

O que é melhor?

- 1) Um sábio: acerta 95% das vezes
- 2) 10000 goiabas: acerta 51% das vezes

$X_i$ : resposta da pessoa  $i$

$X_i \sim_{ind} \text{Bernoulli}(p = 0,51)$

Democracia das goiabas:  $Y$ : contagem de sucessos em  $n = 10000$   
tentativas

$Y \sim \text{Binomial}(n = 10000, p = 0,51)$

Decisão das goiabas: acertam se 5001 <sup>ou mais</sup> goiabas acertarem

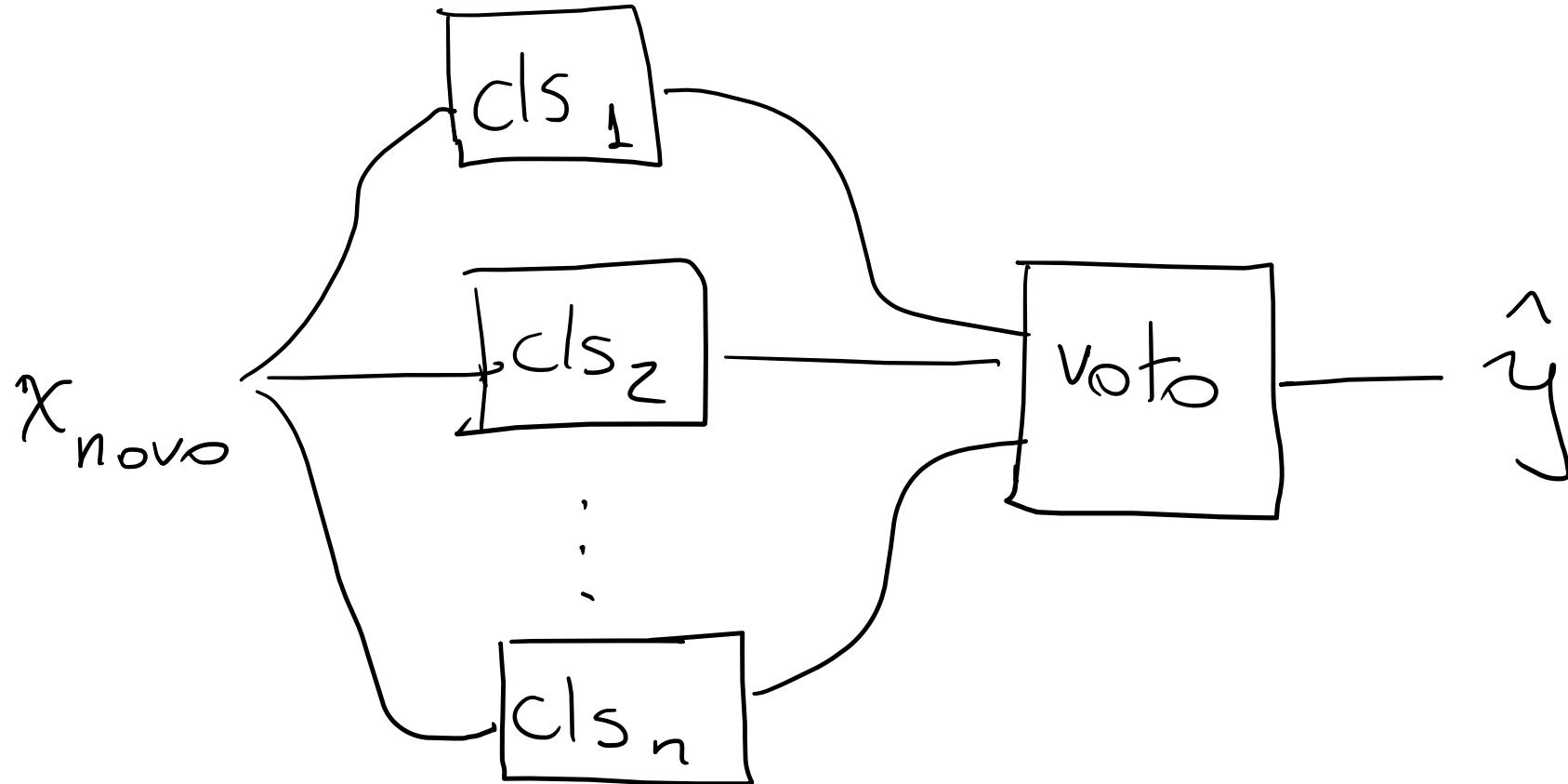
$$P(Y > 5000) = ?$$

$$P(Y > 5000) = 1 - P(Y \leq 5000) \rightarrow \text{Python} \rightarrow 97,7\%$$

## Princípios da democracia: em machine learning

- $p > 50\%$ , pelo menos melhor que totalmente random
    - "Aprendizes" fracos (weak learners) → modelos fracos
  - $n$  grande, para cancelar os erros
    - número grande de modelos
      - Algoritmos diferentes, mesmos dados
      - mesmo algoritmo, dados diferentes
  - independencia
    - Algoritmos diferentes, mesmos dados
    - mesmo algoritmo, dados diferentes
- mesmo algoritmo,  
subamostras do  
mesmo dado

# Voting Classifier



- Bootstrap:
- distribuição  $F$  = desconhecida
  - Coleta  $n$  amostras
    - $X_i \sim_{ind} F$
    - Cada valor observado é  $x_i \leftarrow$  amostra de  $X_i$

$(x_1, x_2, \dots, x_n) \leftarrow$  conjunto de observações

- Quero estimar  $\bar{F} = E_F[X]$

- 1)  $\bar{F} = \frac{1}{n} \sum_{i=1}^n x_i$  media amostral

*Como cada  $x_i$  veio de uma V.A.  $X_i$ , então  $\bar{F}$  por si só é uma V.A.!*

*Em outras palavras: cada vez que eu repetir o processo (obter amostras, calcular  $\bar{F}$ ), vai dar um resultado diferente!*

$\bar{F}$  é v. A.  $\Rightarrow$  tem média, variância, histograma, etc.

Como obter um histograma de  $\bar{F}$  só com as amostras  $(x_1, x_2, \dots, x_n)$ ?

## BOOTSTRAP!

“to pull oneself up by one's own bootstraps.”  
se erguer sozinho, a perdir do nada

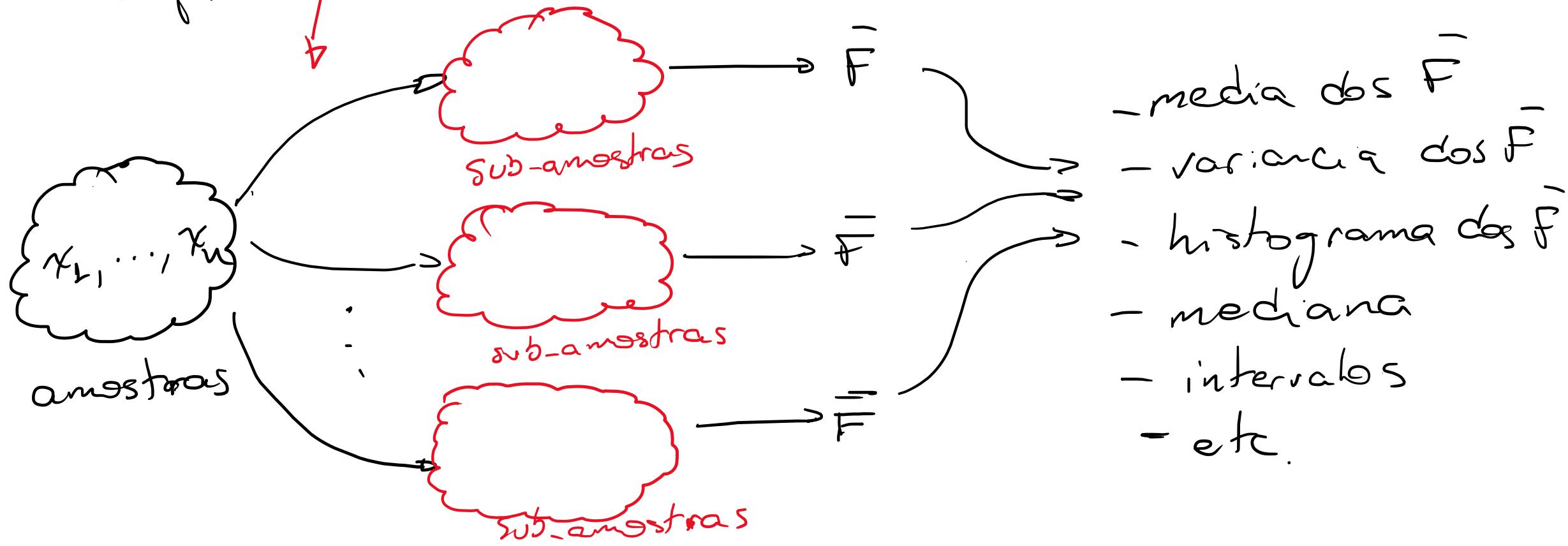
Bradley Efron (1979)

Bootstrap  $\rightarrow$  generalizações do “jackknife”  
 $\hookrightarrow$  convete

$\bar{F}$  é v.a.  $\rightarrow$  tem média  $\rightarrow$  se eu tivesse  
varias amostras  
de  $\bar{F}$  tirava  
a média delas!

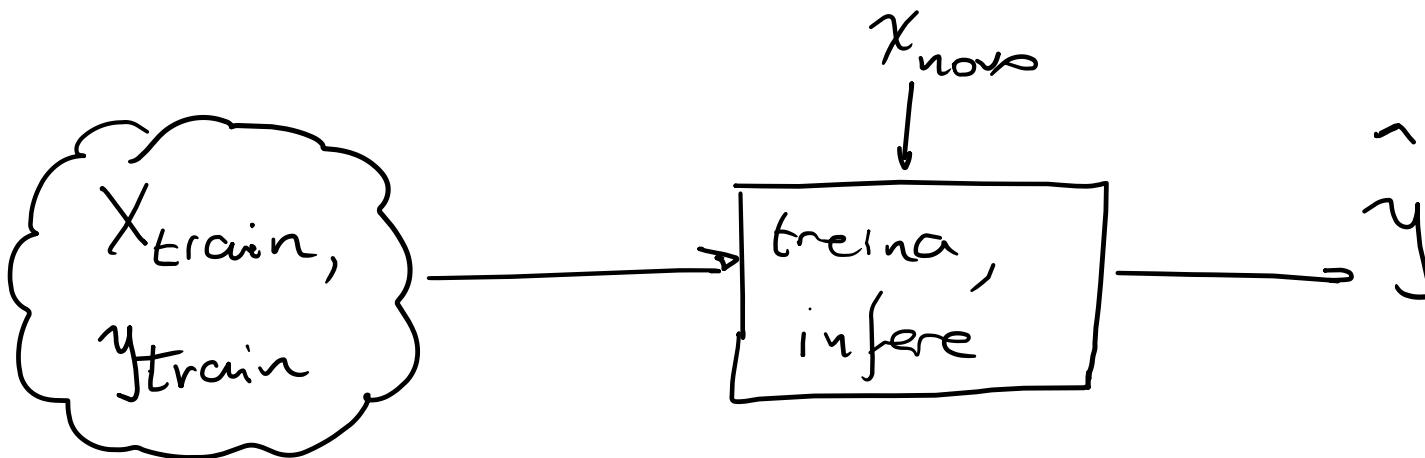
Como obter  
amostras de  
 $\bar{F}$ ?

bootstrap:  
subamostras aleatórias  
com repetição

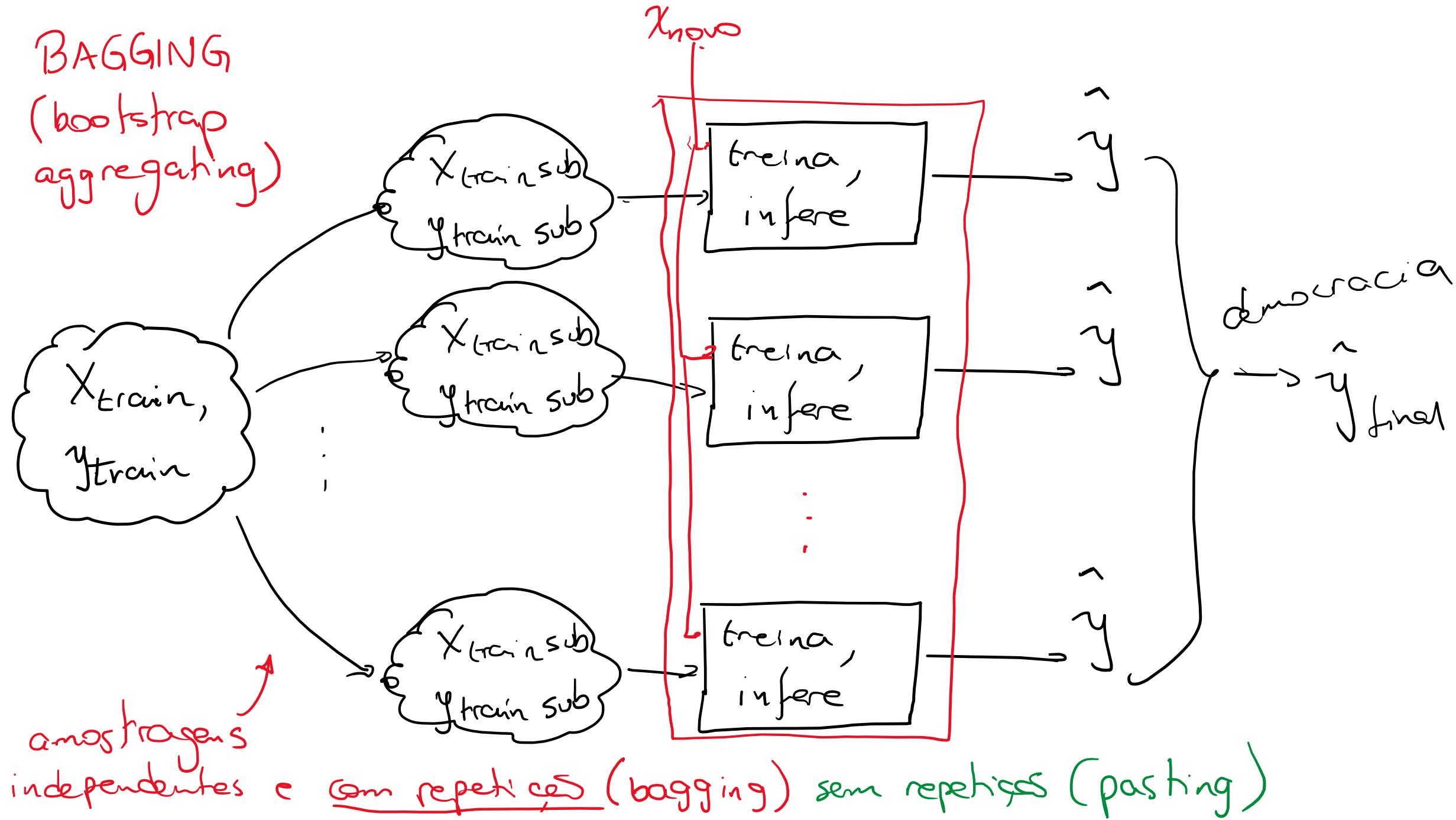


O processo de decidir classe de  $x_{novo}$ :

- Treina modelos com  $X_{train}, y_{train} \rightarrow$  modelos
- Inferência:  $\hat{y} = \text{modelo.predict}(x_{novo})$

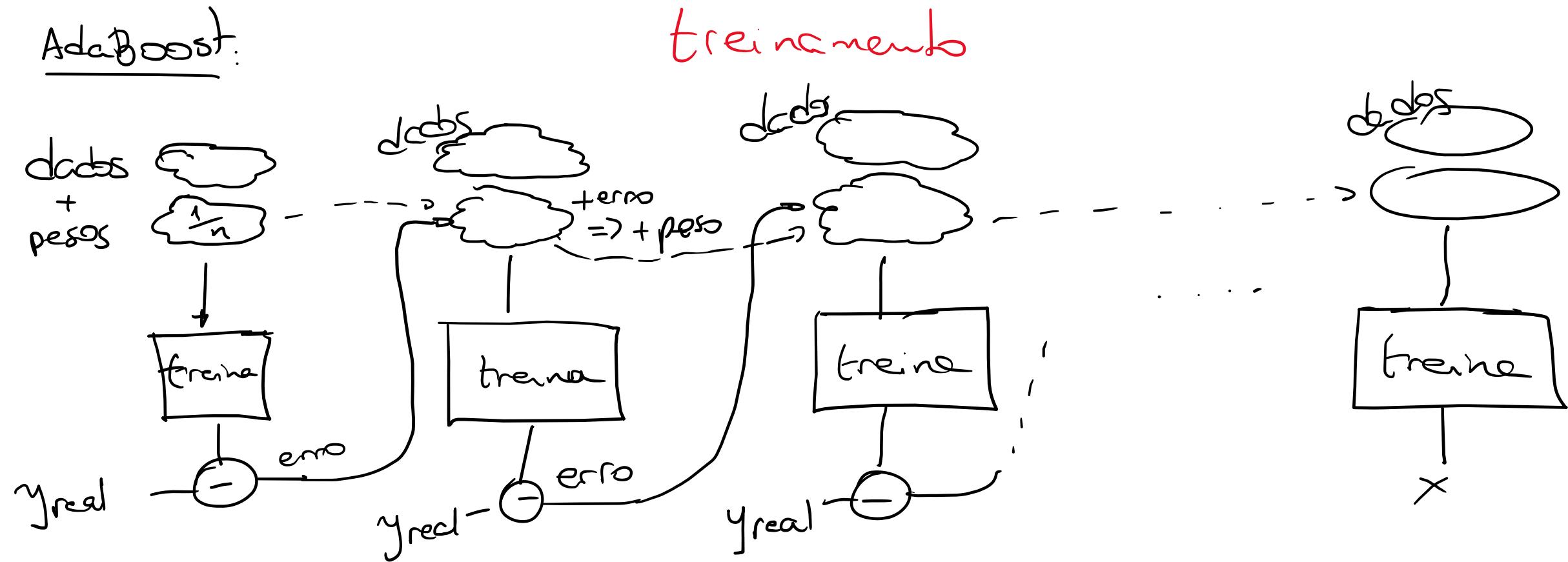


# BAGGING (bootstrap aggregating)



Boosting: aprendizado sequencial de uma coleção de modelos

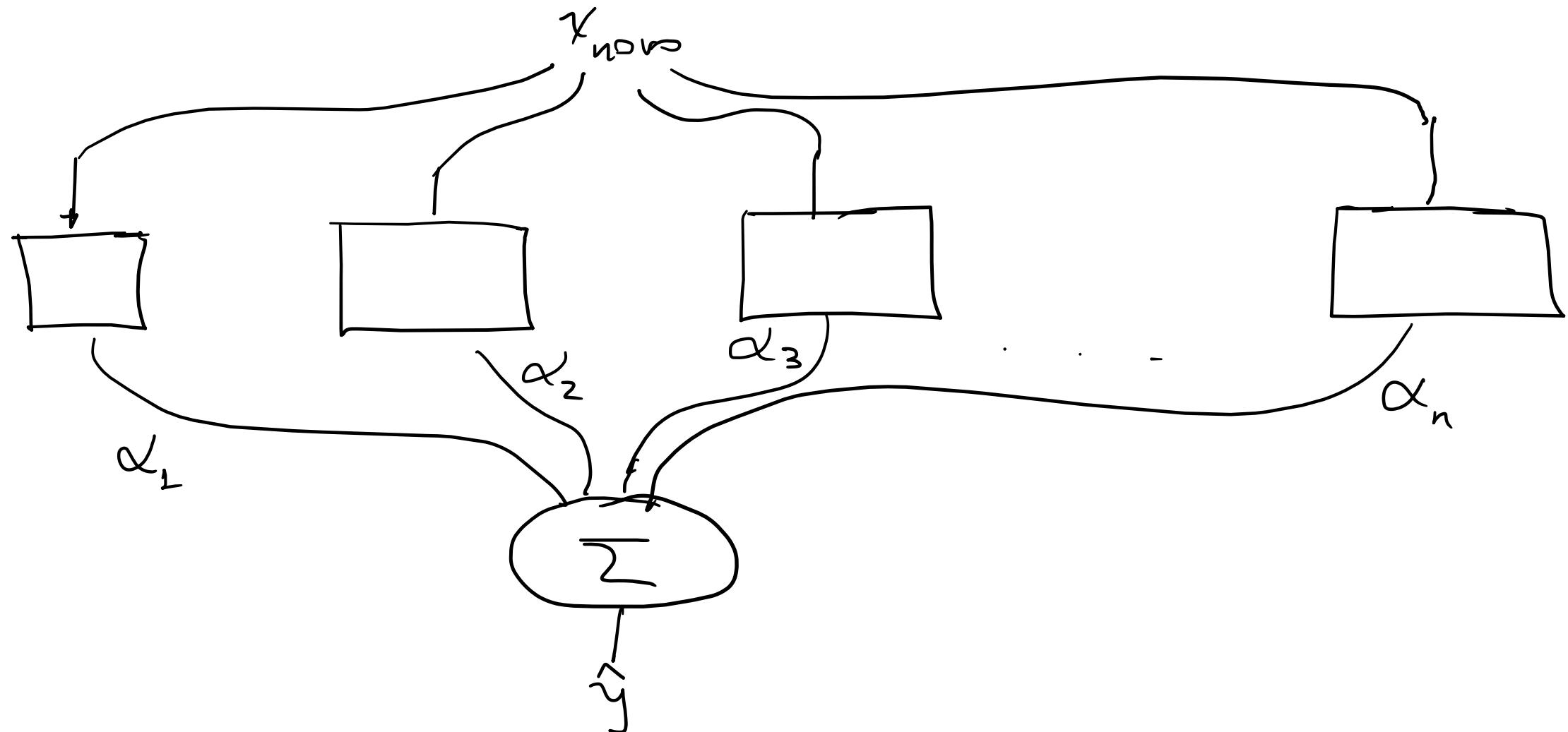
AdaBoost:



Boosting: aprendizado sequencial de uma coleção de modelos

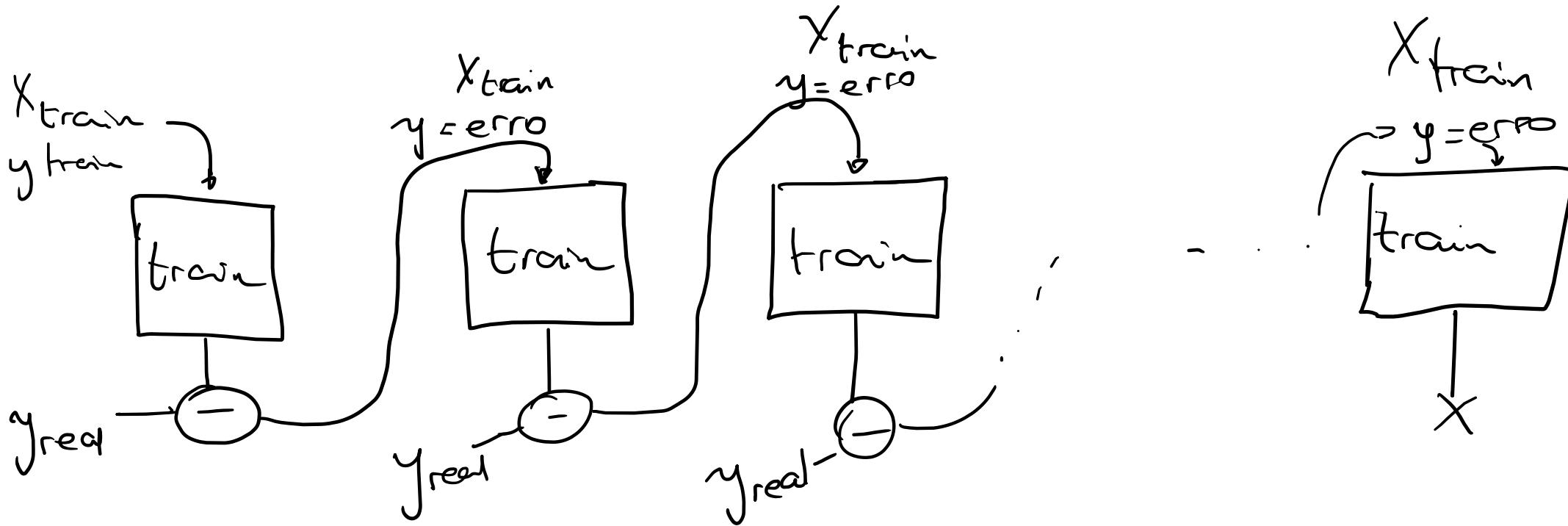
AdaBoost:

Inferência



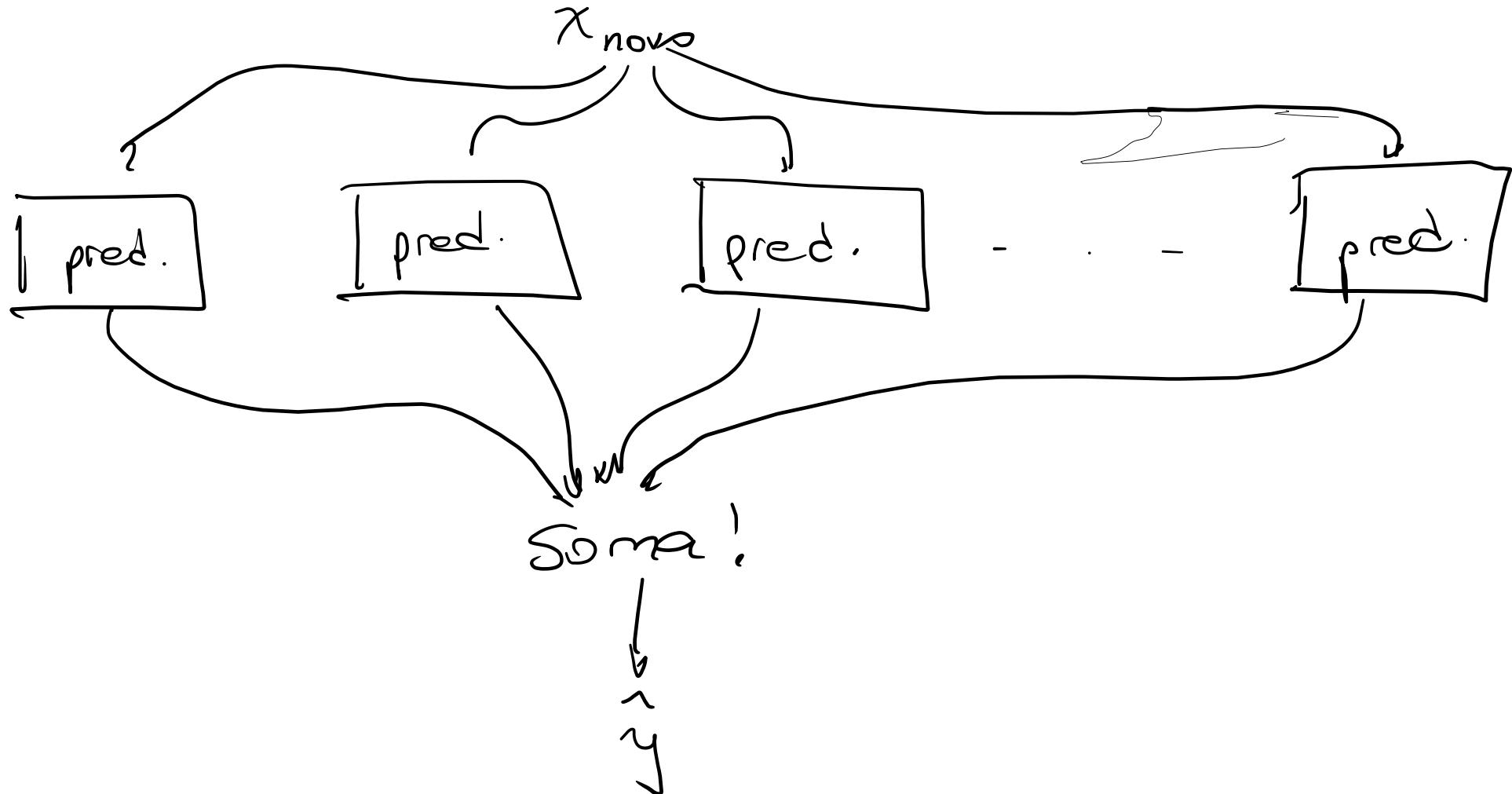
# Gradient Boosting

trainments



# Gradient Boosting

inferência

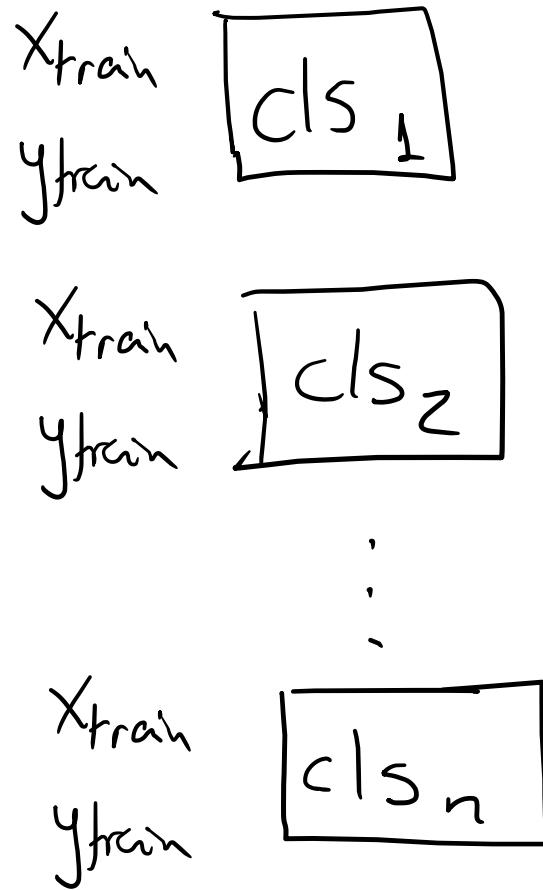


# Stacking

Fase 1

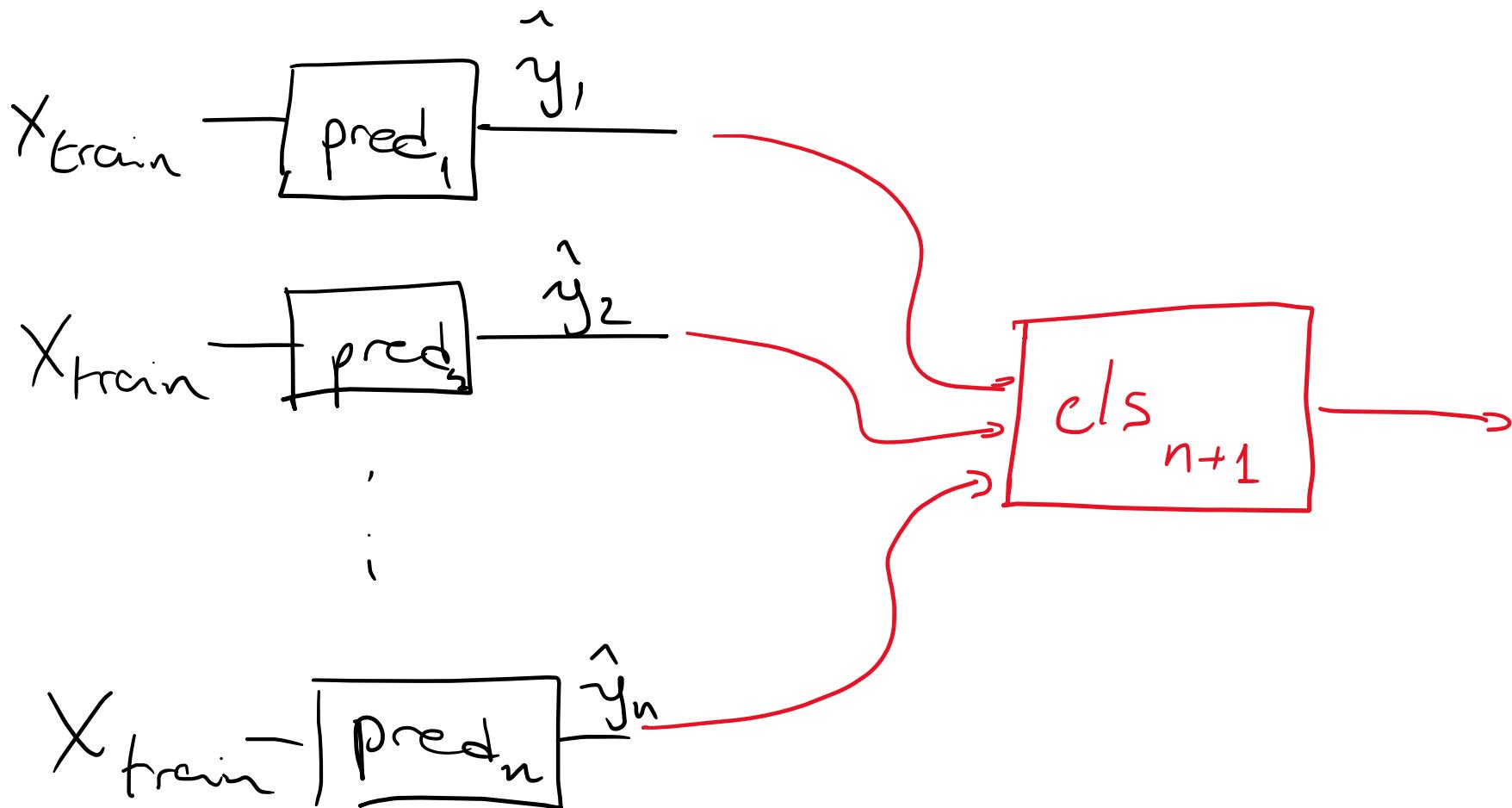
treina n cls

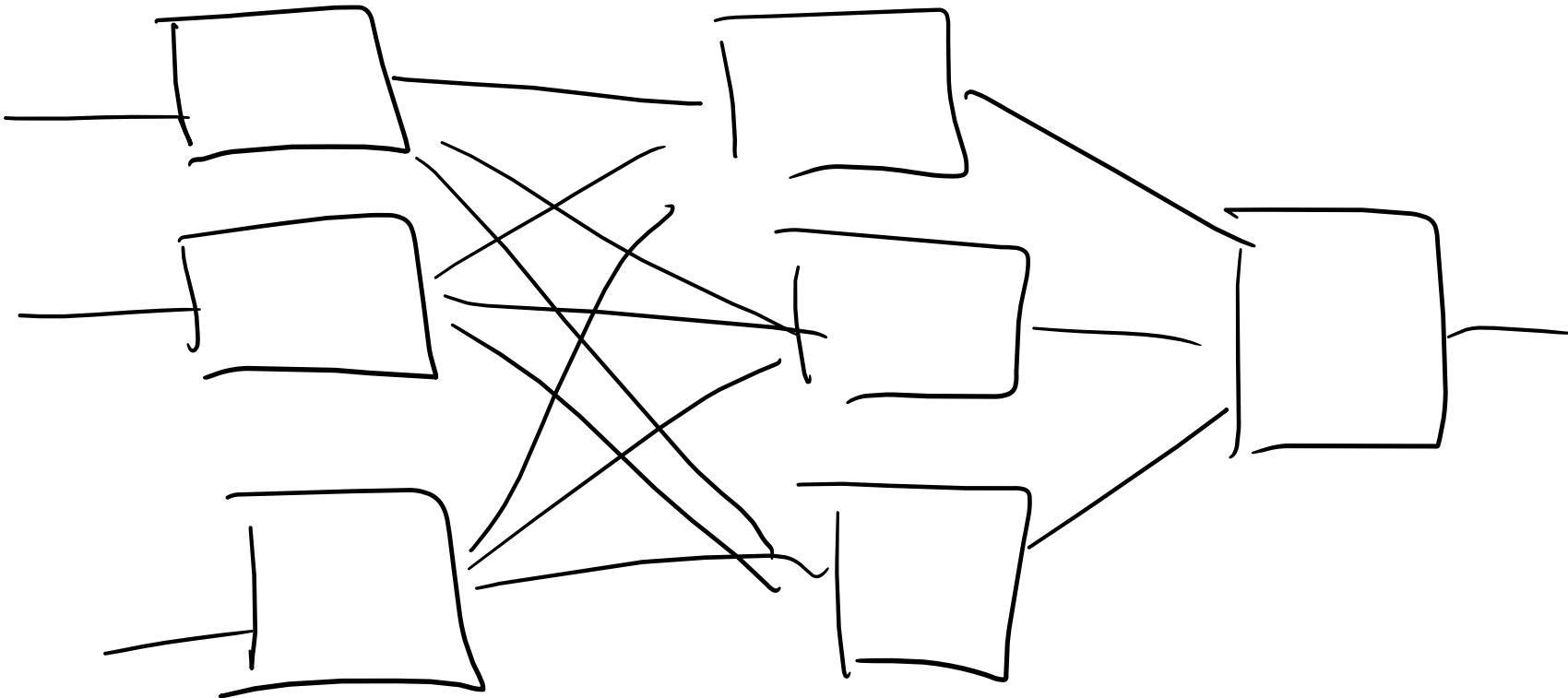
Separable



## Stacking

Fase 2 : training  $\text{cls}_{n+L}$





Stacking de reg. logísticas  $\approx$  rede  
neural