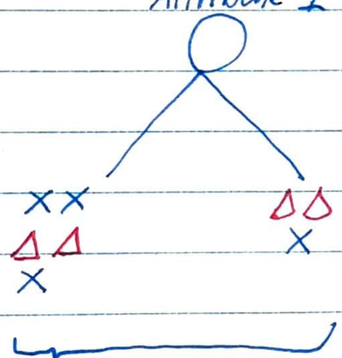
Tutorial 2.

1 Describe how a decision tree could be learnt.

→ A key element in decision trees is to find the attributes that are most discriminative. Or the attributes that better separate the ~~wt~~ instances that belong to different classes

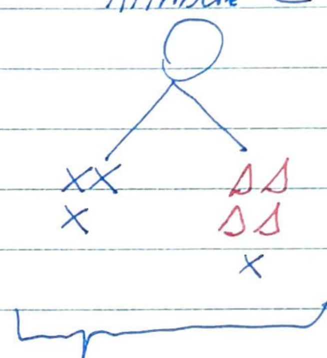
Examples

Attribute 1



this attribute does not separate well the classes

Attribute 2



This attribute seems that separates the classes better than attribute 1

But we have to do this numerically.....

b) Show how the idea of entropy could be used to pick the first node in the decision tree ...

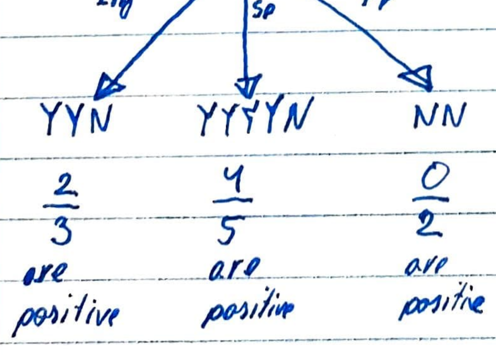
* As we mentioned before, we have to find the attributes that are more discriminative. Entropy can help us with this.

* the higher the entropy of a split, the less discriminative the attribute is. Or what is the same, the higher the entropy the less helpfull is that attribute to take good decisions.

* The lower the entropy of a split, the more discrimination produced by the attribute.

→ here, discrimination is a good thing....
It means that the capacity of the classifier to differencitate items of different clases/labels is higher.

→ Let's calculate the entropy of every split !!:
Visual example for language attribute



1 Step: Calculate the entropy of each split:

$$\text{Entropy}\left(\frac{2}{3}\right) = B\left(\frac{2}{3}\right) = -\underbrace{\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)}_{0.91} =$$

$$\text{Entropy}\left(\frac{4}{5}\right) = -\underbrace{\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right)}_{0.72}$$

$$\text{Entropy}\left(\frac{0}{2}\right) = -\underbrace{\left(\frac{0}{2}\log\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right)}_{0}$$

→

2 b) Now, once we have calculated the ~~split~~ weighted Entropy or total entry based on the number/ proportion of items that go to each split.

$$Entropy\,(lang) = \frac{3}{10} \times 0.91 + \frac{5}{10} \times 0.72 + \frac{2}{10} \times 0 \Rightarrow 0.63$$

So the entropy of the language split is this.

Now, we have to check the entropy of other splits and see which one is lower

$$Entropy\,(Type) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$\frac{4}{10}\,Entropy\,(Action) + \frac{3}{10}\,Entropy\,(Comedy) + \frac{3}{10}\,Entropy\,(Drama)$$

$$-\left(\frac{2}{4} \times \log_2\frac{2}{4} + \frac{2}{4} \times \log_2\frac{2}{4}\right) \qquad -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)$$

0.951 → So the entropy of type is higher than the entropy of lang. So, we will give preference to lang

Finally, we calculate the entropy for the attribute New

$$Entropy\,(New) = 0.846$$ → So this one is still higher than language

$\boxed{\text{We choose language because has the lowest entropy}}$

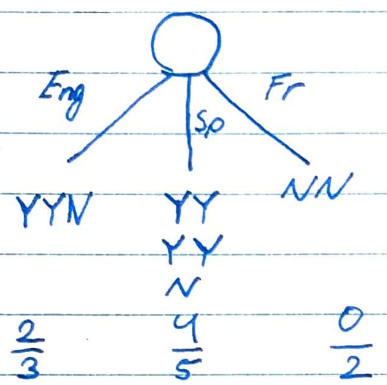
Gini works in a similar way as Entropy. The lower the value the better the discrimination of that attribute.

Tutorial 2

Question 3.

Now we have to use $\overbrace{\text{Gini impurity}}$ rather than entropy to decide the splits of our tree.

As we did before → Let's calculate the gini Impurity of each split. Then, let's calculate the total one.



* $Gini(Eng) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right) = 0.444$
* $Gini(Sp) = 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2 \right) = 0.32$
* $Gini(Fr) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2 \right) = 0$

Now, we have the gini impurity of each split. We have to calculate to total Gini

* $Gini(Language) = \frac{3}{10} \times 0.444 + \frac{5}{10} \times 0.32 + \frac{2}{10} \times 0 = 0.29$

* $Gini(Type) =$

$$\frac{4}{10} \times \underbrace{Gini(Action)}_{1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)} + \frac{3}{10} \times \overbrace{Gini(Comedy)}^{1-\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right)} + \frac{3}{10} \times \underbrace{Gini(Drama)}_{1-\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right)}$$

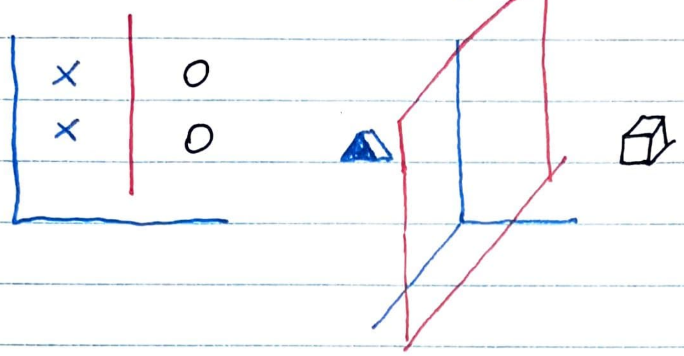
$= \boxed{0.46}$

* $Gini(News) = \boxed{0.4}$

The lowest Gini is the one of the language attribute

Tutorial 02
Question 3

a) a $\overbrace{\text{dataset}}^{\text{of n dimension}}$ is linearly separable if we can perfectly separate the classes of the datase with a n-1 dimensional plane



b) Examples: .....

- Extreme cases such as defferenciating the characteristics of a flower and the characteristics of a non-biological entity..... if we chose the righ parameters.

c) We will choose a classifier that is used in scenarios where the data is not linearly separated .... we will see some examples.

Tutorial 2

Question 4.

batch gradient descent

→ So we update the weights with the following formula.

$$w_0 \leftarrow w_0 + \alpha \sum_j (y_j - h_w(x_j))$$

$$w_1 \leftarrow w_1 + \alpha \sum_j (y_j - h_w(x_j))x_j$$

learning rate

This indicates that goes over The entire dataset once...

Rather than explain this with a formula, Let's do an example. $\boxed{\text{INITIAL WEIGHTS} \rightarrow w_0 = 0}$

Instance	x	y	prediction (bias)	error
E_1	1.5	1	$1 \times 0 + 1.5 \times 0 = 0$	1-0
E_2	3.5	3	$1 \times 0 + 3.5 \times 0 = 0$	3-0
E_3	3	2	$1 \times 0 + 3 \times 0 = 0$	2-0
E_4	5	3	$1 \times 0 + 5 \times 0 = 0$	3-0
E_5	2	2.5	$1 \times 0 + 2 \times 0 = 0$	2.5-0

Total error = 11.5

$\boxed{\text{UPDATE TIME}}$ → 1 Batch or epoch

$w_0 \leftarrow w_0 + \alpha \cdot \text{Total error} \rightarrow 0 + 0.01 \times 11.5 = 0.115$

$w_1 \leftarrow w_1 + \alpha \cdot \boxed{\text{Total error (weighted)}} \rightarrow 0 + 0.01 \times 38 = 0.38$

$$\sum_j (y_j - h_w(x_j))x_j$$

E_1 error $= 1 \times 1.5 = 1.5$

E_2 error $3 \times 3.5 = 10.5$

E_3 error $2 \times 3 = 6$

E_4 error $3 \times 5 = 15$

E_5 error $2.5 \times 2 = 5$

weighted error = 38

b) Now a couple of updates with stochastic gradient descent.

$w_0 = 0$, $w_1 = 0$ $\alpha = 0.01$

	x_i	y	Prediction	Error
→ E_1	1.5	1	$1 \times 0 + 1.5 \times 0 = 0$	$1 - 0 = 1$

UPDATE TIME (here we do not calculate the error of the entire dataset before updating. We update after every instance)

$w_0 \rightarrow w_0 + \alpha \cdot \text{error} \Rightarrow 0 + 0.01 \cdot 1 = \boxed{0.01}$

$w_i \Rightarrow w_1 + \alpha \cdot \text{error} \cdot x_1 \Rightarrow 0 + 0.01 \cdot 1 \cdot 1.5 = \boxed{0.015}$

new weights to use in the next iteration

	x_i	y	prediction	error
E_2	3.5	3	$1 \times 0.01 + 0.015 \times 3.5 = 0.0625$	$3 - 0.0625 = 2.9375$

UPDATE TIME

$w_0 \leftarrow 0.01 + 0.01 \cdot 2.9375 = \boxed{0.0393}$

$w_i \leftarrow 0.015 + 0.01 \cdot 2.9375 \times 3.5 = \boxed{0.118}$

new weights to use in the next iteration.

Tutorial 2

Exercise 5

$$w_0 \leftarrow w_0 + \alpha\left(y - \left(w_0 + x_{1j} w_1 + x_{2j} w_2\right)\right)$$

$$w_1 \leftarrow w_1 + \alpha\left(y - \left(w_0 + x_{1j} w_1 + x_{2j} w_2\right)\right) x_{1j}$$

$$w_2 \leftarrow w_2 + \alpha\left(y - \left(w_0 + x_{1j} w_1 + x_{2j} w_2\right)\right) x_{2j}$$