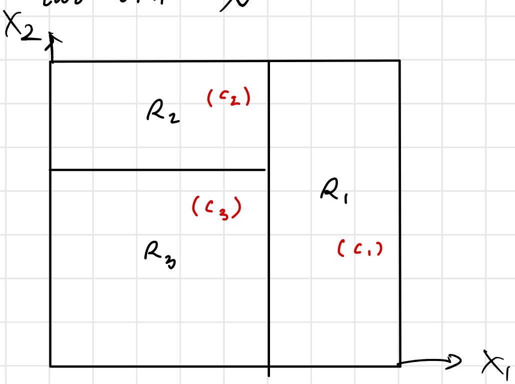
Regression



one-dim X



two-dim X



Piecewise constant function

$$s = f(x) = \sum_{m=1}^{M} c_m \mathbb{1}(x \in R_m)$$

$$\mathbb{1}(x \in R) = \begin{cases} 1 & \text{if } x \in R \\ 0 & \text{if } x \notin R \end{cases}$$

- Least - squares loss : guide recursive partitioning

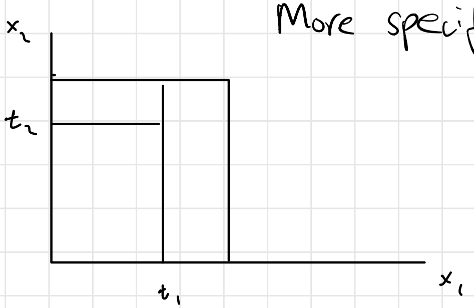
$$\text{Loss} = \sum_{i=1}^{n} (y_i - s_i)^2$$

$$= \sum_{m=1}^{M} \sum_{i: x_i \in R_m} (y_i - c_m)^2$$

Note : $\hat{c}_m = \dfrac{\sum_{i: x_i \in R_m} y_i}{n_m}$, n_m = # of example in R_m

$$\sum_{i: x_i \in R_m} (y_i - \hat{c}_m)^2 = n_m \cdot \text{variance of } y_i \text{ in } R_m, \text{ purity}$$

More specific form

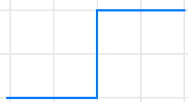


$$s = f(x) = c_1 \mathbb{1}(x_1 \le t_1) + c_2 \mathbb{1}(x_1 > t_1)$$ (first cut)

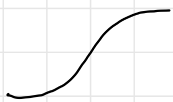
$$s = f(x) = c_{11} \mathbb{1}(x_1 \le t) \mathbb{1}(x_2 \le t_2) + c_{12} \mathbb{1}(x_1 \le t_1) \mathbb{1}(x_2 > t_2) + c_2 \mathbb{1}(x_1 > t_1)$$ (second cut)

Notice:

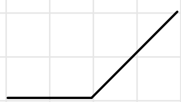
$\mathbb{1}(x_j > t)$



Sigmoid



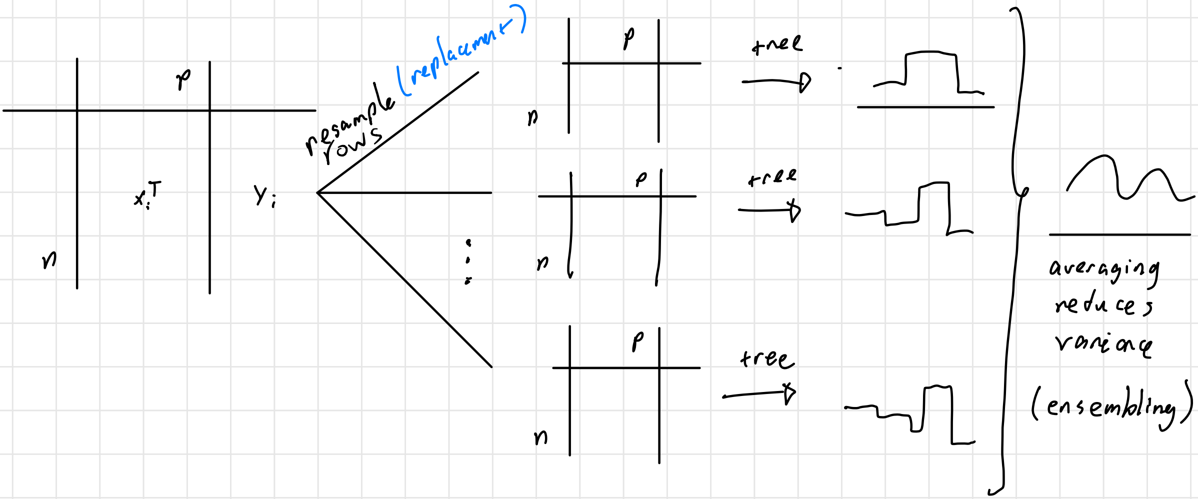
ReLU



- Multivariate adaptive regression spline (MARS)

- Trees are unstable (Add an observation changes the tree)
  High variance

- Forest :



  - each iteration
    choose a variable to split
    $\in$ {subset of $\sqrt{p}$ variables}

- Boosting :

  - Regression Tree :

$$s = f(x) = \sum_{k=1}^{K} \underbrace{h_k(x)}_{\text{a tree}} \quad \text{ensemble / comitee}$$

  - Sequentially adding trees:

$$s = \underbrace{\sum_{k=1}^{t-1} h_k(x)}_{\hat{s}} + \underbrace{h_t(x)}_{\Delta s}$$

current committee of trees, fixed — grow a new tree

- Regression, ℓ_2 boosting

  - Loss $= \sum_{i=1}^{n} (y_i - s_i)^2$

$$= \sum_{i=1}^{n} (y_i - (\hat{s}_i + \Delta s))^2$$

$$= \sum_{i=1}^{n} (\underbrace{y_i - \hat{s}_i}_{\hat{e}_i} - h_t(x))^2$$

grow a new tree for $\hat{e}_i$

→ Classification, extreme gradient boosting, logistic regression

	p	
n	x_i^T	$y_i \in \{0,1\}$

$$\text{Log-like} = \sum_{i=1}^{n} \underbrace{\left(y_i s_i - \log(1+e^{s_i}) \right)}_{\ell(s_i)}$$

$$\ell(s_i) \doteq \ell(\hat{s}_i) + \ell'(\hat{s}_i)\Delta s_i + \ell''(\hat{s}_i)\Delta s_i^2$$

$$\ell'(s_i) \doteq y_i - p_i = e_i$$

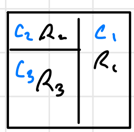
$$\ell''(s_i) \doteq -p_i(1-p_i) = -w_i$$

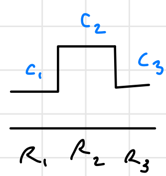
So $\ell(s_i) \doteq \text{const} + \hat{e}_i \Delta s_i - \frac{1}{2}\hat{w}_i \Delta s_i^2$

$$= \text{const} - \frac{1}{2}\hat{w}_i\left(\Delta s_i^2 - 2\frac{\hat{e}_i}{\hat{w}_i}\Delta s_i\right)$$

$$= \text{const} - \frac{1}{2}\hat{w}_i\left(\Delta s_i - \underbrace{\frac{\hat{e}_i}{w_i}}_{\hat{y}_i}\right)^2$$

$$\text{Loss} = \frac{1}{2}\sum_{i=1}^{n} \hat{w}_i\left(\hat{y}_i - h_t(x_i)\right)^2$$





$$\text{Loss} = \frac{1}{2}\sum_{m=1}^{M} \underline{\sum_{i:x_i\in R_m} \hat{w}_i\left(\hat{y}_i - c_m\right)^2} + \underbrace{\frac{1}{2}\gamma\sum_{m=1}^{M} c_m^2}_{\text{small correction}} + \underbrace{\lambda M}_{\text{small \# of regions}}$$

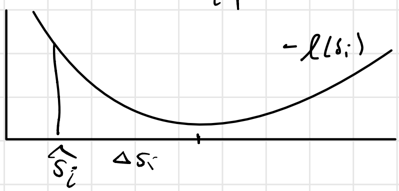
$$\text{Loss} = \frac{1}{2}\sum_{m=1}^{M} \underline{\sum_{i:x_i \in R_m} \hat{w}_i (\hat{y}_i - c_m)^2} + \underbrace{\frac{1}{2}\gamma \sum_{m=1}^{M} \underline{c_m}^2}_{\text{regularization}} + \underbrace{\lambda M}_{\text{regularize}}$$

$$\frac{\partial}{\partial c_m} : \quad = -\sum_{i:x_i \in R_m} \hat{w}_i (\hat{y}_i - c_m) + \gamma c_m = 0$$

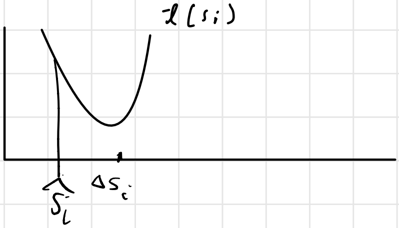
$$\hat{c}_m = \frac{\sum_{i:x_i \in R_m} \hat{w}_i \, \hat{y}_i}{\sum_{i:x_n \in R_m} \hat{w}_i + \underbrace{\gamma}_{\text{shrinkage}}}$$

Recall IRLS

$$\begin{array}{ccc} \hat{s}_i & \xrightarrow{\Delta s_i} & s_i \\ \| & & \| \\ x_i^T \beta_t & \xrightarrow{x_i^T \Delta\beta} & x_i^T \beta \end{array}$$



small $\hat{w}_i$



big $\hat{w}_i$