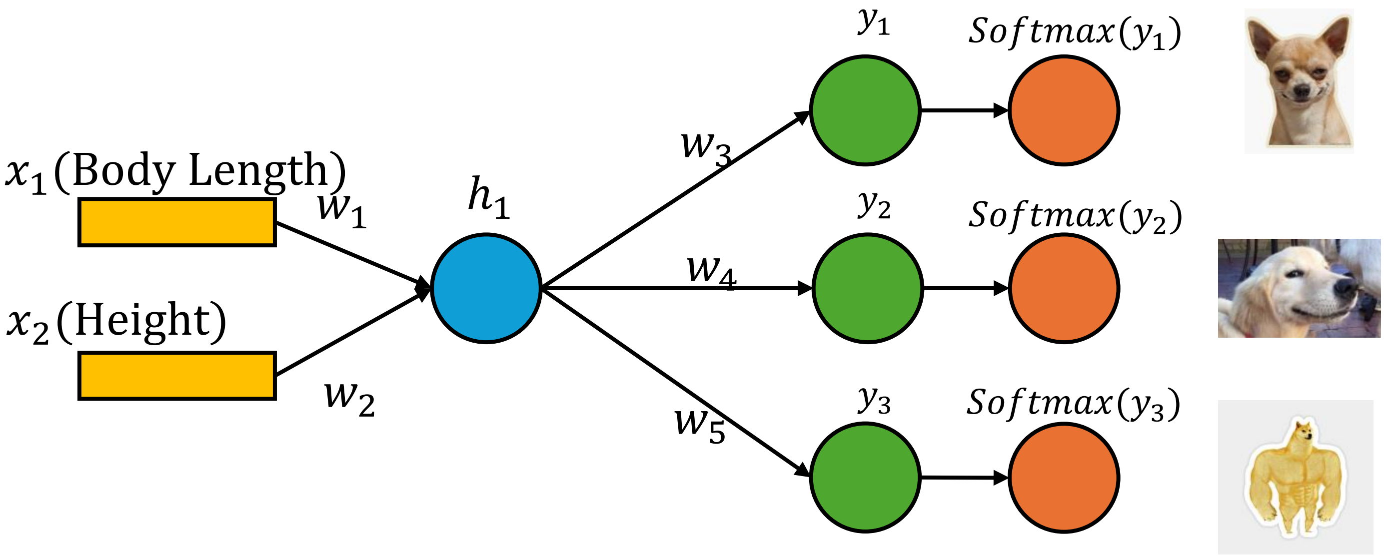
Softmax
Cross Entropy

$$Softmax = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

$$CE = -\sum_{i=1}^{n} Observed \cdot \log(P_i)$$

Classification Neural Network



Why Softmax + Cross Entropy is works with classification???

Conditional Probabilities

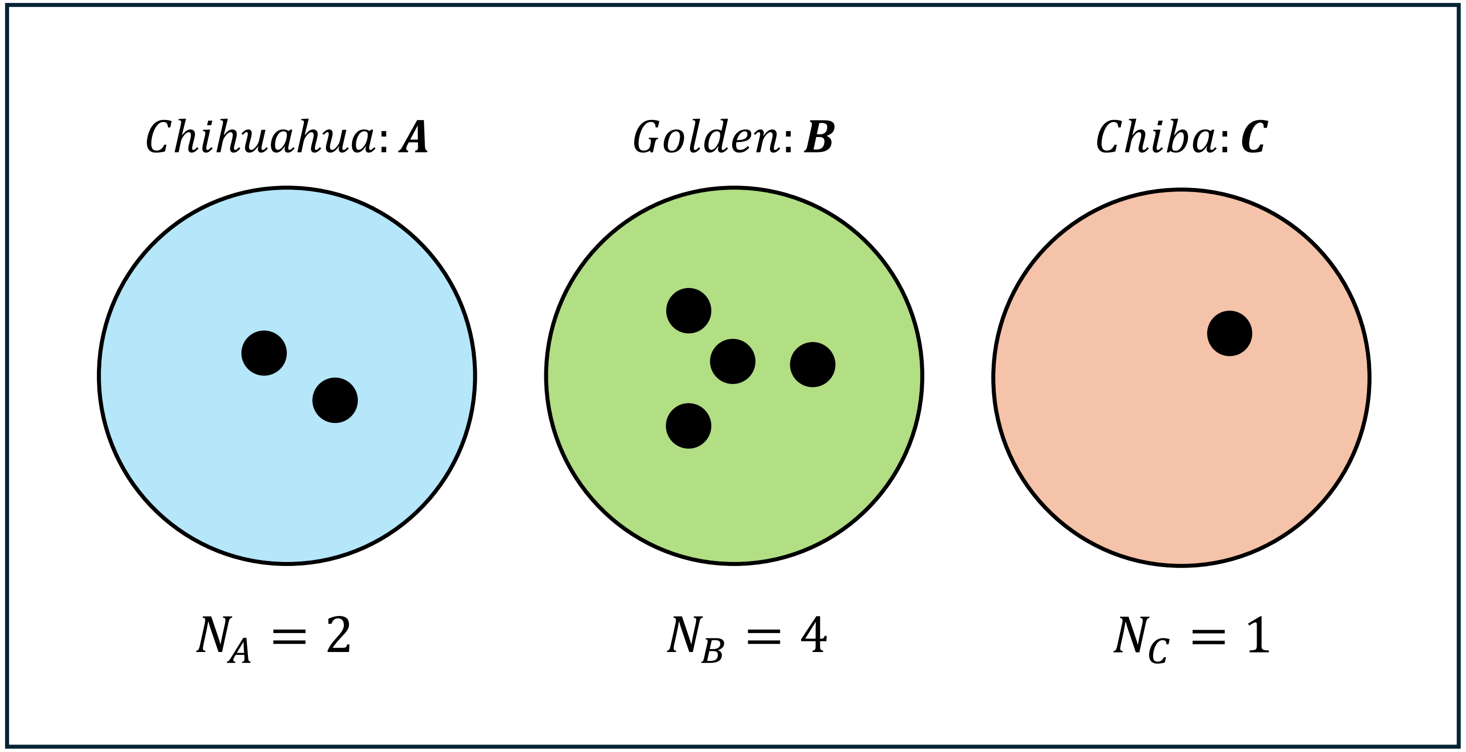
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$Numbers\ of$
$somethings\ you\ interested$

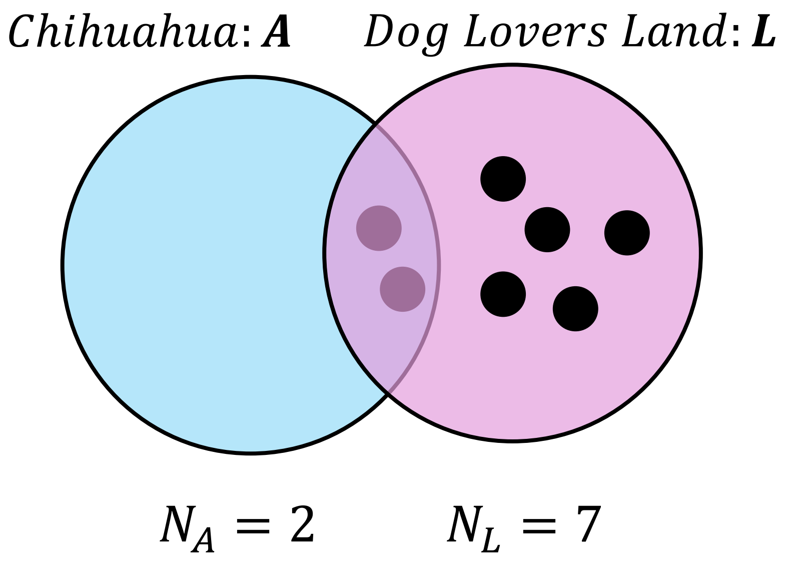
$Numbers\ of$
$what\ we\ know$

Conditional Probabilities

$Dog\ Lovers\ Land.\ (Short\ in\ \boldsymbol{L})$



$$Total\ Populations = N_A + N_B + N_C = 7$$



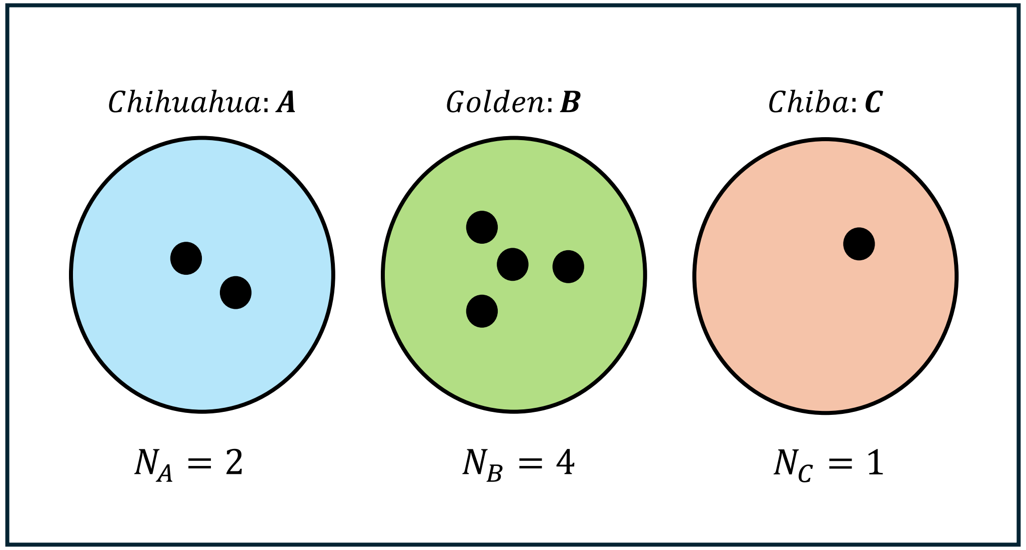
$$P(A|L) = \frac{P(A \cap L)}{P(L)}$$

$$P(A|L) = \frac{Number\ of\ A \cap L}{Number\ of\ L}$$

$$P(A|L) = \frac{2}{7} = 0.29$$

$$Range\ of\ probability: [0, 1]$$

$Dog\ Lovers\ Land.\ (Short\ in\ \boldsymbol{L})$



$$Total\ Populations = N_A + N_B + N_C = 7$$

$$Chihuahua \rightarrow P(A|L) = \frac{2}{7} = 0.29$$

$$Golden \rightarrow P(B|L) = \frac{4}{7} = 0.57$$

$$Chiba \rightarrow P(C|L) = \frac{1}{7} = 0.14$$

$$P(A|L) + P(B|L) + P(C|L) = 0.29 + 0.14 + 0.57$$

$$P(A|L) + P(B|L) + P(C|L) = 1$$

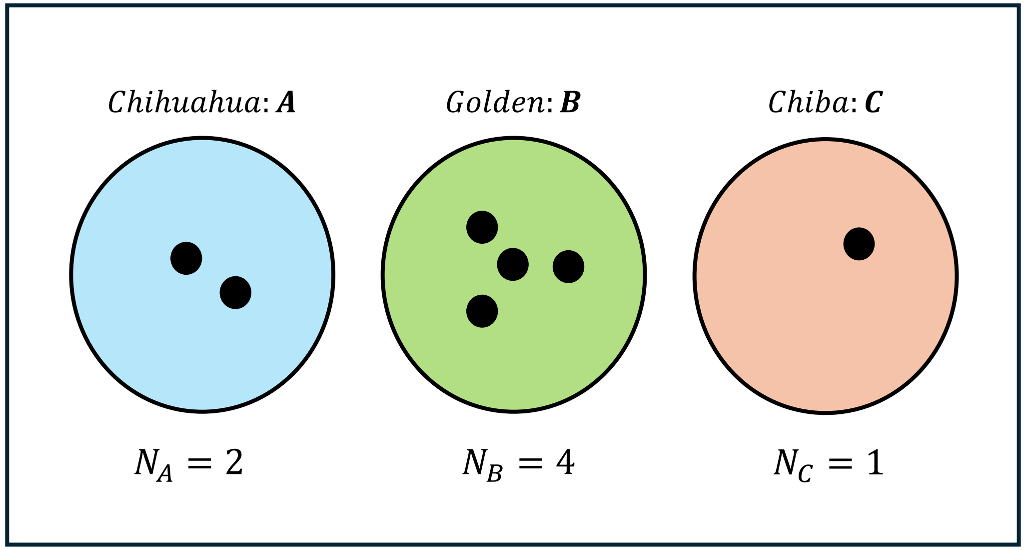
Why we need Softmax ???

Softmax

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

It can normalize outputs to scale within the range of [0, 1].

$Dog\ Lovers\ Land.\ (Short\ in\ \boldsymbol{L})$



$$Total\ Populations = N_A + N_B + N_C = 7$$

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

$$Softmax(x_A) = \frac{e^{x_A}}{e^{x_A} + e^{x_B} + e^{x_C}}$$

$$Softmax(x_B) = \frac{e^{x_B}}{e^{x_A} + e^{x_B} + e^{x_C}}$$

$$Softmax(x_C) = \frac{e^{x_C}}{e^{x_A} + e^{x_B} + e^{x_C}}$$

$$Softmax(x_A) = \frac{e^{x_A}}{e^{x_A} + e^{x_B} + e^{x_C}} = \frac{e^{0.29}}{e^{0.29} + e^{0.57} + e^{0.14}} = 0.31$$

$$Softmax(x_B) = \frac{e^{x_B}}{e^{x_A} + e^{x_B} + e^{x_C}} = \frac{e^{0.57}}{e^{0.29} + e^{0.57} + e^{0.14}} = 0.42$$

$$Softmax(x_C) = \frac{e^{x_C}}{e^{x_A} + e^{x_B} + e^{x_C}} = \frac{e^{0.14}}{e^{0.29} + e^{0.57} + e^{0.14}} = 0.27$$

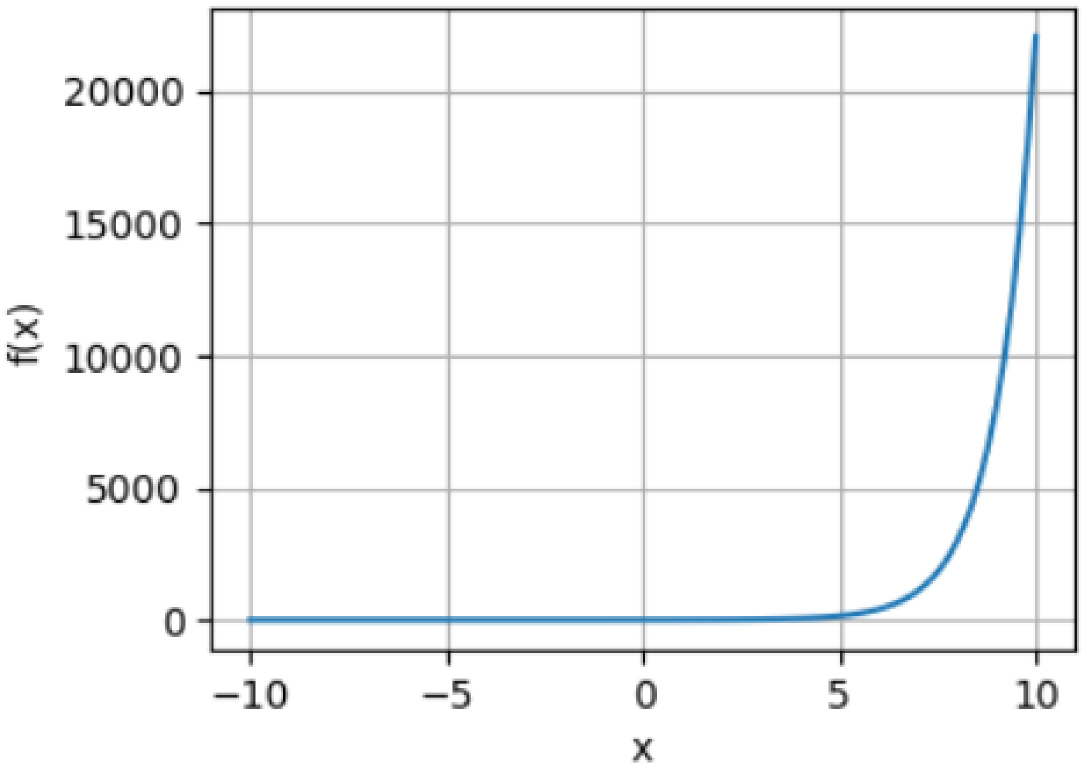
$$Softmax(x_A) + Softmax(x_B) + Softmax(x_C) = 1.0$$

If you just need to scale it into range of [0,1]

Why not we just use normalization ???

The output from sum of node
in neural network maybe
negative number

$$f(x) = e^x$$



$$Range\ of\ output\ always\ be: (0, +\infty)$$

Furthermore, the **Softmax** can boost the value of output to closer of the expectation.

Example

$$\boldsymbol{Output} = [10, 20]$$

$$\boldsymbol{Normalization} = \left[\frac{10}{10+20}, \frac{20}{10+20}\right] = [0.333 \ldots, 0.666 \ldots]$$

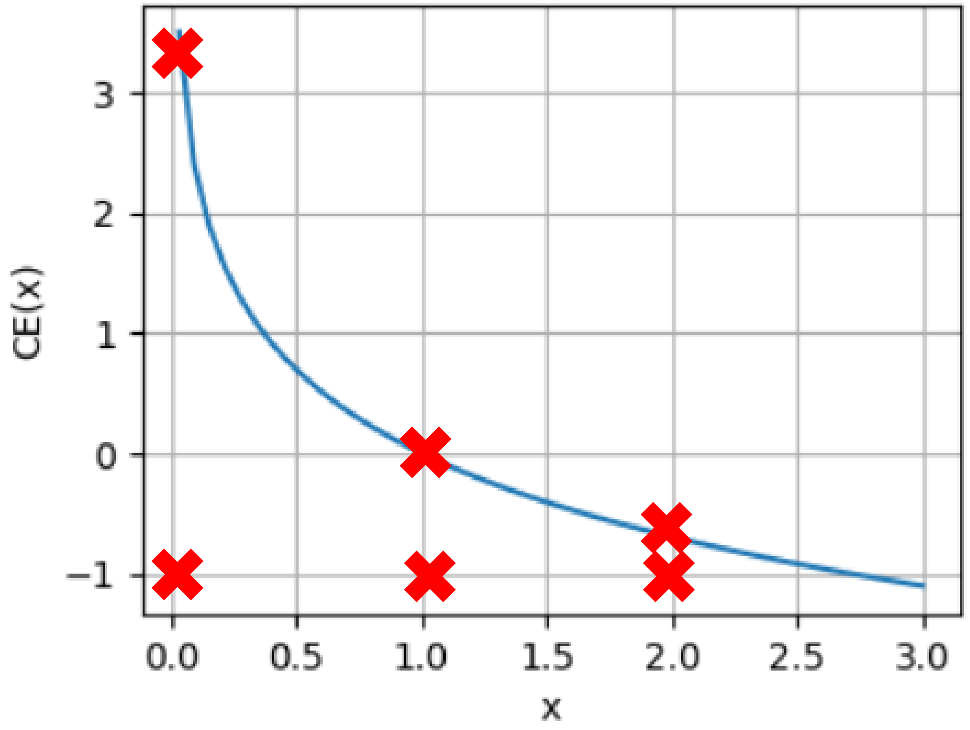
$$\boldsymbol{Softmax} = \left[\frac{e^{10}}{e^{10}+e^{20}}, \frac{e^{20}}{e^{10}+e^{20}}\right] = [0.00 \ldots, 0.999 \ldots]$$

Next, Why we always use **Cross Entropy** with **Softmax** ???

Cross Entropy

$$CE = -\sum_{i=1}^{n} Observed \cdot \log(P_i)$$

$$CE = -\sum_{i=1}^{n} Observed \cdot \log(Softmax_i)$$



Therefore, If we try to minimize the **Cross Entropy**
The error is going to less.