Lecture 2

Recap

Shannon Information

$$h(x) = -\log_2 P(x)$$

(shannon) bits

Entropy

$$H(X) = -\sum_{x \in A_X} P(x) \log_2 P(x)$$

x is an event

X is R.V.

A_X is alphabet for X

E.g. 2.1.

Coin Toss

H = Heads
T = Tails.

$A_x = \{H, T\}$

Fair Coin

$h(H) = -\log_2 \frac{1}{2} = 1 \text{ bit.}$
$h(T) = -\log_2 \frac{1}{2} = 1 \text{ bit.}$

$H(x) = \frac{1}{2} \times 1 + \frac{1}{2} \times 1 = \underline{\underline{1 \text{ bit}}}$

Biased Coin

$P(h) = 0.25$

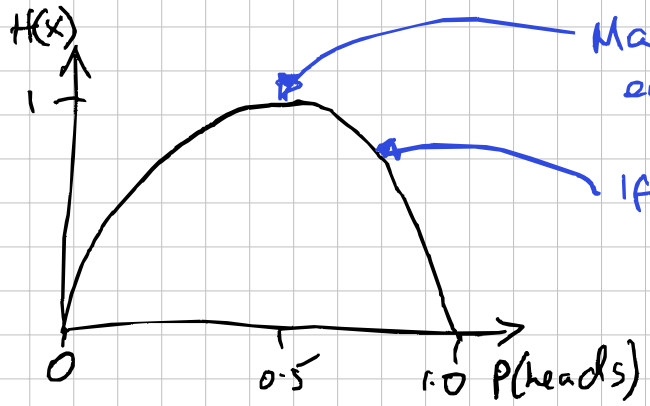
$h(H) = -\log_2 \frac{1}{4} = 2 \text{ bits}$
$h(T) = -\log_2 \frac{3}{4} = 0.415 \text{ bits.}$

$H(x) = \frac{1}{4} \times 2 + \frac{3}{4} \times 0.415 = \underline{\underline{0.811 \text{ bits}}}.$

When not equiprobable $\Rightarrow$ Less information on average.

Generalise

$$H(X) = -P(H)\log_2 P(H) - P(T)\log_2 P(T)$$
$$= -P(H)\log P(H) - (1-P(H))\log_2 (1-P(H))$$



Max entropy occurs when equiprobable

If you can't be equiprobable get as close as you can

Q Do these rules generalise to larger $|A_x|$ problems?

Generalising

$$L(x,\lambda) = H(x) - \lambda\left(\sum P_i - 1\right)$$

$$\frac{\partial L}{\partial P_i} = \frac{\partial H(x)}{\partial P_i} - \lambda = \frac{\partial}{\partial P_i}\left(-P_i \log P_i\right) - \lambda$$

$$= -\log_2 P_i - \frac{P_i}{P_i} - \lambda = 0 \quad \text{for stationary points}$$

$$\left.\begin{aligned} -\log_2 P_0 &= \lambda + 1 \\ -\log_2 P_1 &= \lambda + 1 \\ -\log_2 P_2 &= \lambda + 1 \\ &\vdots \end{aligned}\right\} \quad \text{Only possible when } P_0 = P_1 = P_2 = P_i$$

If $P_i = P_j \ \forall\, i,j$ and $\sum_i P_i = 1$ then $P_i = \frac{1}{|A_x|}$

Stationary point at $P_i = \frac{1}{|A_x|}$

Getting close

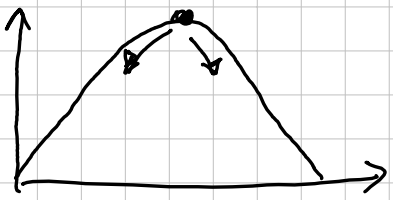
$$\lambda = \log_2 \frac{1}{|A_x|} - 1$$

$$\frac{\partial L}{\partial P_i} = -\log_2 P_i - \left(-\log_2 \frac{1}{|A_x|} - 1\right) - 1$$

$$= -\log_2 \left(P_i \cdot |A_x|\right)$$

$$P_i < \frac{1}{|A_x|} \Rightarrow \frac{\partial L}{\partial P_i} = -\log\left(P_i |A_x|\right) > 0$$

$$P_i > \frac{1}{|A_x|} \Rightarrow \frac{\partial L}{\partial P_i} = -\log\left(P_i |A_x|\right) < 0$$



$\therefore$ Entropy is maximum for equiprobable P_i

Entropy falls monotonically as you move away from equiprobable

E.g. 21 Solution

E.g. 2.2

Information Gain

Imagine building a decision tree for cats vs dogs.

Your training set has 12 cats and 8 dogs.

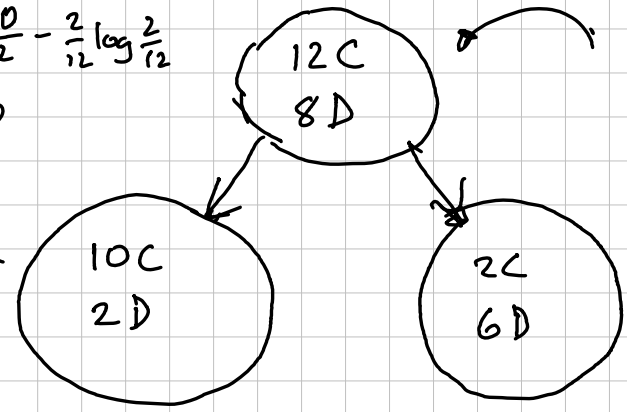
You have 2 features: length and weight.

Q. which do you split on first?

Split by length first

$H(x) = -\frac{10}{12}\log\frac{10}{12} - \frac{2}{12}\log\frac{2}{12}$

$= 0.650$



$H(x) = -\frac{12}{20}\log\frac{12}{20} - \frac{8}{20}\log\frac{8}{20}$

$= 0.971$

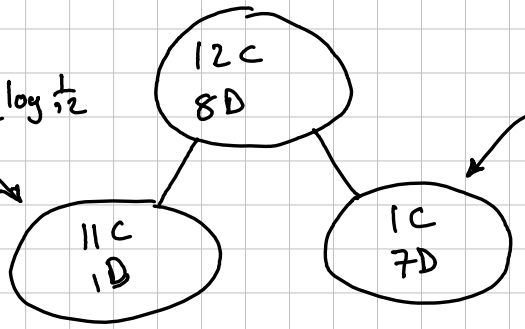
$H(x) = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8}$

$= 0.811$

Average $H(x)$ after $= \frac{12}{20} \times 0.65 + \frac{8}{20} \times 0.811 = 0.714$

Expected info. gain $= 0.971 - 0.714 = \underline{0.257}$ bits

Split by weight first



$H(x) = -\frac{11}{12}\log\frac{11}{12} - \frac{1}{12}\log\frac{1}{12}$

$= 0.414$

$H(x) = -\frac{1}{8}\log\frac{1}{8} - \frac{7}{8}\log\frac{7}{8}$

$= 0.568$

Expected entropy after $= \frac{12}{20} \times 0.414 + \frac{8}{20} \times 0.568$

$= 0.476$

Expected info gain $= 0.971 - 0.476$

$= \underline{0.495}$

∴ Splitting by weight is more informative as the first split.