Entropy

Now

Consider a set $S = \{y_1 \cdots y_n\}$ n elements

Let K be the no: of different classes

Now let $P_S(i)$ be the probability that if you randomly select a label y' from S it will belong to class 'i'

So $P_S(i) = \frac{\text{No: of elements of class } i}{\text{Total no: of elements in } S}$

Now

Consider the worst possible case.

It will be $P_S(1) = P_S(2) \cdots P_S(K) = 1/K$

i.e equal probability to all classes

it is the most impure set

We want our set to be as far as possible from this set.

Now

let the worst possible set with 'K' classes in set Q be $P_Q(1) = P_Q(2) = \cdots = P_Q(K) = 1/K$

Let our set be P with 'K' classes

and probabilities = $P_P(1), P_P(2) \ldots, P_P(K)$

Now

consider a function

~~$\sum_{i=1}^{k} P_i \log \frac{P_i}{Q_i}$~~

$$f(P,Q) = \sum_{i=1}^{K} P_p(i) \log \left(\frac{P_p(i)}{P_Q(i)} \right)$$

$\Rightarrow$ Let us not go into the details of why this particular function was chosen

$\Rightarrow$ Let us understand what this functions tells us

If $P_p(i)$ are equal to $P_Q(i)$ then $f(P,Q) = 0$

So the more different $P_p(i)$ are from $P_Q(i)$ the larger the function will be.

$\Rightarrow$ Also it can mathematically shown that $f(P,Q) \geq 0$ (always) but let us not go too deep.

$\Rightarrow$ Now our aim is to maximize $f(P,Q)$

$$\Rightarrow \max \left(\sum_{i=1}^{K} P_p(i) \log \left(\frac{P_p(i)}{P_Q(i)} \right) \right)$$

$$\Rightarrow \max \left(\sum_{i=1}^{K} P_p(i) \log P_p(i) - P_p(i) \log (P_Q(i)) \right)$$

$$\Rightarrow \max \left(\sum_{i=1}^{K} P_p(i) \log P_p(i) - \log \sum_{i=1}^{K} P_p(i) \log(1/k) \right)$$

$$\Rightarrow \max \left(\sum_{i=1}^{K} P_p(i) \log P_p(i) - \underbrace{\log(1/k)}_{\text{constant}} \right)$$

$$\Rightarrow \max \left(\sum_{i=1}^{K} P_p(i) \log P_p(i) \right)$$

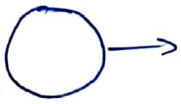
$$\Rightarrow \quad \min \left(-\sum_{i=1}^{k} P_p(i) \log\left(P_p(i)\right) \right)$$

Let for set P

$$E(P) = -\sum_{i=1}^{K} P_p(i) \log\left(P_p(i)\right)$$

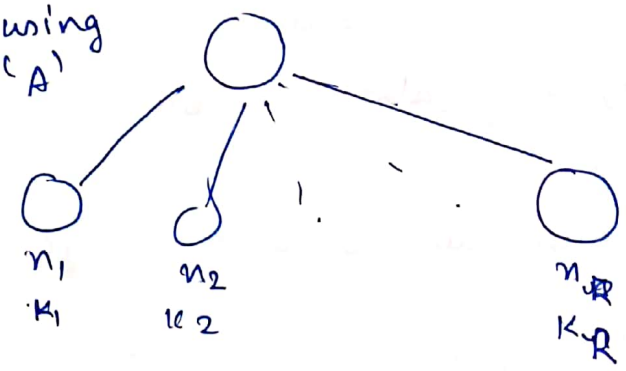
This is our entropy as we want to minimize it.

Now information Gain (IG)

let  node with n samples and K diff types of outcomes

we can split it in many ways depending on our features of our dataset.

Let us split using feature 'A'



Now $E(n)$ ⟶ entropy of parent node

$E(n_1) \ldots E(n_R)$ → entropy of child nodes

Now

IG of this split due to feature 'A' is given as

$$IG(A) = E(n) - \sum_{i=1}^{R} \frac{n_i}{n} E(n_i)$$

Weight of given child node.

Now

if gain is high

it means, ~~entropy of~~ weighted average of entropies of child nodes is much lower than parent node

So we choose feature with max gain and split it.

Note → Gain can be +ve, -ve or zero also.