Decision Trees

Information = # bits needed to encode the probability of an event

$$= -\log_2 p$$

e.g. coin flip from fair coin ($p = 1/2$) is

$$-\log_2 \tfrac{1}{2} = 1 \text{ bit}$$

e.g. certain event ($p = 1$)

$$-\log_2 1 = 0 \text{ bits}$$

Want to define I so that:

- $I(p) \geq 0$
  $I(1) = 0$
- Two indep events w prob p_1 & p_2
  $I(\text{both events}) = I(p_1 \cdot p_2) = I(p_1) + I(p_2)$
- $I(p)$ should be continuous

  Together: $I(p^2) = 2I(p)$
  $I(p^n) = n\,I(p) \rightsquigarrow -\log_b(p)$

  $b = 2$ for "bits"

e.g. coin flip gives $-\log_2 1/2 = 1$

e.g. biased coin, heads with $p = .99$ gives $-\log_2 .99 = 0.0145$ bits

$p = .01$ $\quad -\log_2 .01 = 6.643$ bits

Entropy ~ expected information of a set of events

events	V_1	---	V_J
probs	p_1	---	p_J
information	$I(p_1)$	---	$I(p_J)$

$$\text{Entropy} = p_1 I(p_1) + p_2 I(p_2) + \cdots + p_J I(p_J)$$

$$= \sum_j p_j I(p_j) = -\sum_j p_j \log_2(p_j) =: H(p_1 \cdots p_J)$$

↑ entropy

So if

heads	tails
p	$1-p$
$I(p)$	$I(1-p)$

$$H(p, 1-p) = p\,I(p) + (1-p)\,I(1-p)$$

$$= -p \log_2 p - (1-p)\log_2(1-p)$$

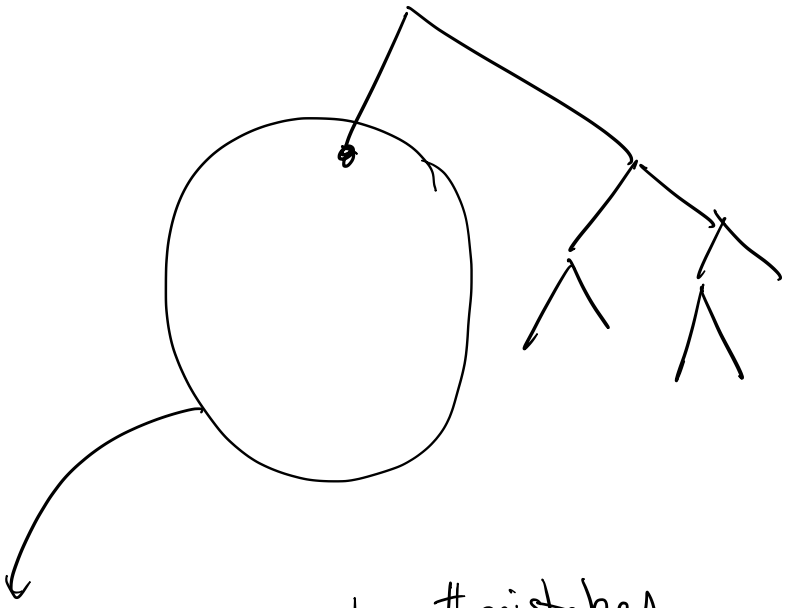
if $p = .5$ $\quad H(.5, .5) = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 = 1$ bit

if $p = .99$ $\quad H(.99, .01) = \cdots = .08$ bits

prune

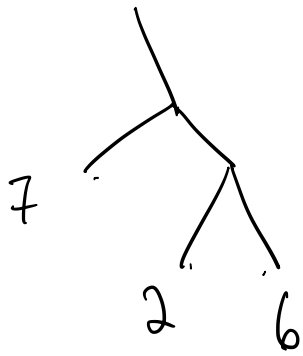
CART "minimal cost complexity pruning"

Each subtree is assigned a cost



$$\text{Cost}(\text{subtree}) = \frac{1}{n}\,(\#\text{mistakes made by subtree}) + C\,[\#\text{leaves in subtree}]$$

C: 0.01



$$\text{Cost}(\text{subtree}) = \frac{1}{100}\,[7+2+6] + \frac{1}{100}\cdot 3$$

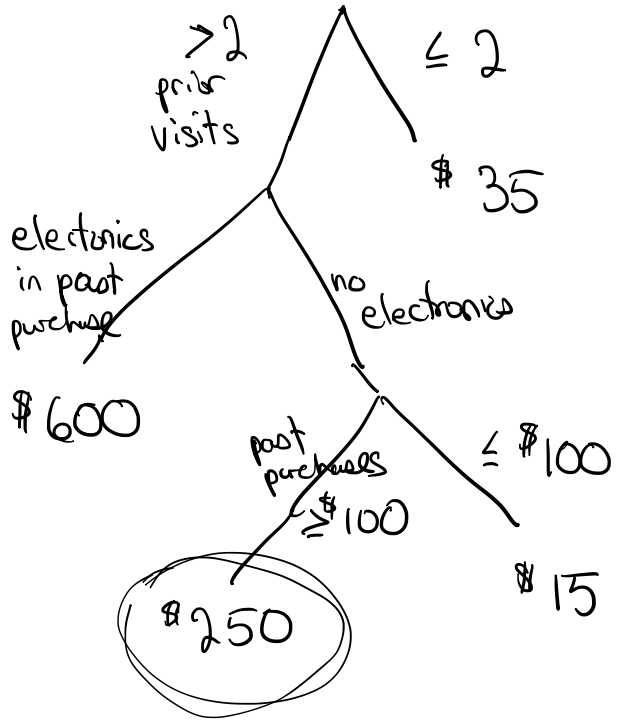
$$= \frac{15}{100} + \frac{3}{100} = \frac{18}{100}$$

Why did I choose $C = 0.01$? Adding one leaf $\cong$ 1% error decrease

Should I add one more leaf that reduces error by .005% ?

Regression Trees

How much will the customer spend at an online store?

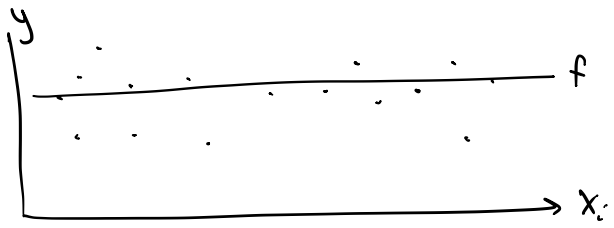


In regression, want to minimize

$$\sum_i \left(f(x_i) - y_i \right)^2$$

(f: 26 bits, y_i: 57 bits)

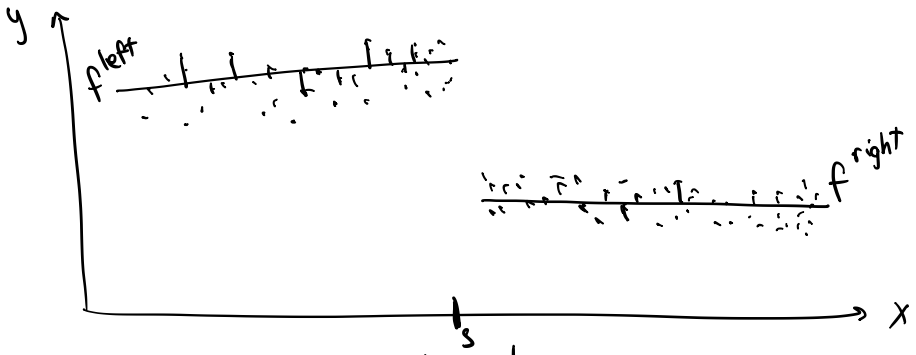
What if f does not depend on x_i?



f is average of y_i's, $f = \frac{1}{n} \sum_i y_i$

> In every leaf of regression tree, choose f in the leaf to be the average of the y_i's

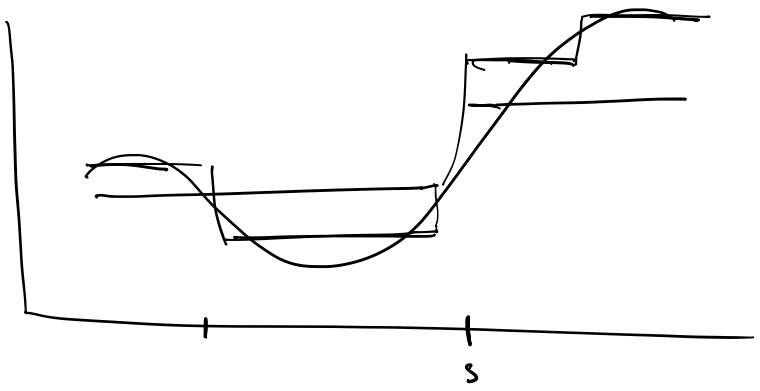
How to split?



Want to choose split point s.

$$\text{minimize}_{s} \sum_{i \text{ where } x_i < s} (f^{left} - y_i)^2 + \sum_{i \text{ where } x_i \geq s} (f^{right} - y_i)^2$$

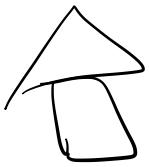
- i where $x_i < s$: left
- f^{left}: average of y_i's on the left
- i where $x_i \geq s$: right
- f^{right}: average of y_i's on the right



Nearest neighbors - can be done well or badly

+ interpretable
+ accuracy
- can be misleading if done wrong

- like real estate comps

	location	sqft	bdrms	bath	sold price
	***	3000	4	3	$324K
	***	2600		2	$310K
	****	3200	5	3	$3301

KNN - find k "nearest neighbors" and take
- majority vote (classif)
- average (regression)

What distance metric to use for comparison?

$$\text{dist}\left(\text{house}_1, \text{house}_2\right) = \begin{cases} \infty \text{ if not in same nhbd} \\ (\text{sqft}_1 - \text{sqft}_2)^2 + C_{\text{bed}}(\text{bdrm}_1 - \text{bdrm}_2)^2 \\ \quad + C_{\text{bath}}(\text{bath}_1 - \text{bath}_2)^2 \end{cases}$$

Can also do "weighted" knn

lot of work lately on learning distances for knn