Shannon Entropy, Information Gain and Tree Construction

(Shannon) Entropy originates from the field of information theory. The basic intuition behind information theory is, that learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.

Let us denote the amount of information of an event with probability p as $i(p)$. and It should fulfill:

(1) $i(p) \geq 0$ decreasing (more info for unlikely events)

(2) $i(pq) = i(p) + i(q)$ (independent events)

It can be shown that [Shannon]:

$$\boxed{i(p) = \log\left[\frac{1}{p}\right]}$$

Note: The base of the logarithm can be chosen freely! If the base 2 is used, then the units of information are bits:

$$\boxed{i(p) = \log_2\left[\frac{1}{p}\right] \quad \text{in } \underbrace{\text{bits}}_{\text{binary digit}}}$$

[$\log_{10} \Rightarrow$ dit, $\log_e \rightarrow$ nat]

Now consider a random variable X with probability distribution $p(x)$. The amount of information of an elementary event $X = x$ is $i(p(x)) = \log\left[\frac{1}{p(x)}\right]$

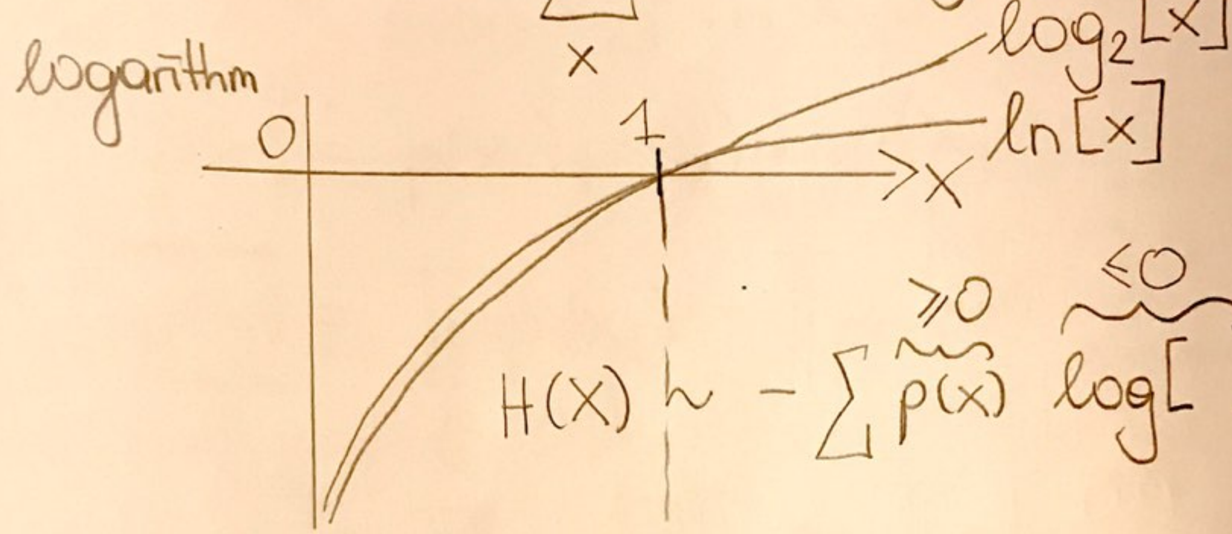
The average amount of information about X is given by the expectation value:

$$\boxed{\underbrace{H(X)}_{\substack{\text{not a function} \\ \text{of } X \text{ but } p(x): \\ H(p(x))}} = \sum_x p(x) \log\left[\frac{1}{p(x)}\right]}$$

Shannon's entropy of the random variable X with distribution $p(x)$.

We arrive at: $\left(\log\frac{1}{x} = -\log x\right)$

$$H(X) = -\sum_{x} p(x)\log[p(x)]$$



$$H(X) \sim -\sum \underbrace{p(x)}_{\geq 0}\ \underbrace{\log[\ ]}_{\leq 0} \geq 0$$

$$\Rightarrow \quad H(X) \geq 0$$

Side remark: It can be shown that Boltzmann's entropy S of statistical thermodynamics (= the disorder of a system increases) is formally identical with H (Shannon entropy).

Example: X takes M equally probable values:

$x = 0, 1, \ldots, M-1 \Rightarrow p(x) = \frac{1}{M}$

$$H(X) = -\sum_{i=1}^{M} \frac{1}{M}\log\left[\frac{1}{M}\right] = \underbrace{\log[M]}_{\text{log of number of possible values}}$$

How does $H(X)$ arise from a given distribution $p(x)$:

$$x = [x_1, x_2, \ldots x_n] \quad \underbrace{\text{i.i.d}}_{\text{independent and identically distr.}}$$

$$p(x) = p(x_1)p(x_2)\ldots p(x_n)$$

$$\boxed{n(x) = \begin{cases} 1 & x = x_i \\ 0 \end{cases}}$$

number of x_i

$$\boxed{p(x)} = \prod_i p(x_i) = \prod_x p(x)^{n(x)}$$

$$= \prod \left[p(x)^{\frac{n(x)}{n}}\right]^n \approx \prod \left[p(x)^{p(x)}\right]^n$$

$$\boxed{\text{Asymptotic equipartition property}} \quad \stackrel{!}{=} \boxed{e^{-nH(X)}}$$

for very large n the value of the probability of a sequence $x = (x_1 \ldots x_n)$ is $\approx e^{-nH(X)}$

Proof:

$$e^{-nH(X)} = e^{-n\sum_x p(x)\log\left[\frac{1}{p(x)}\right]} = e^{+n\sum p(x)\log[p(x)]}$$

$$= e^{+\sum_x \log\left([p(x)]^{p(x)n}\right)}$$

$$= e^{+[\log[f(x_1)] + \log[f(x)] + \ldots]}$$

$$= e^{+\log\left[\prod[p(x)]^{p(x)n}\right]}$$

$$= + \prod\left[p(x)^{p(x)}\right]^n$$

$$\boxed{\begin{array}{l} p(x) \overset{n \text{ large}}{\approx} e^{-nH(X)} \\ \text{for very large } n \\ p \text{ close to a} \\ \text{constant } e^{-nH}! \end{array}}$$

Continuous Entropy:

$$h(X) = \mathbb{E}\left(\log\left[\frac{1}{p(X)}\right]\right)$$

$$= \int p(x) \log\left[\frac{1}{p(x)}\right] dx$$

Example: Uniform distribution U in interval (a, b)

$$p(x) = \frac{1}{b-a}:$$

$$h(U) = \int_a^b \frac{1}{b-a} \log[b-a]\, dx$$

$$= \log[b-a] \frac{1}{b-a} \underbrace{\left[\int_a^b dx\right]}_{1\big|_a^b = b-a}$$

$$= \log[b-a]$$

If $b-a < 1 \Rightarrow h(U) < 0$

continuous entropy can become negative!

(one cannot assign an "amount of information" to continuous entropy)

Exercise: $h(X)$ of the normal distribution?

Answer: $h(X) = \log[\sqrt{2\pi e}] + \log\sigma$

Conditional Entropy: (measure of information provided by Y about X)

Mutual Information/Information Gain

$$\underbrace{I(X;Y)}_{\text{Random Variables}} = \mathbb{E}\left(\log\left[\frac{p(X|Y)}{p(X)}\right]\right)$$

with $p(x|y)$ conditional distrib. of X knowing $Y=y$.

$$I(X;Y) = \mathbb{E}\left(\log[p(X|Y)] - \log[p(X)]\right)$$

$$= \underbrace{\mathbb{E}\left(\log\left[\frac{1}{p(X)}\right]\right)}_{H(X)} - \underbrace{\mathbb{E}\left(\log\left[\frac{1}{p(X|Y)}\right]\right)}_{H(X|Y)}$$

$$= H(X) - H(X|Y)$$

$$H(X|Y) = \sum_x \sum_y p(x,y) \log\left[\frac{1}{p(x|y)}\right]$$

Let us rewrite $H(X|Y)$:

$$H(X|Y) = \sum_y p(y) \sum_x p(x|y) \log\frac{1}{p(x|y)}$$

$$H(X|Y) = \sum_{y} p(y)\, H(X|Y=y)$$

$H(X|Y)$ is the average uncertainty about X when Y is known

$$\boxed{H(X|Y) \leq H(X)}$$ knowledge reduces uncertainty!

---

2. Entropy Calculation and Decision Trees

id	h [m] (height)	l (legs)	class
1	0.1	0	Fish
2	0.2	2	Bird
3	1.8	2	Human
4	0.2	4	Cat
5	2.1	4	Horse
6	1.7	2	Human
7	0.1	4	Cat
8	1.6	2	Human

Entropy of the class information:

$$H(X) = -\sum_{x} p(x) \log_2 [p(x)]$$

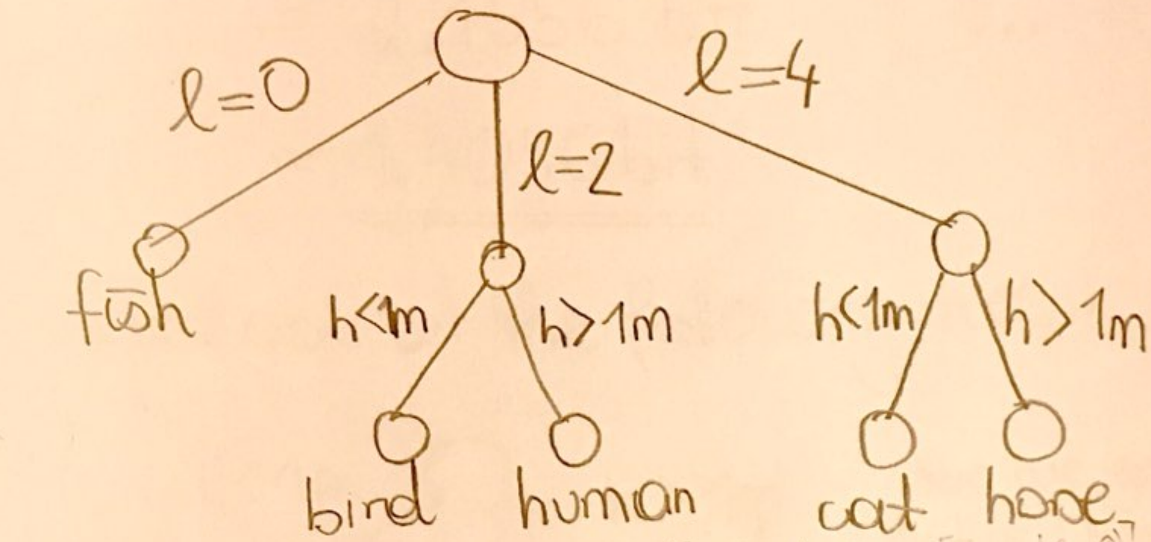
$$= -\sum_{Class=\{F, B, H, C, HO\}} p(Class) \log_2 [p(Class)]$$

$$= -\Big\{ \underbrace{\frac{1}{8} \log_2 \left[\frac{1}{8}\right]}_{F = Fish} + \underbrace{\frac{1}{8} \log_2 \left[\frac{1}{8}\right]}_{B = Bird}$$

$$+ \underbrace{\frac{3}{8} \log_2 \left[\frac{3}{8}\right]}_{H = Human} + \underbrace{\frac{2}{8} \log_2 \left[\frac{2}{8}\right]}_{C = Cat}$$

$$+ \underbrace{\frac{1}{8} \log_2 \left[\frac{1}{8}\right]}_{HO = Horse} \Big\} \simeq 2.1556 \text{ bit}$$

Now let us construct a tree with root node l:



$$-\sum (Class|l=0)\log[Class|l=0]$$

$$H(Class|l=0) = -1\cdot\log_2[1] = \underline{0}$$

$$H(Class|l=2) = -\sum_{class=\{B,H\}} p(Class|l=2)\log_2(Class|l=2)$$

$$= -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \approx \underline{0.8113\,bit}$$

$$H(Class|l=4) = -\sum_{Class=\{C,HO\}} p(Class|l=4)\log(Class|l=4)$$

$$= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = \underline{0.9183\,bit}$$

$$H(Class|l) = \frac{1}{8}\cdot H(Class|l=0) + \frac{4}{8}H(Class|l=2) + \frac{3}{8}H(Class|l=4)$$

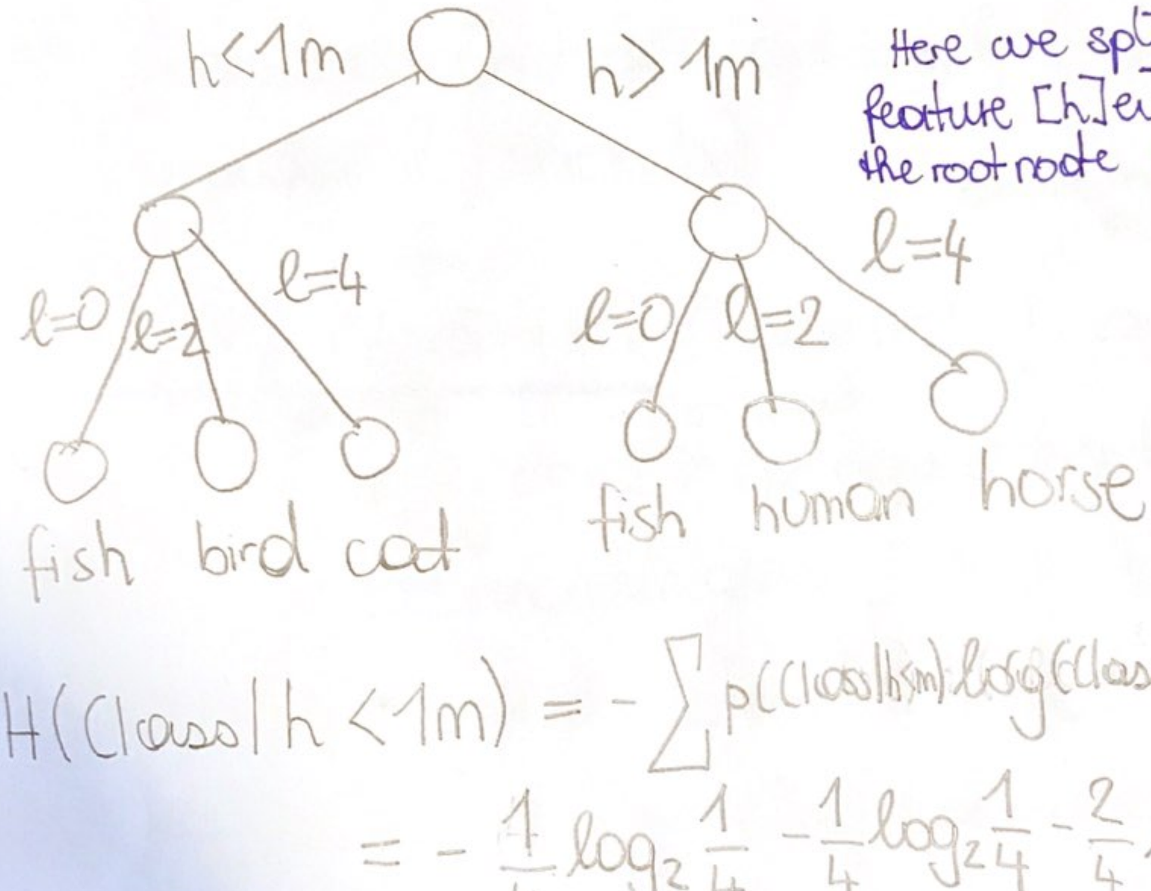
Expected Information Gain:

$$I(Class, \ell) = \underbrace{H(Class)} - H(Class|\ell)$$

$$= 2.1556 \text{ bit} - \dots =$$

$$\approx \underline{\underline{1.4056 \text{ bit}}}$$

Now look at the following tree:



Here we split by feature [h]eight at the root node

$$H(Class|h<1m) = -\sum p(Class|h<m) \log(Class|h<m)$$

$$= -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$\approx 1.5 \text{ bit}$$

$$H(Class|h>1m) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$\approx 0.8113 \text{ bit}$$

$$I(\text{Class};h) = \underbrace{H(\text{Class})} - H(\text{Class}|h)$$

$$= 2.1556\,\text{bit} - \underbrace{\left[\frac{4}{8}H(\text{Class}|h<m) + \frac{4}{8}H(\text{Class}|h \geq m)\right]}_{\approx 1\,\text{bit}}$$

$$\approx 1\,\text{bit}$$

$\Rightarrow I(\text{Class};\ell) > I(\text{Class};h)$

$\Rightarrow$ first decision tree is optimal (= first consider ℓ, then h)

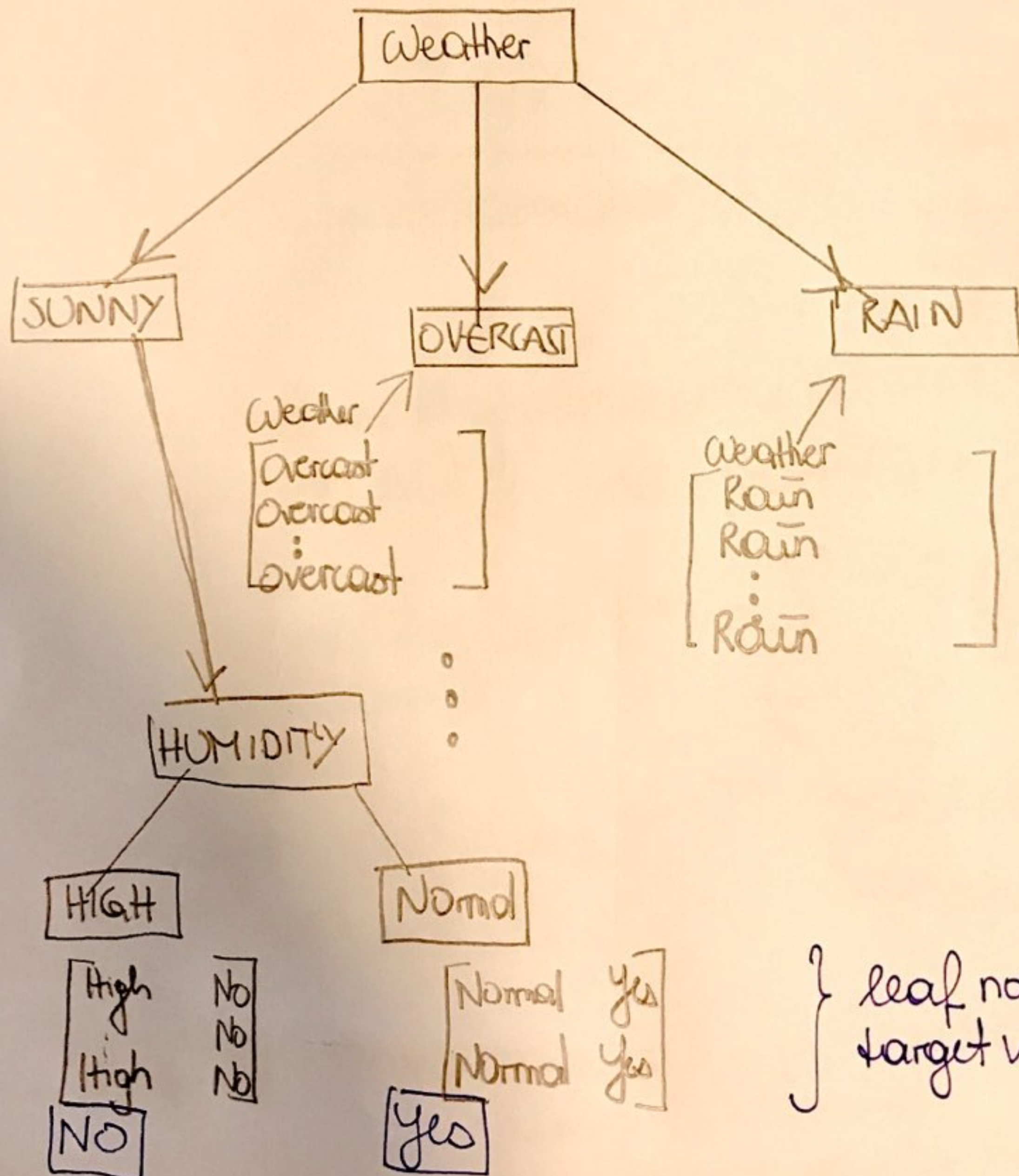
$\Downarrow$

divide into opposite things

The ID3 algorithm (Iterative Dichotomiser 3)

1. If $\underset{\text{feature set}}{I}$ is empty or all $\underset{\text{target}}{Y}$ are equal then terminate
2. Compute Information gain $I_i(X;Y)$ $\forall i \in I$
3. determine winner feature $j = \operatorname{argmax}\{I_i(X,Y)\}$
4. partition X, Y into disjoint subsets for each root note
$$X_i = \{x_k \in X \mid x_k^{(j)} = i\} \quad i = 1, \dots v_j$$
5. create new node N_i
6. back to 1.

DAY	Weather	Humidity	Wind	Play
D1	sunny	high	weak	No
	⋮	⋮	⋮	yes
				⋮
				yes



splitting X,Y into disjoint subsets

leaf node = all target values equal

ID3 has bottlenecks, like requiring discrete features and a tendency to overfit. ← $I(X;Y)$ maximal if feature has all attributes different

An improvement is C4.5. Here instead of $I(X;Y)$ a normalized version is used:

$$\frac{I(X;Y)}{\text{SplitInformation}(X;Y)} = \text{GainRatio}(X;Y)$$

~expected reduction in entropy by knowing attribute Y

information generated by splitting the training set X into i partitions

$$\underset{(\text{Intrinsic Value})}{\text{SplitInformation}(X;Y)} = -\sum_{X} \underbrace{\frac{|X_i|}{|X|}}_{\text{fraction of examples}} \log_2\left[\frac{|X_i|}{|X|}\right]$$

for example: $14 \cdot \left[-\frac{1}{14}\log_2\left[\frac{1}{14}\right]\right] \simeq 3.807$

day	class
1.6	C0
2.6	C0
⋮	⋮
14.6	C1

⇒ reduce bias towards multi-valued attributes

⇒ more reliable than Information Gain

Day	Wind	Class
1	Weak	No
2	Strong	No
3	Weak	yes
4	Weak	yes
5	Weak	yes
6	Strong	No
7	Strong	yes
8	Weak	No
9	Weak	yes
10	Weak	yes
11	Strong	yes
12	Strong	yes
13	Weak	yes
14	Strong	No

Example:

$$H(X) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$$

$$= 0.940 \text{ bits}$$

$$H(\text{Class}|\text{Wind}=\text{weak}) = -\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8}$$

$$= 0.811 \text{ bits}$$

$$H(\text{Class}|\text{Wind}=\text{strong}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

$$= 1 \text{ bits}$$

$$H(\text{Class}|\text{Wind}) = \frac{8}{14}H(\text{Class}|\text{Wind}=\text{weak}) + \frac{6}{14}H(\text{I st})$$

$$\Rightarrow I(X;Y) = 0.940 \text{ bits} - \dots = \underline{0.049 \text{ bit}}$$

$$\text{SplitInfo}(\text{Class};\text{Wind}) = -\underbrace{\boxed{\frac{8}{14}}}_{}^{\text{8 weak wind}}\log_2\frac{8}{14} - \boxed{\frac{6}{14}}^{\text{6 x strong wind}}\log_2\frac{6}{14}$$

$$\Rightarrow \text{GainRatio}(X;Y) = \frac{0.049}{0.985} \simeq 0.049$$

CART (Classification and regression tree) uses Gini index as an attribute selection measure to build a decision tree: Idea: $p_i = 1 \Rightarrow 0$; $p_i = 0 \Rightarrow 0 \Rightarrow 0$ for pure splits

$$\sum_{i=1}^{m} [p_i(1-p_i)] = \sum (p_i - p_i^2) =$$

$$= \sum_{i=1}^{m} p_i - \sum_{i=1}^{m} p_i^2 = 1 - \sum_{i=1}^{m} p_i^2 =$$

$$= \text{Gini Impurity}(X, Y)$$

$$\text{Gini Index} = \sum \frac{|X_i|}{|X|} \text{Gini Impurity}(X, Y)$$

Attribute	Class
M	C0
M	C0
F	C1
F	C0
M	C1

for M:

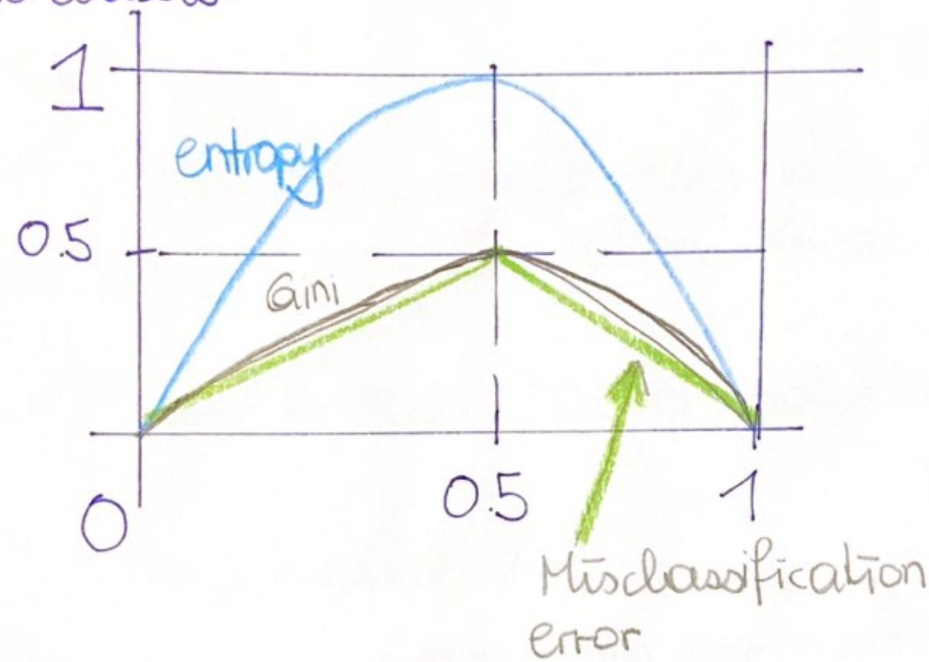
$$1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = A \text{ Impurity}$$

for F:

$$1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = B$$

Gini Index: $\overset{3\times C0}{\boxed{\frac{3}{5}}} A + \overset{2\times C1}{\boxed{\frac{2}{5}}} B$

two classes:



Sources:

[1] O. Rioul, "This is IT: A primer on Shannon's entropy and information." Séminaire Poincaré XXIII (2018) 43-77.

[2] Runkler, T.A. (2020) "Data Analytics"

[3] aitimejournal.com

[4] sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/