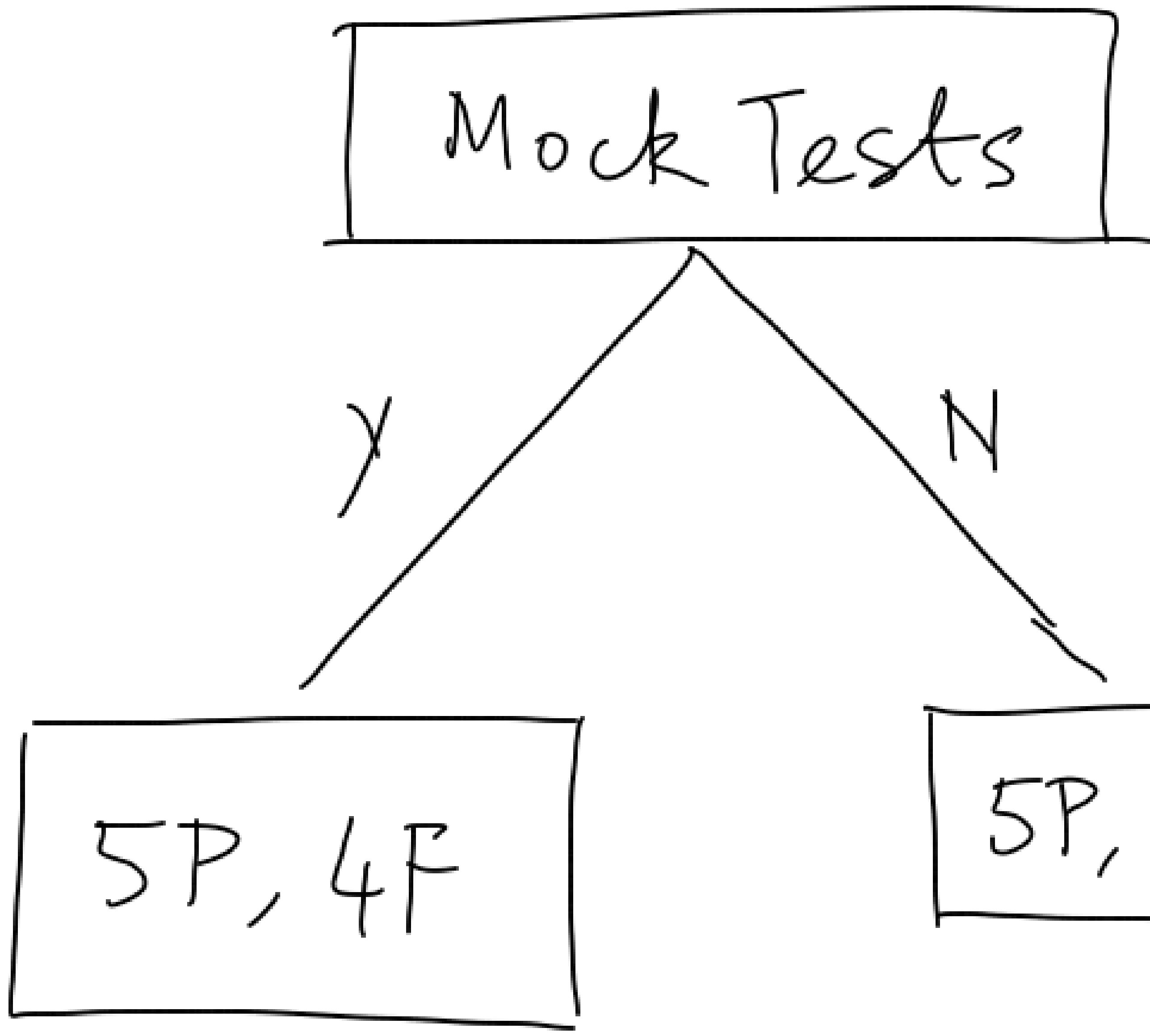
# Lec 10: Decision Trees

Toy example:

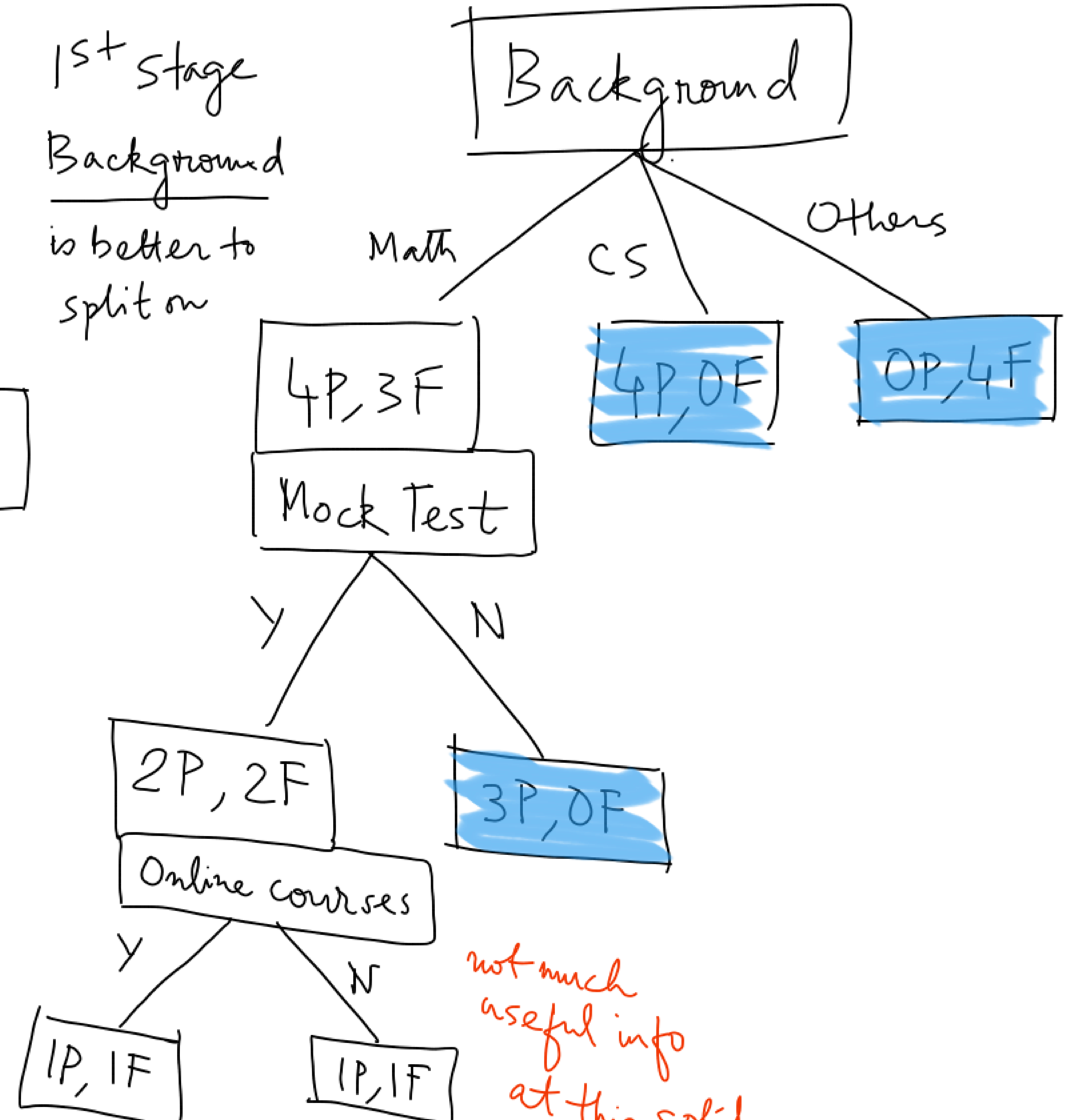
| Exam result ($y$) | Online Courses? | Background | Mock Tests |
|---|---|---|---|
| P | Y | Math | N |
| F | N | M | Y |
| F | Y | M | Y |
| P | Y | CS | N |
| ⋮ | ⋮ | ⋮ | ⋮ |

$x$ = (Online Courses?, Background, Mock Tests)

Goal: create a classifier on whether a given student will pass/not depending on the data



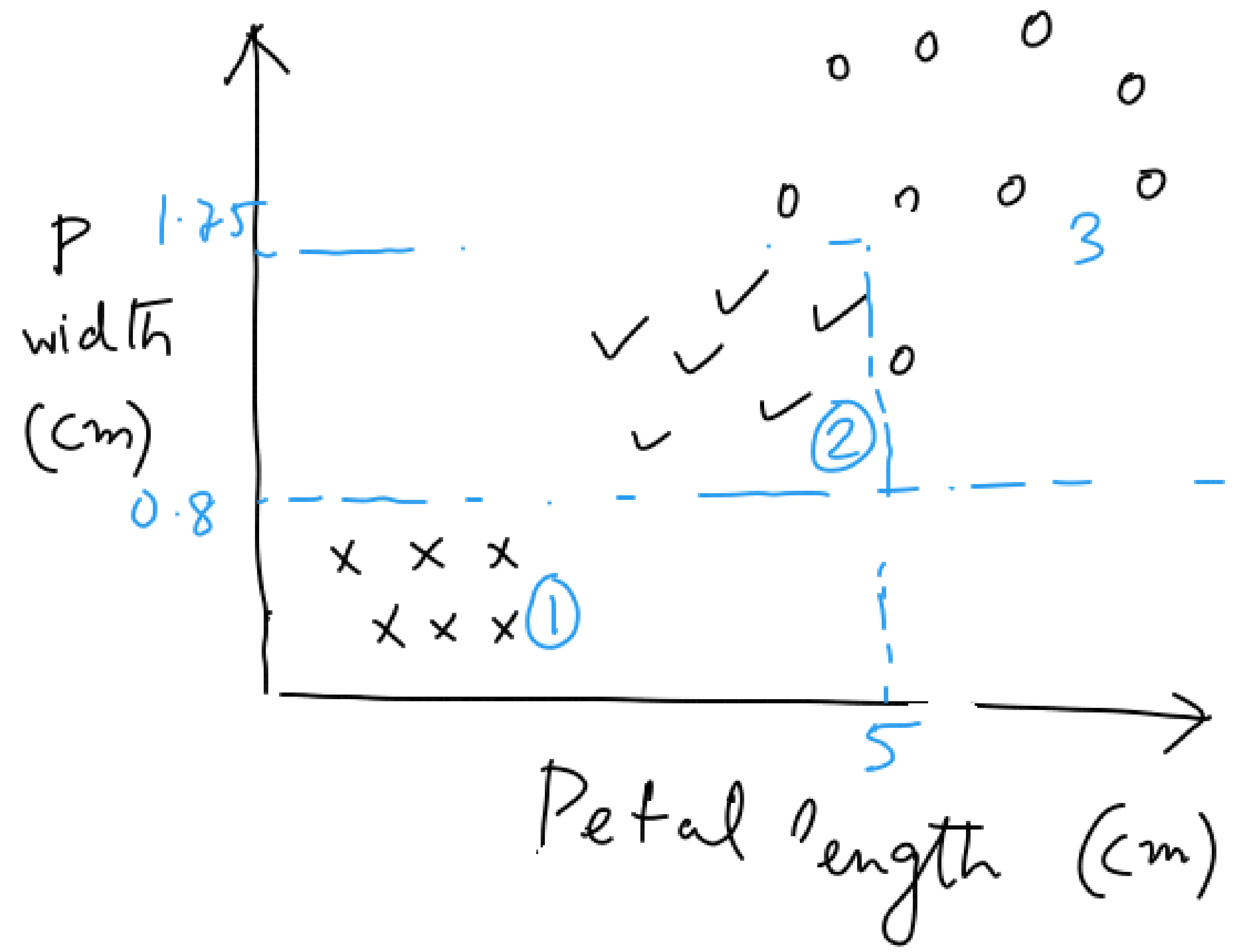

1st stage
Background
is better to
split on

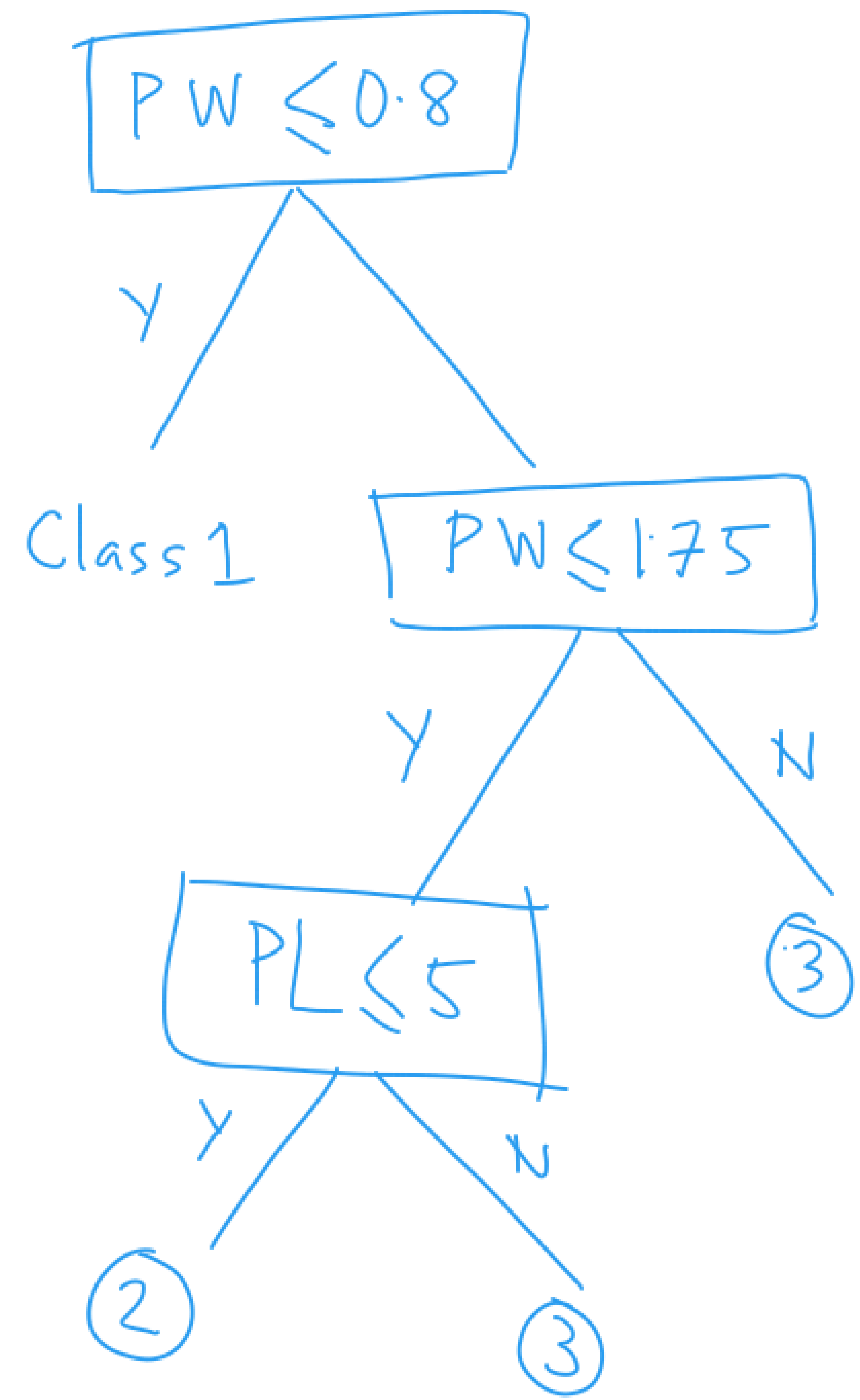



not much
useful info
at this split.

Q: how can we make
such a splitting scheme
more systematic ?

# Ex. 2 Iris dataset



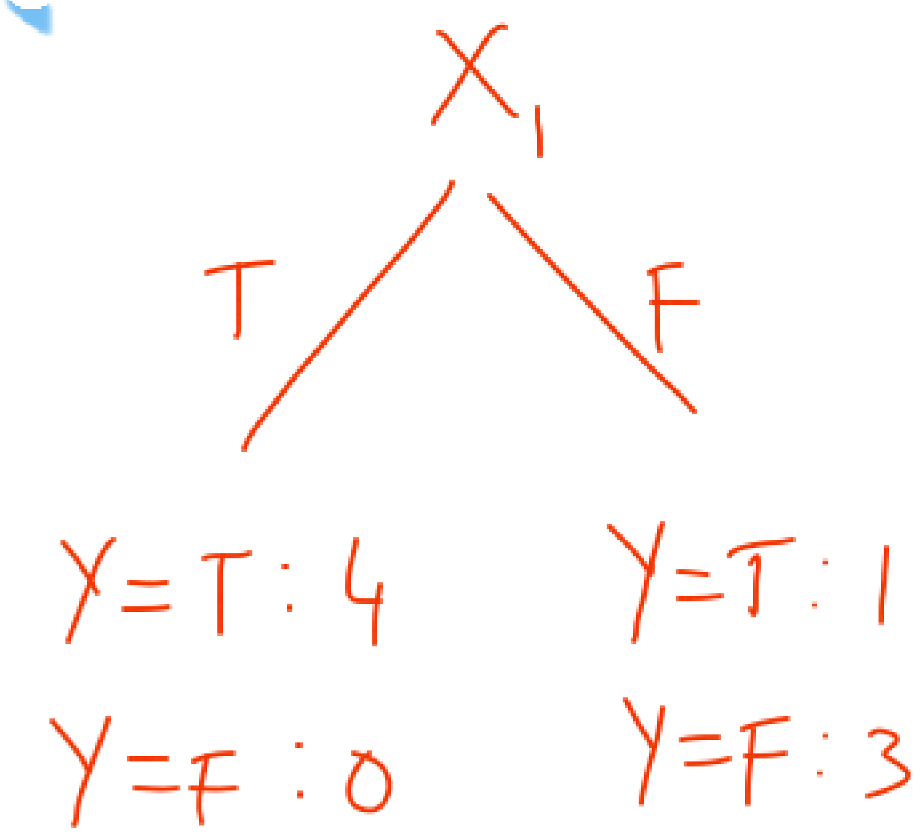



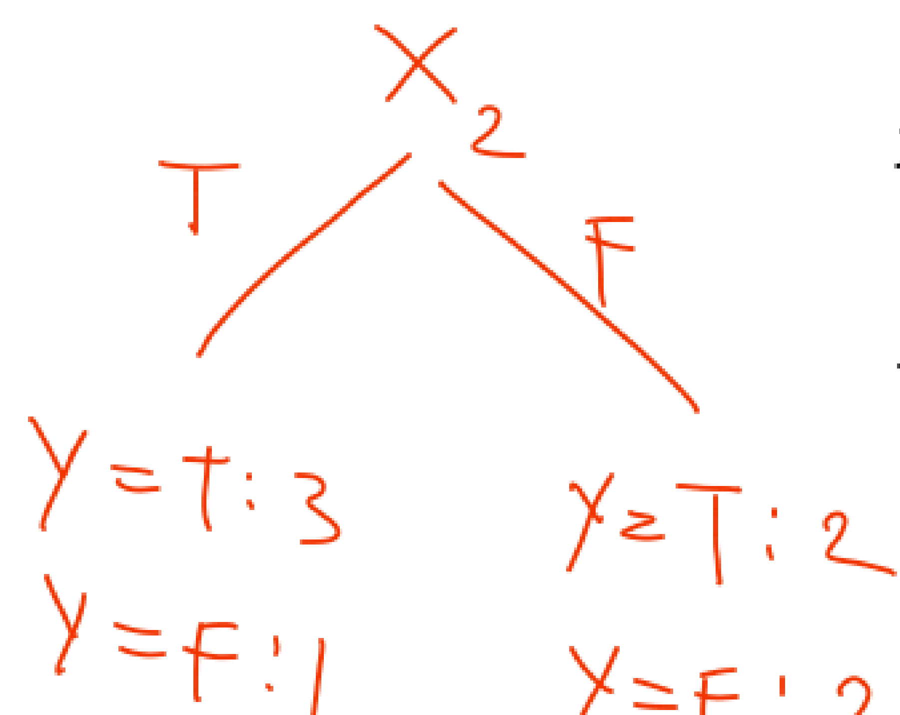

Q1: How to build the tree?

Q2: Where to stop?

## Ex 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 |



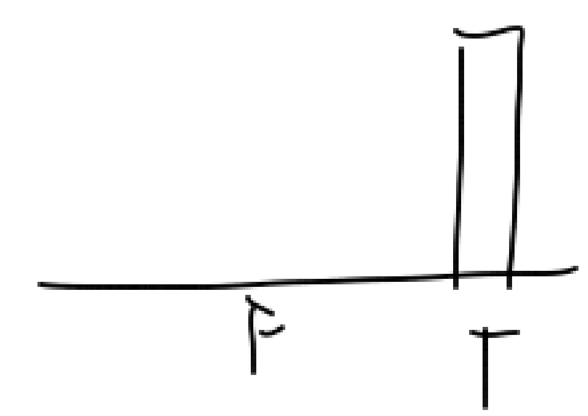

$I(Y;X_1)$ vs $I(Y;X_2)$

$H(Y) - H(Y|X_1)$ $\qquad$ $H(Y) - H(Y|X_2)$

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

$$= p(X_1 = T)\, \cancel{H(Y|X_1 = T)}^{\,0} + p(X_1 = F)\, \underbrace{H(Y|X_1 = F)}$$

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




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

$$= 0.4056$$

$$H(Y|X_2) = 0.9056$$

If we divide wrt $X_1$ or $X_2$ what can say about the classification and with what "certainty"? → entropy

measurement of certainty

**Entropy:** measurement of randomness of a RV

$X$ be a categorical RV, $p(x) = P(X=x)$, $\forall x \in \mathcal{X}$

$$H(X) = -\underbrace{\sum_{x \in \mathcal{X}} p(x) \log_{|\mathcal{X}|} p(x)}_{\mathbb{E}\left[\log_{|\mathcal{X}|} p(X)\right]}$$

, $\mathcal{X} = \{0,1\} \to X$ is a Binary RV and $H$ is measured in bits.

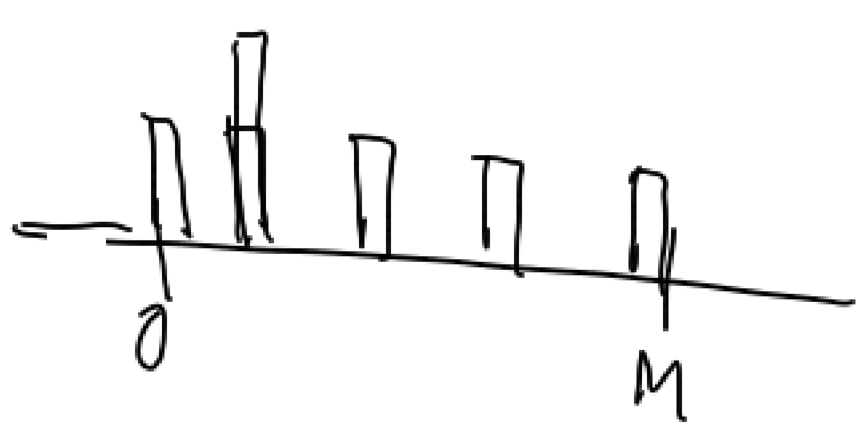
① $H(X) \geq 0$

② $H(X) \leq 1 \to$ Jensen's ineq.

$f$ is convex $\quad f(X)$

$$\mathbb{E} f(x) \geq f(\mathbb{E}X)$$

concave $\to \leq$



Conditional Entropy : Observe $Y$, a proxy of $X$

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

$p(x|y) \to P(X=x \mid Y=y)$

$$= \sum_y p(y) \underbrace{\left( -\sum_x p(x|y) \log p(x|y) \right)}_{H(X|Y=y)} = \sum_y p(y) H(X|Y=y)$$

if $X \perp\!\!\!\perp Y$

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

$$\underset{\text{mutual information}}{I(X;Y)} = H(X) - H(X|Y) \underset{\text{HW}}{=} H(Y) - H(Y|X)$$

## Algorithm for decision tree building

- Repeat until stopping criteria not met
  - find the feature that yields max information gain (min conditional entropy)

**Remark:** Other Metric used : Gini index.

## Where to stop?

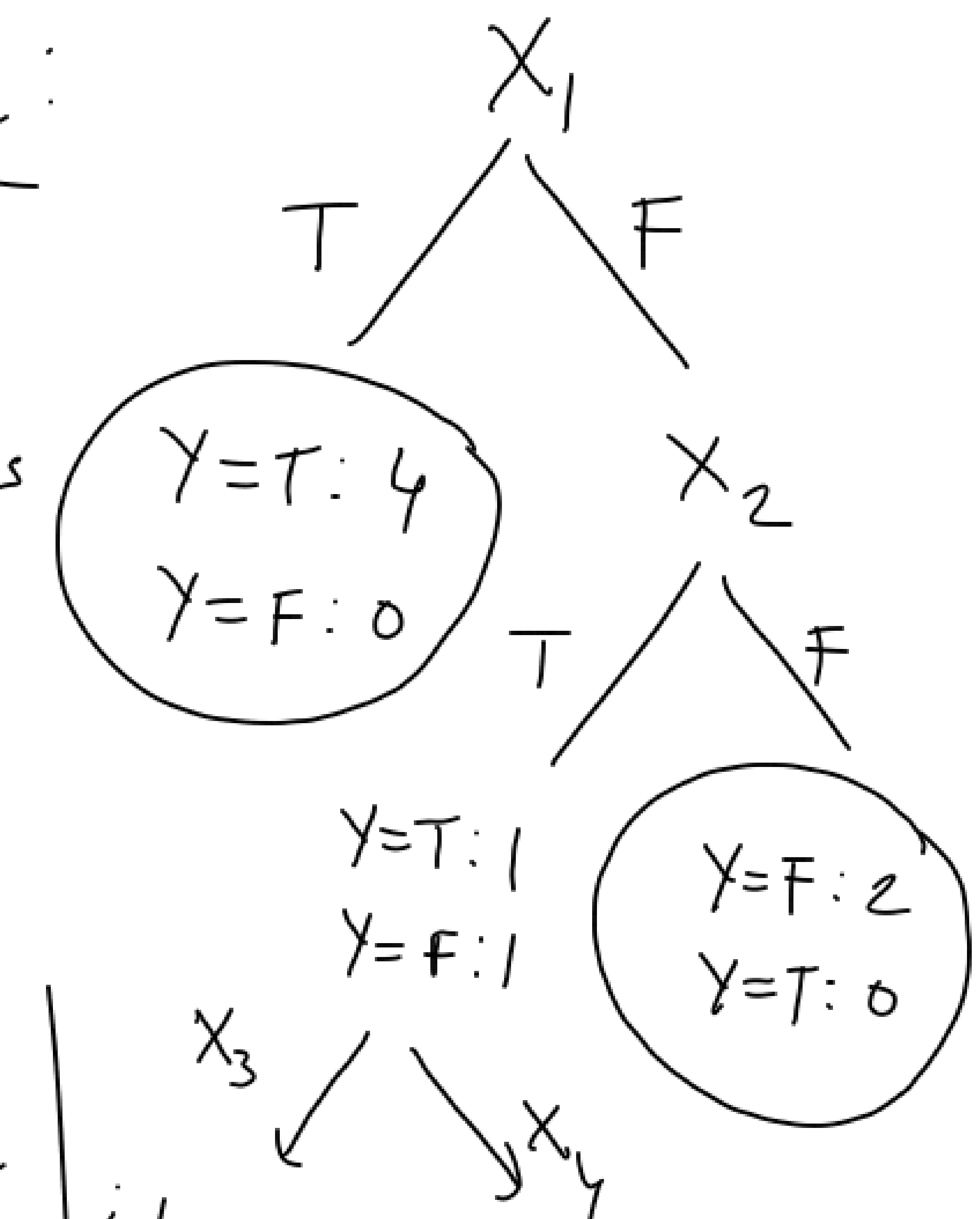
**Base case 1:**

node with atomic distributions

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

**Base case 2:**

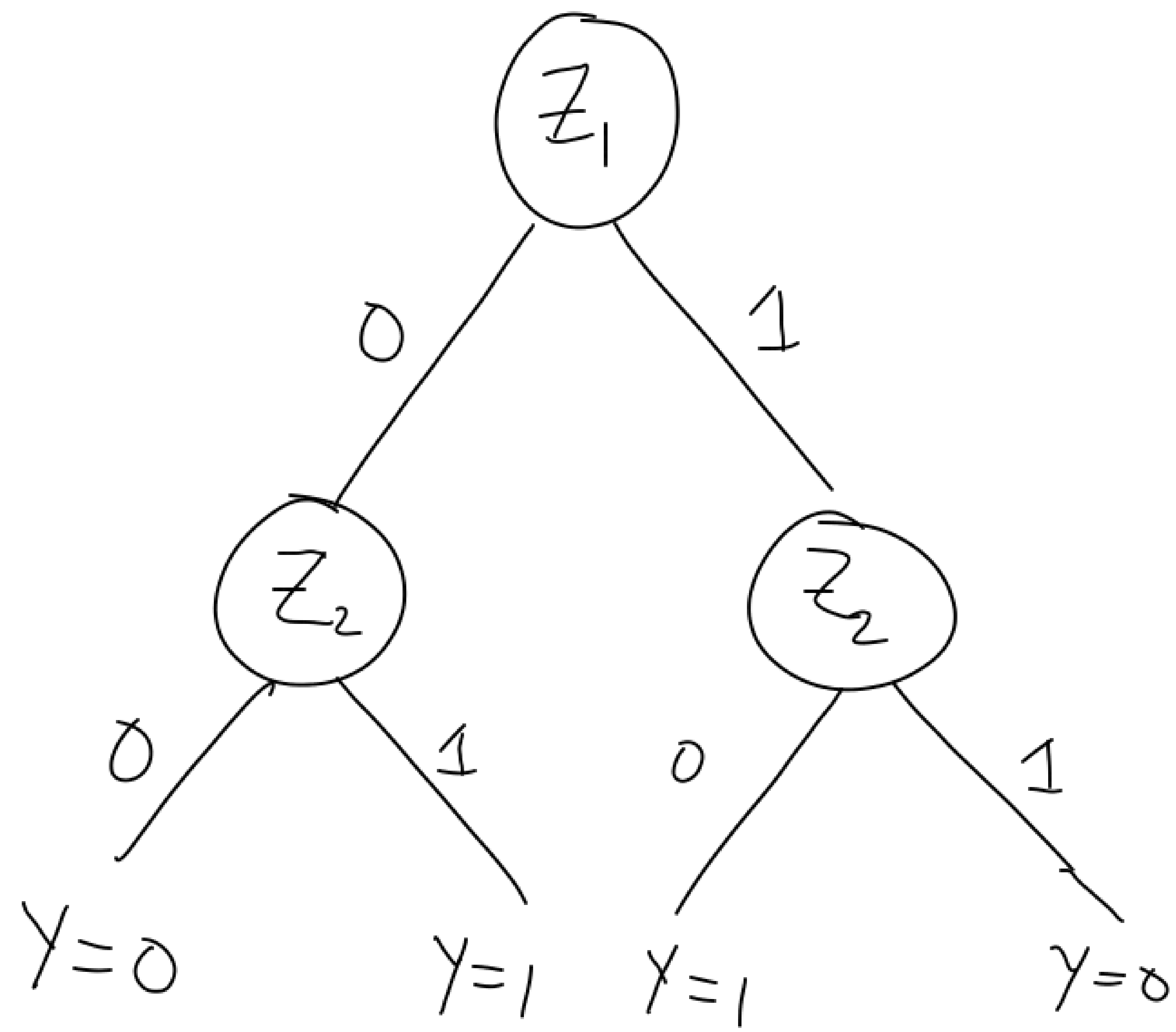
When all remaining features give identical info gain.




info gain is same for all remaining variables

Ex 4:

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




$H(Y) = 1$

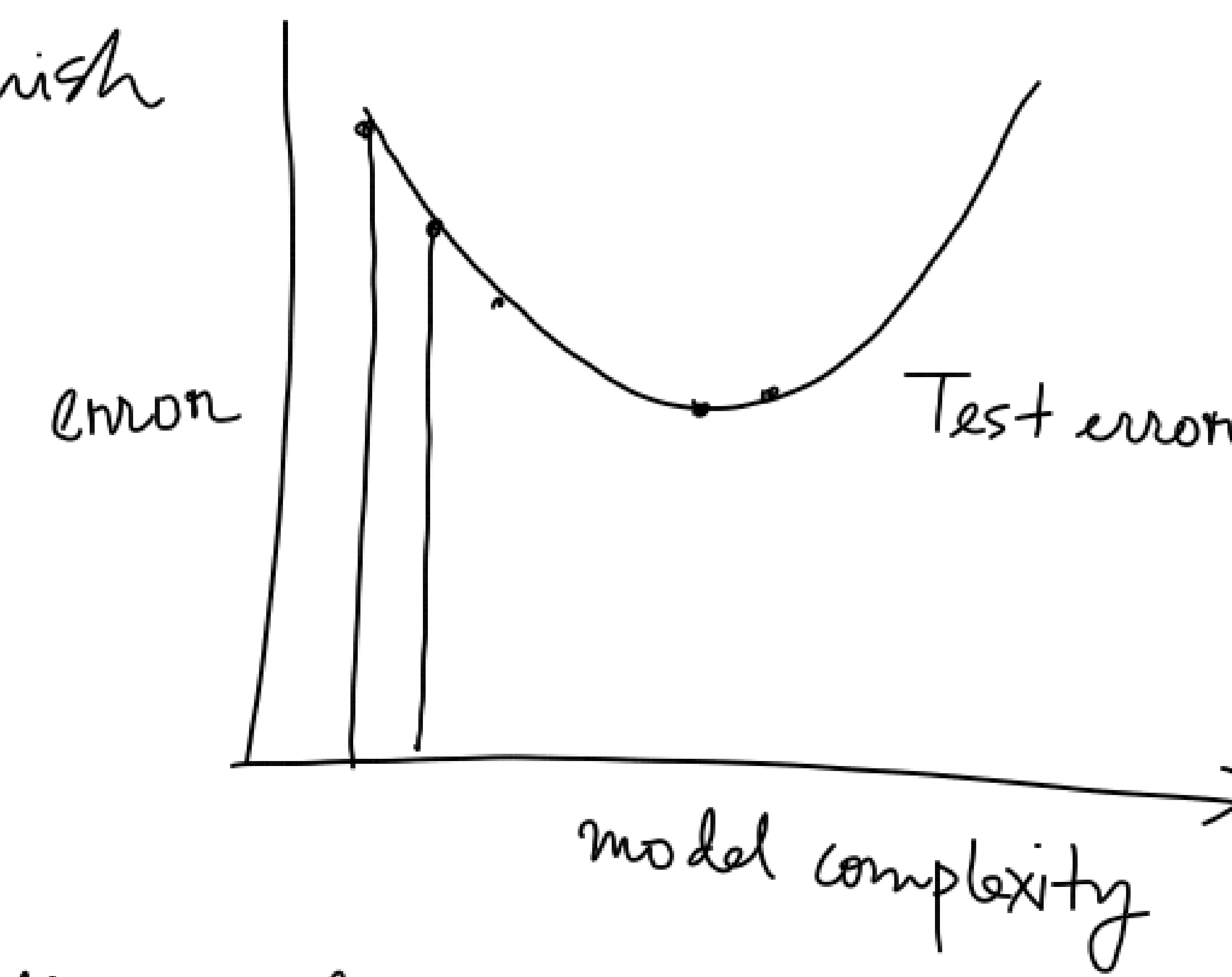
$H(Y|Z_1) = 1 = H(Y|Z_2)$

according to basecase 2, this shouldn't be split

# Overfitting in decision trees

Shallow tree → not enough power to distinguish

deep tree → specific to training examples




## Three methods

- Pre-pruning / Early stopping: hold a validation set
  Keep on creating the tree until test error goes up again.
- Post-pruning: allow the tree to fully grow and then reduce some of its branches.
- Ensemble method: using averages of various models