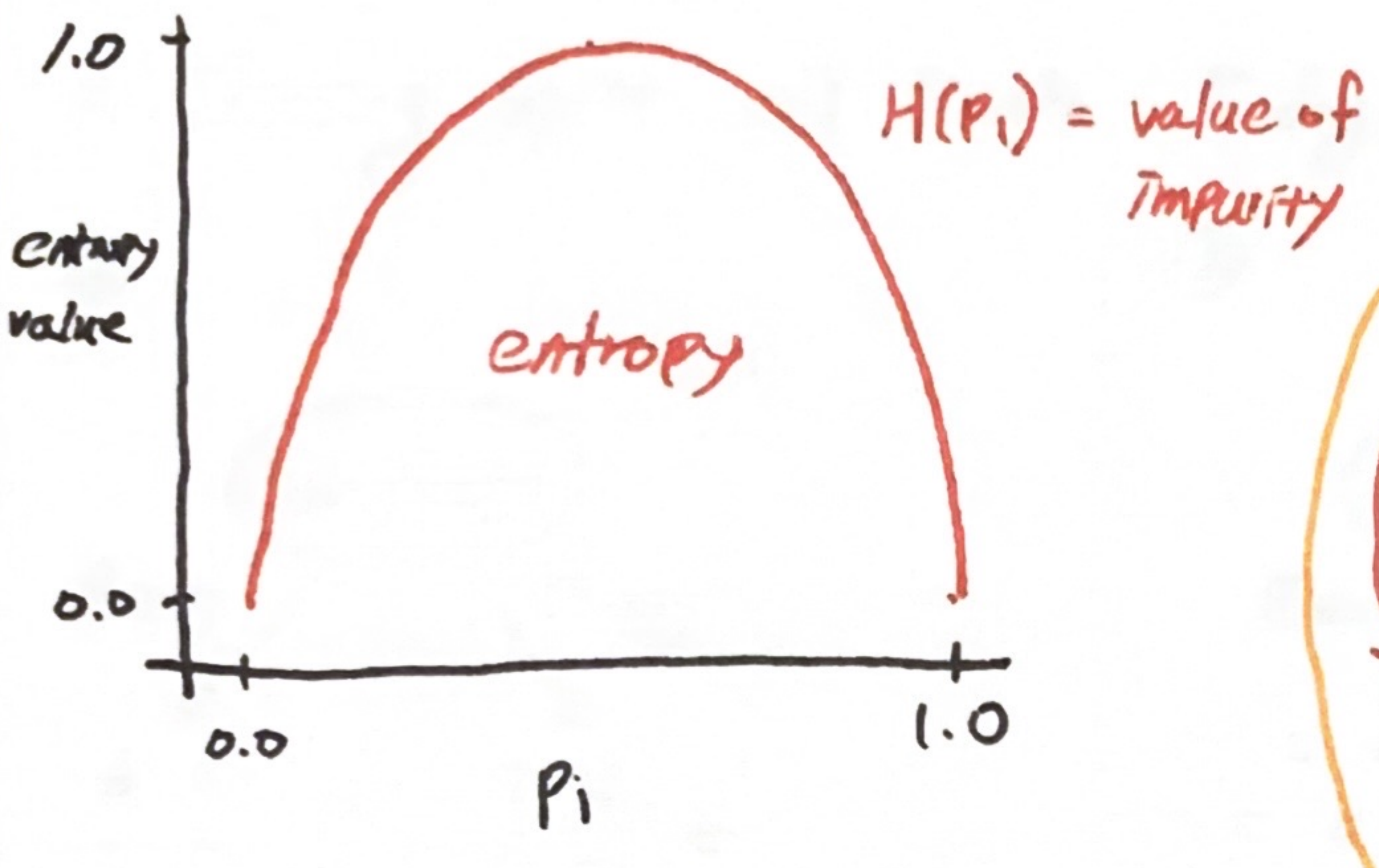
〈Decision Tree - Measuring purity)

- entropy : a measure of the impurity of a set of data

ex) 3 cats, 3 dogs

P_1 = fraction of examples that are cats



① dog dog dog dog dog dog ⇒ $P_1 = 0$ $H(P_1) = 0$
② cat cat dog dog dog dog ⇒ $P_1 = 2/6$ $H(P_1) = 0.92$
③ cat cat cat dog dog dog ⇒ $P_1 = 3/6$ $H(P_1) = 1$
④ cat cat cat cat cat dog ⇒ $P_1 = 5/6$ $H(P_1) = 0.65$
⑤ cat cat cat cat cat cat ⇒ $P = 6/6$ $H(P_1) = 0$

entropy (value of impurity) is the highest when set of examples is 50 : 50 = "most impure"

entropy is the lowest when set of examples is either all positive or all negative = "most pure"

* Actual equation for the entropy function

P_1 = fraction of positive examples

$P_0 = 1 - P_1$

$$H(P_1) = -P_1 \log_2(P_1) - P_0 \log_2(P_0)$$

$$= -P_1 \log_2(P_1) - (1 - P_1) \log_2(1 - P_1)$$

if P_1 or P_0 $= 0$

$\because 0 \log(0) = 0 \Rightarrow$ entropy $= 0$