10/21

Applications of Decision Trees

- Medical Diagnostics
- Credit risk analysis
- Calender scheduling preferences

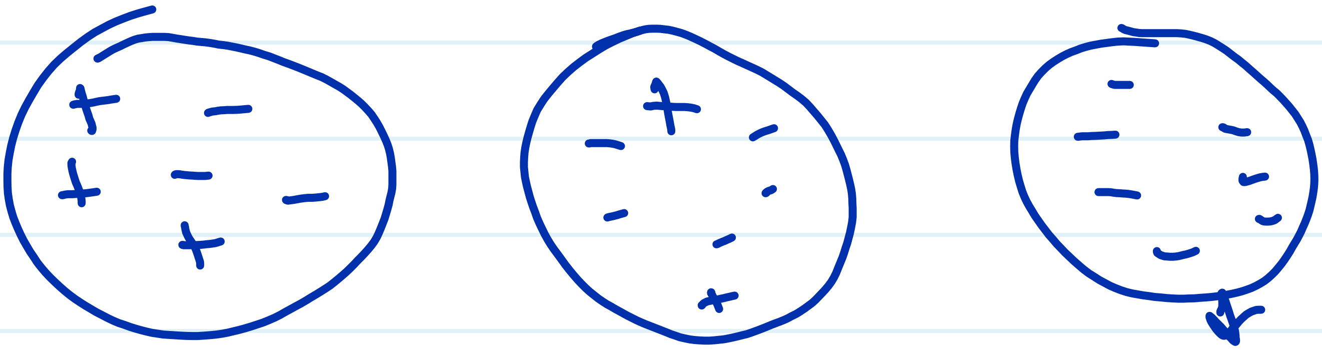
Q: How do we choose the best feature?

ROC curve (Lab 4)
What feature gives most info abt label (Lab 6)

Entropy

Def: Average # of bits needed to send one data point.

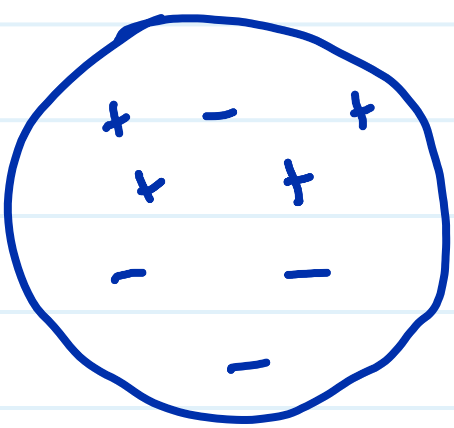
Poisonous & edible mushrooms



High ⟷ Low

Equation:

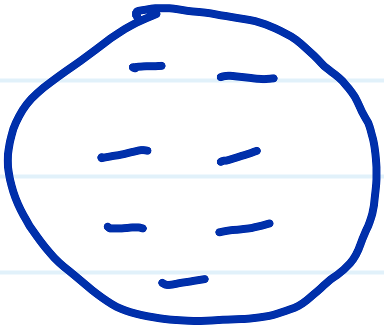
$$\boxed{H(y) = -\sum_{c \in vals(y)} p(y=c) \underbrace{\log_2(p(y=c))}_{\#\text{ of bits}}}$$



$$H(y) = -p(y=+)\log_2(p(y=+)) - p(y=-)\log_2(p(y=-))$$

$$= -0.5\log_2(0.5) - 0.5\log_2(0.5)$$

$$= 0.5 + 0.5 = 1$$



$$H(y) = 0 - 1\log_2(1)$$
$$= 0$$

Encoding Data

Class Year	Fixed-length encoding
senior	0 0
junior	0 1
sophomore	1 0
freshman	1 1

Shannon Encoding

class	prob (p)	cumulative	Binary
senior	0.5	0	0.000..
junior	0.25	0.5	0.100..
sophomore	0.125	0.75	0.110..
freshman	0.125	0.875	0.111..

sorted high to low

Same example:

ceiling

$\lceil -\log_2 p \rceil$	encoding
1	0
2	10
3	110
3	111

# of binary bits

H(class year) =

$0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.125 \cdot 3 = 1.75$

Information Gain for Selecting Features

Conditional Entropy:
Quantifies the amount of info needed to describe outcome Y given X.

$$\boxed{\begin{aligned} H(Y|X) &= \sum_{v \in Vals(X)} P(X=v) H(Y|X=v) \\ H(Y|X=v) &= \sum_{c \in Vals(Y)} P(Y=c|X=v) \log_2 (P(Y=c|X=v)) \end{aligned}}$$

Information Gain:
Reduction in entropy/uncertainty given information.

$$\boxed{G(Y,X) = H(Y) - H(Y|X)}$$

Want High