

Aplicações de Inteligência Artificial

Parte 3.5

Hitoshi Nagano, Ph.D.

referencias

- Max Kuhn, Kjell Johnson: Applied Predictive Modeling, Cap 8
- Kevin P. Murphy: Machine Learning, A Probabilistic Perspective, Cap 16
- Trevor Hastie, Robert Tibshirani, Jerome Friedman: Elements of Statistical Learning, Caps 9 e 10
- Jerome H. Friedman: Greedy Function Approximation: A Gradient Boosting Machine
- Cheng Li: A Gentle Introduction to Gradient Boosting. Disponível em http://www.chengli.io/tutorials/gradient_boosting.pdf

ALGORITMOS BASEADOS EM ENSEMBLE

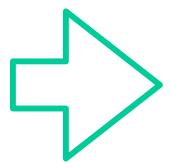
ensemble

- que problemas?
 - viés
 - variância
 - ruído
- ensemble tenta amenizar viés e variância
- ... através da coletividade de vários preditores

Evolução dos ensembles



boosting: também é um ensemble



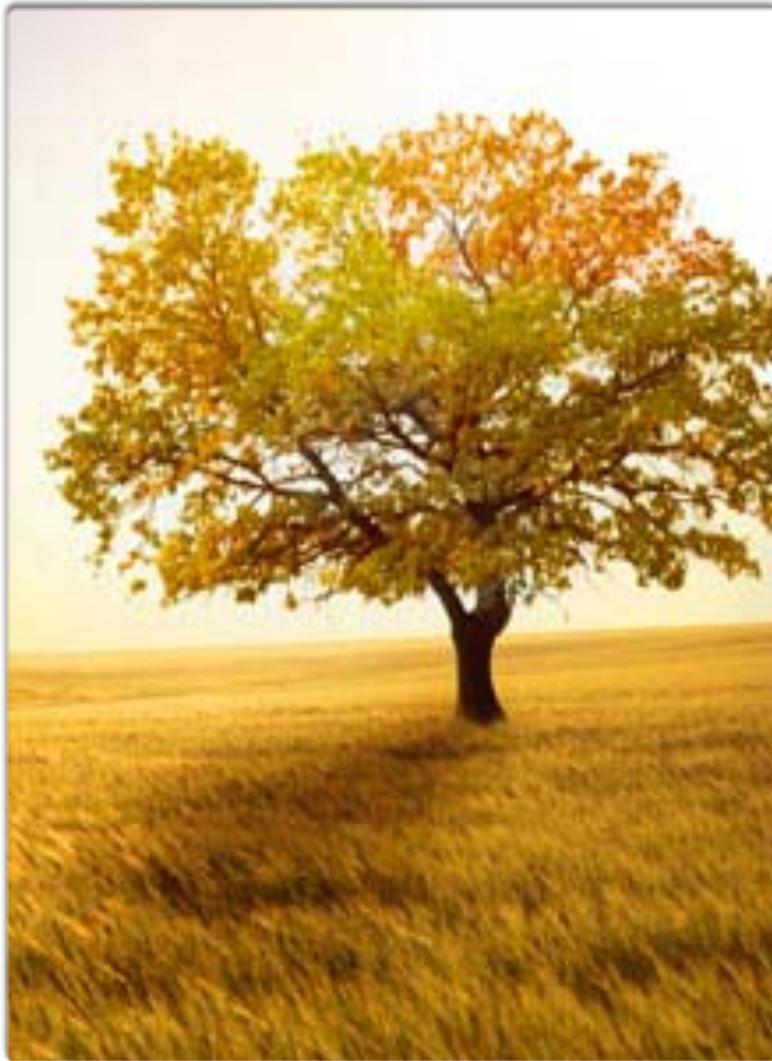
boosting: também é um ensemble



preditores sequenciais

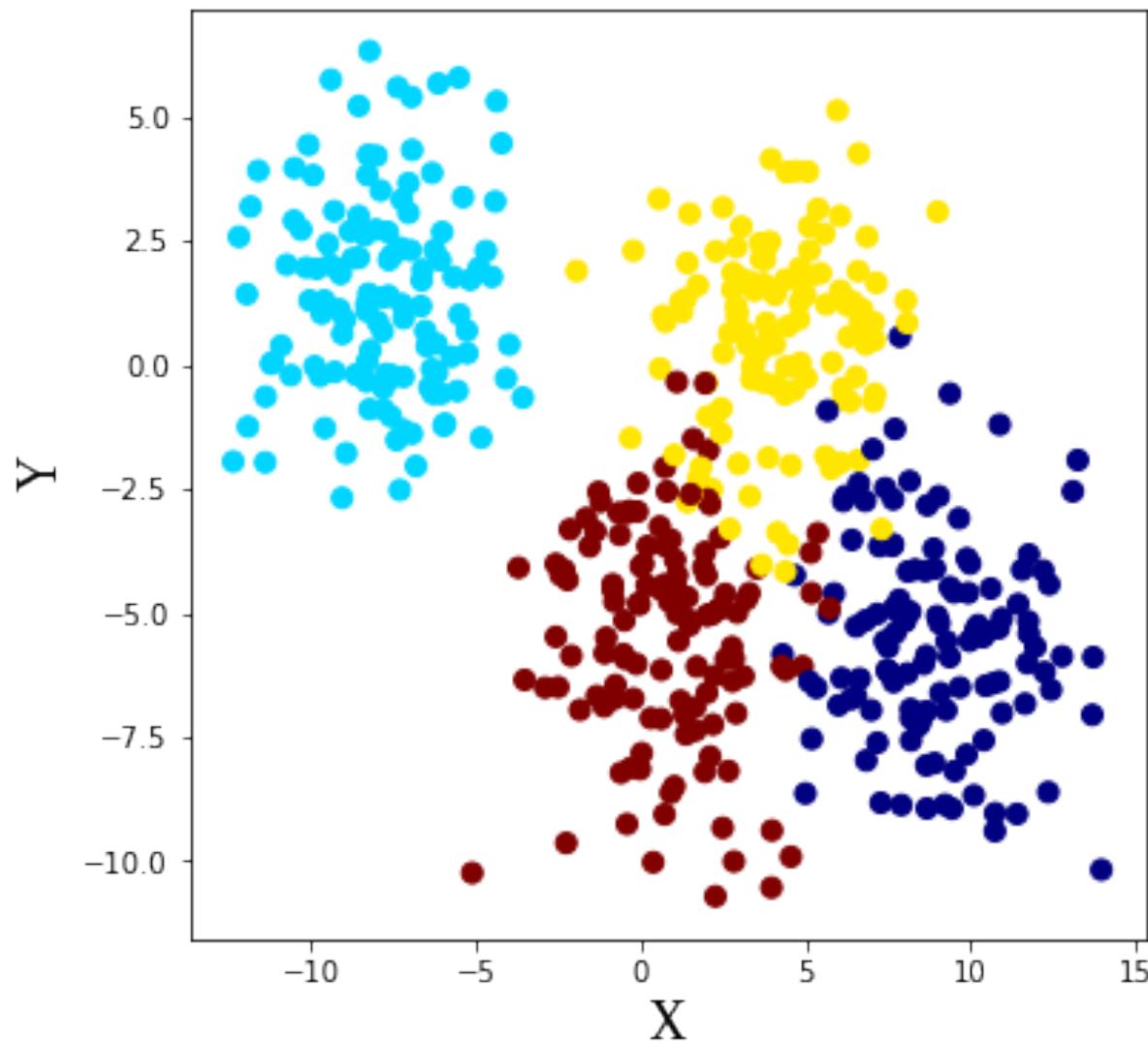


preditores independentes

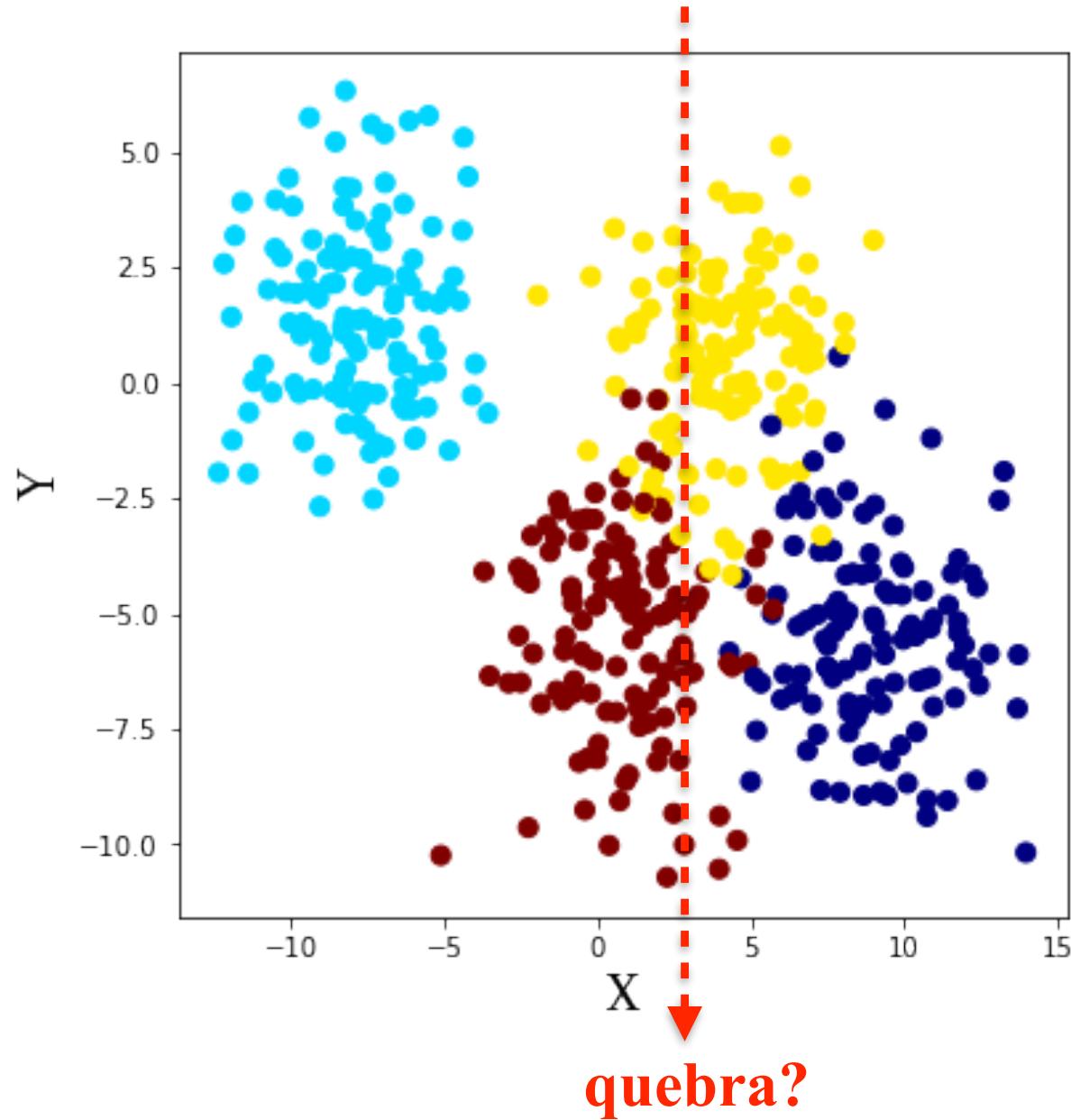


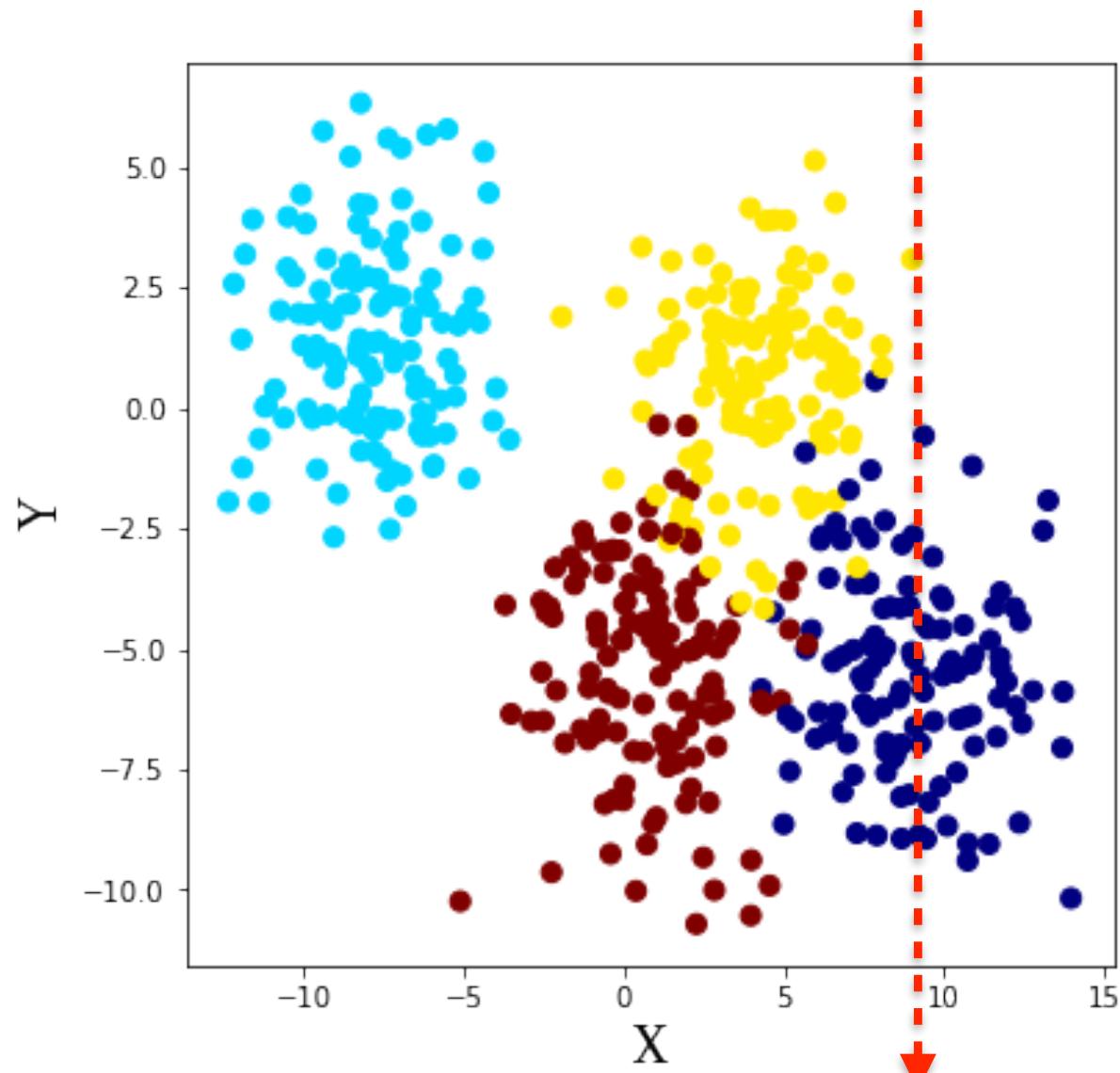
Arvores de decisão

Dataset 4 labels e 2 features

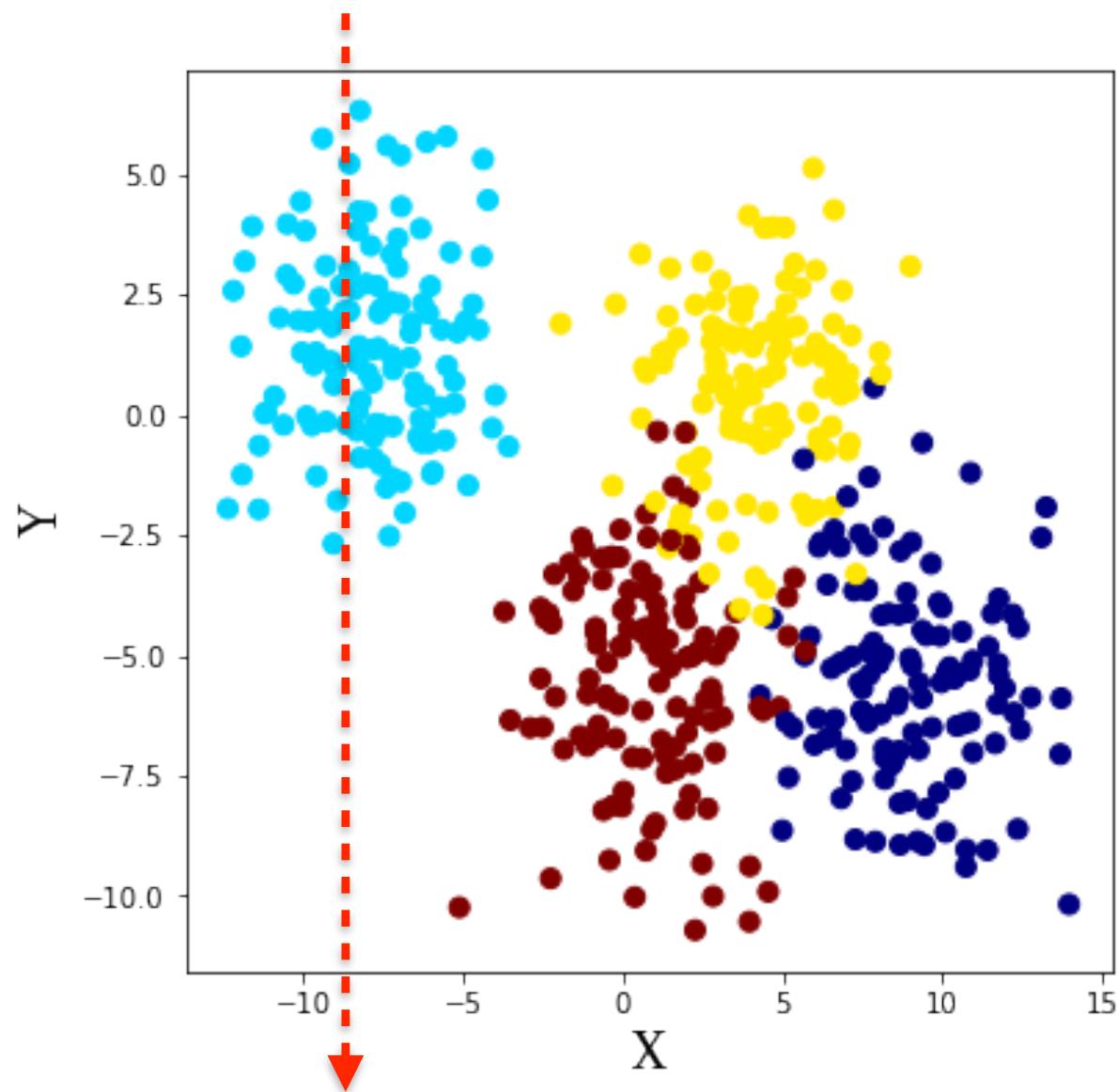


Como dividir?

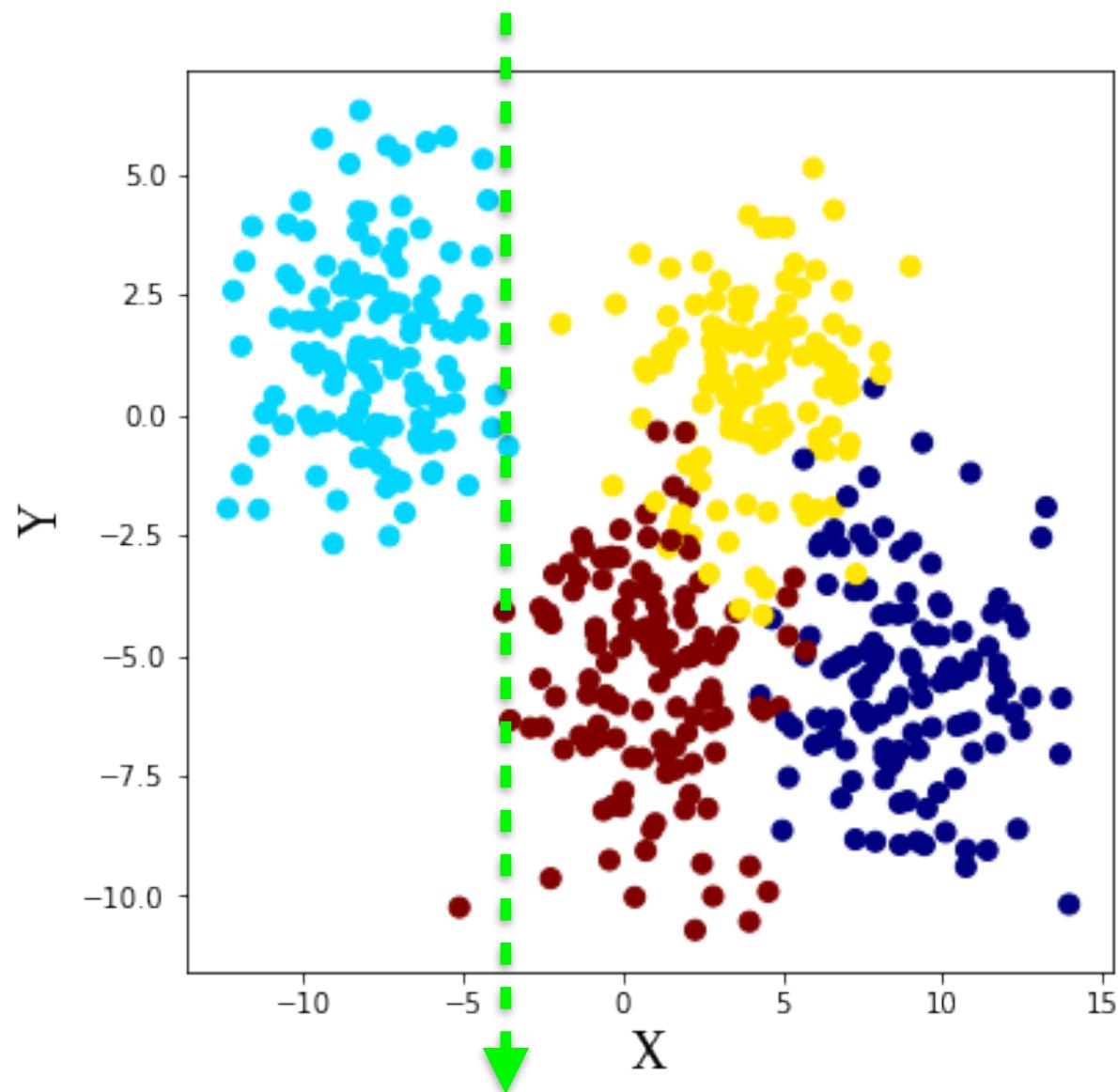




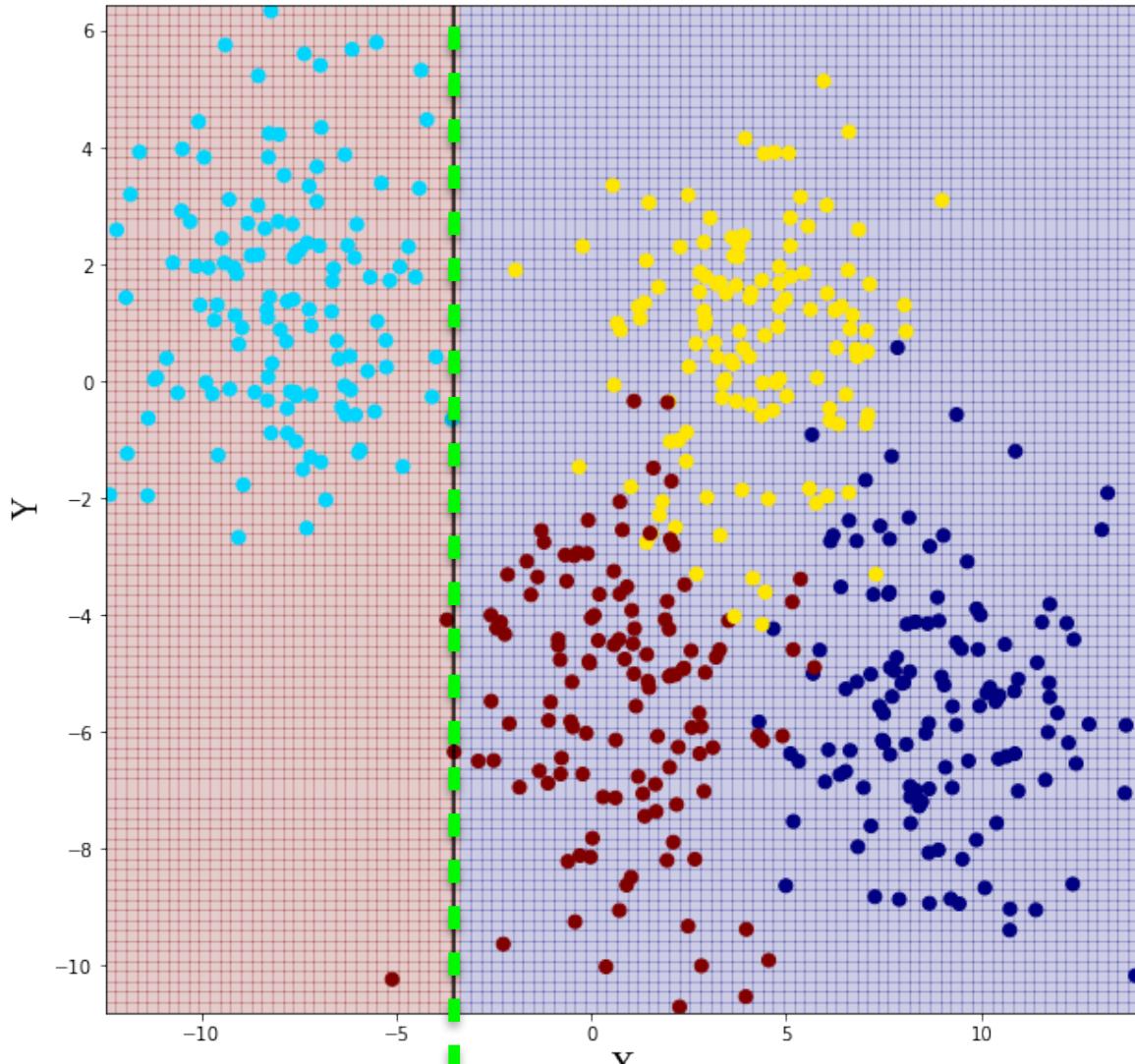
quebra?



quebra?



quebra...OK



$X \leq -3.543$
gini = 0.75
samples = 500
value = [125, 125, 125, 125]

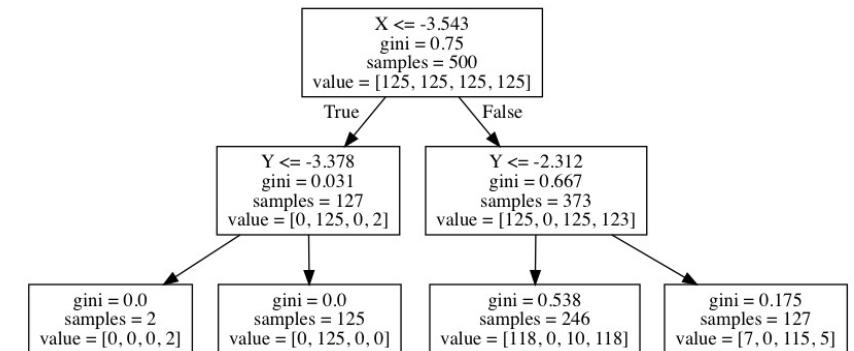
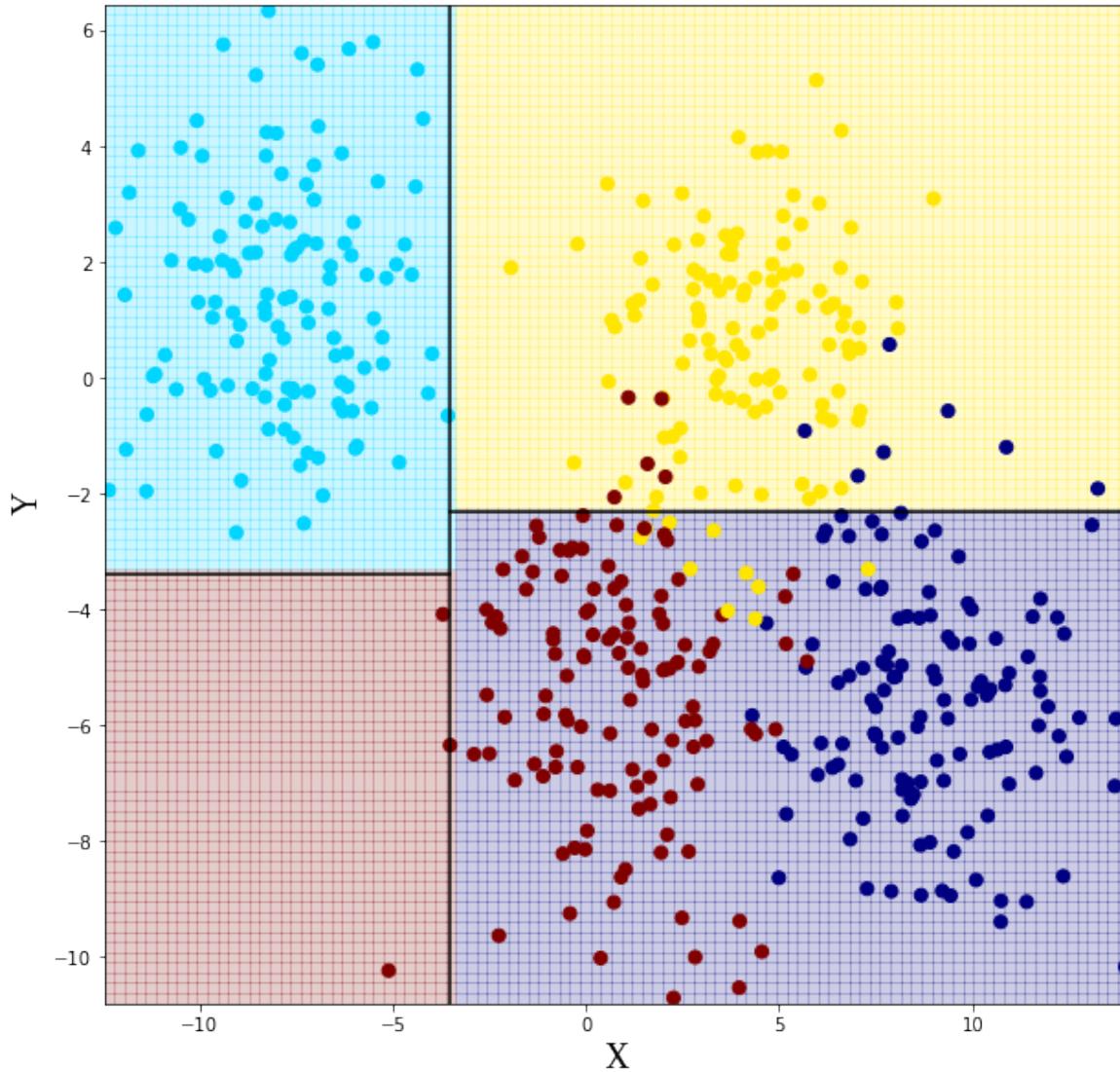
True

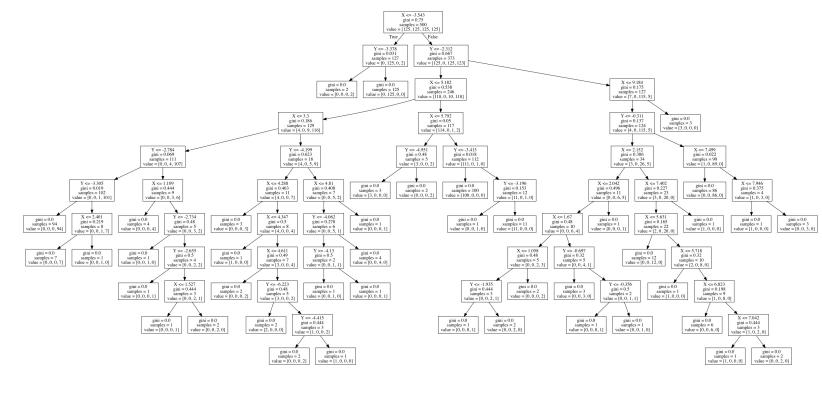
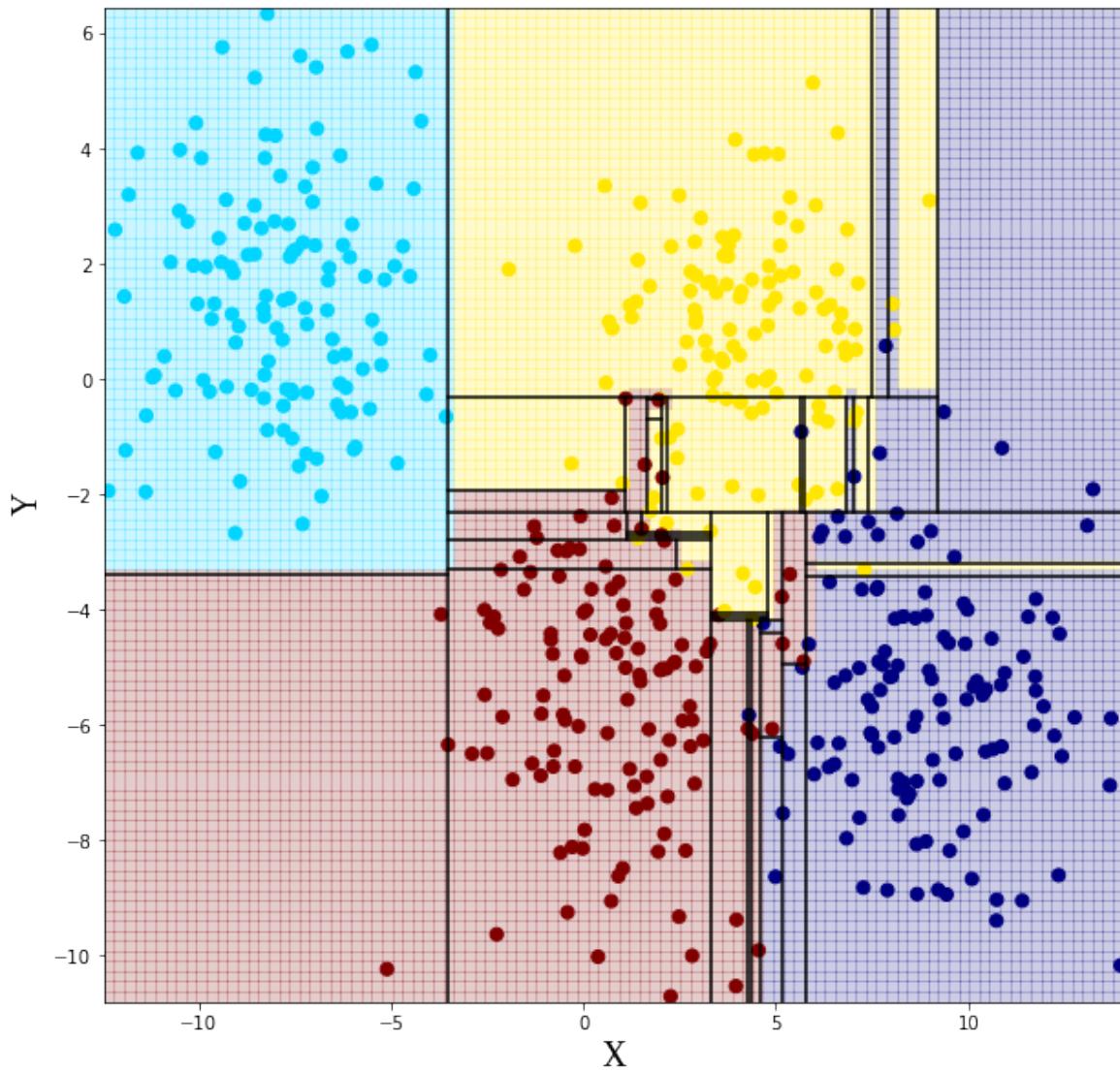
gini = 0.031
samples = 127
value = [0, 125, 0, 2]

False

gini = 0.667
samples = 373
value = [125, 0, 125, 123]

quebra...OK





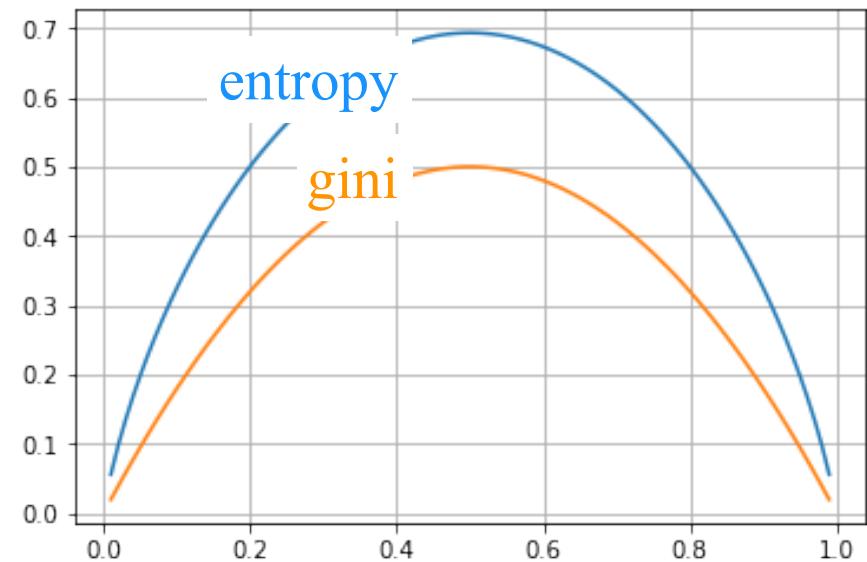
critérios de pureza: gini vs entropy

- Em nó intermediário ou terminal:
 - p_k : probabilidade (estimada) da classe k
 - K : quantidade total de classes
- Como medir a pureza?
 - gini

$$G = \sum_{k=1}^K p_k(1 - p_k)$$

- entropy

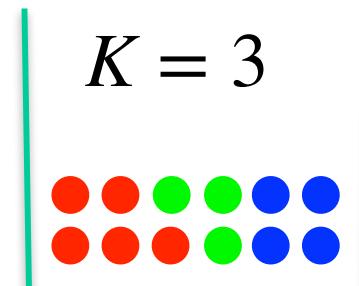
$$D = - \sum_{k=1}^K p_k \log(p_k)$$



gini vs entropy

$$G = \sum_{k=1}^K p_k(1 - p_k)$$

$K = 3$



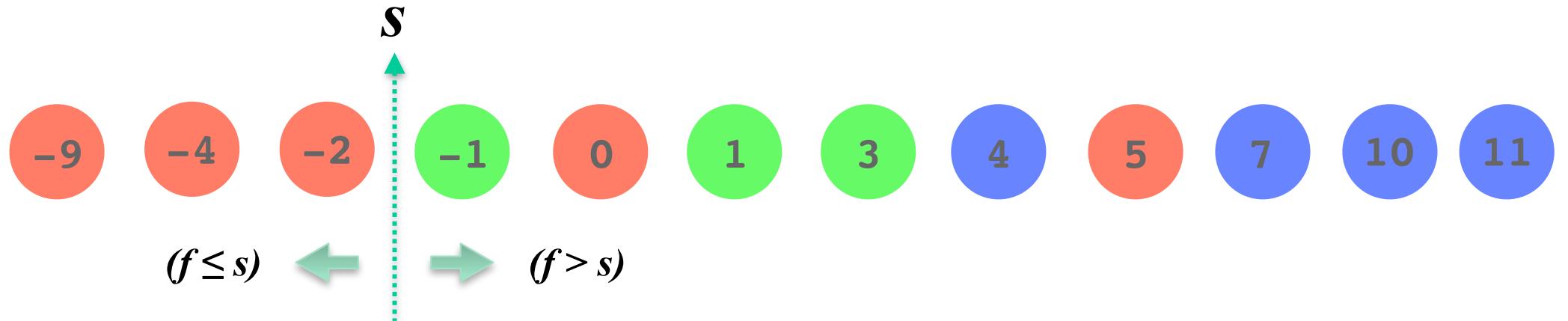
k	#	p_k	$1 - p_k$	$p_k(1 - p_k)$
blue	4	0,33	0,67	0,22
green	3	0,25	0,75	0,19
red	5	0,42	0,58	0,24
gini =				0,65

$$D = - \sum_{k=1}^K p_k \log(p_k)$$



k	#	p_k	$\log(p_k)$	$p_k \log(p_k)$
blue	6	0,50	-1,00	-0,50
green	3	0,25	-2,00	-0,50
red	4	0,33	-1,58	-0,53
entropy =				1,53

uma feature f

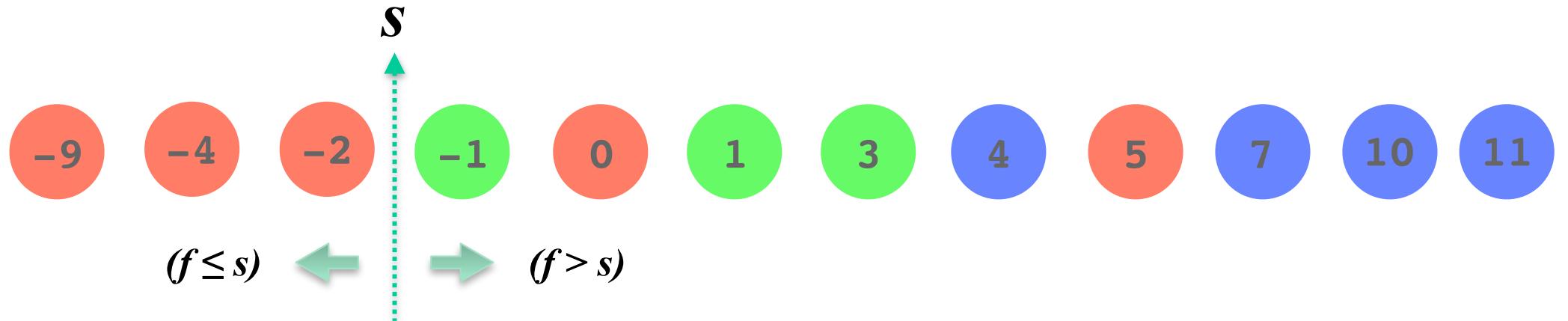


<i>ESQUERDA</i> ($f \leq s$)			$n_e =$	1
k	#	p_k	$1 - p_k$	$p_k(1 - p_k)$
blue				
green				
red				
gini =				

<i>DIREITA</i> ($f > s$)			$n_d =$	0
k	#	p_k	$1 - p_k$	$p_k(1 - p_k)$
blue				
green				
red				
gini =				

	<i>ESQUERDA</i> ($f \leq s$)	<i>DIREITA</i> ($f > s$)
n / N		
<i>gini</i>		
overall gini =		

uma feature f



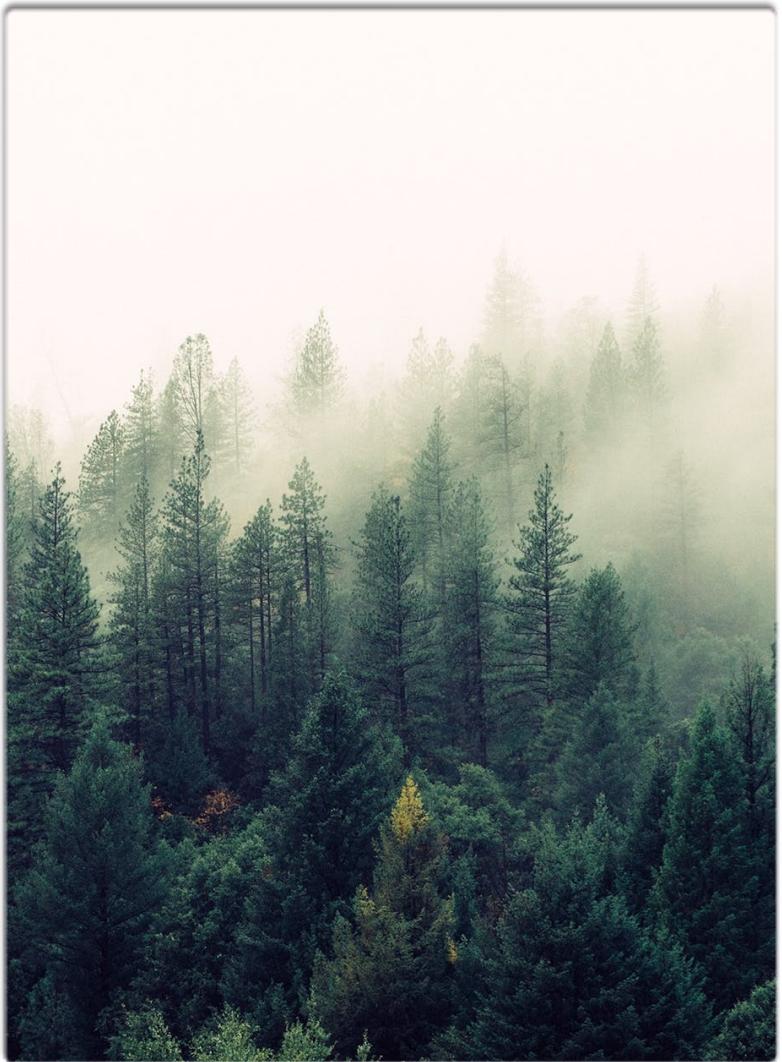
ESQUERDA ($f \leq s$)			$n_e =$	3
k	#	p_k	$1 - p_k$	$p_k(1 - p_k)$
	0	0,00	1,00	0,00
	0	0,00	1,00	0,00
	3	1,00	0,00	0,00
gini =			0,00	

DIREITA ($f > s$)			$n_d =$	9
k	#	p_k	$1 - p_k$	$p_k(1 - p_k)$
	4	0,44	0,56	0,25
	3	0,33	0,67	0,22
	2	0,22	0,78	0,17
gini =			0,64	

	ESQUERDA ($f \leq s$)	DIREITA ($f > s$)
n / N	0,25	0,75
$gini$	0,00	0,64
overall gini =		0,48

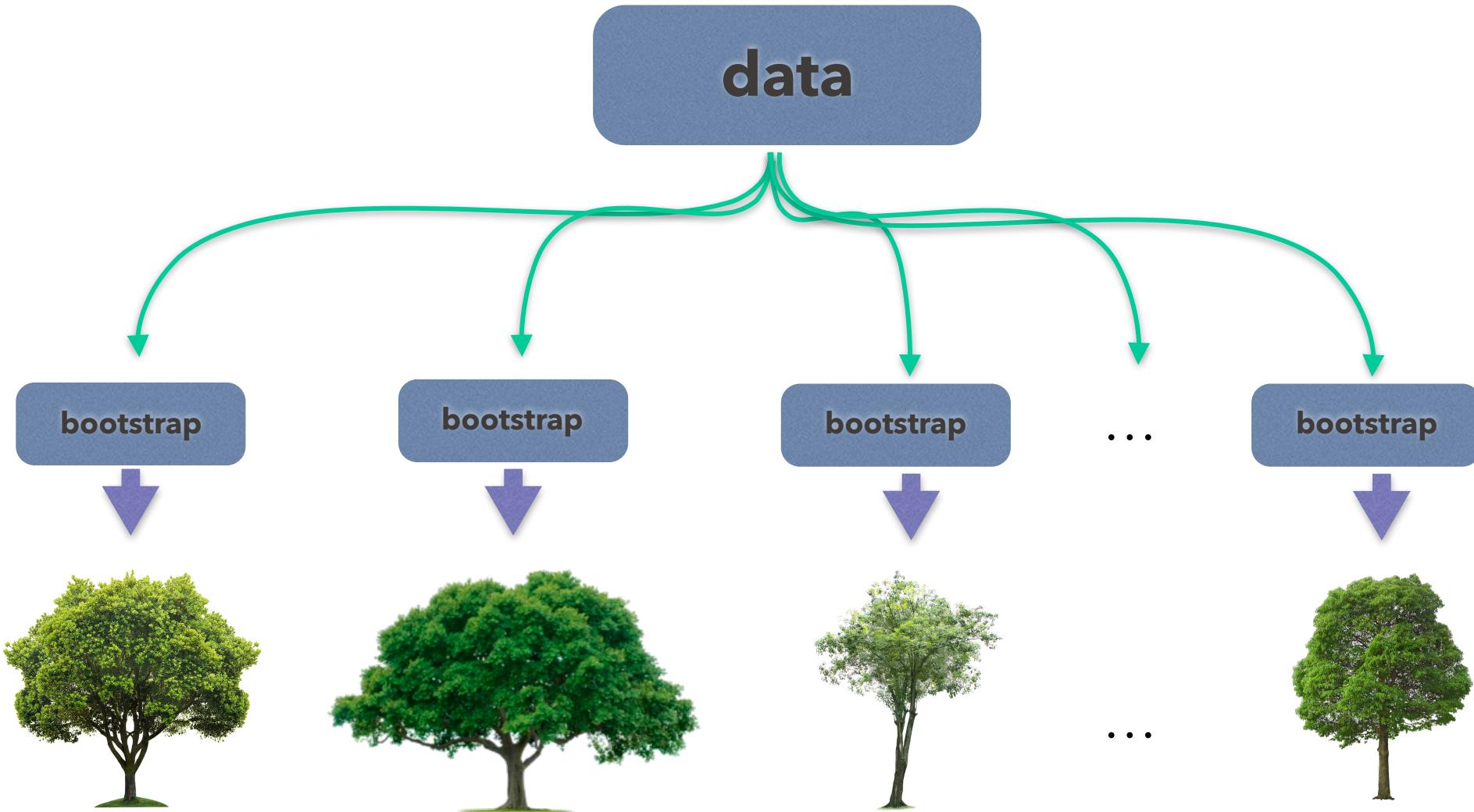
arvores de regressão

- arvores de classificação:
 - pureza: gini/entropia
 - critério do split: diminuição do gini/entropia
 - predição: classe majoritária nó terminal
 - proba: probabilidade das classes nó terminal
- arvores de regressão:
 - pureza: variância
 - critério do split: diminuição da variância
 - predição: média da variável resposta nó terminal



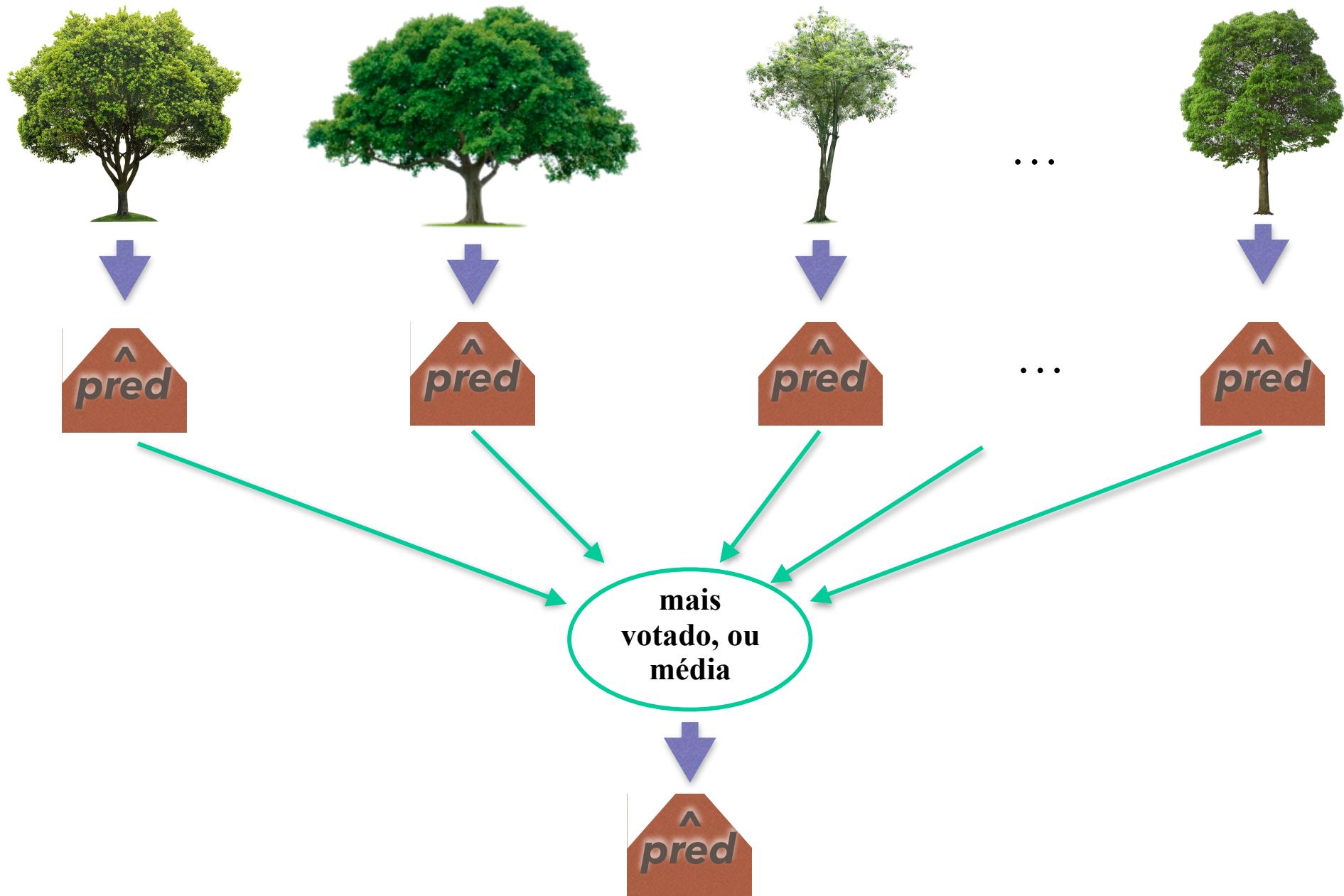
Bagging Trees

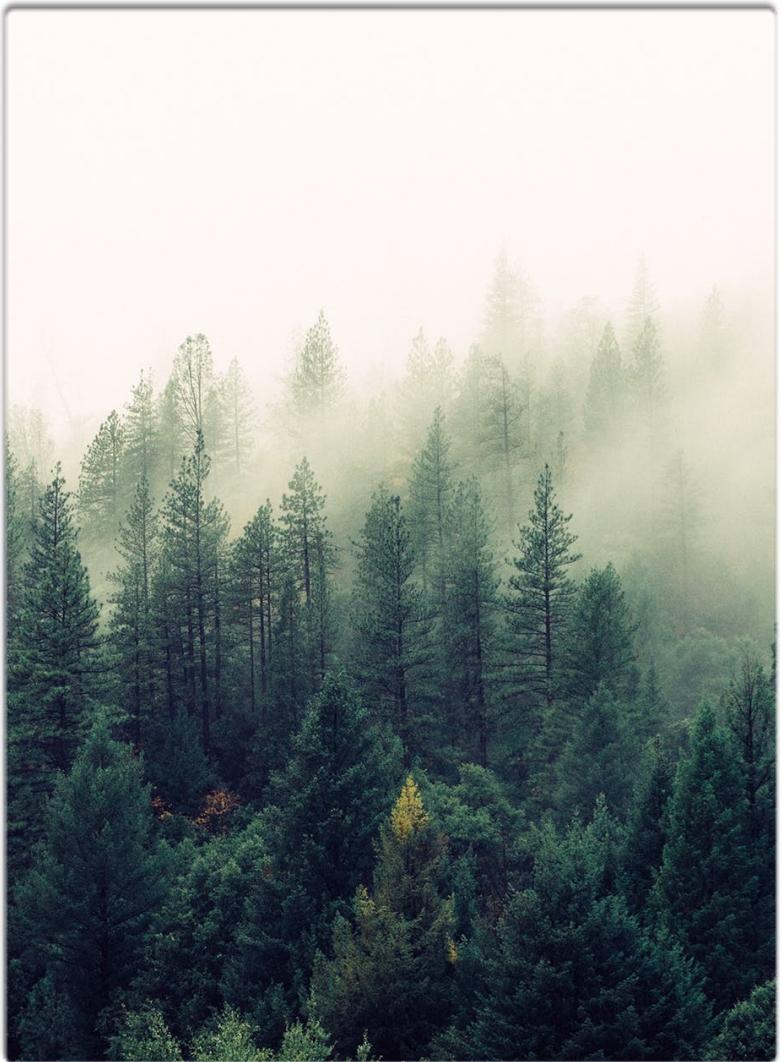
bagging - treino



**Diversas árvores são treinadas segundo amostras
bootstrap a partir de um único dataset**

bagging - predição

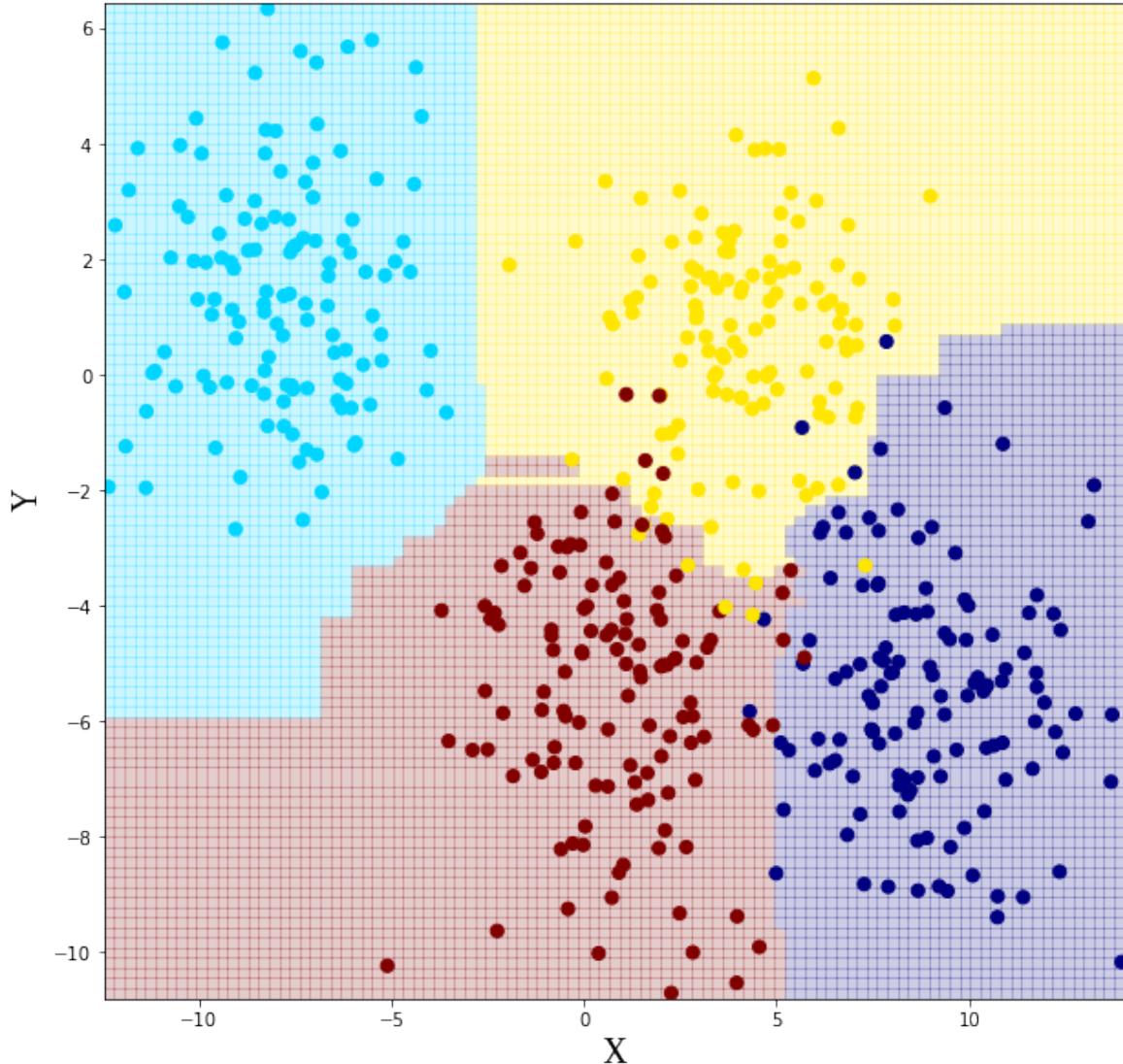




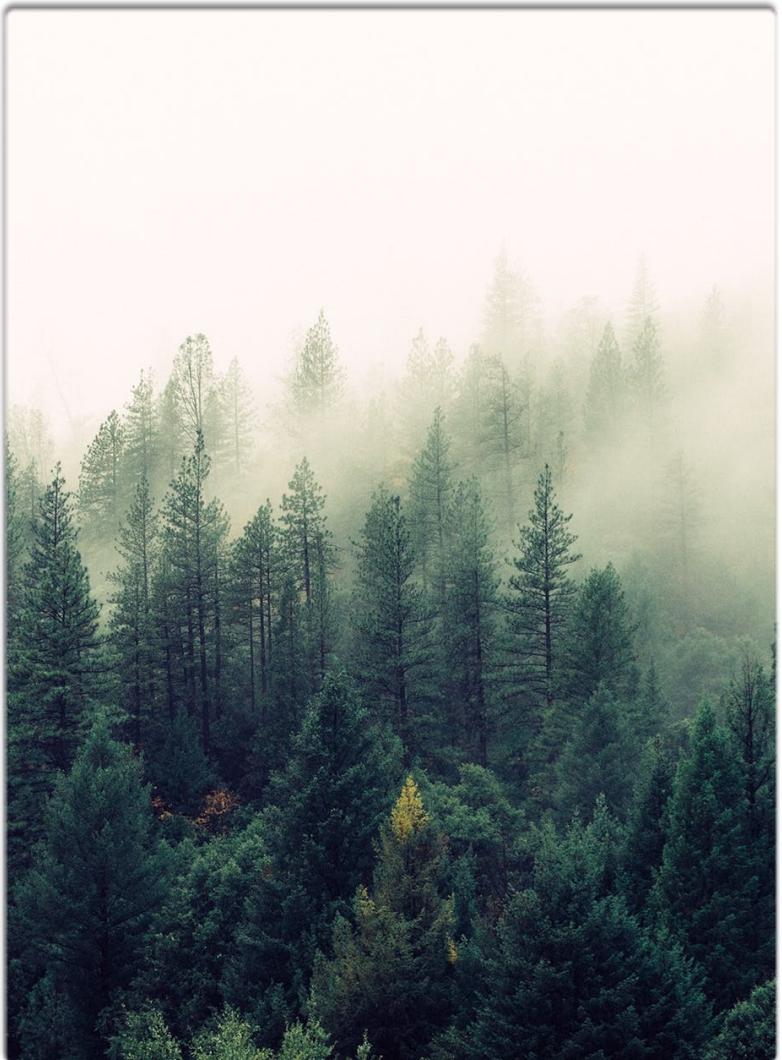
Random Forests (Florestas Aleatórias)

random forests : racional

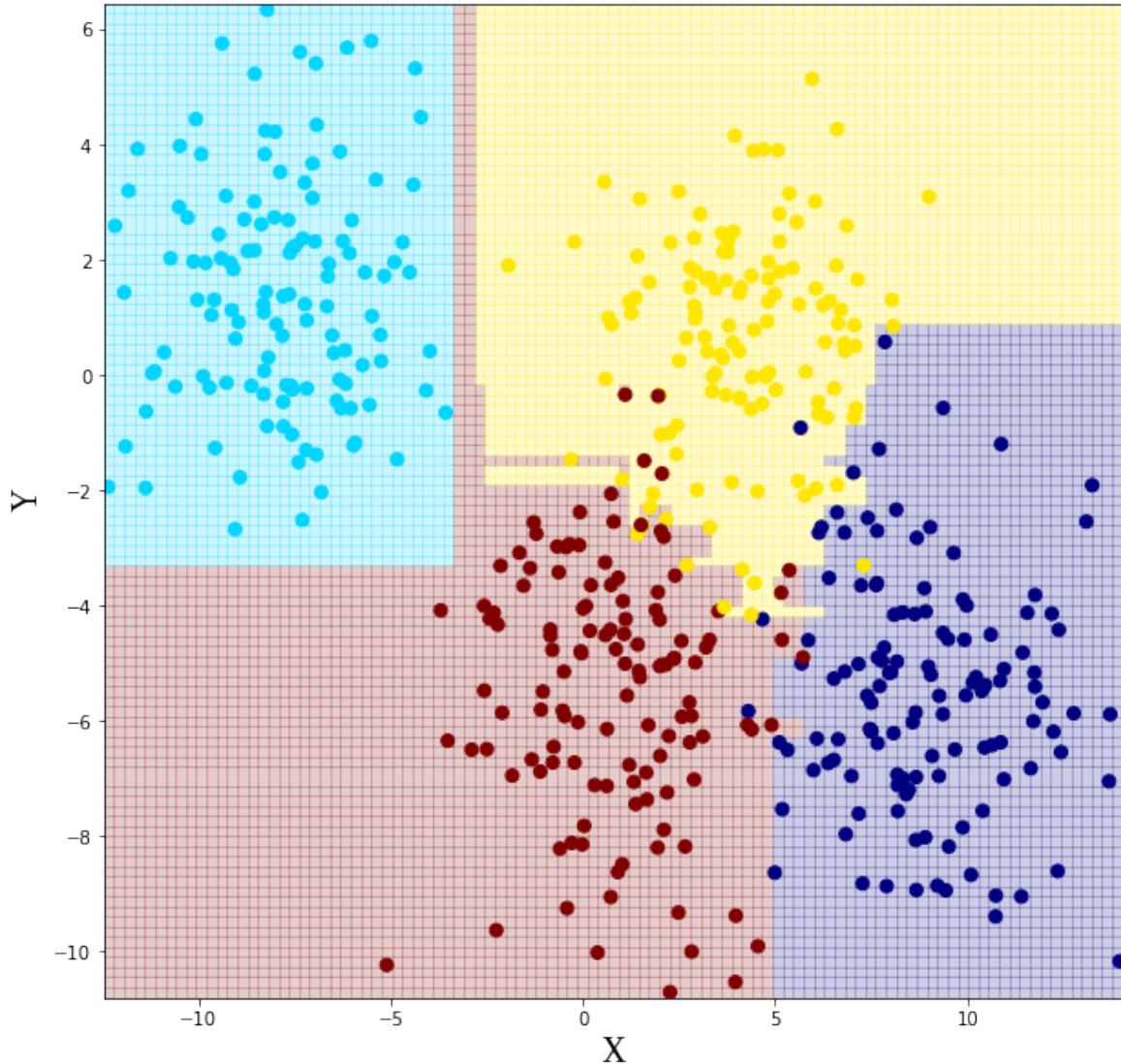
- escolha aleatória de features ****em cada split**** do nó
- attribute bagging ou feature bagging
- no bagging trees, a aleatoriedade vem do bootstrap das linhas
- no RF, aleatoriedade vem também das colunas



`n_estimators=100,
max_depth=20,
min_samples_split=10`



Gradient Boosting Trees



`n_estimators=100,
max_depth=2,
min_samples_split=50`

boosting

- observações são os erros dos preditores das iterações passadas...
- ... ao invés da geração sobre amostras *bootstrapped*
- mais rápido, ou seja menos preditores,
- PERIGO: atenção ao overfitting

boosting: stagewise

- stagewise additive model

$$F(x) = \sum_{m=1}^M \beta_m b(x; \gamma_m)$$

- b é o 'weak learner' segundo seus splits γ_m
- β_m são os pesos
- ao invés de aprender todos os parâmetros em conjunto, no boosting os parâmetros são aprendidos *stagewise*

boosting:

- derivar o gradiente da função de custo em relação aos scores

softmax

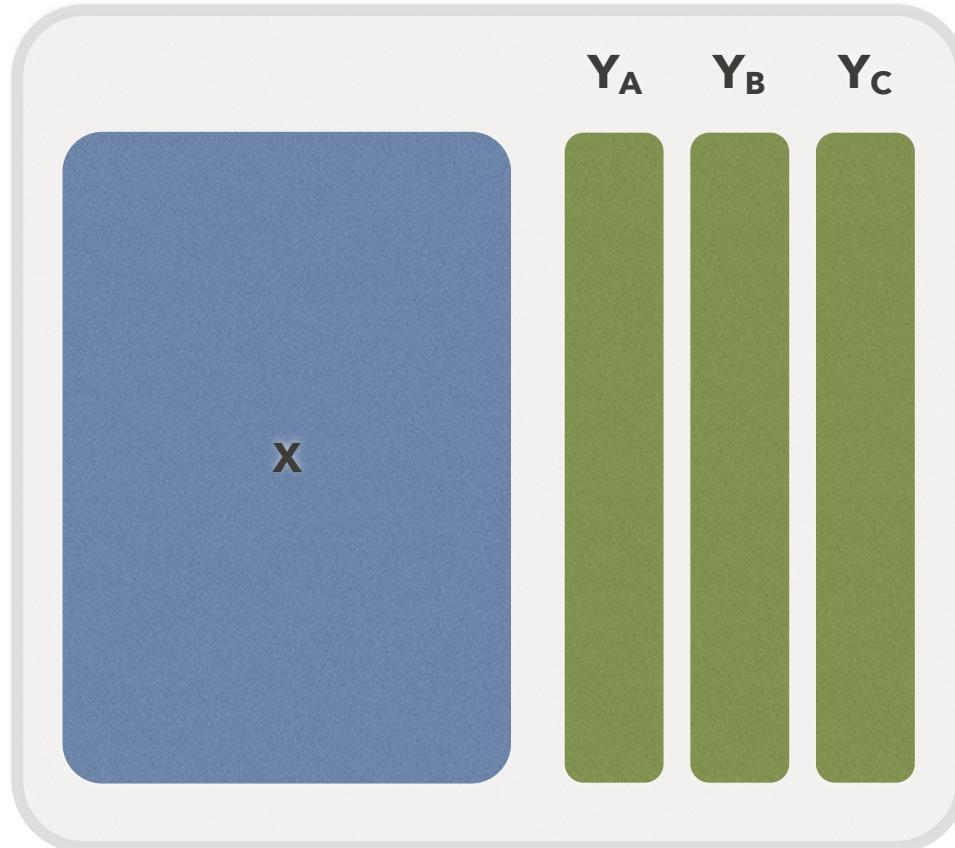


neuralnetworksanddeeplearning.com/chap3.html#softmax

	scores (F)	e^scores	probas (\hat{Y})	true (Y)
y_0	2	7,389	0,67	1
y_1	1	2,718	0,25	0
y_2	-0,1	0,905	0,08	0
fator de normalização →			11,012	

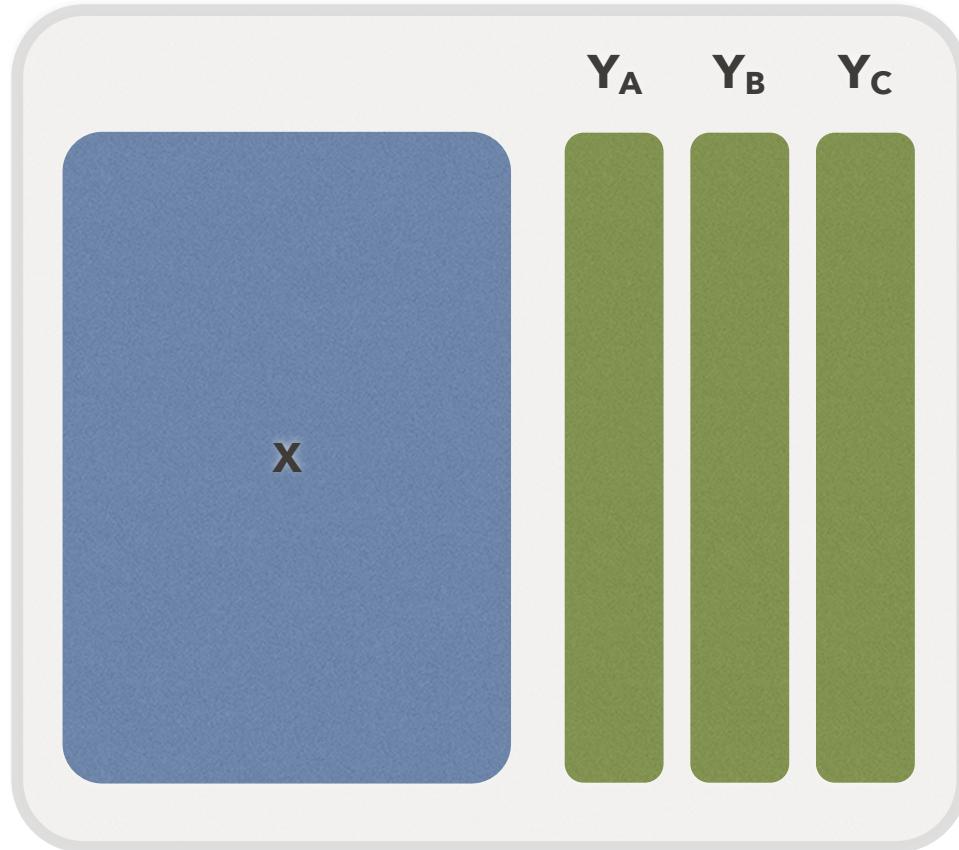
boosting:

- loop até convergir:
 - cálculo do gradiente g da função de custo em relação aos F 's
 - fit de uma arvore de regressão h cuja resposta é o negativo do gradiente $-g$
 - $F \leftarrow F + \rho h$



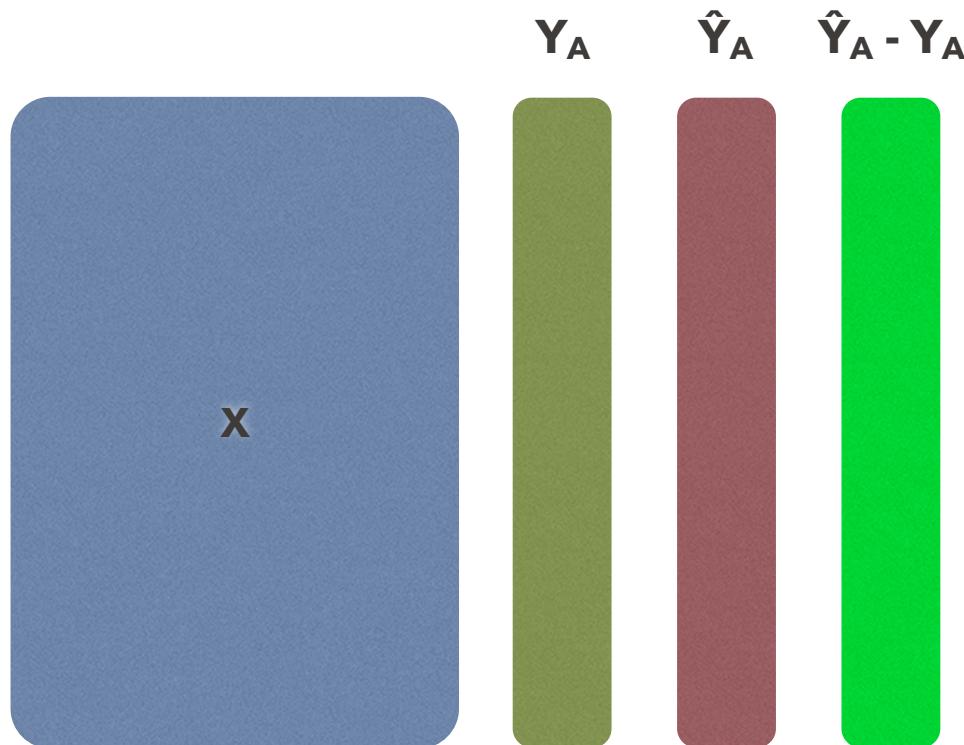
boosting:

- loop até convergir:
 - cálculo do gradiente g da função de custo em relação aos F 's
 - fit de uma arvore de regressão h cuja resposta é o negativo do gradiente $-g$
 - $F \leftarrow F + \rho h$



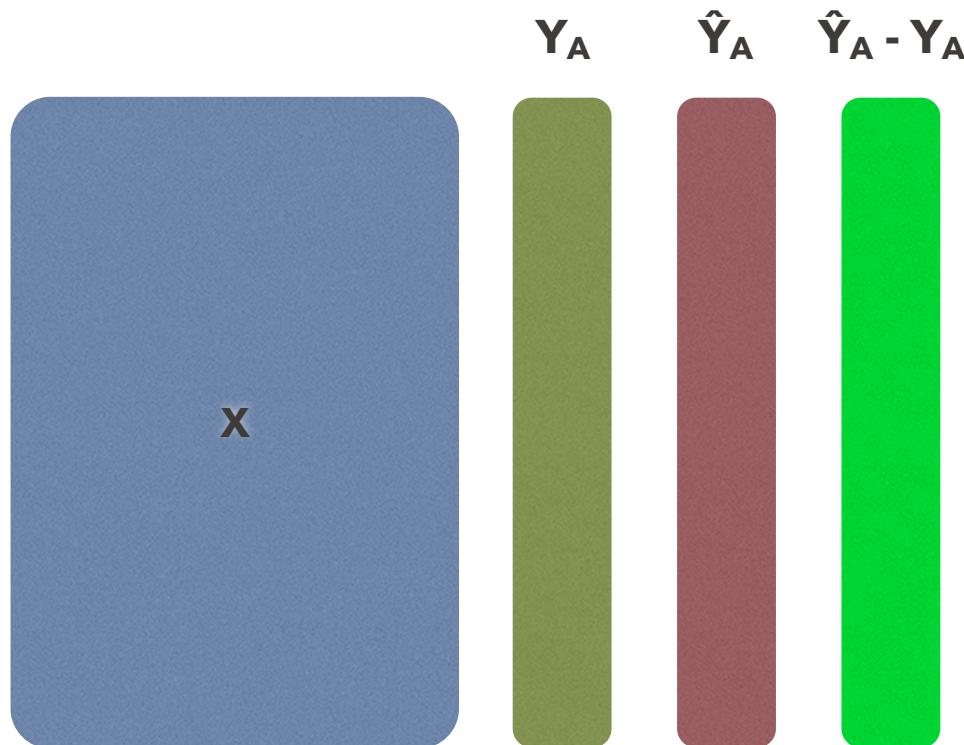
boosting:

- loop até convergir:
 - cálculo do gradiente g da função de custo em relação aos F 's
 - fit de uma arvore de regressão h cuja resposta é o negativo do gradiente $-g$
 - $F \leftarrow F + \rho h$



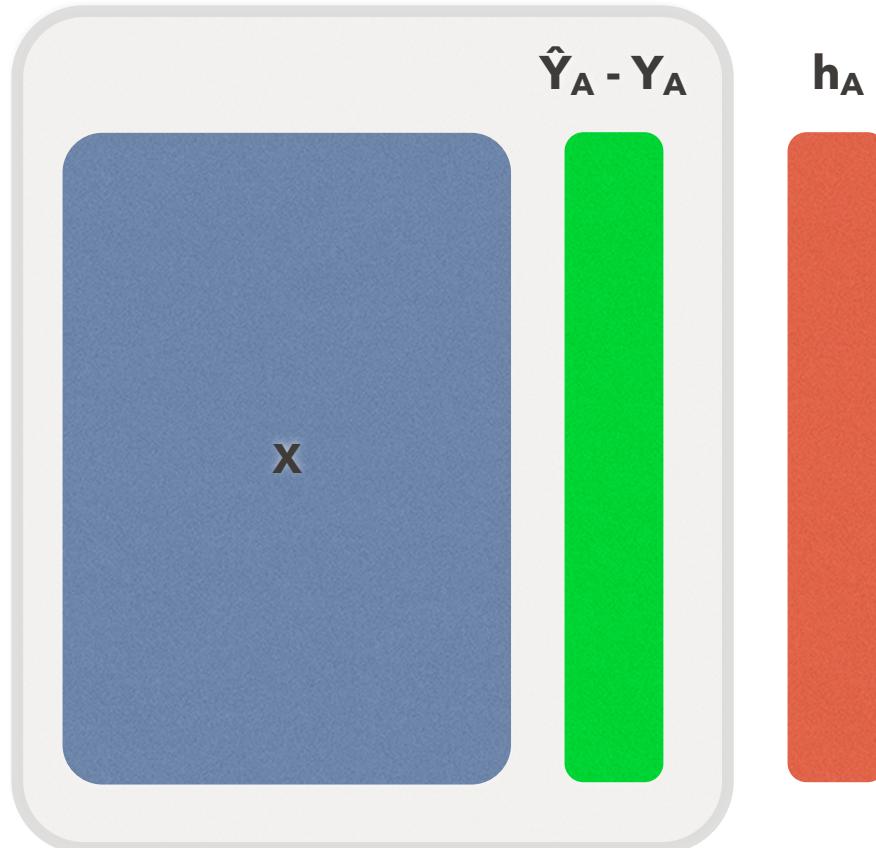
boosting:

- loop até convergir:
 - cálculo do gradiente g da função de custo em relação aos F 's
 - fit de uma arvore de regressão h cuja resposta é o negativo do gradiente $-g$
 - $F \leftarrow F + \rho h$



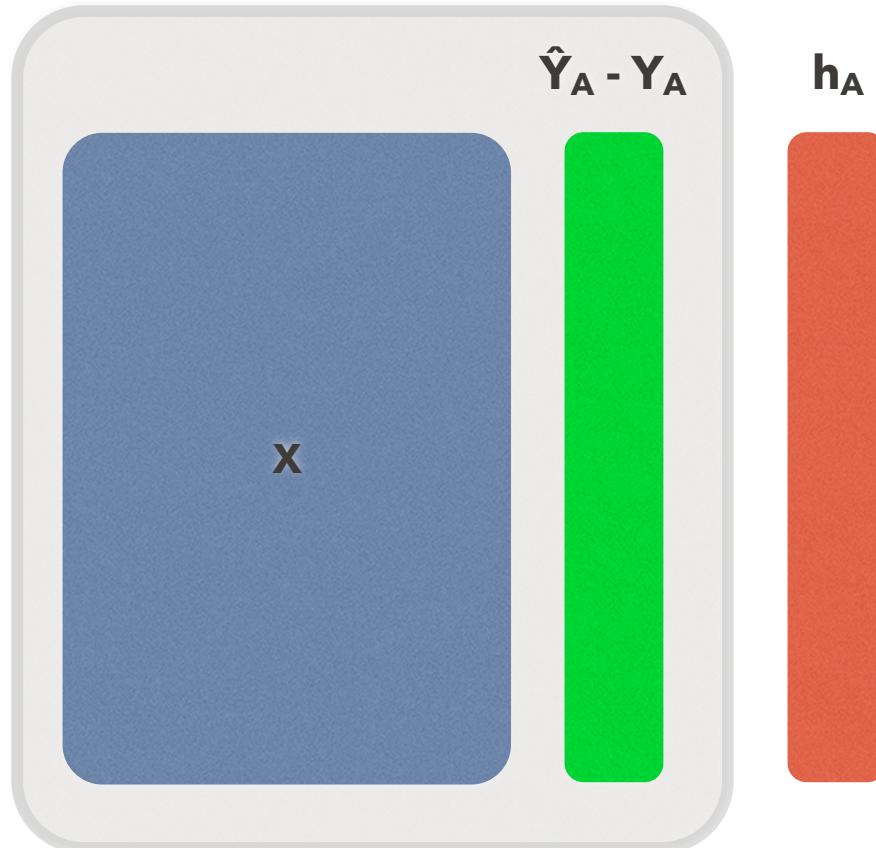
boosting:

- loop até convergir:
 - cálculo do gradiente g da função de custo em relação aos F 's
 - fit de uma arvore de regressão h cuja resposta é o negativo do gradiente $-g$
 - $F \leftarrow F + \rho h$



boosting:

- loop até convergir:
 - cálculo do gradiente g da função de custo em relação aos F 's
 - fit de uma arvore de regressão h cuja resposta é o negativo do gradiente $-g$
 - $F \leftarrow F + \rho h$



boosting:

- loop até convergir:
 - cálculo do gradiente g da função de custo em relação aos F 's
 - fit de uma arvore de regressão h cuja resposta é o negativo do gradiente $-g$
 - $F \leftarrow F + \rho h$

