Last class (Sept 13)

1) Quizzes + Recap
2) Recap of splitting based on numerical & categorical features (encoding them)
3) Recap of max_depth for overfit vs underfit
4) Hyperparameter tuning
5) Visualizing DT
6) High Dimensional Data
7) Data Imbalance
8) Feature Importance
9) Regression using DT
10) ~~Overview of Bagging and Random Forests~~

Today's class

1) Quizzes
2) Pruning & recap of best depth selection
3) Ensembles & bagging
4) Random Forest & Combining decision trees.
5) Randomness in model

6) Validating RF (Random Forest)

7) Overall performance

8) OOB Score (Out-of-Bag)

9) Bias Variance Trade off

10) Reducing Variance

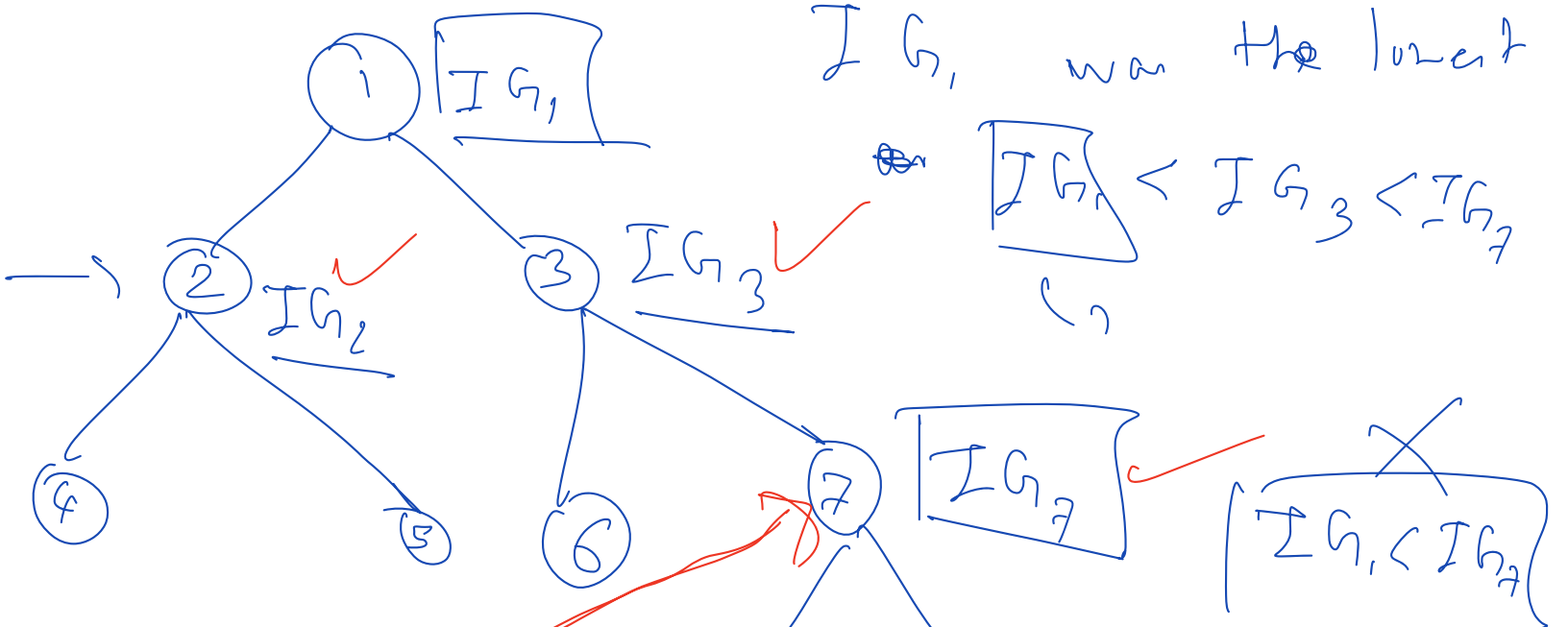
11) Code

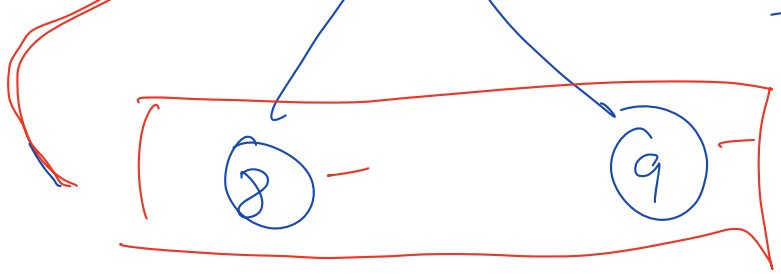
12) Optimizing RF

13) Hyper-parameter tuning

14) Computing Feature Importance

Pruning

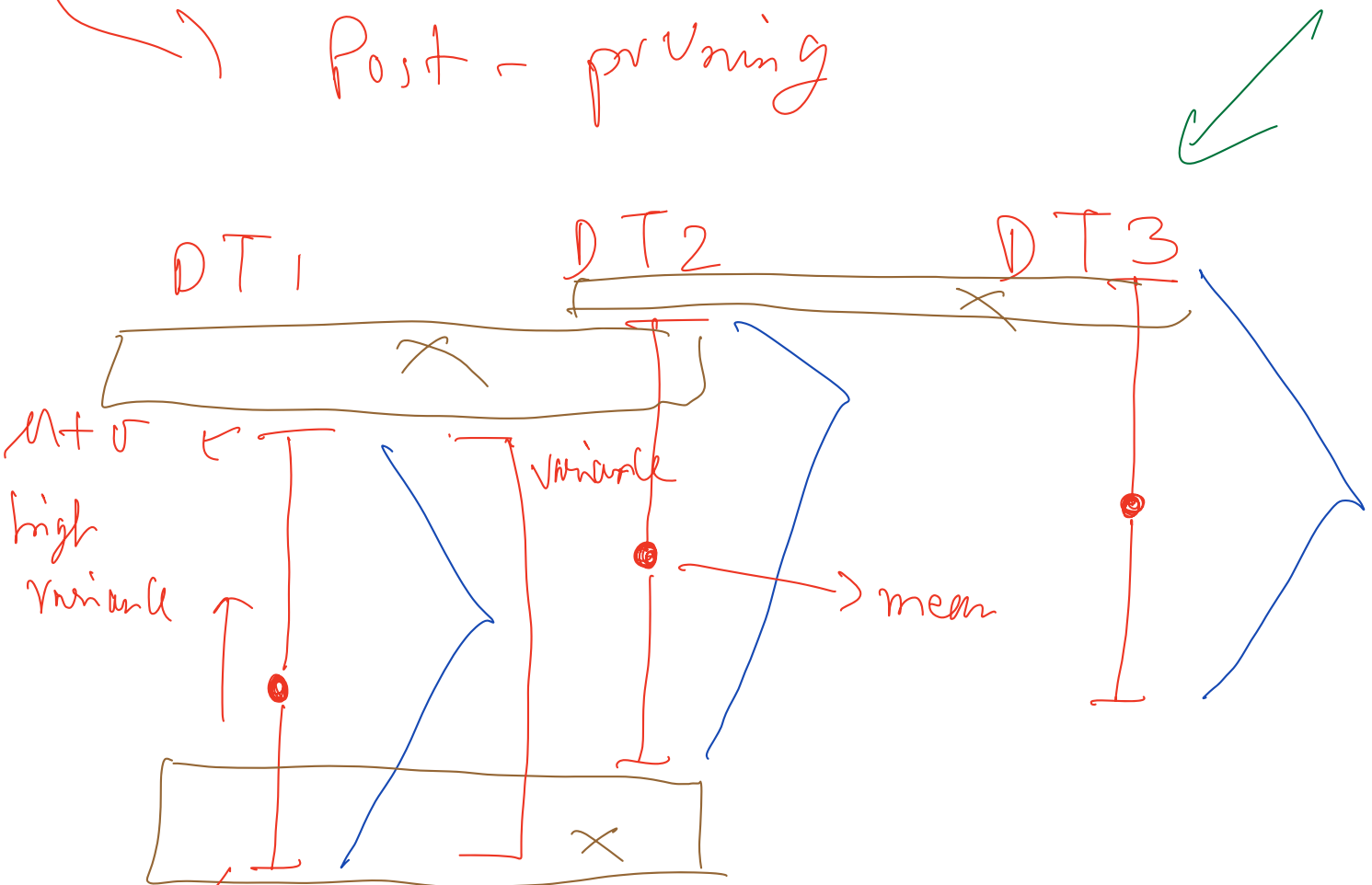


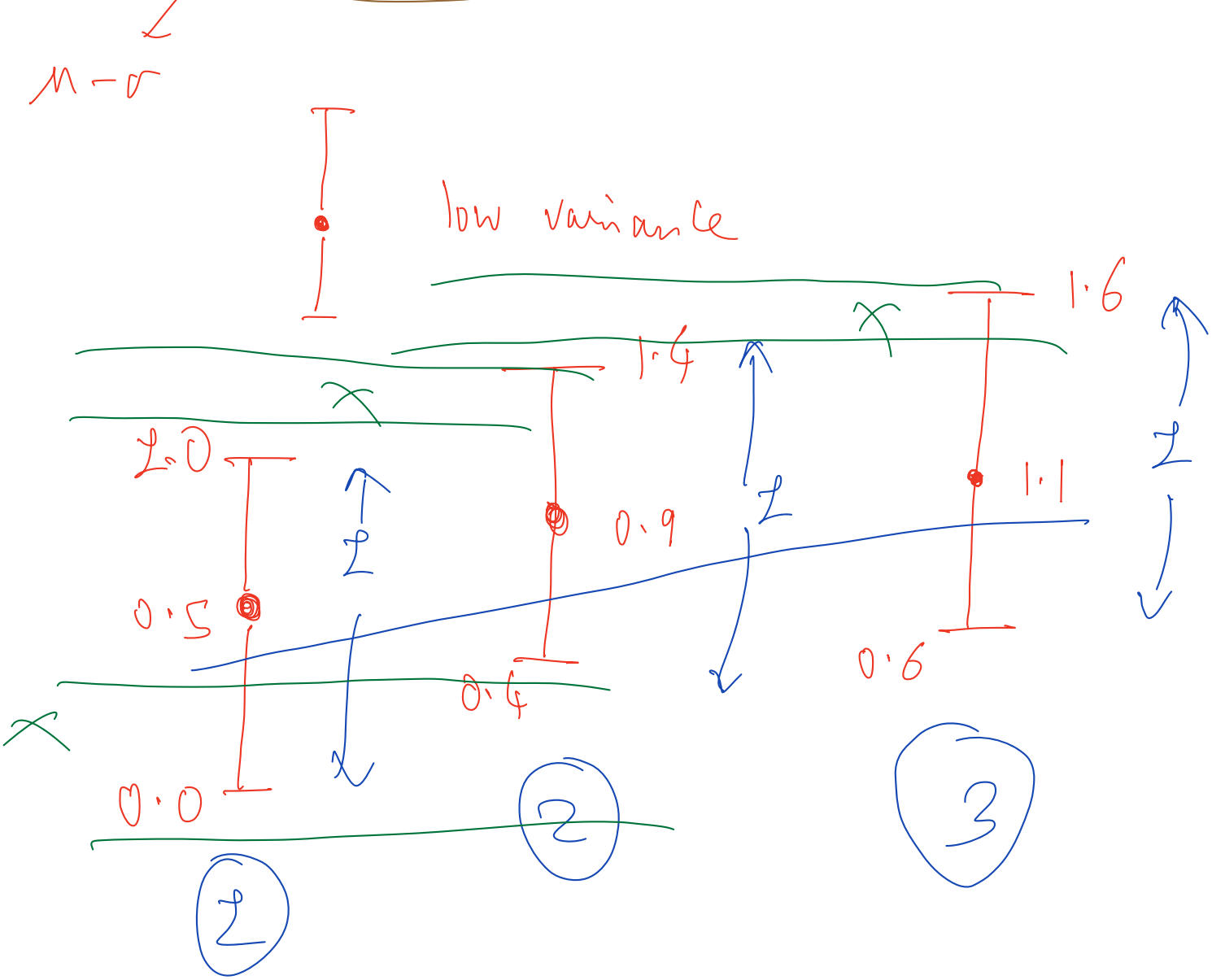


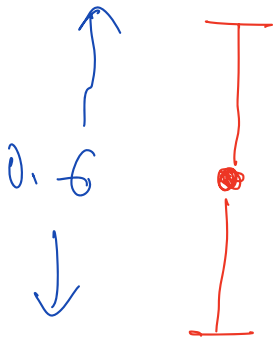
Find the split with lowest IG.
Merge if back to parent node
check for overfitting.

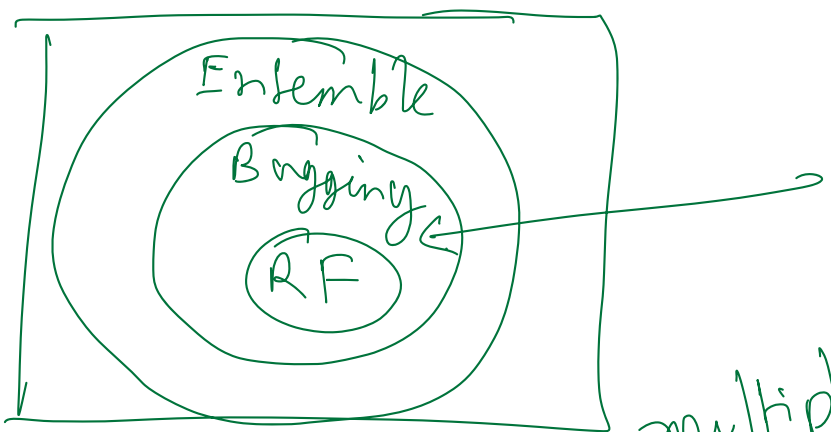
IG_7 , IG_3 , IG_2

→ Post-pruning

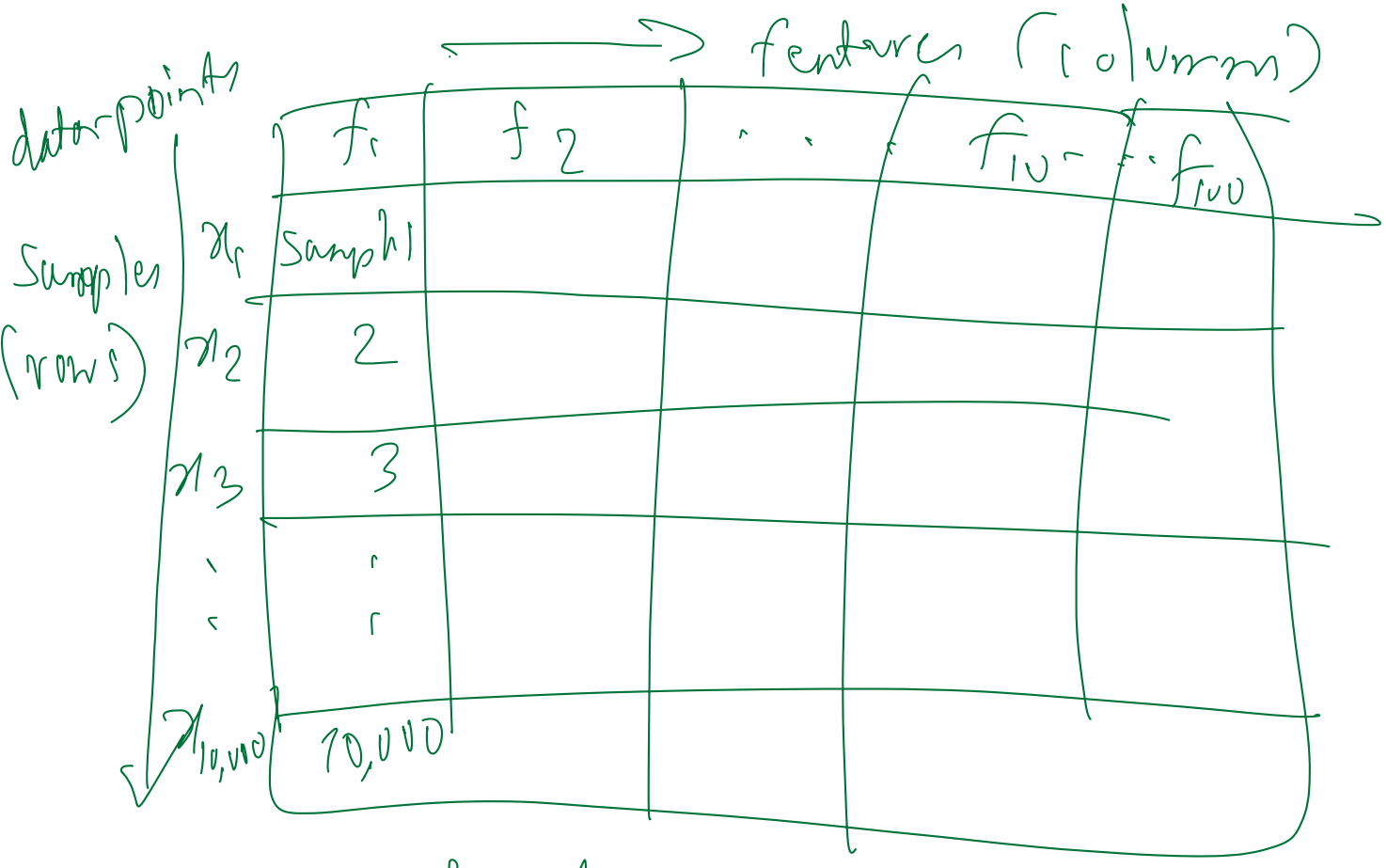








Bagging
↓
combination of
multiple log-reg



100 out of 10,000

$x_{100}, x_{1000}, x_1, x_2, x_{9999}, \ldots$

Row sampling 100 Samples

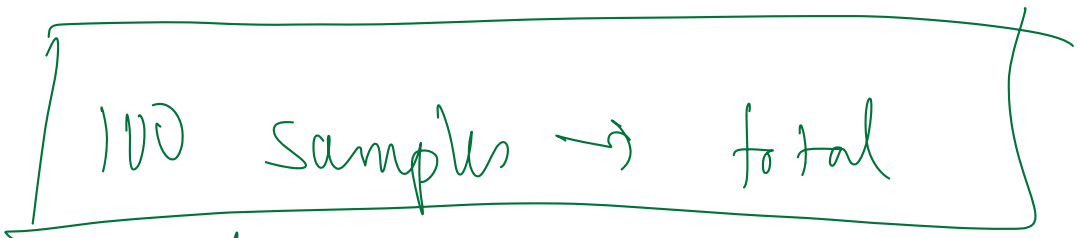
$f_1, f_2, f_3, f_4, \ldots f_{100}$

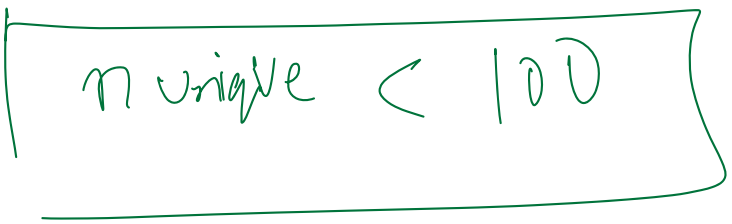
10 features out of 100

$f_5, f_{94}, f_{80}, f_{20}, \ldots$

10 features

column sampling





↳ 100 samples

still a sub-set of these 100 samples

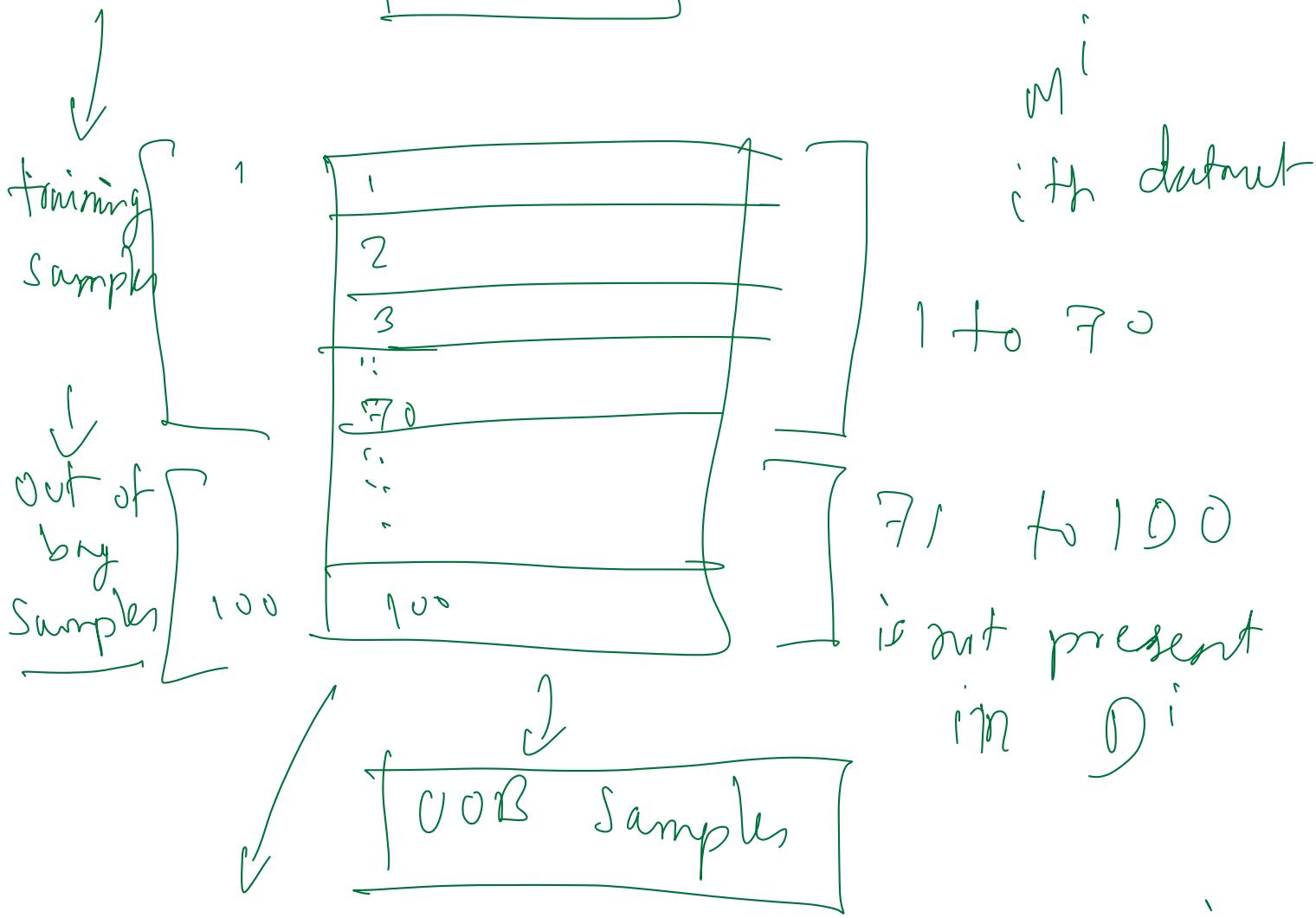
N samples

row sampling ratio → [0.0, 1.0]

1.0

with replacement

10:25 → Resume



OOB samples

OOB^i is used for tuning M^i

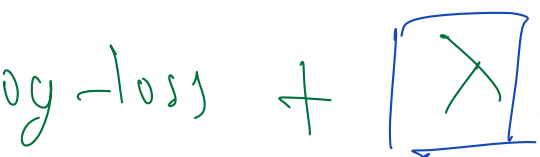
Regularization

Loss fn = log-loss/t + $\left[\sum_{i=1}^{d} w_i^2\right]$

MSE

L2 regularization values

D.T

loss fn = log-loss + 

ccp_alpha