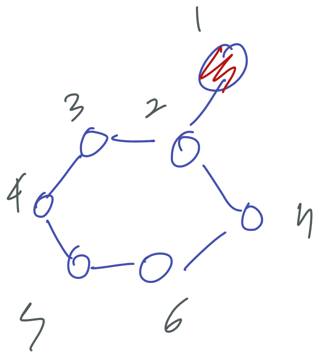
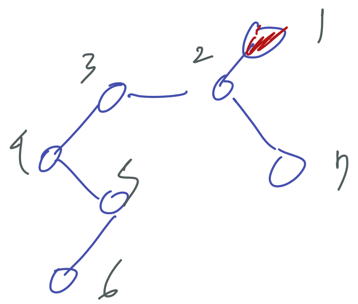
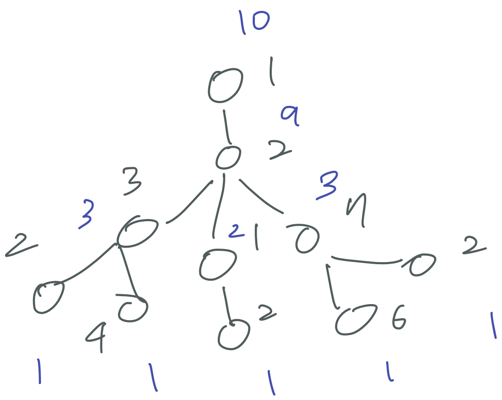
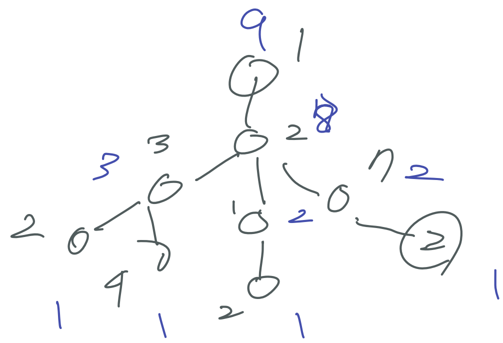
1 - a)









⇒ 2 layer 이면 node 1에 대해서 10과 9로 다르게 embedding 됨

1 - b)

$$M(h_v^k) = \begin{cases} 1, & \text{if } h_v^k = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$h_{N(v)}^{k+1} = \max_{u \in N(v)} \left(M(h_k^u), 0 \right)$$

$$h_v^{k+1} = \max\left(M(h_v^k), h_{N(v)}^{k+1} \right)$$

2-a)

i) $1-(p^2+p^2-2p+1) = -2p^2+2p$ = GINI

$$\frac{d^2\,GINI}{dp^2} = \frac{d}{dp}(-4p+2) = -4 < 0 \Rightarrow \text{concavity}$$

ii) Entropy $\quad p\log_2\left(\frac{1}{p}\right) + (1-p)\log_2\left(\frac{1}{1-p}\right)$

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

$$\frac{d^2\,Entropy}{dp^2} = \frac{d}{dp}\left(\frac{d\,Entropy}{dp}\right) = \frac{d}{dp}\left(-\log_2 p - \frac{1}{\ln 2} + \log_2(1-p) - (1-p)\cdot\frac{1}{\ln 2}\cdot\frac{-1}{1-p}\right)$$

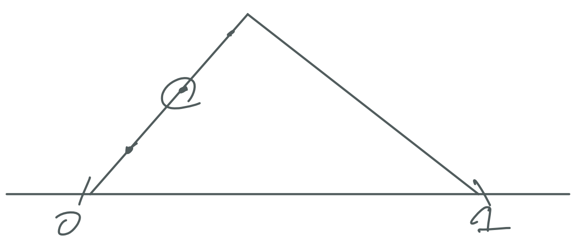
$$= \frac{d}{dp}\left(-\log_2 p - \frac{1}{\ln 2} + \log_2(1-p) + \frac{1}{\ln 2}\right)$$

$$= \frac{d}{dp}\left(\log_2(1-p) - \log_2 p\right)$$

$$= \frac{1}{\ln 2}\cdot\frac{-1}{1-p} - \frac{1}{\ln 2}\cdot\frac{1}{p} < 0$$

iii) $\underline{p<0.5}$ $\qquad\qquad$ $p \geq 0,5$

$1-\max(p,1-p) = p$ $\qquad\qquad$ $1-\max(p,(1-p)) = 1-p$



$x = 0,1 \qquad y = 0,5, \qquad z = 0,3$

$z=0,3 \quad f(z) = 0,3$

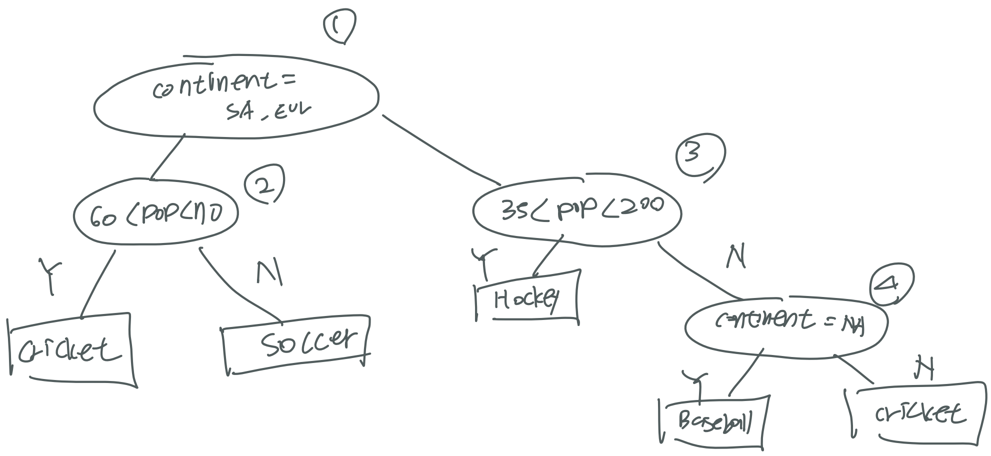
$x=0,1 \quad f(x) = 0,1$

$y=0,5 \quad f(y) = 0,5$

$f(z) = 0,3$ $\qquad$ same

$$\frac{y-z}{y-x}f(x) + \frac{z-x}{y-x}f(y) = \frac{0,2}{0,4}\,0,1 + \frac{0,2}{0,4}\,0,5 = \frac{1}{2}\cdot(0,6) = 0,3$$

2-b)



1 - GINI Index

$P_{soccer} = \frac{5}{12}$, $P_{cricket} = \frac{3}{12}$, $P_{Hockey} = \frac{2}{12}$, $P_{baseball} = \frac{2}{12}$

$$1 - \left(\frac{25}{144} + \frac{9}{144} + \frac{4}{144} + \frac{4}{144} \right) = 1 - \frac{42}{144} = \frac{102}{144}$$

1 - accuracy

$$1 - \max\left(\frac{5}{12}, \frac{3}{12}, \frac{2}{12}, \frac{2}{12} \right) = \frac{7}{12}$$

2 - GINI Index

$P_{soccer} = \frac{5}{6}$, $P_{cricket} = \frac{1}{6}$

$$1 - \left(\frac{25}{36} + \frac{1}{36} \right) = 1 - \frac{26}{36} = \frac{10}{36} = \frac{5}{18}$$

2 - accuracy

$$1 - \max\left(\frac{25}{36}, \frac{1}{36} \right) = \frac{11}{36}$$

3 - GINI Index

$P_{hockey} = \frac{2}{6}$, $P_{baseball} = \frac{2}{6}$, $P_{cricket} = \frac{2}{6}$

$$1 - \left(\frac{4}{36} + \frac{4}{36} + \frac{4}{36} \right) = 1 - \frac{12}{36} = \frac{24}{36} = \frac{2}{3}$$

(3) - accuracy

$$1 - \frac{2}{6} = \frac{2}{3}$$

(4) - GINI Index

$P_{baseball} = \frac{2}{4}$ $P_{cricket} = \frac{2}{4}$

$$1 - \left(\frac{1}{4} + \frac{1}{4}\right) = \frac{1}{2}$$

(4) - accuracy

$$1 - \frac{1}{2} = \frac{1}{2}$$

3 - a)

i) $\left(1 - e^{-km/n}\right)^k$

$k=3$, $n = 8$ billion, $m = 1$ billion. $\left(1 - e^{-3 \cdot \frac{1}{8}}\right)^3 = 0.030$

$k=4$, $n = 8$ billion, $m = 1$ billion. $\left(1 - e^{-4 \cdot \frac{1}{8}}\right)^4 = 0.024$

ii) $1 - k/n$

$\frac{k}{n} = \frac{1}{x}$

$x = \frac{n}{k}$

$\left(1 - \frac{k}{n}\right)^m$ $\quad 1 - \left(1 - \frac{k}{n}\right)^m = 1 - \left(1 - \frac{k}{n}\right)^{m \cdot \frac{n}{k} \cdot \frac{k}{n}}$

$= 1 - \left(1 - \frac{1}{x}\right)^{m \cdot x \cdot \frac{k}{n}}$

$= 1 - e^{-m \cdot \frac{k}{n}}$ for each array

for every k array

$\frac{d}{dx}\left(f(x)\right)^x$

$\left(1 - e^{-m \cdot \frac{k}{n}}\right)^k$

iii) optimal value of $k = \frac{n}{m} \ln 2$ $\quad \left(1 - e^{-m \cdot \frac{k}{n}}\right)^k \cdot \ln.$

3-b)

i)

a) $h(x) = 2x+1 \mod 32$

elem	hash	bit	tail length
3	7	00111	0
1	3	00011	0
4	9	01001	0
1	3	00011	0
5	11	01011	0
9	19	10011	0
2	5	00101	0
6	13	01101	0
5	11	01011	0

$\Rightarrow$ estimated distinct elem

$\Rightarrow 2^0 = 1$

b) $h(x) = 3x+7 \mod 32$

elem	hash	bit	tail length
3	16	10000	4
1	10	01010	1
4	19	10011	0
1	10	01010	1
5	22	10110	1
9	2	00010	1
2	13	01101	0
6	25	11001	0
5	22	10110	1

$\Rightarrow$ number of distinct elements

$\Rightarrow 2^4 = 16$

c) $h(x) = 4x \mod 32$

elem	hash	bit	tail length
3	12	01100	2
1	4	00100	2
4	16	10000	4
1	4	00100	2
5	20	10100	2
9	4	00100	2
2	8	01000	3
6	24	11000	3
5	20	10100	2

$\Rightarrow$ number of distinct elements

$\Rightarrow 2^4 = 16$

i)

hash function이 $ax+b \mod 2^k$ 일때, 이 숫자는 k-bit length로 나타내진다.

전체 element를 담아낼수 있는 정도의 k bit가 있어야 한다. 위 예시를 보았을때 k=5는 충분했다.

하지만 hash function의 output 2^k로 측정하는게 너무 fluctuation이 심하므로

일단 적절한 hash function을 사용해야한다.

또한 각 hash function의 결과를 평균을 내서 distinct eleme