ที่มาของ $Gini\ Impurity$

$$\bar{x} = \sum_{i=1}^{n} p_i x_i$$

$$\sigma^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$

Show $\bar{x} = \sum_{i=1}^{n} p_i x_i$

$X = \{1, 3, 3, 5, 5, 5, 9, 9\}$

แบบที่เราคุ้นเคย $\bar{x} = \frac{1+3+3+5+5+5+9+9}{8} = \frac{40}{8} = 5$

ลองปรับมุมมอง $\bar{x} = \frac{1}{8} + \frac{3+3}{8} + \frac{5+5+5}{8} + \frac{9+9}{8}$

$$\bar{x} = \frac{1 \cdot 1}{8} + \frac{2 \cdot 3}{8} + \frac{3 \cdot 5}{8} + \frac{2 \cdot 9}{8}$$

$$\bar{x} = \frac{1}{8} \cdot 1 + \frac{2}{8} \cdot 3 + \frac{3}{8} \cdot 5 + \frac{2}{8} \cdot 9$$

$$\bar{x} = p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4$$

$$\bar{x} = \sum_{i=1}^{4} p_i x_i$$

Show $\sigma^2 = \sum_{i=1}^{n} p_i(x_i - \bar{x})^2$

แบบที่เราคุ้นเคย

$$\sigma^2 = \frac{(1-\bar{x})^2 + (3-\bar{x})^2 + (3-\bar{x})^2 + (5-\bar{x})^2 + (5-\bar{x})^2 + (5-\bar{x})^2 + (9-\bar{x})^2 + (9-\bar{x})^2}{8}$$

ลองปรับมุมมอง

$$\sigma^2 = \frac{(1-\bar{x})^2}{8} + \frac{(3-\bar{x})^2 + (3-\bar{x})^2}{8} + \frac{(5-\bar{x})^2 + (5-\bar{x})^2 + (5-\bar{x})^2}{8} + \frac{(9-\bar{x})^2 + (9-\bar{x})^2}{8}$$

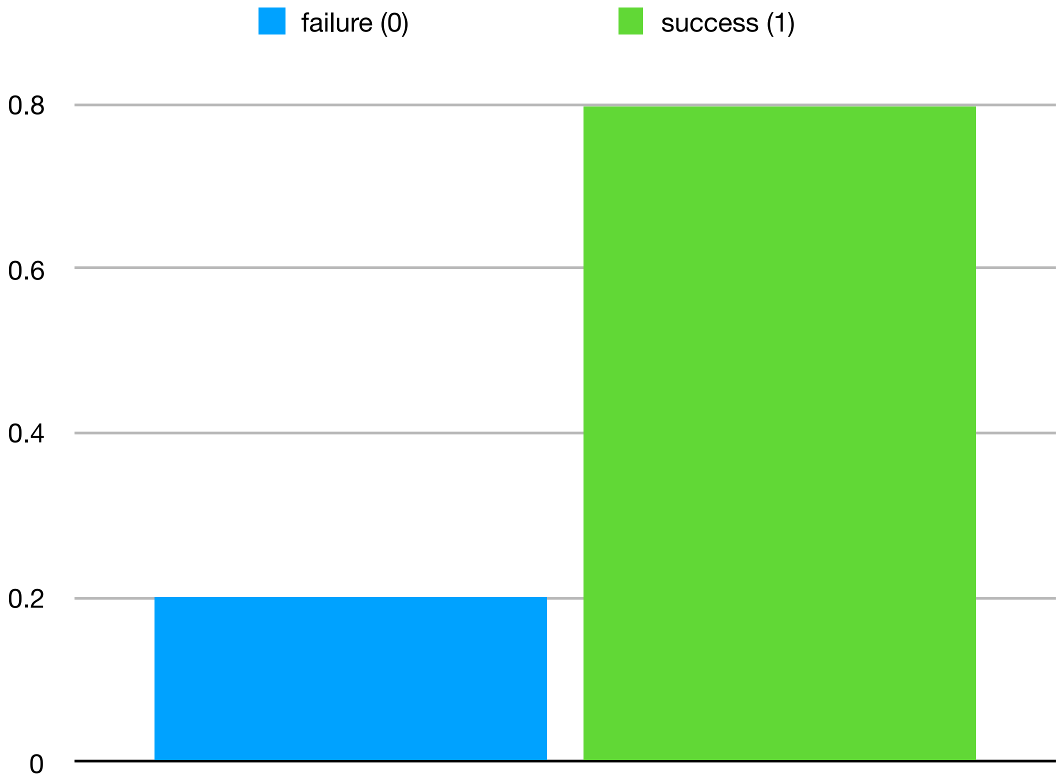
$$\sigma^2 = \frac{1 \cdot (1-\bar{x})^2}{8} + \frac{2 \cdot (3-\bar{x})^2}{8} + \frac{3 \cdot (5-\bar{x})^2}{8} + \frac{2 \cdot (9-\bar{x})^2}{8}$$

$$\sigma^2 = \frac{1}{8} \cdot (1-\bar{x})^2 + \frac{2}{8} \cdot (3-\bar{x})^2 + \frac{3}{8} \cdot (5-\bar{x})^2 + \frac{2}{8} \cdot (9-\bar{x})^2$$

$$\sigma^2 = p_1(x_1-\bar{x})^2 + p_2(x_2-\bar{x})^2 + p_3(x_3-\bar{x})^2 + p_4(x_4-\bar{x})^2$$

$$\sigma^2 = \sum_{i=1}^{n} p_i(x_i - \bar{x})^2$$

Bernoulli Distribution Mean and Variance



ให้ p คือ ความน่าจะเป็นของ success

$1 - p$ คือ ความน่าจะเป็นของ failure

จาก $\bar{x} = \sum_{i=1}^{n} p_i x_i$

$$\bar{x} = (1-p)0 + p(1) = p$$

จาก $\sigma^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$

$$\sigma^2 = (1-p)(0-p)^2 + p(1-p)^2$$

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

$$= p^2 - p^3 + p(1 - 2p + p^2)$$

$$= p^2 - p^3 + p - 2p^2 + p^3$$

$$= p - p^2$$

$$= p(1-p)$$

$Gini\ Impurity$

$S = \{A, A, A, A, B, B, B, C, C, C, C, C\}$

ความแปรปรวนที่จะได้ A หรือ ไม่ได้ A สามารถหาได้จาก $\sigma^2 = p(1-p)$

$$\begin{aligned}\sigma_A^2 &= p_A(1-p_A)\\ &= \frac{4}{12}(1-\frac{4}{12})\end{aligned}$$

ความแปรปรวนที่จะได้ B หรือ ไม่ได้ B สามารถหาได้จาก $\sigma^2 = p(1-p)$

$$\begin{aligned}\sigma_B^2 &= p_B(1-p_B)\\ &= \frac{3}{12}(1-\frac{3}{12})\end{aligned}$$

ความแปรปรวนที่จะได้ C หรือ ไม่ได้ C สามารถหาได้จาก $\sigma^2 = p(1-p)$

$$\begin{aligned}\sigma_C^2 &= p_C(1-p_C)\\ &= \frac{3}{12}(1-\frac{3}{12})\end{aligned}$$

$Gini\ Impurity$ คือการหาผลรวมความแปรปรวนของทุกกรณีที่เกิดขึ้น

$$\begin{aligned}Gini\ Impurity &= \sigma_A^2 + \sigma_B^2 + \sigma_C^2\\ &= p_A(1-p_A) + p_B(1-p_B) + p_C(1-p_C)\\ &= p_A - p_A^2 + p_B - p_B^2 + p_C - p_C^2\\ &= p_A + p_B + p_C - p_A^2 - p_B^2 - p_C^2\\ &= 1 - \sum_{k\in\{A,B,C\}} p_k^2\end{aligned}$$

$$\boxed{Gini\ Impurity = 1 - \sum_{k\in K} p_k^2}$$

ตัวอย่างการคำนวณ $Gini\ Impurity$

<u>ตัวอย่าง 1</u>

$S_1 = \{A, A, A, A, A, A, A, A, A, A\}$

จาก $Gini = 1 - \sum_{k \in K} p_k^2$

$$Gini = 1 - 1^2$$

$$= 0$$

<u>ตัวอย่าง 2</u>

$S_1 = \{A, A, A, A, A, B, B, B, B, B\}$

จาก $Gini = 1 - \sum_{k \in K} p_k^2$

$$Gini = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2$$

$$= 1 - \frac{1}{4} - \frac{1}{4}$$

$$= 0.5$$