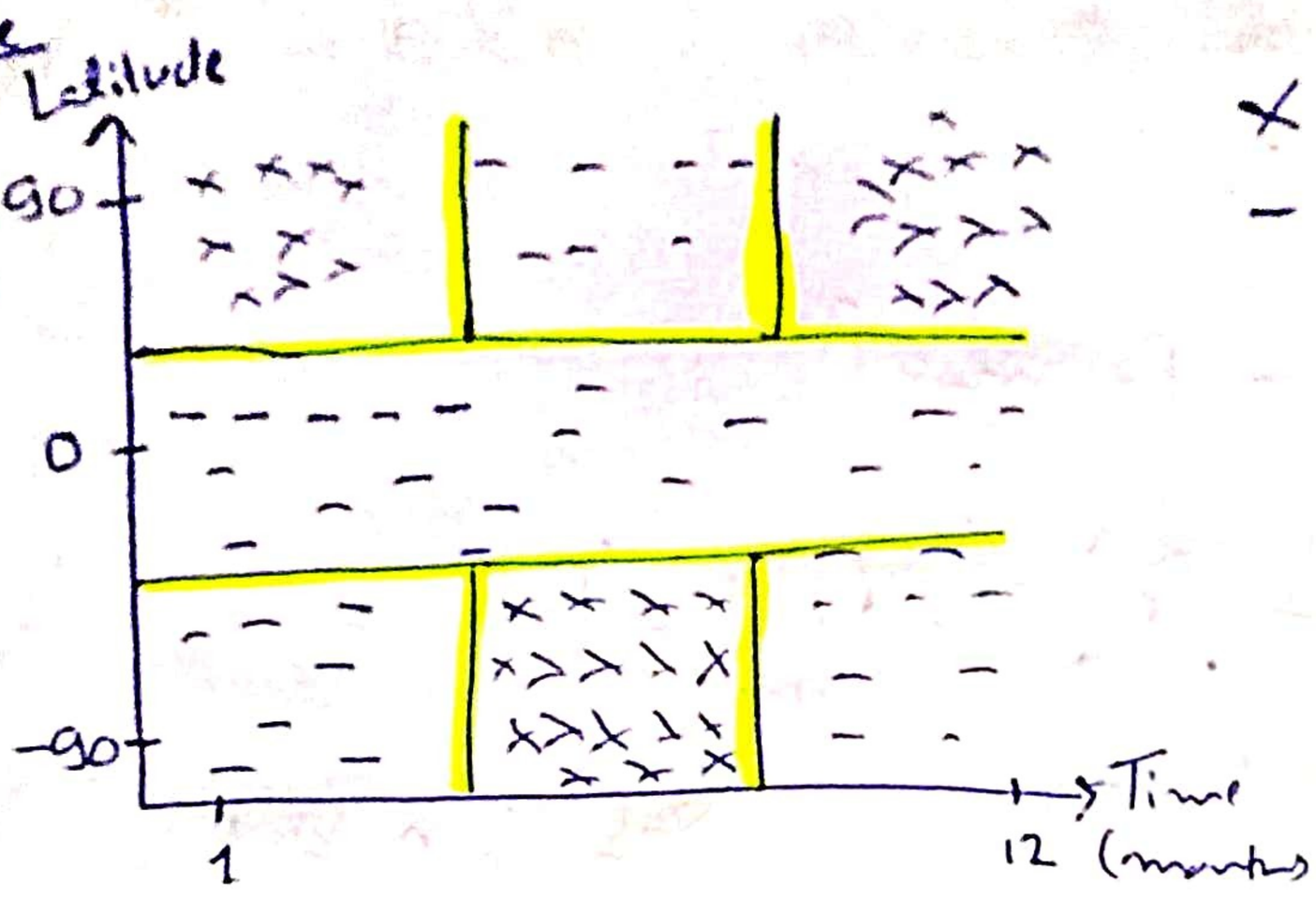
(10)

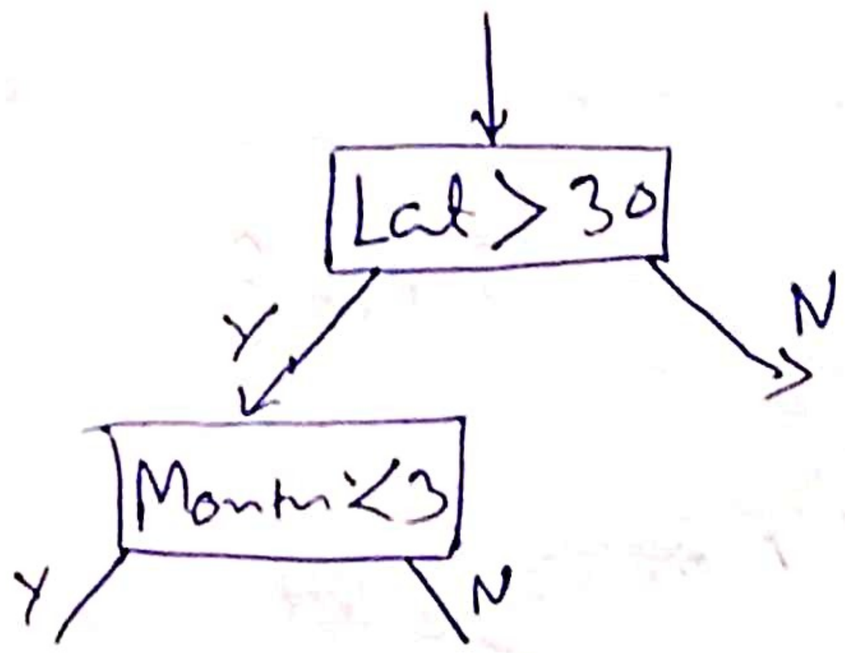
Decision Trees & Ensemble Methods

* Decision Trees

Example



⇒ Greedy, Top-Down, Recursive Partitioning.



⇒ Region R

⇒ Looking for a split S_p

$$S_p(j, t) = \left(\{X \mid X_j < t, X \in R_p\}, \{X \mid X_j \geq t, X \in R_p\} \right)$$

$\{X \mid X_j < t, X \in R_p\} \rightarrow R_1$

$\{X \mid X_j \geq t, X \in R_p\} \rightarrow R_2$

j: Feature number

t: Threshold

⇒ How to choose splits?

⇒ Define $L(R)$: loss on R

⇒ Given C classes, define $\hat{P}_c$ to be the Proportion of examples in R that are of class c.

$$L_{misclass} = 1 - \max_c \hat{P}_c$$

$$\max_{j,t} \; L(R_p) - \left(L(R_1) + L(R_2)\right)$$

$L(R_p)$ → Parent loss

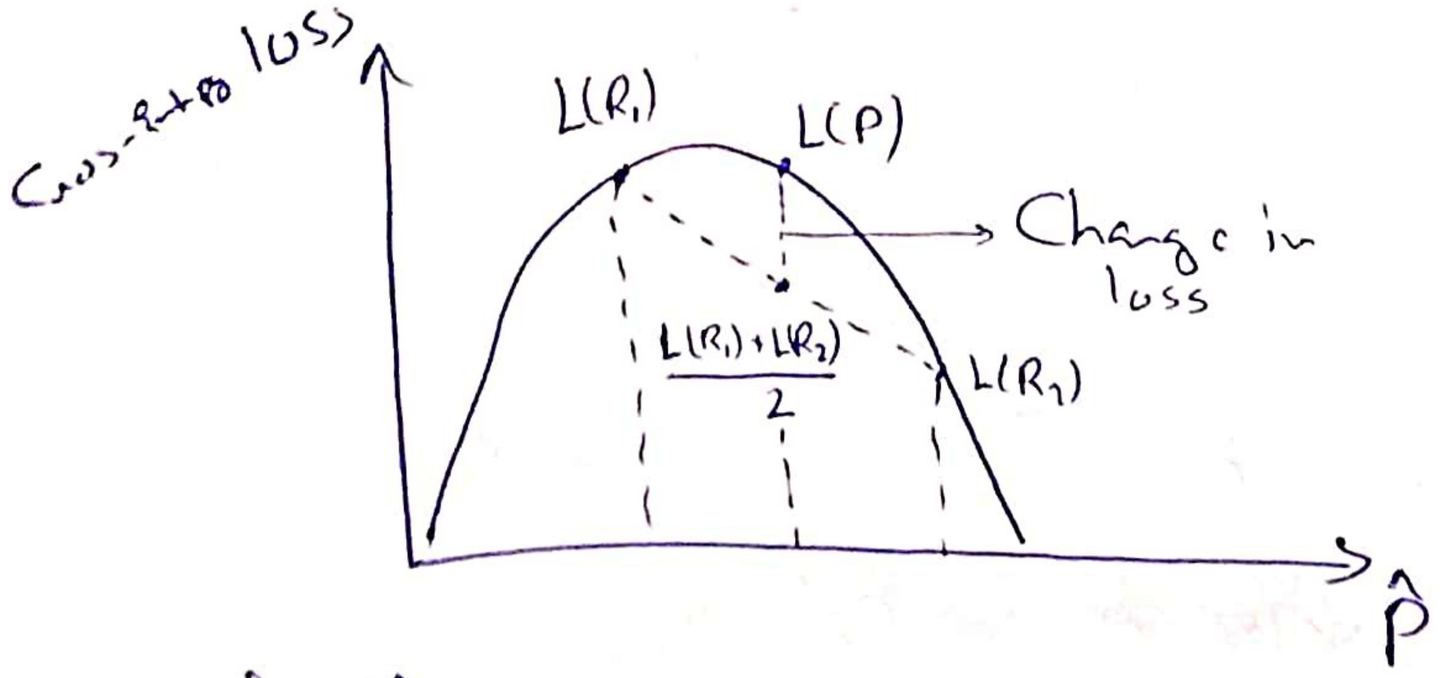
$L(R_1) + L(R_2)$ → Children loss

Misclassification loss

⇒ Misclassification Loss has issues!!

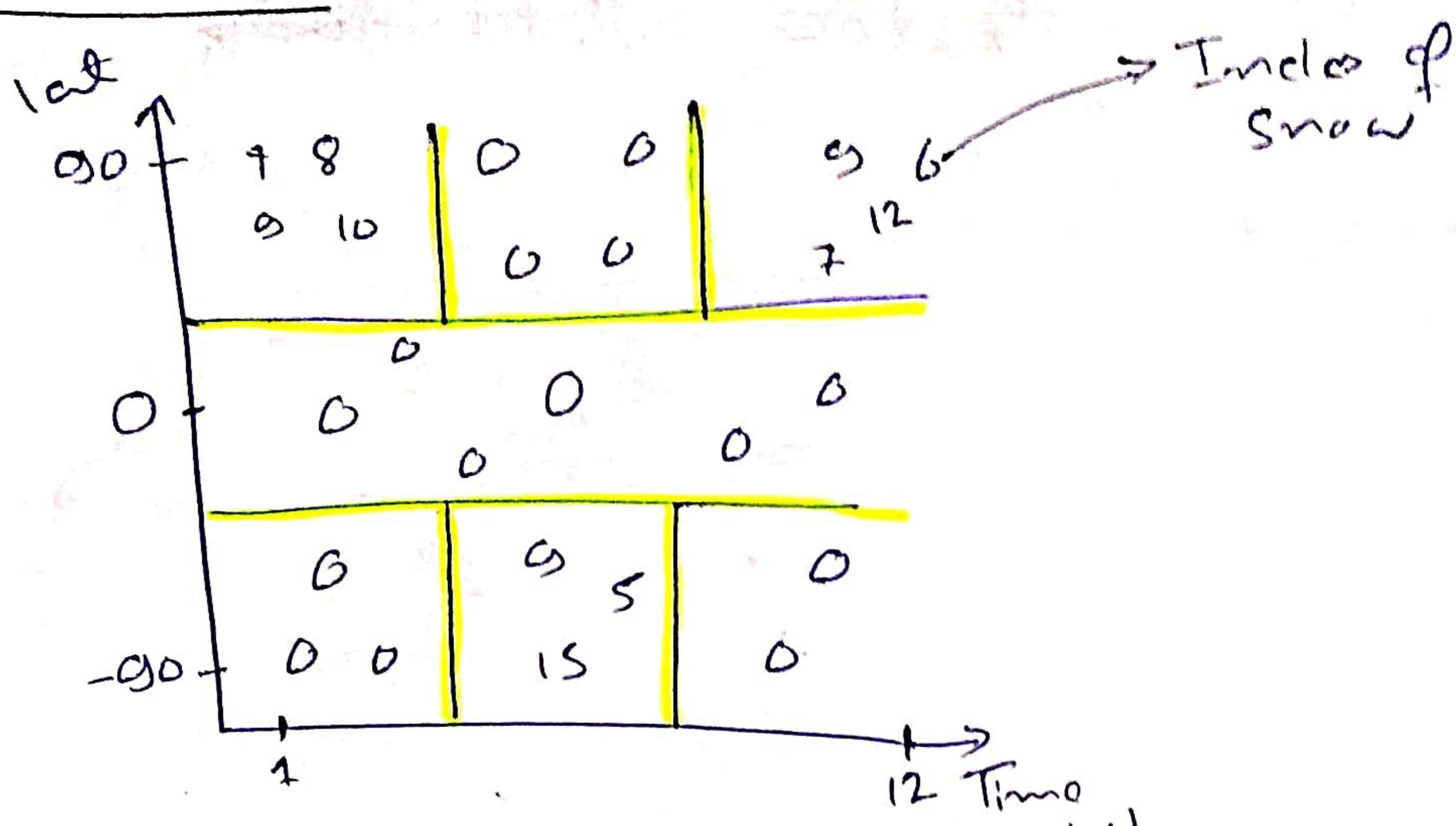
⇒ Instead, define Cross-entropy loss

$$L_{Cross} = \sum_c \hat{P}_c \log_2 \hat{P}_c$$



$$\text{Gini Loss} = \sum_c \hat{P}_c(1-\hat{P}_c)$$

* Regression Trees



$$\text{Predict } \hat{Y}_m = \frac{\sum_{i \in R_m} Y_i}{|R_m|}$$

$$L_{squared} = \frac{\sum_{i \in R_m} (Y_i - \hat{Y}_m)^2}{|R_m|}$$

* Regularization of DTs { Heuristics }

1. min leaf size
2. max depth
3. max number of nodes
4. min decrease in loss
5. Pruning (misclassification with validation set)

⇒ Down side of decision trees:

- No additive structure
- High Variance model
- Generally Low predictive accuracy

★ Ensembling

⇒ Take X_i s which are RV and are IID.

$$Var(X_i) = \sigma^2$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\Sigma X_i\right) = \frac{\sigma^2}{n}$$

⇒ Drop the independence assumption

So now X_is are id

X_i^s correlated by ρ

$$Var(\bar{X}) = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$$

⇒ Ways to ensemble :-

1) different algorithms
2) different training set
3) Bagging (Random Forest)
4) Boosting (Adaboost, Xgboost)

⊞ Bagging (Bootstrap Aggrigation)

→ Have a true population P
→ Training Set S ~ P
→ Assume P = S
→ Bootstrap sample $Z \sim S$

"We will take a bunch of bootstrap samples, train seperate models on each and then average their output"

- Bootstrap Samples $Z_1 \cdots Z_{pM}$

Train model G_m on Z_m

$$G(m) = \frac{\sum_{m=1}^{M} G_m(x)}{M}$$

Random Forests

↳ At each split, Consider only a fraction of your total features.

★ Boosting

↳ Decreasing bias
↳ Additive

⇒ It increases the weights on the mistake of previous model.