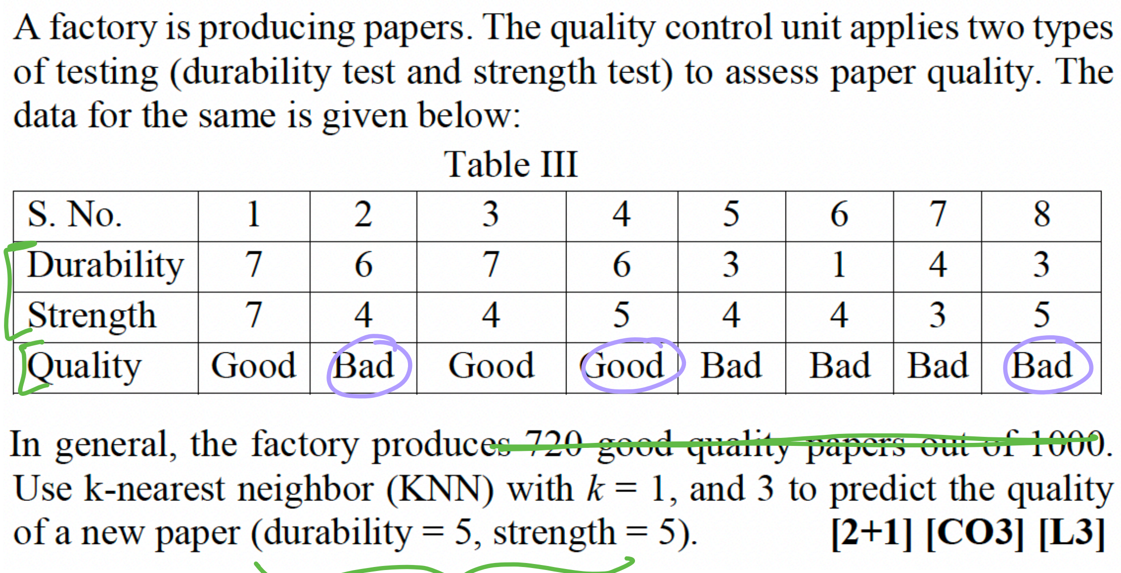


A factory is producing papers. The quality control unit applies two types of testing (durability test and strength test) to assess paper quality. The data for the same is given below:

Table III

	S. No.	1	2	3	4	5	6	7	8
features	Durability	7	6	7	6	3	1	4	3
	Strength	7	4	4	5	4	4	3	5
y	Quality	Good	Bad	Good	Good	Bad	Bad	Bad	Bad

In general, the factory produces ~~720 good quality papers out of 1000~~. Use k-nearest neighbor (KNN) with $k = 1$, and 3 to predict the quality of a new paper $\underbrace{\text{(durability = 5, strength = 5)}}_{\text{test datapoint ?}}$. **[2+1] [CO3] [L3]**

D	S	labels	Distance	K=1	k=3
7	7	G	$\sqrt{(7-5)^2+(7-5)^2} = 2\sqrt{2}$		
6	4	B	$\sqrt{(6-5)^2+(4-5)^2} = \sqrt{2}$		✓ B
7	4	G	$\sqrt{(7-5)^2+(4-5)^2} = \sqrt{5}$		
6	5	G	$\sqrt{(6-5)^2+(5-5)^2} = 1$	✓ G	✓ G
3	4	B	$\sqrt{(3-5)^2+(4-5)^2} = \sqrt{5}$		
1	4	B	$\sqrt{17}$		
4	3	B	$\sqrt{5}$		
3	5	B	2		✓ B
				Good	Bad

Naive Bayes Classifier:

Supervised

Conditional Probability:

$$P(A|B) = \frac{P(A\cap B)}{P(B)}$$

$$P(A\cap B) = P(A|B)\cdot P(B)$$

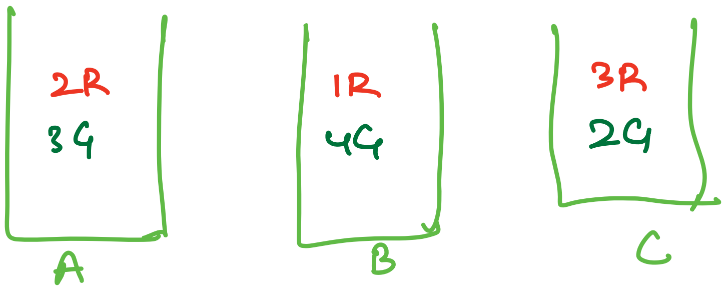
$$P(B|A) = \frac{P(B\cap A)}{P(A)}$$

$$P(B\cap A) = P(B|A)\cdot P(A)$$

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B\cap A)}{P(B)} = \frac{P(B|A)\cdot P(A)}{P(B)}$$

$$\boxed{P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)}}$$

Eg:



Q: Prob. of getting a red ball given that A box is chosen

$$P(R|A) = \frac{2}{5}$$

Q: Prob. of getting a red ball?

$$P(R) = P(R\cap A) + P(R\cap B) + P(R\cap C)$$

Q: Prob. that bag A is chosen given that Red ball is drawn

$$P(A|R) = \frac{P(A\cap R)}{P(R)}$$

$$= \frac{P(R|A)\cdot P(A)}{P(R)}$$

$$= \frac{\overset{2/5}{P(R|A)}\cdot \overset{1/3}{P(A)}}{P(R\cap A) + P(R\cap B) + P(R\cap C)}$$

Naive Bayes Classifier:

(Likelihood)
Conditional Probability

Prior Prob.

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B|A)\; P(A)}{P(B)}$$

Posterior Probability

→ Prob. of B

mushroom Dataset:

count:

Class: 1 2 3

features: shape, color, radius, weight ....

$$P(y=1) = \frac{n_1}{n_1+n_2+n_3}$$

$$P(y=2) = \frac{n_2}{n_1+n_2+n_3}$$

$$P(y=3) = \frac{n_3}{n_1+n_2+n_3}$$

Test Mushroom

class → 1
→ 2
→ 3
?

$P(y=1|x)$ → 0.25

$P(y=2|x)$ → 0.15

$P(y=3|x)$ → 0.6

max is 0.6

Test Mushroom belongs to Class 3.

Sum of all these
prob. should be 1

$$P(\overset{A}{\underline{y=1}}|\overset{B}{\underline{x}}) = \frac{P(x|y=1) * P(y=1)}{P(x)}$$

$$\boxed{P(A|B) = \frac{P(B|A) * P(A)}{P(B)}}$$

$$= \frac{P(x|y=1) * P(y=1)}{\underline{P(x \cap y=1)} + \underline{P(x \cap y=2)} + \underline{P(x \cap y=3)}}$$

$$P(y=1|x) = \frac{P(x|y=1) * P(y=1)}{\underline{P(x|y=1) \cdot P(y=1)} + \underline{P(x|y=2) \cdot P(y=2)} + \underline{P(x|y=3) \cdot P(y=3)}}$$

$$P(y=2|x) = \frac{P(x|y=2) * P(y=2)}{P(x|y=1) \cdot P(y=1) + P(x|y=2) \cdot P(y=2) + P(x|y=3) \cdot P(y=3)}$$

$$P(y=3|x) = \frac{P(x|y=3) * P(y=3)}{P(x|y=1) \cdot P(y=1) + P(x|y=2) \cdot P(y=2) + P(x|y=3) \cdot P(y=3)}$$

$$P(y=1|x) = \frac{P(x|y=1) \cdot P(y=1)}{\text{Denom}}$$

$$P(y=2|x) = \frac{P(x|y=2) \cdot P(y=2)}{\text{Denom}}$$

$$P(y=2|x) = \frac{P(x|y=3) \cdot P(y=3)}{}$$

Denom

$$P(y=1|x) \propto P(x|y=1) \cdot P(y=1)$$

$$P(y=2|x) \propto P(x|y=2) \cdot P(y=2)$$

$$P(y=3|x) \propto P(x|y=3) \cdot P(y=3)$$

$$P(x|y=1) = P(x_1|y=1) \cdot P(x_2|y=1) \cdot P(x_3|y=1) \cdots\cdots P(x_n|y=1)$$

x is a test data point

shape, color, radius, weight

$$P(x|y=1) = \prod_{i=1}^{n} P(x_i|y=1)$$

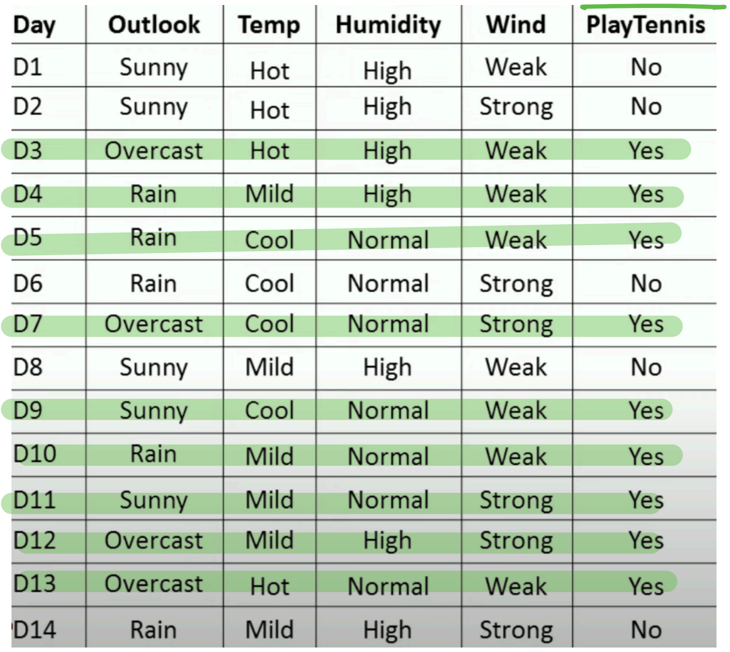
$$P(y=1|x) \propto \prod_{i=1}^{n} P(x_i|y=1) \cdot P(y=1)$$

$c \in \{1,2,3\}$

$$\underbrace{P(y=c|x)}_{\text{Posterior Probability}} \propto \prod_{i=1}^{n} \underbrace{P(x_i|y=c)}_{\text{Conditional Probability / Likelihood}} \cdot \underbrace{P(y=c)}_{\text{Prior}}$$

Eg:

x_1 x_2 x_3 x_4 y



Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$c \in \{\text{Yes}, \text{No}\}$

$$P(y=\text{Yes}) = \frac{9}{14}$$

$$P(y=\text{No}) = \frac{5}{14}$$

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

$$P(\underline{\text{Outlook} = \text{Sunny}} \mid \overset{9}{\underline{y = \text{Yes}}})$$

$$P(\text{temp} = \text{cool} \mid y = \text{Yes})$$

Temp

	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity

	Yes	No
High	2/9	4/5
Normal	6/9	1/5

Windy

	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

Test Datapoint

x [Outlook = Sunny, Temp = Cool, Humidity = high, wind = Strong

$P(y=\text{Yes}\mid x) = P(\text{Outlook}=\text{Sunny}\mid y=\text{Yes})\,\cdot$ ✓
$P(\text{temp}=\text{Cool}\mid y=\text{Yes})\,\cdot$
$P(\text{humidity}=\text{high}\mid y=\text{Yes})\,\cdot$
$P(\text{wind}=\text{Strong}\mid Y=\text{Yes})\,\cdot$
$P(y=\text{Yes})$

$$= \frac{\cancel{2}}{9}\cdot\frac{\cancel{3}}{\cancel{9}}\cdot\frac{\cancel{2}}{\underset{\cancel{3}}{\cancel{9}}}\cdot\frac{\cancel{3}}{\underset{3}{\cancel{9}}}\cdot\frac{\overset{3}{\cancel{9}}}{\underset{7}{\cancel{14}}} = \frac{1}{9\cdot3\cdot7} = \frac{1}{189}$$

$= 0.0053$

$P(y=\text{No}\mid x) = P(\text{Outlook}=\text{Sunny}\mid y=\text{No})\,\cdot$
$P(\text{temp}=\text{Cool}\mid y=\text{No})\,\cdot$
$P(\text{humidity}=\text{high}\mid y=\text{No})\,\cdot$
$P(\text{wind}=\text{Strong}\mid Y=\text{No})\,\cdot$
$P(y=\text{No})$

$$= \frac{3}{5}\cdot\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{3}{5}\cdot\frac{5}{14} = \frac{18}{875} = 0.0206$$

Test Datapoint belongs to class No.

Decision Trees

↳ Supervised Algo

↳ Classification & Regression

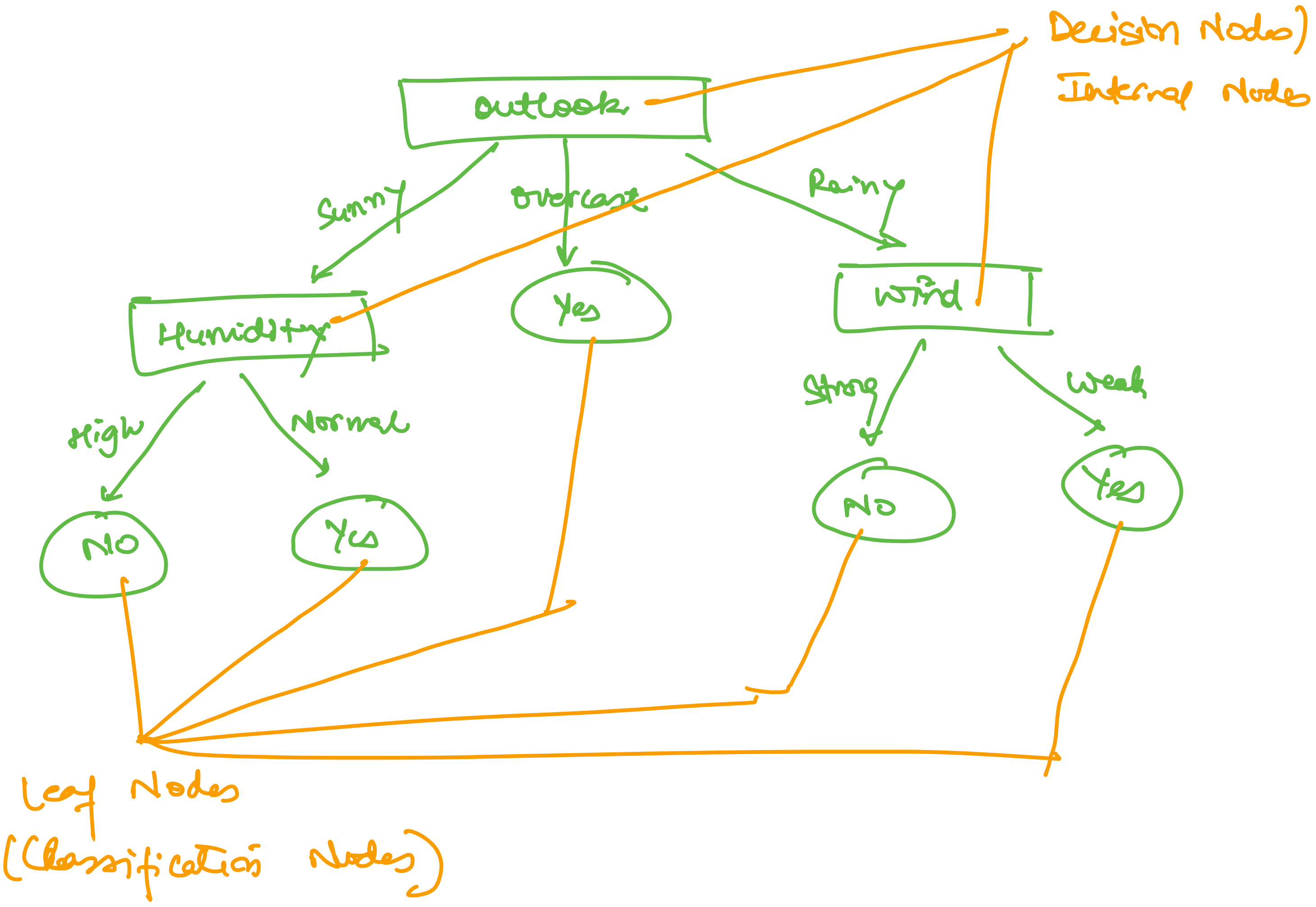
$y \in$ discrete set $y \in R$

x_1 : outlook $\in$ {Sunny, overcast, Rainy}

x_2 : humidity $\in$ {High, Normal}

x_3 : wind $\in$ {strong, weak}

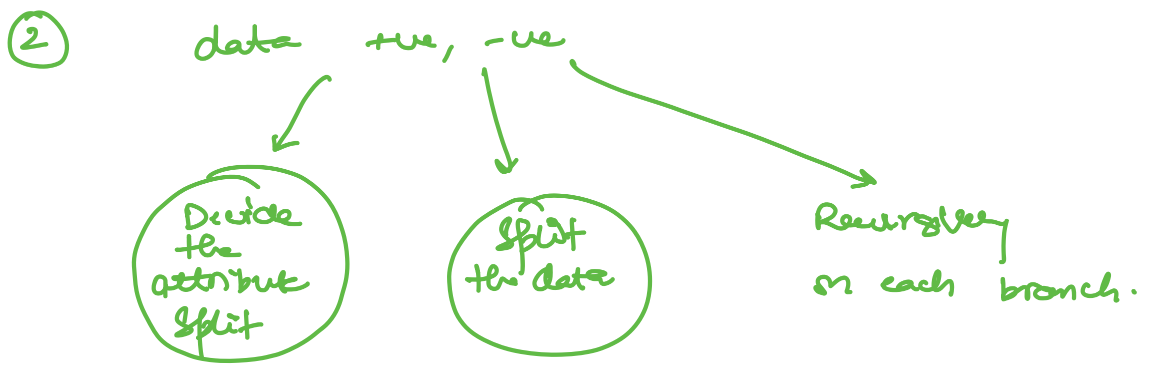
x_4 : temperature $\in$ { Hot, Moderate, cold }

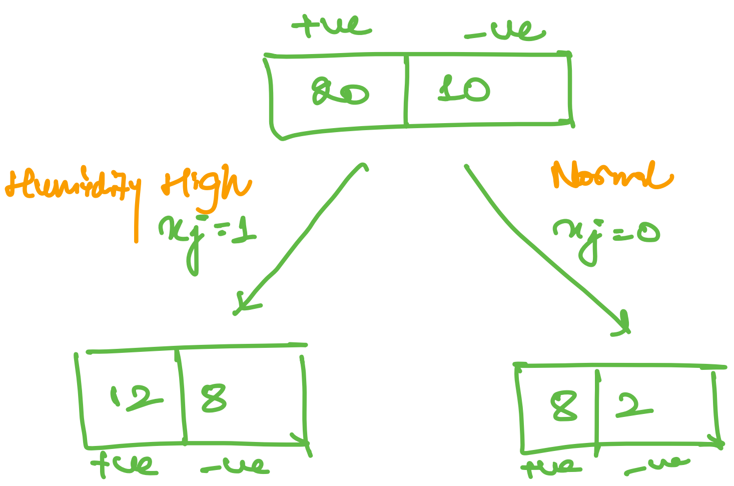


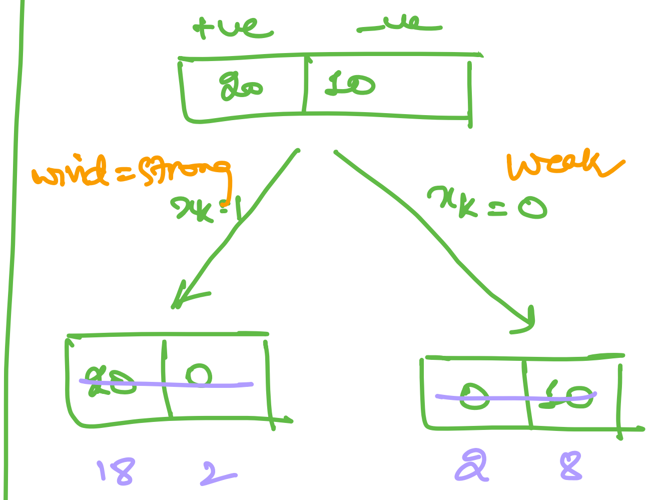
BUILD A DECISION TREE?

Come at a node

① ↳ data is "pure" make it a leaf Node







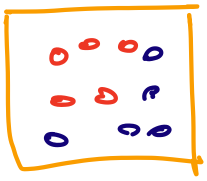
Entropy   Randomness

$$\overset{\text{(label)}}{y} \in \{\overbrace{1 \cdots\cdots r}^{\text{classes}}\}$$

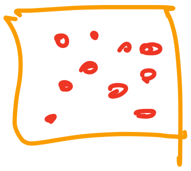
$$P(y=k) = p_k$$

$$H(y) = -\sum_{i=1}^{r} p_k \log p_k$$

$$= -\left(p_1 \log p_1 + p_2 \log p_2 + \cdots\cdots p_r \log p_r\right)$$



: max



min

Boolean case $\quad y \in \{0, 1\}$

p_0, p_1

$p_0 + p_1 = 1$

$$H(y) = -\left(p_0 \log p_0 + p_1 \log p_1\right)$$

$$= -\left(p_1 \log p_1 + (1-p_1)\log(1-p_1)\right)$$

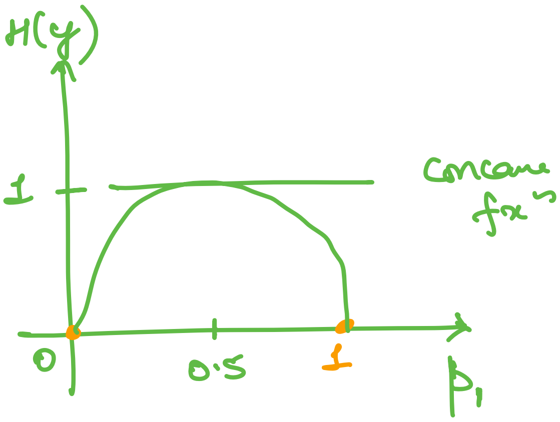
$p_1 = 0$ $\quad \underbrace{p_1 \log p_1}_{0} + \underbrace{(1-p_1)}_{1}\ \underbrace{\log 1}_{0}$

$$\underbrace{\qquad\qquad\qquad}_{0}$$

$p_1 = 1$ $\quad \underbrace{\log 1}_{0} + 0$

Class: Cat $\quad$ Dog

$$\begin{tikzcd}
\text{Cat} \arrow[d] & & \\
P(\text{Cat}=1) \arrow[r] & \text{Centropy}=0 \\
P(\text{Dog}=1) \arrow[ur] &
\end{tikzcd}$$



$$\frac{d}{dp_1}\left(-\left(p_1 \log p_1 + (1-p_1)\log(1-p_1)\right)\right)$$

$$-\left(\log p_1 + \frac{p_1}{p_1} - \frac{(1-p_1)}{(1-p_1)} - \log(1-p_1)\right)$$

$$\log p_1 + \cancel{1} - \cancel{1} - \log(1-p_1) = 0$$

$$\log p_1 = \log(1-p_1)$$

$$p_1 = 1 - p_1$$

$$2p_1 = 1 \qquad \underline{\underline{p_1 = 0.5}} = \tfrac{1}{2}$$

$y \in \{1, 2, \cdots \cdot r\}$

$H(Y) = -\sum_{k=1}^{r} p_k \log p_k$ is maximum at $p_k = 1/r \;\; \forall k$