Decision Tree Splitting Heuristics

1) Gini Impurity
2) Entropy

Our goal: Find feature that creates splits that are as homogenous as possible

we can measure heterogeneity of data with Gini impurity or entropy

Gini Impurity: what is the probability of getting two different classes when sampling twice with replacement?

Exc: 7 red, 3 blue

P(red) = 0.7

$P(\text{blue}) = 0.3$

$P(\text{blue}, \text{blue}) = 0.3^2$

$P(\text{red}, \text{red}) = 0.7^2$

$P(\text{same}, \text{same}) = 0.3^2 + 0.7^2$

$P(\text{different draws}) = 1 - (0.3^2 + 0.7^2)$

$\therefore$ Gini Impurity $= 1 - \sum_{i \in C} p_i^2$

If $P(\text{red}) = 1$

Then $GI = 1 - (1^2 + 0^2)$

$= 0$

No impurity, makes sense

$$\text{Entropy} = -\sum_{i \in C} p_i \log p_i$$

Measures chaos / disorder

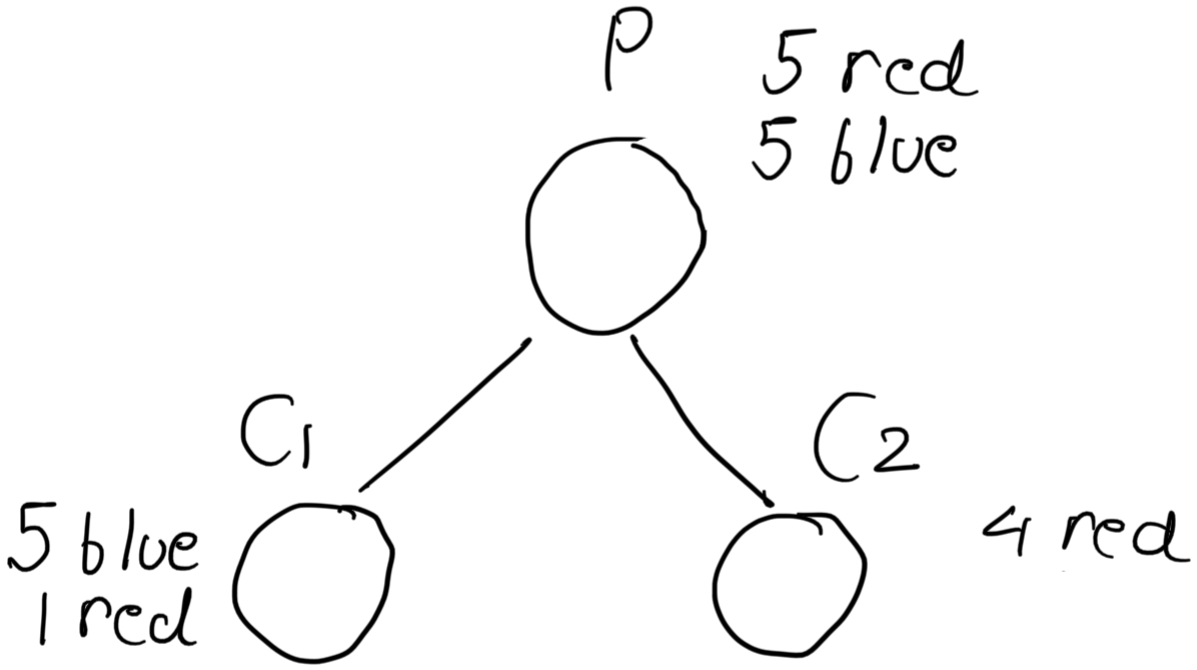
E(sample) is max when

$p = 0.5$ (very mixed)

If $P(\text{red}) = 1$

$$E = -1\log 1 - 0\log 0$$
$$= 0$$

0 chaos, makes sense!



Let H = GI or E

want to calculate "gain" for feature split using H

Heterogeneity before = H(P)

Heterogeneity before = H(P)

Heterogeneity after = H(C)

= weighted combination of H(C1), H(C2)

where weight = proportion of samples that ended up at that child

$$\boxed{\text{Gain} = H(P) - H(C)}$$

$$= H(P) - (0.6\,H(C_1) + 0.4\,H(C_2))$$

Find feature split that produces highest gain, and recursively and greedily continue building tree that way!