02/14/2023

Entropy:

$$-\sum_i prob_i \log(prob_i)$$

$$prob_i = \frac{N(i, D)}{N(D)}$$

$H(D)$

Gini:

$$\sum_i prob_i (1 - prob_i)$$

information gain

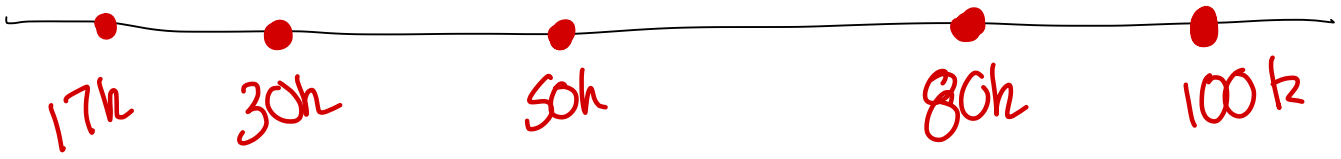
$$H(P) - \left[\frac{N(P_l)}{N(P)} H(P_l) + \frac{N(P_r)}{N(P)} H(P_r)\right]$$

impurity of split:

$$\frac{N(P_l)}{N(P)} H(P_l) + \frac{N(P_r)}{N(P)} H(P_r)$$

Choose threshold:

1) consider threshold value between each data point (N-1 thresholds to consider)



2) large # of continuous values
   - ↳ randomly choose subset of data
   - ↳ consider thresholds as midway points of data in subset

employed 0

unemployed 1

self-emobyed 2

---

red T

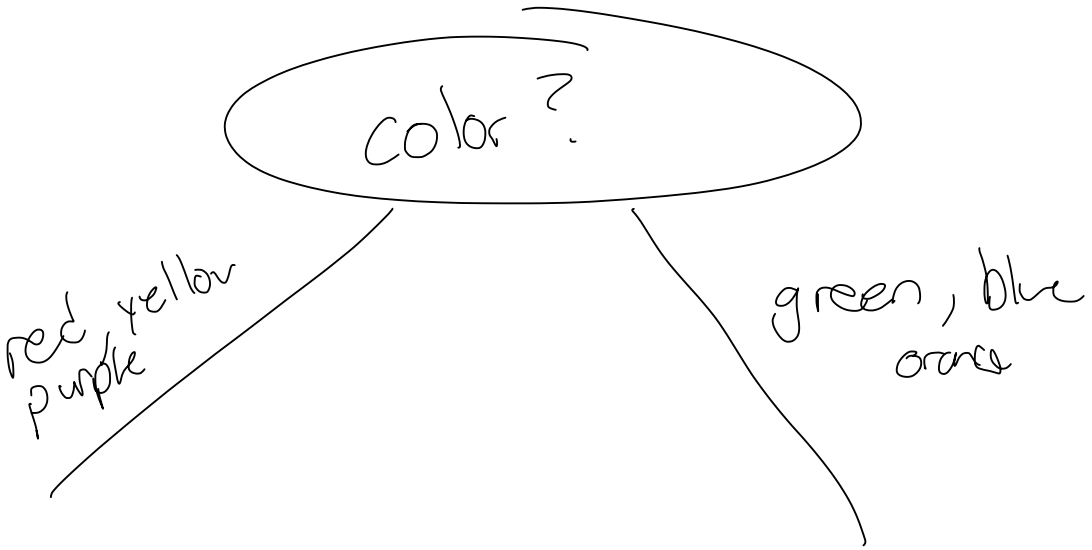
green T

blue T

yellow H

purple T

orange H



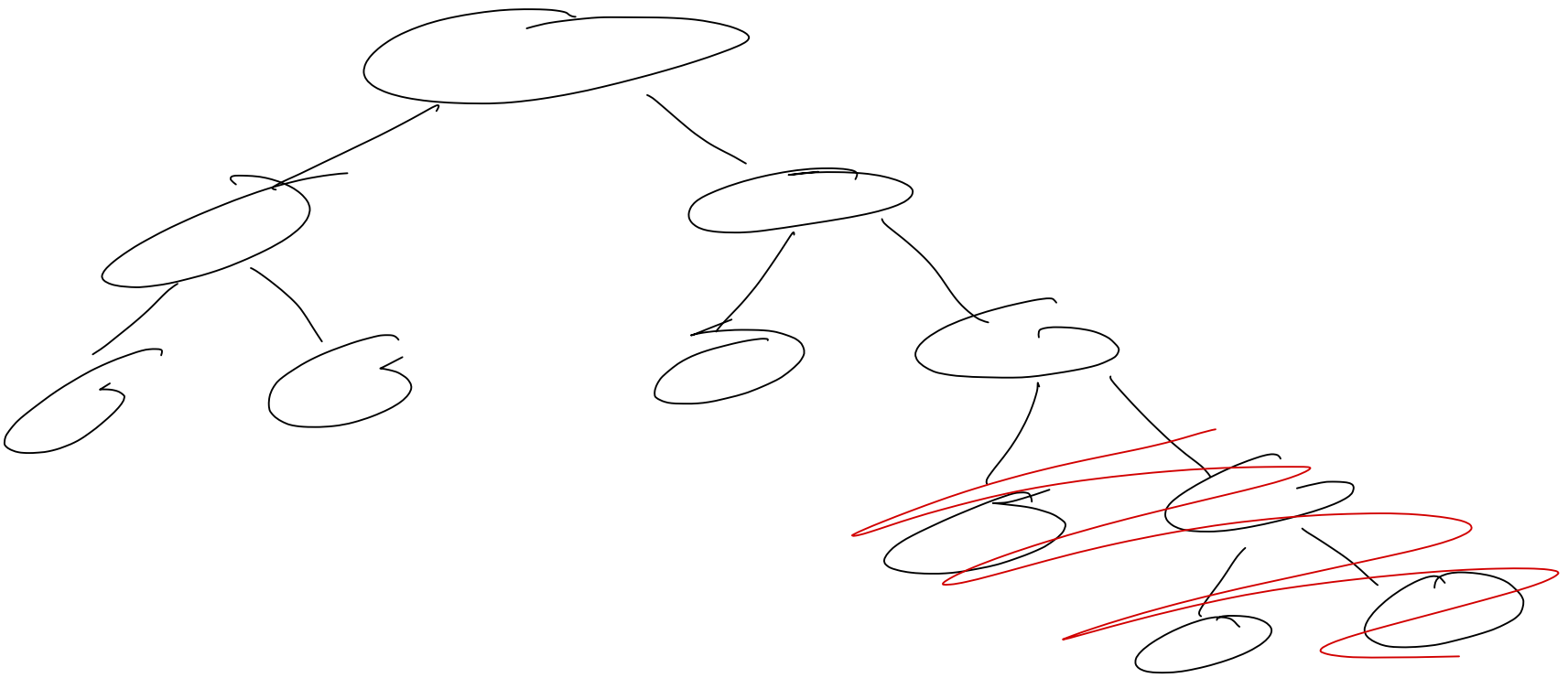
weight	length	crocodile or snake?
2200	18	crocodile
300	17	snake
500		
2200	18	snake

When to stop:

1) if pool is too small

2) if all items in a pool have same class

3) if depth has reached limit



Random Forest

1) at leaf, record a vote

↳ choose class with most votes

2) at leaf, record N_e votes for each label that occurs

N_e: # of times label appears in data at leaf

↳ choose label with votes