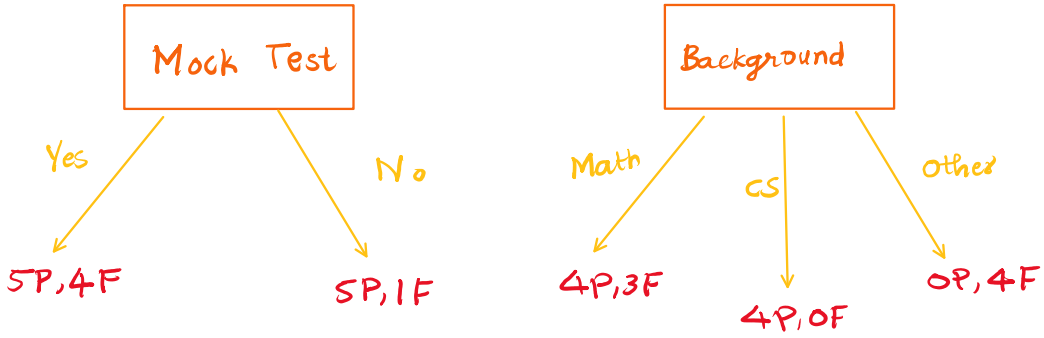
# Lecture 10: Decision Tree

14 February 2024 11:03

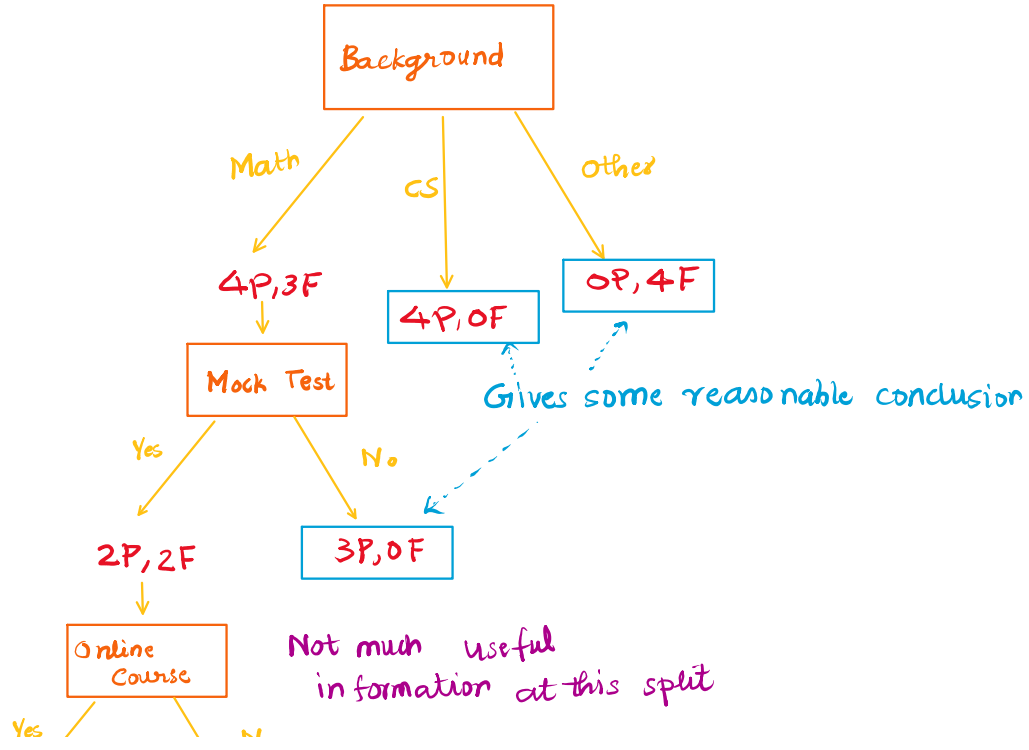
Consider an example of an exam result:

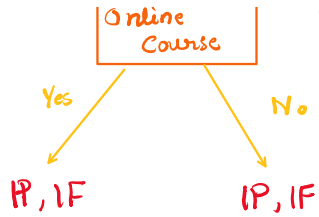
| Exam result | Online course? | Background | Mock Test |
|---|---|---|---|
| P | Y | Math | N |
| F | N | Math | Y |
| F | Y | Math | Y |
| P | Y | CS | N |
| ⋮ | ⋮ | ⋮ | ⋮ |

Goal: Create a classifier on whether a given student will pass or not depending on the data



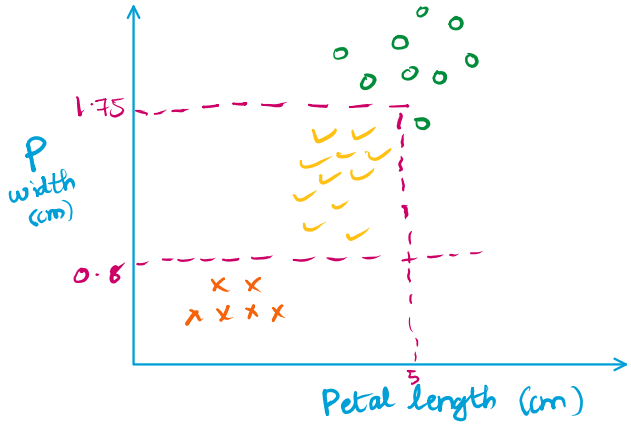

Note: In the first stage, background seems to be a better split.

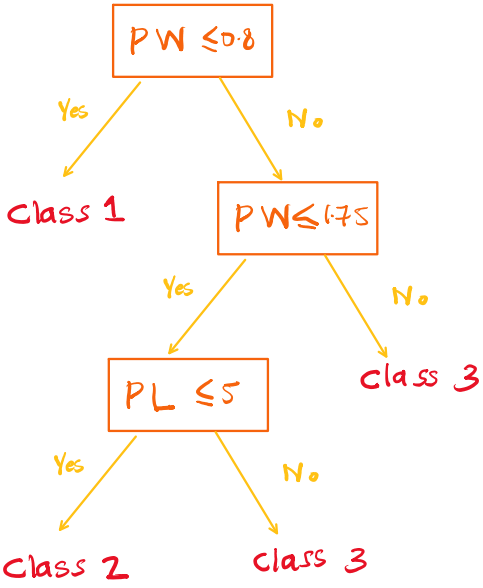





Not much useful information at this split

## Example 2: Iris dataset



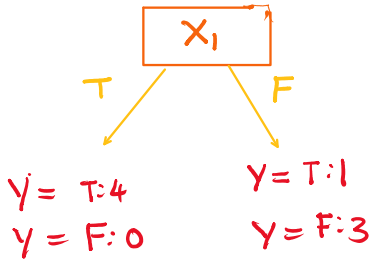

Having seen these examples, we address the following questions:

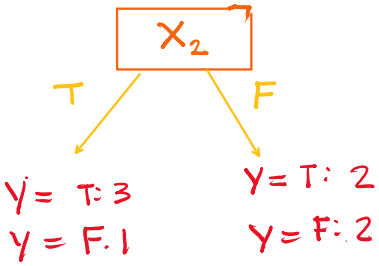
i) How to build the tree?

ii) When and where to stop?

## Example 3

| $X_1$ | $X_2$ | Y |
|---|---|---|
| T | T | T |
| T | F | T |
| T | T | T |
| T | F | T |
| F | T | T |
| F | F | F |
| F | T | F |
| F | F | F |




If we can divide wrt $X_1$ and $X_2$ what can we say about the [illegible] with what "category"?

If we can divide wrt $X_1$ and $X_2$ what can we say about the classification and with what "category"?

" measurement of certainty" (entropy)

**ENTROPY**: Measurement of randomness of a random variables.

Let $X$ be a categorical Random Variable with $p(x) = P(X=x)$ $\forall x \in X$

We define entropy as:

$$H(X) = -\sum_{x \in X} p(x) \log_{|X|} p(x)$$ , $X = \{0,1\} \rightarrow X$ is a binary R.V and its H is measured in bits.

We've following properties:

i) $H(x) \geq 0$

ii) $H(X) \leq 1$ Using Jensen's Inequality:

> If $f$ is a convex fn, $\mathbb{E}(f(x)) \geq f(\mathbb{E}X)$

Observe: $H(X) = \mathbb{E}\left(\log_{|X|} p(x)\right)$

**Conditional Entropy**: We observe $Y$, a proxy of $X$. We've $p(x|y) = P(X=x|Y=y)$

Now, we define conditional entropy as:

$$H(X|Y) = -\sum_x \sum_y p(x,y) \log p(x|y)$$

$\rightarrow H(X|Y=y)$

$$= \sum_y p(y) \left(-\sum_x p(x|y) \log p(x|y)\right)$$

$$= \sum_y p(y) H(X|Y=y)$$

If $X$ and $Y$ are independent, $H(X|Y) = H(X)$ (How?)

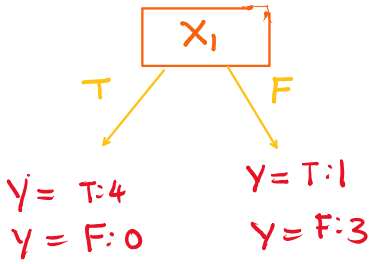
We define Information entropy as, $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

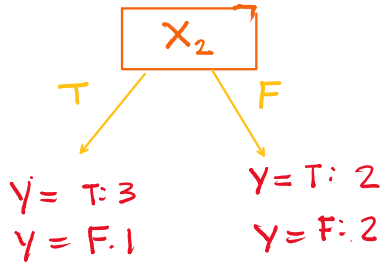
? Mutual information

Lets come back to Example 3:

## Example 3

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| T | T | T |
| T | F | T |
| T | T | T |
| T | F | T |
| F | T | T |
| F | F | F |
| F | T | F |
| F | F | F |




We will compare $I(Y; X_1)$ v/s $I(Y; X_2)$

$$H(Y|X_1) = \sum_{x_1 \in \{T,F\}} p(X_1 = x_1)\, H(Y|X_1 = x_1)$$

$$= p(X_1 = T)\, H(Y|X_1 = T) + p(X_1 = F)\, H(Y|X_1 = F)$$

($p(X_1 = T)$: $1/2$; $H(Y|X_1 = T)$: $0$; $p(X_1 = F)$: $1/2$)

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

Observe that $p(y|X_1 = T) = P(Y = y|X_1 = T) = \begin{cases} 1 & \text{if } y = T \\ 0 & \text{if } y = F \end{cases}$

Thus, $H(Y|X_1 = T) = 0$

$$\therefore H(Y|X_1) = 0.4056$$

Similarly we can find; $H(Y|X_2) = 0.9056$

## Algorithm for building Decision Tree:

- Repeat until Stopping criteria:
  - find the feature that yeilds maximum informational gain (or, minimal condtional entropy)

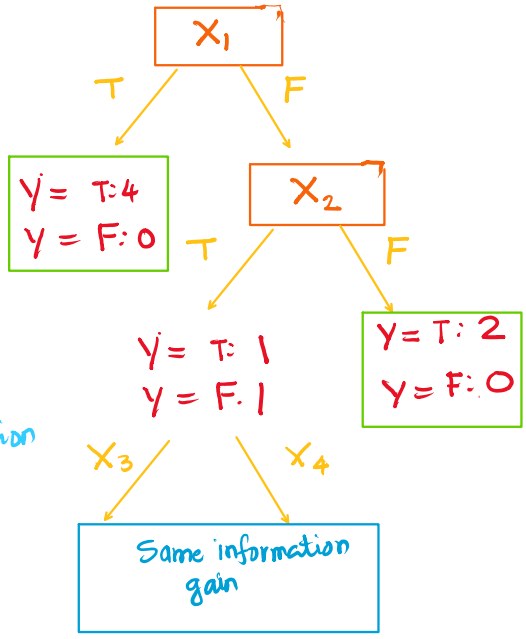
**Remark:** Alternative approach using a metric called 'Gini Index'

## Where to Stop?

### Base Case 1:

Node with atomic distributions:

$H(Y|\text{node}) = 0$




### Base Case 2:

When all remaining features have same information gain

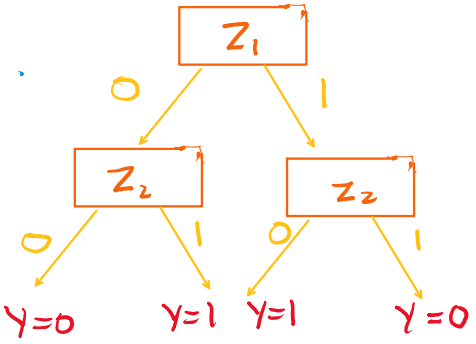
This may not work always :-

**Example 4:**

$H(Y) = 1$

$H(Y|Z_1) = 1 = H(Y|Z_2)$

| $Z_1$ | $Z_2$ | $Y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |




According to base case 2, this should not be split. But we can split as above

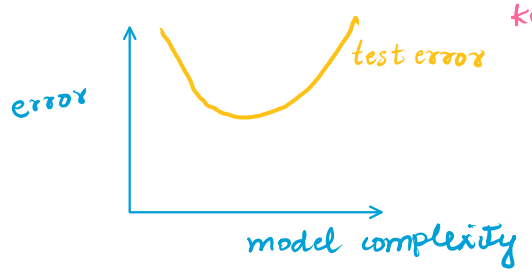
### Overfitting in decision trees:

Shallow tree → not enough power to distinguish

Deep trees → specific to training example

### Three methods

1. Pre-Pruning / Early Stopping: Hold a validation set;




keep on creating the tree until test error goes up again

2. Post-Pruning: allow the tree to fully grow and then reduce some of its branching.

3. Ensemble method: Using averages of various models.