Recap of what we've done

0) ML flow
1) Linear regression (Regression)

2) Naive Bayes (Classification)

Differentiate → Regression vs Classification } Supervised Learning

Student	Math	ML	DB	Salary
1				5000
2				8000
⋮				
1000				100 k
1001	81	79	51	?

new sample → 1001

Student	Math	ML	DB	work at Microsoft?
1				YES
2				No
⋮				⋮
1000				YES
1001	81	79	01	?

Decision Trees

- Can be used for classification AND regression
- But mainly famous because of classification.

x_1 x_2 x_3 x_4 y

$N=14$

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
D15	Rain	Hot	Low	Weak	?

Entropy, Information Gain, Gini Coefficient

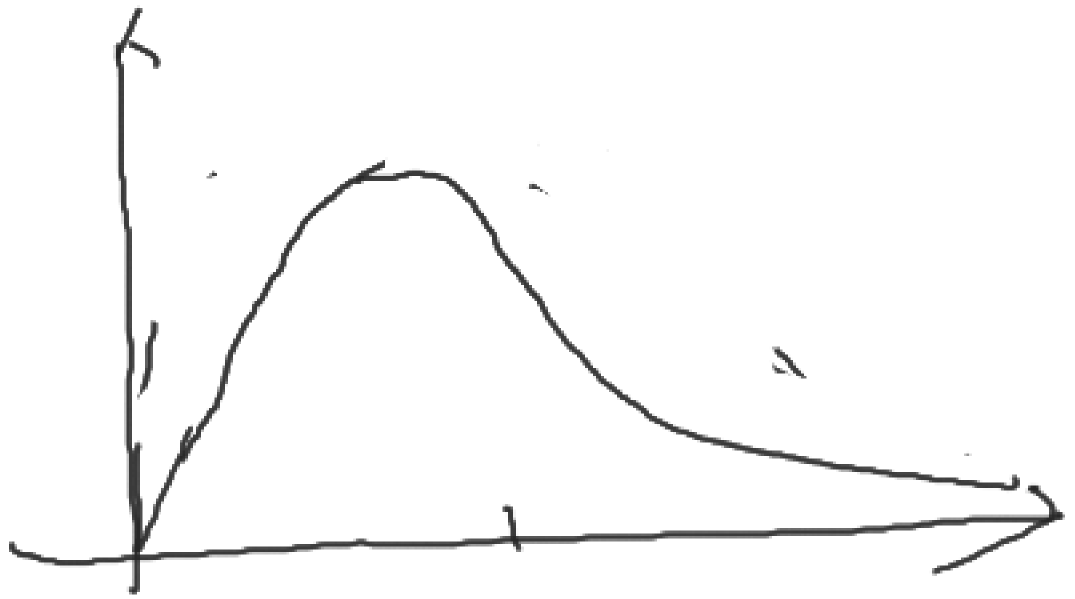
Entropy

$$\boxed{H(X) = -\sum p(x) \log_2 p(x)}$$

Measuring the amount of surprise

P(Married) In college

p(married)



Information Gain IG

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

$$H(x) = -\sum_{k=1}^{K} P(x)\log_2[P(x)]$$

6. We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. (For this problem, you can write your answers using log_2, but it may be helpful to note that $log_2 3 \approx 1.6$.)

Weather = Sunny, Cloudy, Rainy

$$H(\text{Weather}) = -\sum_{k=1}^{k=3} P(\text{Weather})\log_2 P(\text{Weather})$$

	GPA	Studied	Passed
1	L	F	F
2	L	T	T
3	M	F	F
4	M	T	T
5	H	F	T
6	H	T	T

$P(\text{Pass}=T) = 4/6 = 2/3$

$P(\text{Pass}=F) = 2/6 = 1/3.$

$$H(\text{Pass}) = -\sum_{k=1}^{K=2} P(\text{Pass})\log_2 P(\text{Pass})$$

Pass Fail

$$= -\Big[\underbrace{P(\text{Pass}=T)\log_2 P(\text{Pass}=T)}_{T} + \underbrace{P(\text{Pass}=F)\log_2 P(\text{Pass}=F)}_{F}\Big]$$

$$= -\left[\frac{2}{3}\log_2(2/3) + 1/3\log_2(1/3)\right] \approx 0.92$$

6. We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. (For this problem, you can write your answers using log_2, but it may be helpful to note that $log_2 3 \approx 1.6$.)

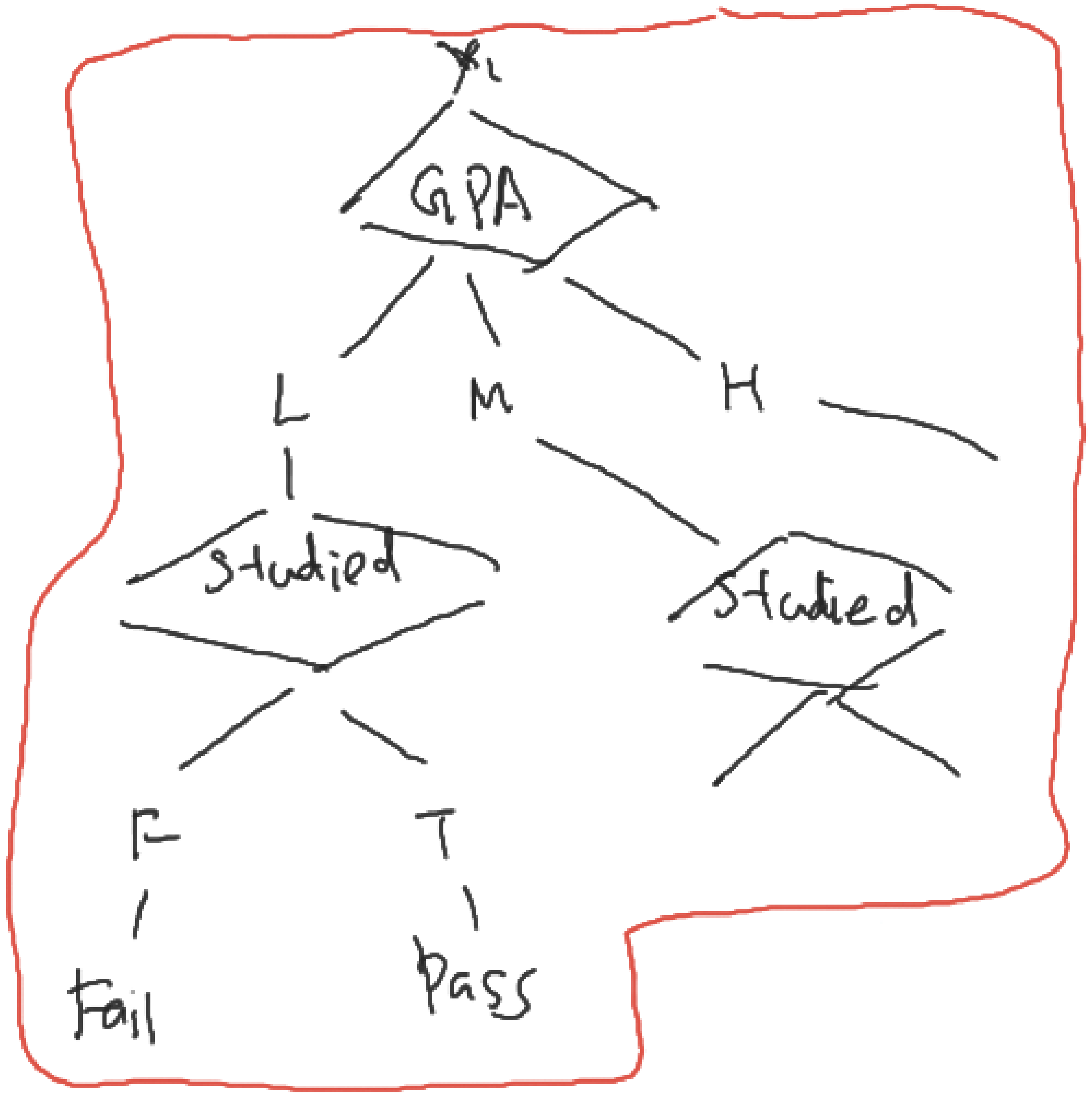
Two features : GPA (x_1), Studied (x_2)

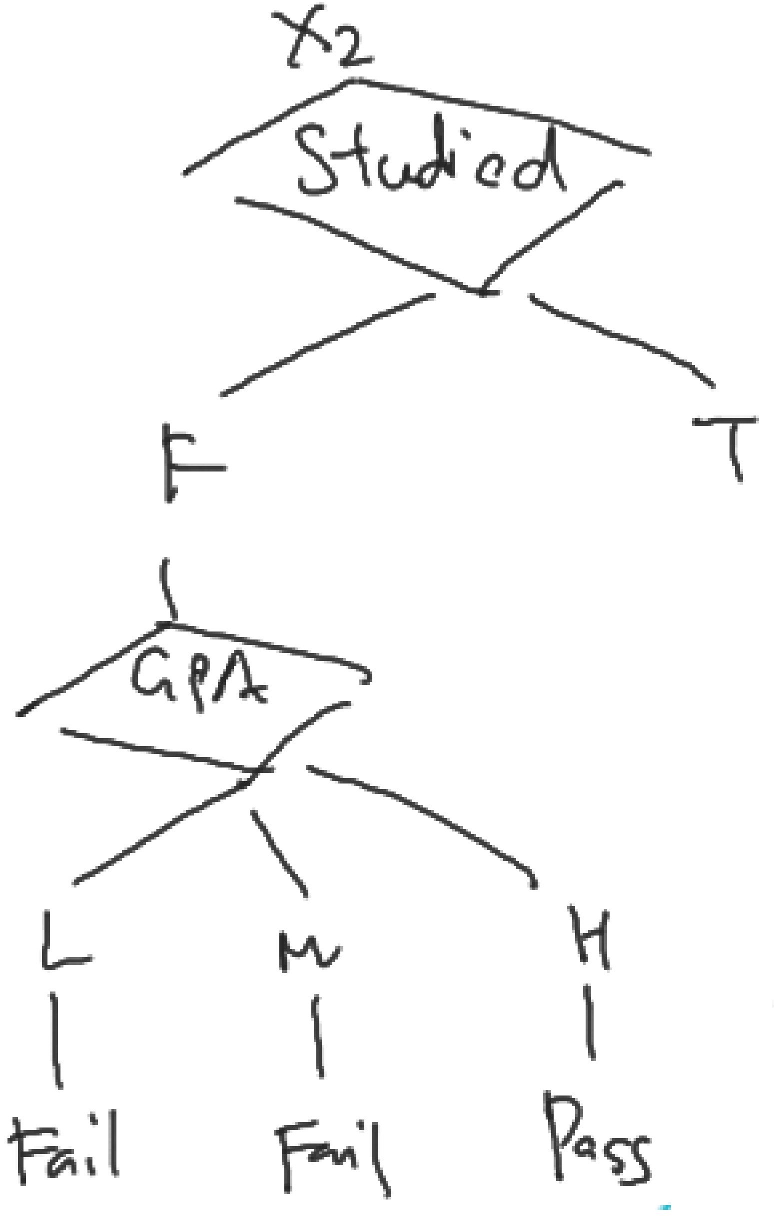
One target output : Passed (y)

GPA (x_1)	Studied (x_2)	Passed (y)
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

Q) How do we expand the tree / Choose the root node?

$H(Pass) = 0.92$





Do we choose GPA or Studied as the root node?

↓

We have to calculate Information Gain

- IG(Pass, GPA)
- IG(Pass, Studied)

6. We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. (For this problem, you can write your answers using log_2, but it may be helpful to note that $log_2 3 \approx 1.6$.)

Feature → X   Y output

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

GPA

L

Studied

F   T

IG(Pass, GPA) > IG(Pass, Studied)

0.26   0.24

Arg max IG(Pass, Features$_i$)

GPA

⇒ GPA is chosen as the root node because it has the highest information gain.

IG: Information Gain

6. We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. (For this problem, you can write your answers using log_2, but it may be helpful to note that $log_2 3 \approx 1.6$.)

we know this already

$$\boxed{IG(Pass, GPA) = H(Pass) - H(Pass|GPA)}$$

$$\boxed{H(Pass|GPA) = -\sum P(GPA) \sum P(Pass|GPA) \log_2 P(Pass|GPA)}$$

X Y

	GPA	Studied	Passed
1	L	F	F
2	L	T	T
3	M	F	F
4	M	T	T
5	H	F	T
6	H	T	T

Pass = TRUE (Pass)

GPA=L $P(Pass = T \mid GPA = L) = 1/2$

GPA=M $P(Pass = T \mid GPA = M) = 1/2$

GPA=H $P(Pass = T \mid GPA = H) = 1$

Pass = FALSE (Fail)

$P(Pass = F \mid GPA = L) = 1/2$

$P(Pass = F \mid GPA = M) = 1/2$

$P(Pass = F \mid GPA = H) = 0$

$P(GPA = L) = 2/6$

$P(GPA = M) = 2/6$

$P(GPA = H) = 2/6$

$H(Pass|GPA) = -\sum P(GPA) \sum P(Pass|GPA) \log_2 P(Pass|GPA)$

6. We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. (For this problem, you can write your answers using log_2, but it may be helpful to note that $log_2 3 \approx 1.6$.)

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

$$H(Pass|GPA) = -\Big\{ \overset{2/6}{P(GPA=L)} \Big[\overset{1/2}{P(Pass=T|GPA=L)} \log_2 \overset{\log_2(1/2)}{P(Pass=T|GPA=L)} + \overset{1/2}{P(Pass=F|GPA=L)} \log_2 P(Pass=F|GPA=L) \Big]$$

$$+ \overset{2/6}{P(GPA=M)} \Big[\overset{1/2}{P(Pass=T|GPA=M)} \log_2 \overset{\log_2(1/2)}{P(Pass=T|GPA=M)} + \overset{1/2}{P(Pass=F|GPA=M)} \log_2 \overset{\log_2(1/2)}{P(Pass=F|GPA=M)} \Big]$$

$$+ \overset{2/6}{P(GPA=H)} \Big[\overset{1}{P(Pass=T|GPA=H)} \log_2 \overset{\log_2(1/2)}{P(Pass=T|GPA=H)} + \overset{0}{P(Pass=F|GPA=H)} \log_2 \overset{\log_2(0)}{P(Pass=F|GPA=H)} \Big] \Big\}$$

$H(Pass|GPA) \approx 0.66$

$IG(Pass, GPA): H(Pass) - H(Pass|GPA)$

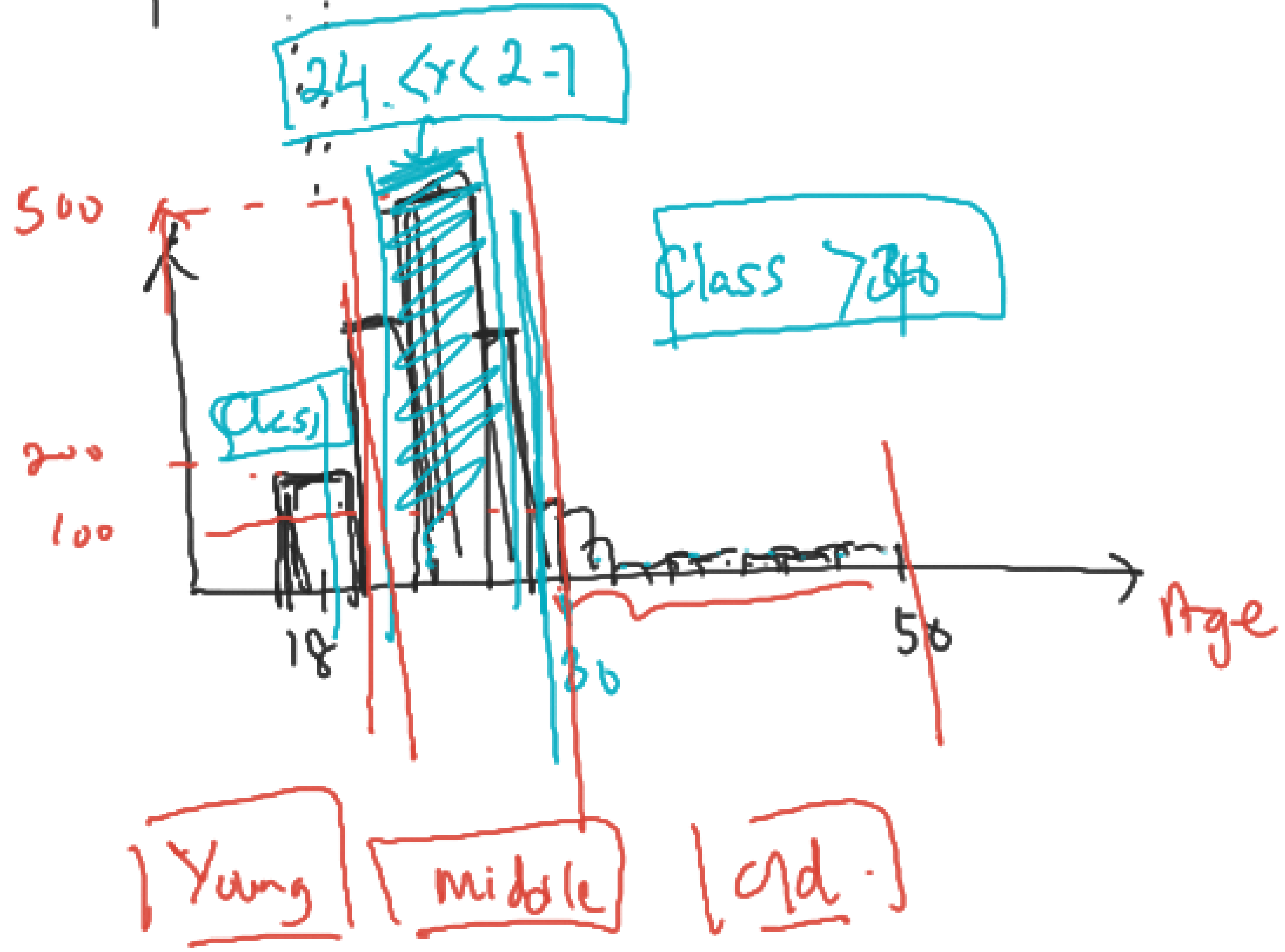
$= 0.92 - 0.66$

$= \underline{\underline{0.26}}$ #

Decision Trees

GPA	Age	Studied	Pass
L	$\boxed{32}$	T	
H	$\boxed{18}$	T	
M		F	
M	50	F	

Age ≤ 32



IG(Pass, Age > 40)

- Number of samples for each 'category'
- Information gain of each category

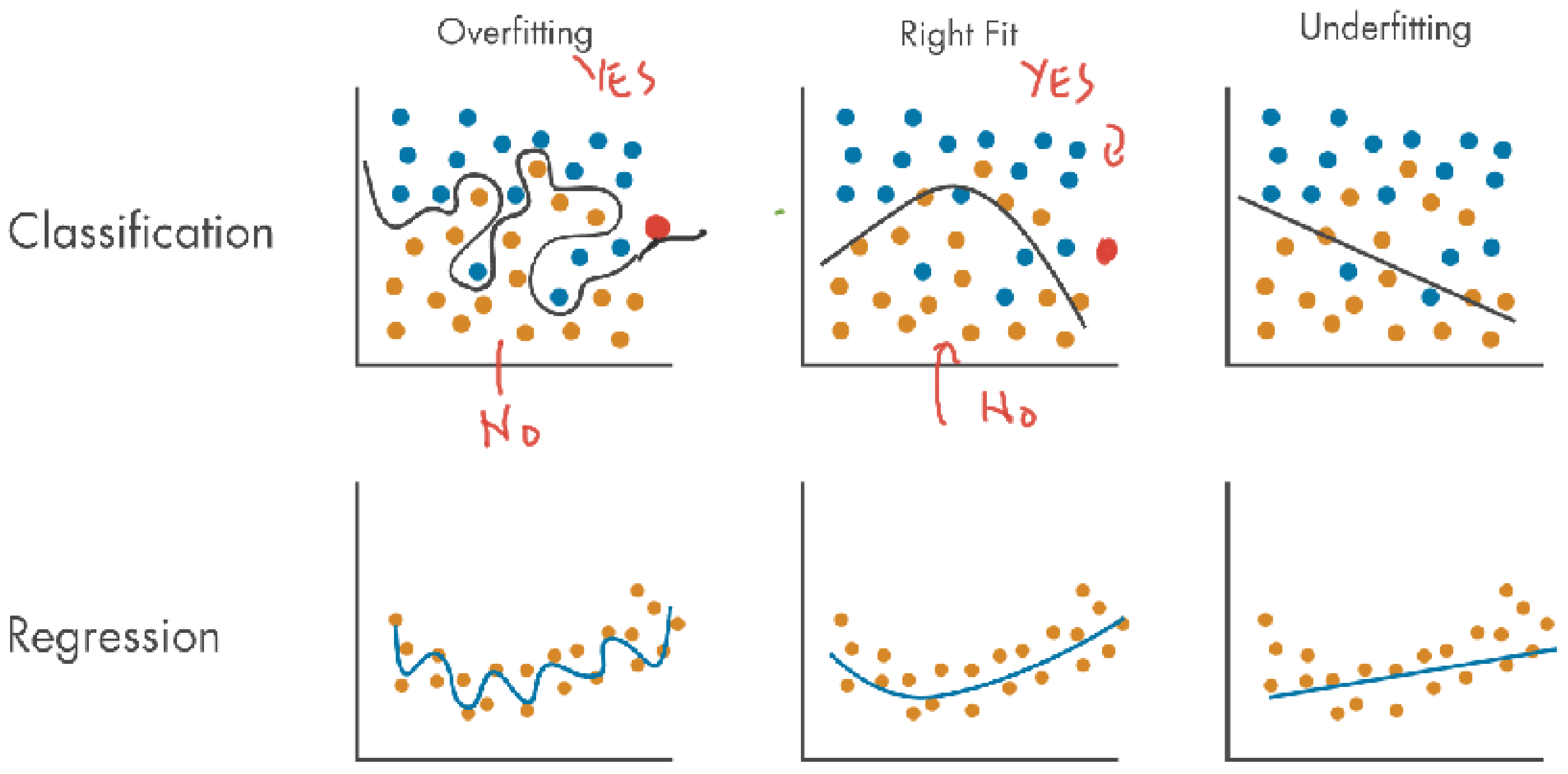
Problems with DT

1) Overfitting

2) overcome by 2 methods

a) Pre-Pruning

b) Post-Pruning



https://medium.datadriveninvestor.com/decision-tree-algorithm-with-hands-on-example-e6c2afb40d38

https://www.kaggle.com/code/prashant111/decision-tree-classifier-tutorial#14.-Decision-Tree-Classifier-with-criterion-entropy-

Ensemble Learning → The foundation of Random Forest

DT + SVM + NB

Why ensemble learning?

RF = DT + DT + DT ......

Example: 3 classifiers, each is correct with 60% accuracy

- Ensemble: combination of this 3 -> take the majority vote

- What is the accuracy of this ensemble?

- Answer: 3 votes - ensemble is wrong if there are at least 2 wrong predictions

F

right: T

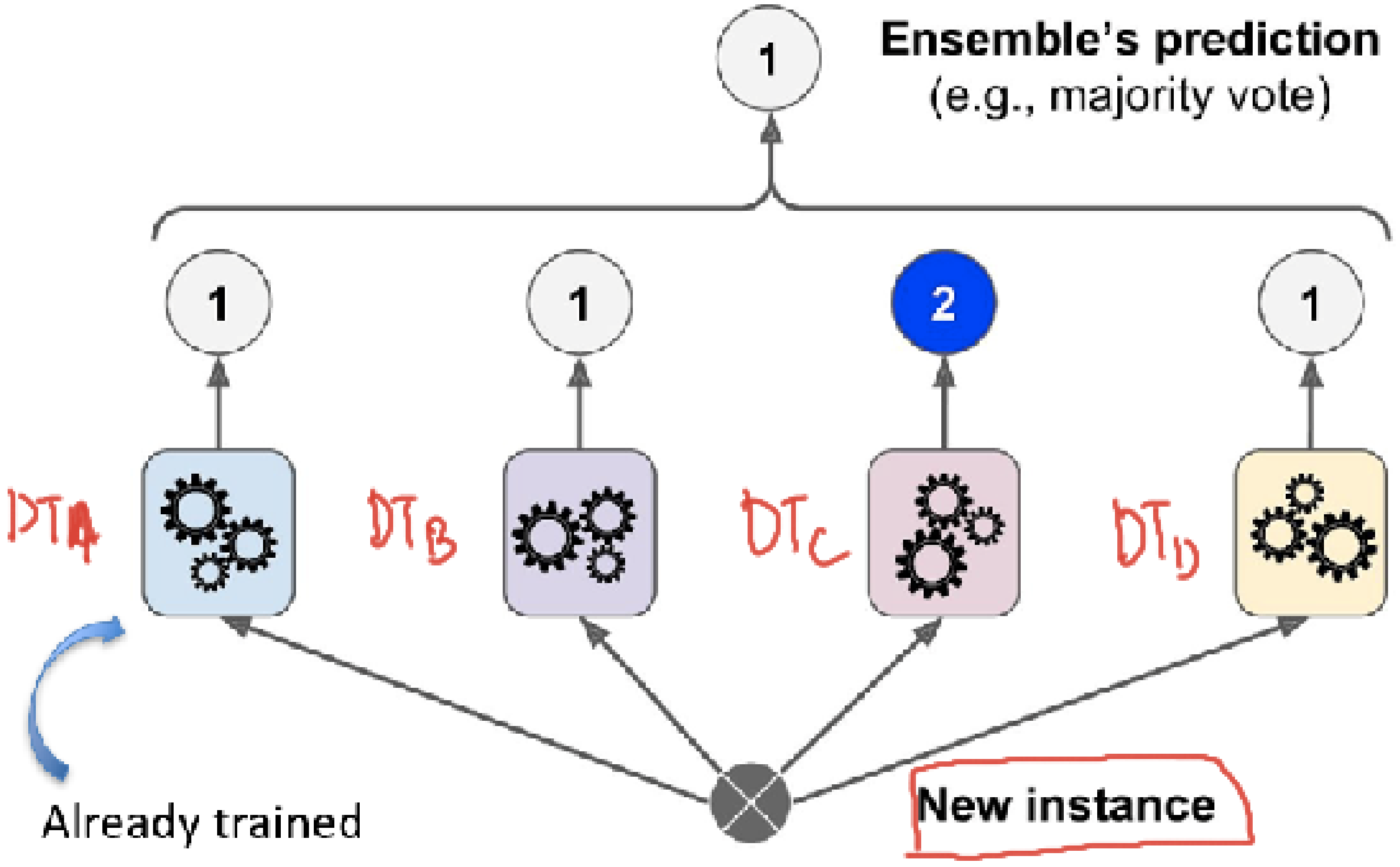
F F T

- Case 1 : Exactly 2 wrong predictions: 3*0.4*0.4*0.6 = 0.288
- Case 2: All 3 are wrong: 0.4^3 = 0.064
- Total: 0.352 ->ensemble is correct with 64.8% accuracy

- Another example: 1000 classifiers with 51% accuracy each -> 75% accuracy

Law of the large numbers

Ensemble learning



Common practices

Classifiers should be independent: DT_A is not dependent on DT_B

- Correlated classifiers don't work well together (why?)
  - E.g., 2 classifiers: both make mistakes at the same data points
  - Their combination will not be better at all
- Ensemble (completely) different classifiers SVM + NB + DT ✓
  SVM + SVM + NB ×
  RF uses mainly
  bagging
- Vary the training data (bagging and pasting)
- Vary the feature sets (random patches & random subspaces)

Bagging and pasting

$D_1, D_2 \ldots D_n$ ~ Sampling ~ $\mathcal{N} \sim (\mu, \sigma)$ = $D_i + \eta$ — $\eta \sim \mathcal{N}(\mu, \sigma)$

Samples are perturbed

- Use combination of same type of classifiers (e.g., all are SVMs)
- Randomly sample a subset of training data to train each classifier
- **Bagging** (= bootstrap aggregating): choose with replacement (**can** sample same data point for the same classifier multiple times)
- **Pasting**: choose without replacement (**cannot** sample the same data point for same classifier multiple times)

Random Forest

Sample randomly with replacement N=800

GPA	Studied	Pass
L	F	T
M	T	F
H	T	T
L	F	F

Train data

subspace

Test data