CS4811
Feb 15, 2017
Wednesday (1)

week 6 now — hw 2 due Friday search

week 7 M ✓
W career fair no class

week 8 M
W exam 1 4^{00}–5^{30}

Spring Break

↓
search Ch 3
Ch 4
Ch 5
neural network

UPE resume clinic
Fri 4:00
career services
Google workshop

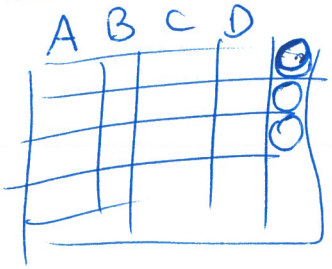
Previous class

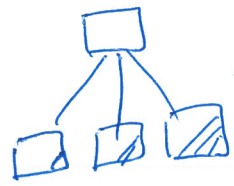
Constructing decision trees

"learning" classification multi-valued concepts

↓

problem data → concise representation





series of operations

minimum depth tree

look for the minimal tree that accurately represents this table

search: for this tree

optimization problem

(goal is not explicitly defined)

goal: > to classify objects "efficiently"

reduce uncertainty

13 examples ~~3~~ : false : 7 true = 6

uncertainty: $\frac{6}{13}$, $\frac{7}{13}$

One question: which one would reduce the uncertainty the most?

heuristic:

originally probability of error, without asking any questions $\frac{6}{13}$

A is true

true: 6
False: 2

class

classification $\frac{2}{8}$

$\frac{8}{13}$ $\frac{2}{8}$

A is false

true: 0
False: 5

$\frac{0}{5}$

$\frac{5}{13}$ $\frac{0}{5}$

$\frac{6}{13}$ $\frac{2}{13}$

B: $\frac{5}{13}$ C: $\frac{4}{13}$ D: $\frac{3}{13}$ E: $\frac{6}{13}$

A

Entropy: measure of uncertainty

how many bits are needed to describe the possible outcomes

one value: 0 bits

2 values: 1 bit

4 values: 2 bits

Random variable V with values v_k each of with probability $P(v_k)$

a variable that can take some values each with a probability

entropy: $H(V) = \sum_k P(v_k) \log_2 \frac{1}{P(v_k)}$

$$= -\sum_k P(v_k) \log_2 P(v_k)$$

Example fair coin flip

$H(\text{fair coin}) = -\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right]$

$\frac{1}{2}\log_2\frac{1}{2} = 0.5 \cdot -1 = -0.5$

$= -[-1]$

$= 1$

$$H(\text{loaded}) = -[(0.99 \underbrace{\log_2 0.99}_{-0.01}) + (0.01 \underbrace{\log_2 0.01}_{-6.64})] \quad (4)$$

$$= 0.08$$

$B(q)$ = entropy of a Boolean variable that is true with probability q.

$$B(q) = -(q \log_2 q + (1-q)\log_2(1-q))$$

p: # of positive examples 6 total: 13

n: # of negative examples 7 all: p+n

entropy before asking any questions

$$B\left(\frac{p}{p+n}\right) \qquad B\left(\frac{6}{13}\right)$$

After picking question A

Remainder(A)

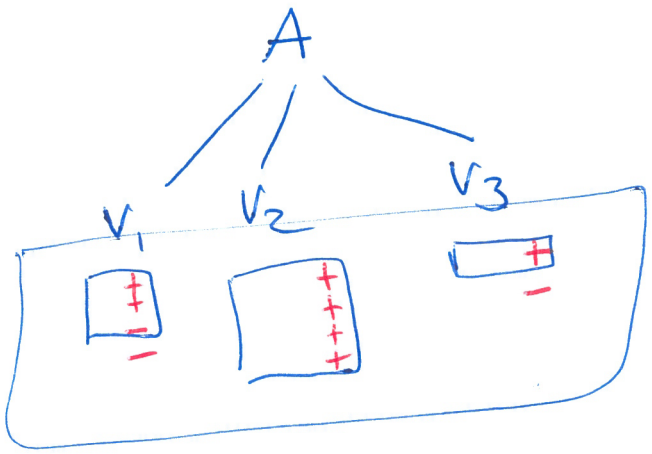
A has d answers

for each: how much uncertainty do I still have?

$$= \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k+n_k}\right)$$

p_k: # of positive examples when A's value is v_i

n_k: # of negative examples when A's value is v_i



"expectation"

lottery		
	0.1	1000
	0.2	10
	0.7	0

expected value of this lottery is

$$0.1 \times 1000 + 0.2 \times 10 + 0.7 \times 0 = 102$$

100 2

Using the entropy formula

gain a reduction in uncertainty when we ask a question

$$\text{gain}(A) = B\left(\frac{p}{p+n}\right) - \text{remainder}(A)$$

The question (attribute) that has the highest gain will win a spot at the root of the tree.

(6)

$$gain(A) = B\left(\frac{6}{13}\right) - \left[\frac{8}{13} B\left(\frac{6}{8}\right) + \frac{5}{13} B\left(\frac{0}{5}\right)\right]$$

$B\left(\frac{6}{13}\right) = 1.00$

$B\left(\frac{6}{8}\right) = 0.81$

$\frac{8}{13} B\left(\frac{6}{8}\right) = 0.62 \cdot 0.81 = 0.50$

$B\left(\frac{0}{5}\right) = 0$

$= 0.50$ for A

B, C, D, E will have lower values for gain

Summary

entropy: measure of uncertainty

random variable V with values v_k each with probability $P(v_k)$

$$H(V) = \sum_k P(v_k) \log_2 \frac{1}{P(v_k)}$$

$$= -\sum_k P(v_k) \log_2 P(v_k)$$

$$B(q) = -(q \log_2 q + (1-q) \log_2 (1-q))$$

$$H(\text{"goal"}) = B\left(\frac{p}{p+n}\right)$$

$$\text{Remainder}(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right)$$

$$\text{gain}(A) = H(\text{goal}) - \text{remainder}(A)$$