

CLASSIFICAÇÃO

TÓPICOS

R

Análise exploratória

Pré-processamento

- Balanceamento
- Outliers
- Missing values
- Normalização
- Seleção de atributos (Filtros, Wrappers, PCA)

Associação:

- Apriori
- FP-Growth
- Eclat

Classificação:

- Regressão logística
- Support Vector Machine (SVM)
- Árvores de Decisão
- Random Forest
- Redes Neurais
- K nearest neighbors

Regressão

- Regressão linear simples
- Regressão linear múltipla
- Regressão não linear simples
- Regressão não linear múltipla

Agrupamento

- Particionamento (K-means, K-medoids)
- Hierárquico (DIANA, AGNES)
- Densidade (DBSCAN)

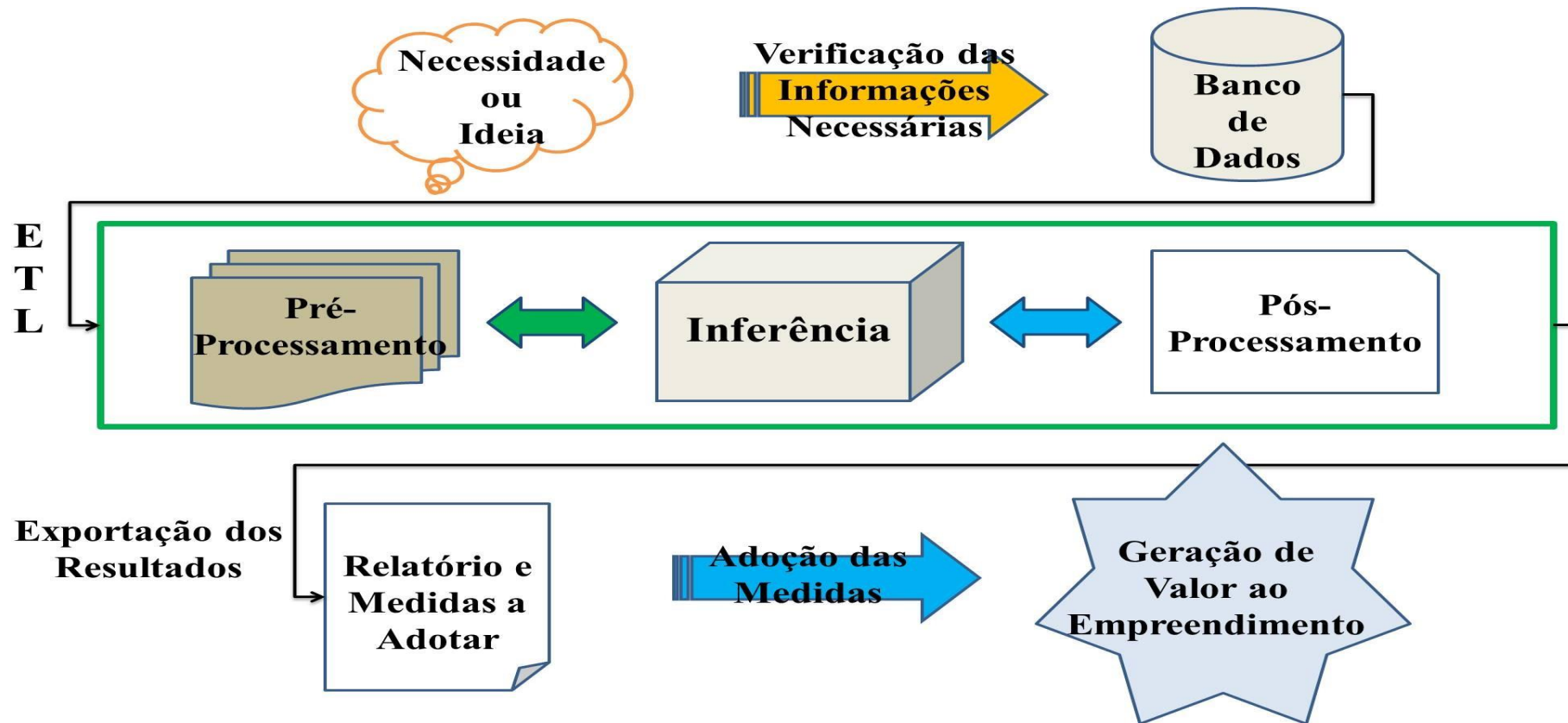
Séries Temporais

- Naive
- Média Móvel
- Amortecimento exponencial
- Auto-regressivo integrados de média móvel
- Auto regressivo não linear

Recapitulação

ETAPAS DE UM PROJETO DE DATA MINING

ESQUEMA BÁSICO DE UM PROJETO DE DM



Análise Exploratória de Dados

Tratamento de Dados

Estudos de Caso

HOJE

CLASSIFICAÇÃO

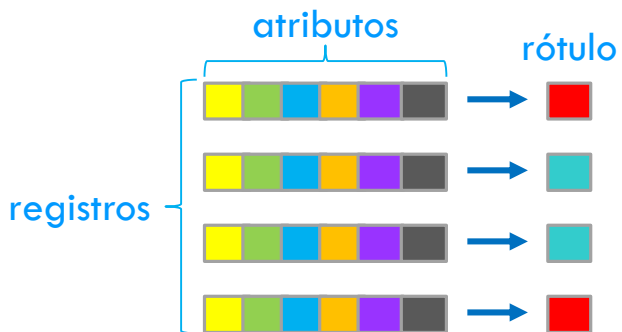
5 CLASSES DE PROBLEMAS DE DM



Machine Learning

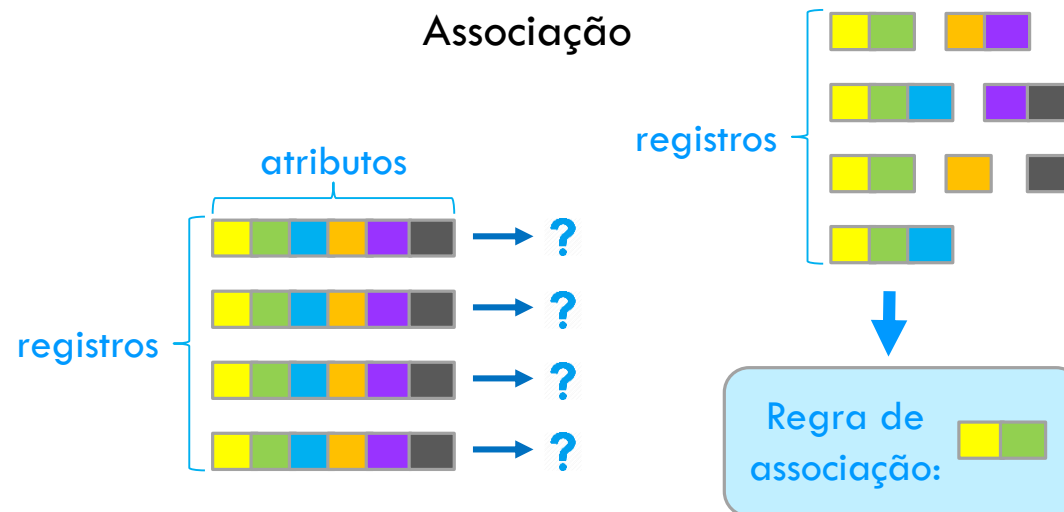
Supervisionado

Classificação
Regressão
Previsão de séries temporais



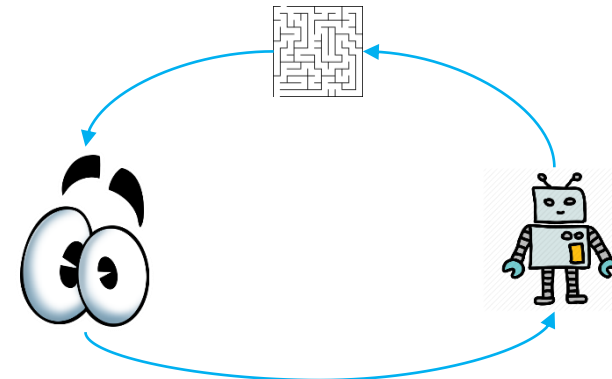
Não Supervisionado

Agrupamento
Associação



Reforço

Aprendizado através da interação de agentes com um ambiente

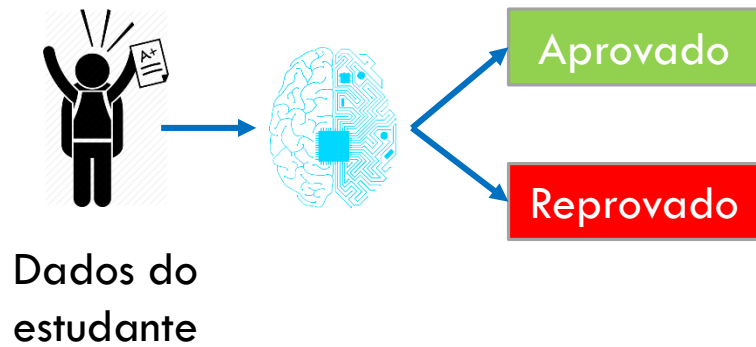


SUPERVISIONADO

- Aproximador: função mapeia entradas e saída.

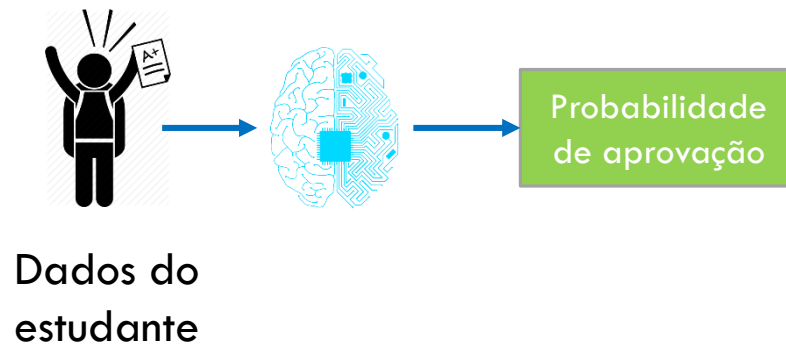
Classificação

Rótulo é categórico.



Regressão

Rótulo é contínuo.



Previsão de Séries

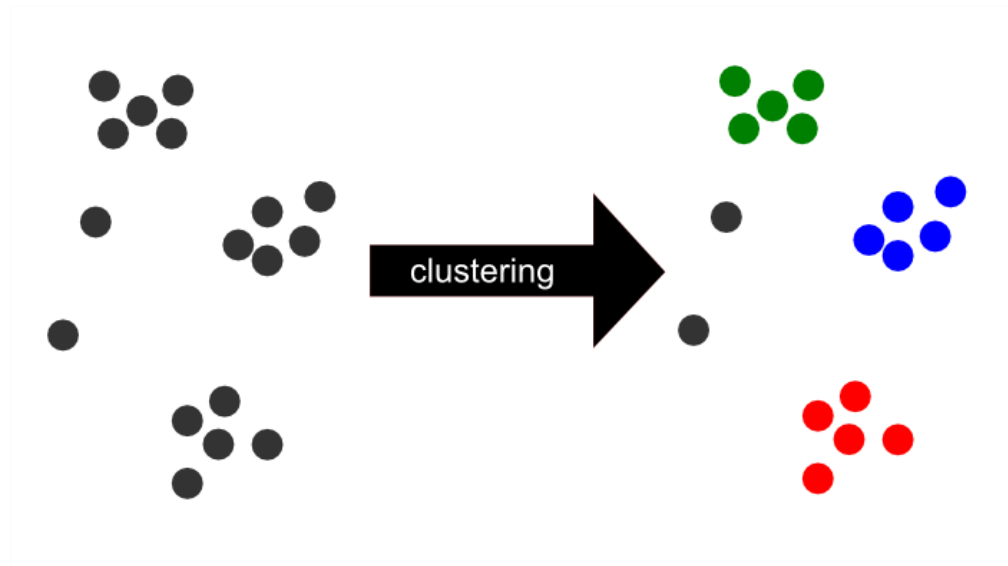
Rótulo é contínuo e dependente do tempo.



NÃO SUPERVISIONADO

Agrupamento

Descoberta de semelhanças e grupos entre registros.



Associação

Descoberta de relações entre variáveis.



EMENTA

TEMAS

Classificação

- Support Vector Machine

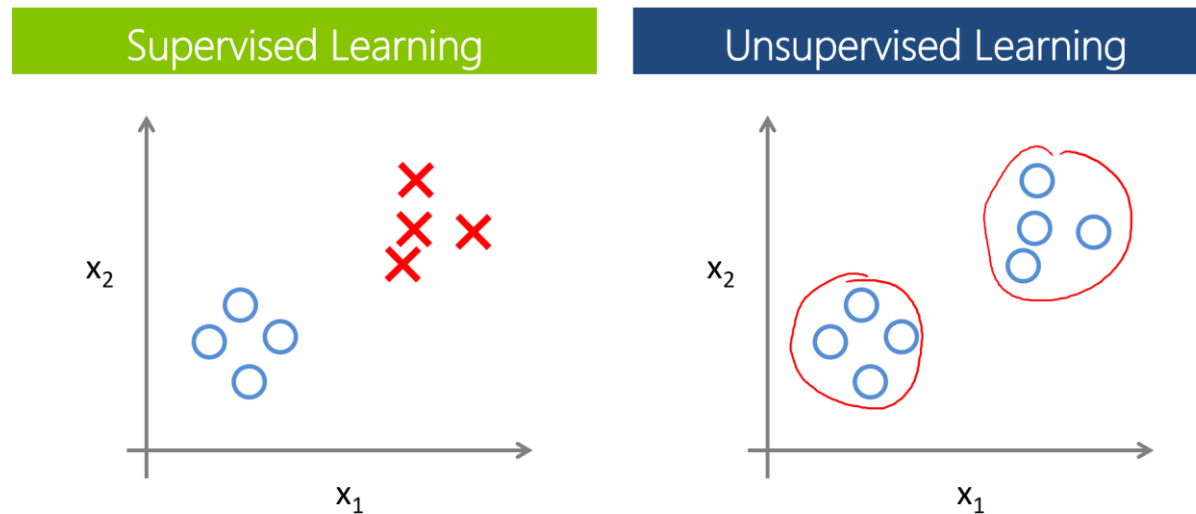
Métodos baseados em árvores

- Árvores de Decisão
- Bagging e Boosting
- Random Forest

CLASSIFICAÇÃO

Treinamento supervisionado;

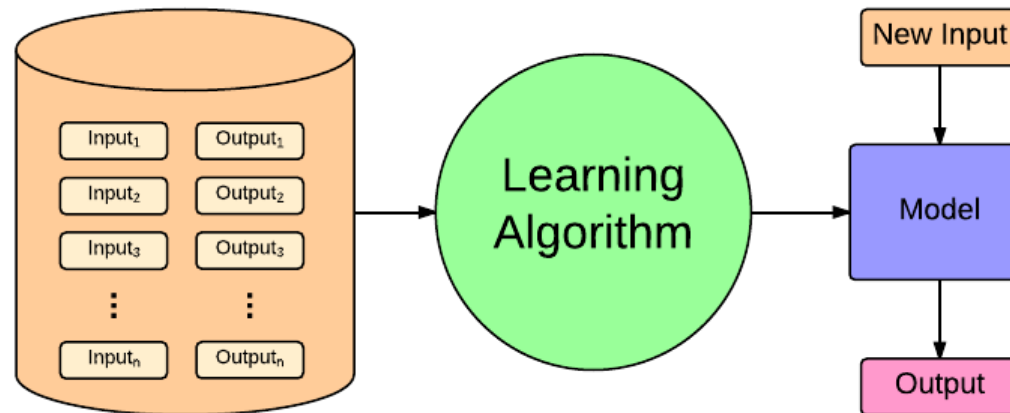
Dados da base possuem indicação da classe a qual pertencem;



CLASSIFICAÇÃO

Treinamento supervisionado;

Dados da base possuem indicação da classe a qual pertencem;



CLASSIFICAÇÃO

APLICAÇÕES:

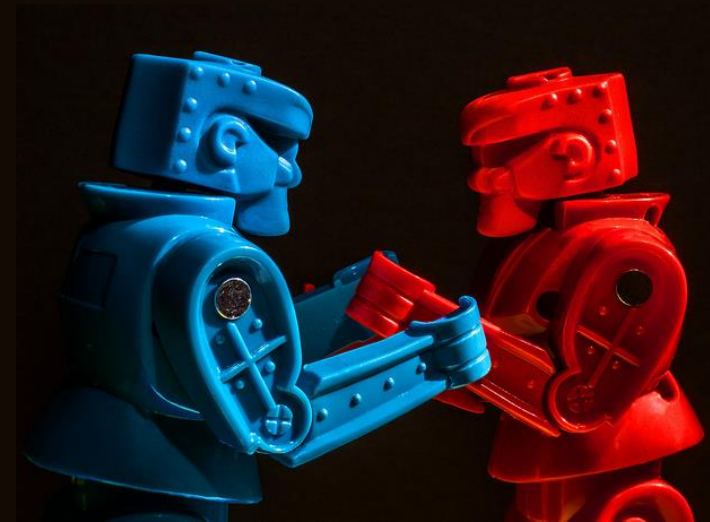
Reconhecimento de padrões (reconhecimento pessoas, de voz, doenças vocais, dígitos, ...)

Contratos em negociação;

Detecção de fraude (energia elétrica, cartão de crédito, ...)

Análise de crédito;

Etc.



SVM - Support Vector Machine

SVM - SUPPORT VECTOR MACHINE

Método supervisionado de aprendizado de máquina;

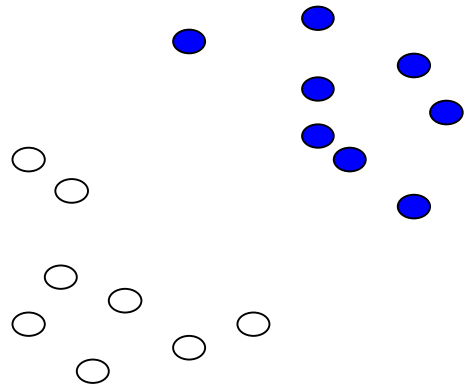
Muitas vezes apresenta resultados melhores que muitos métodos populares de classificação;

O SVM foi originalmente concebido para lidar com **classificações binárias**.

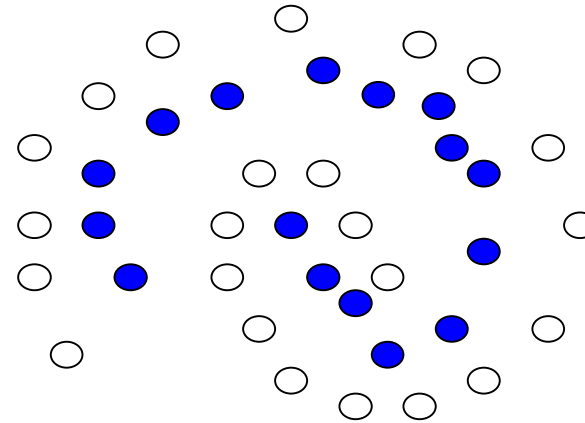
Entretanto, a maior parte dos problemas reais requerem **múltiplas classes**.

SVM

Como separar as classes?

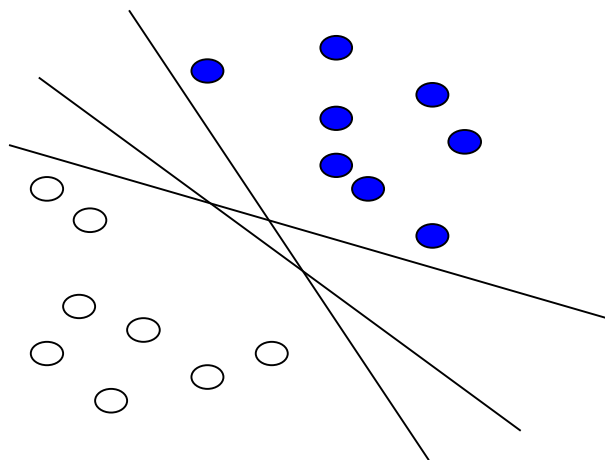


Como separar as classes?



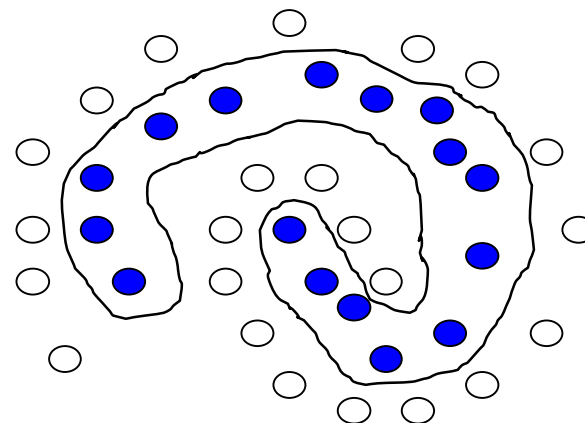
SVM

Reta / Plano / Hiperplano



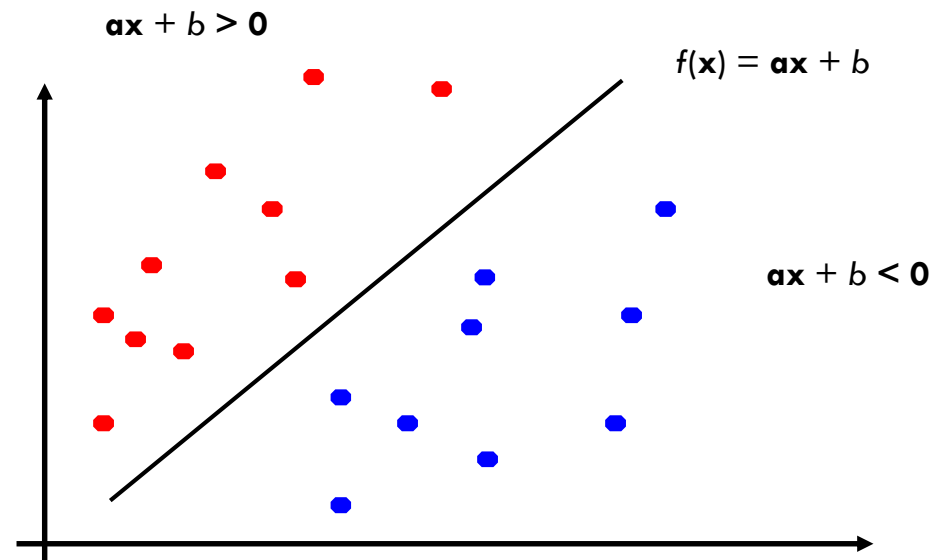
Qual o hiperplano ótimo?
Menor erro de classificação

?



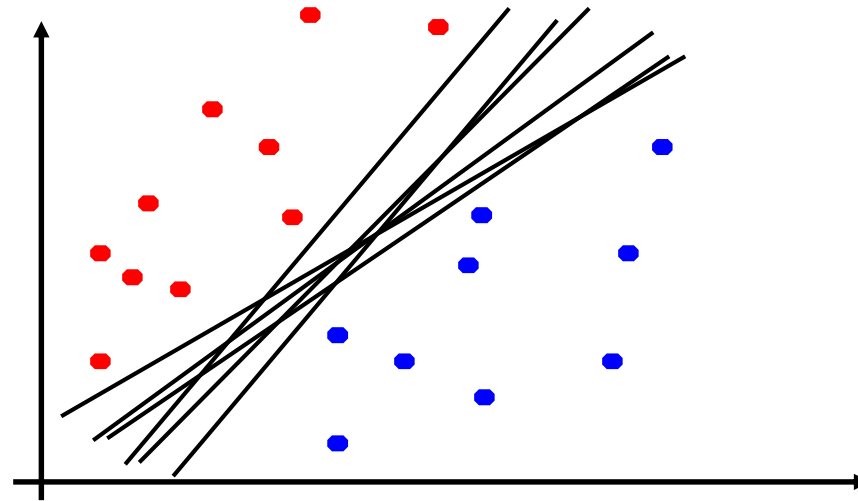
SUPPORT VECTOR MACHINE

SVM para conjuntos separáveis linearmente

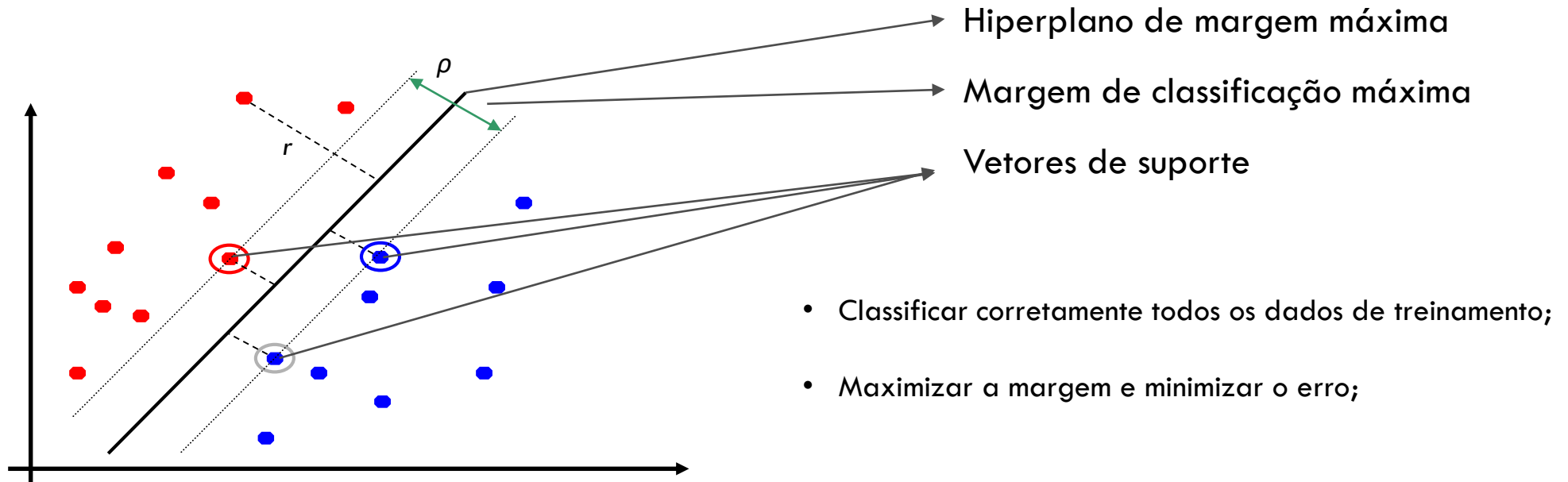


SUPPORT VECTOR MACHINE

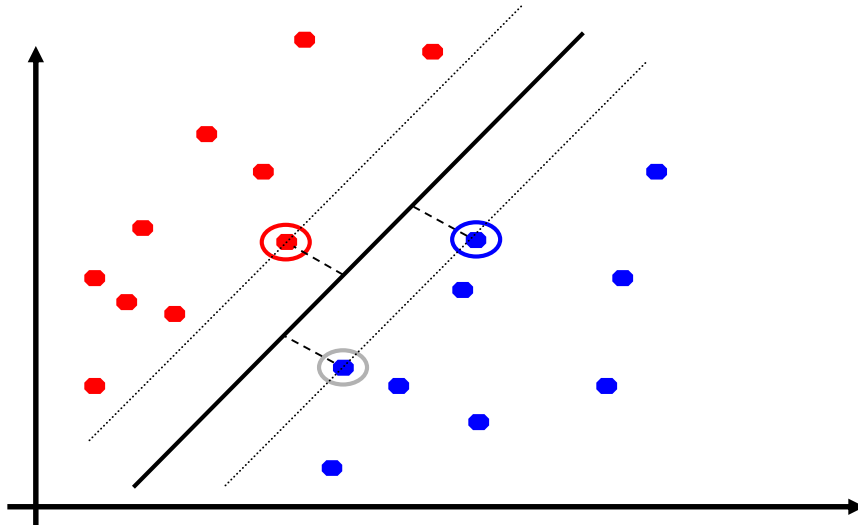
Qual é o separador ótimo?



SUPPORT VECTOR MACHINE



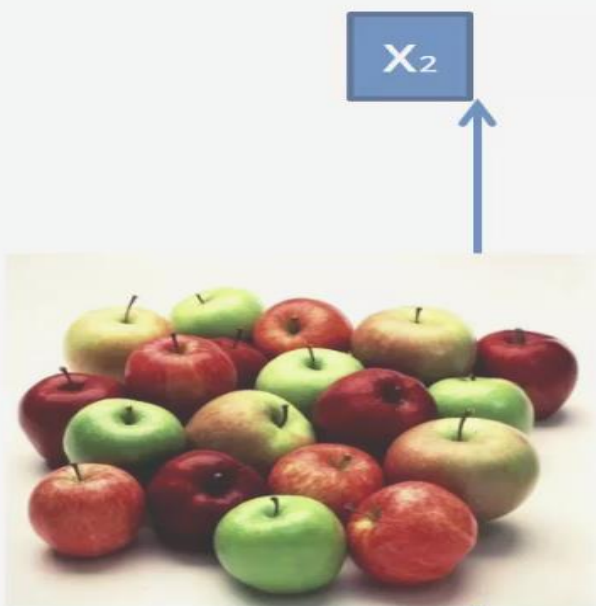
SUPPORT VECTOR MACHINE



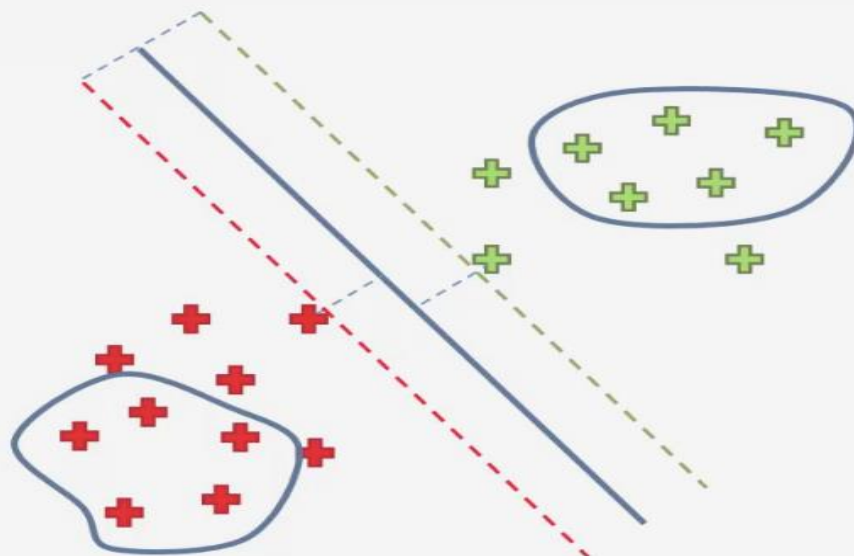
Implica que somente os vetores de suporte interessam: eles definem o hiperplano separador; outros dados do conjunto de treinamento não têm importância depois do modelo criado.

SVM - DIFERENÇAS

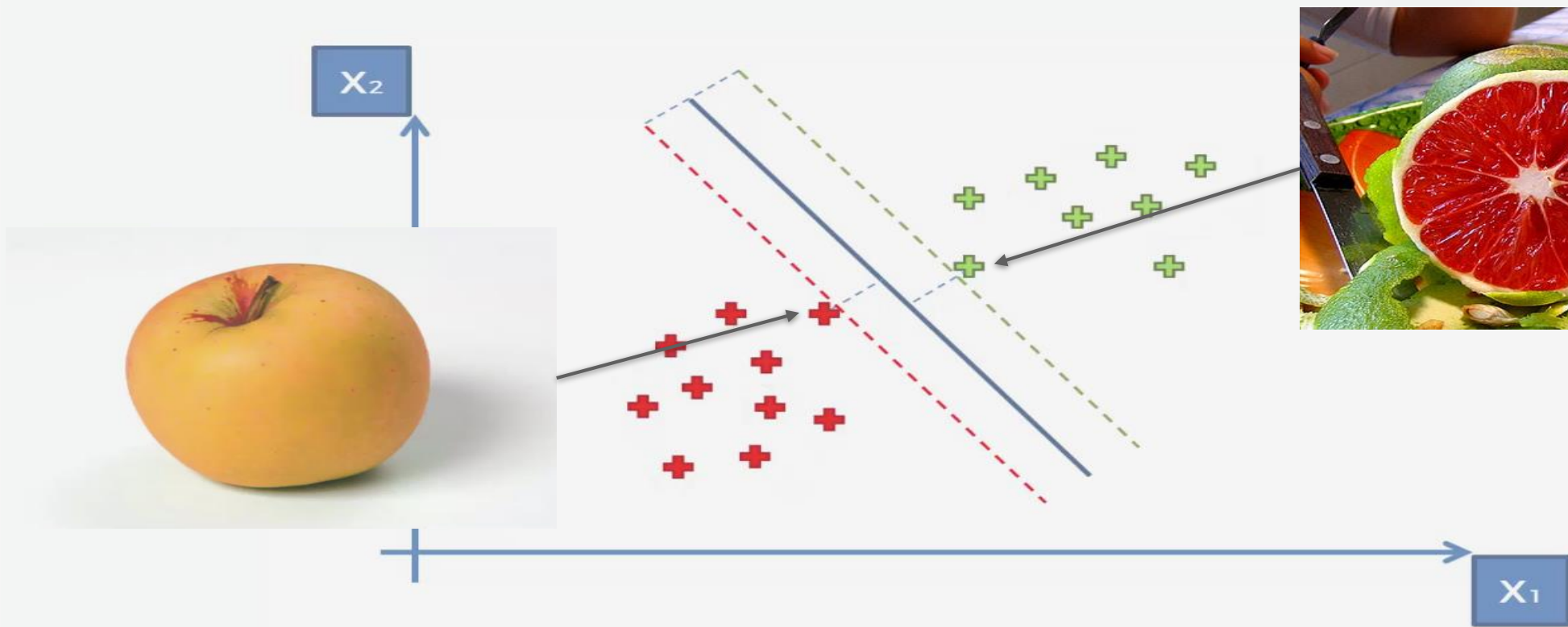
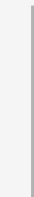




X_2



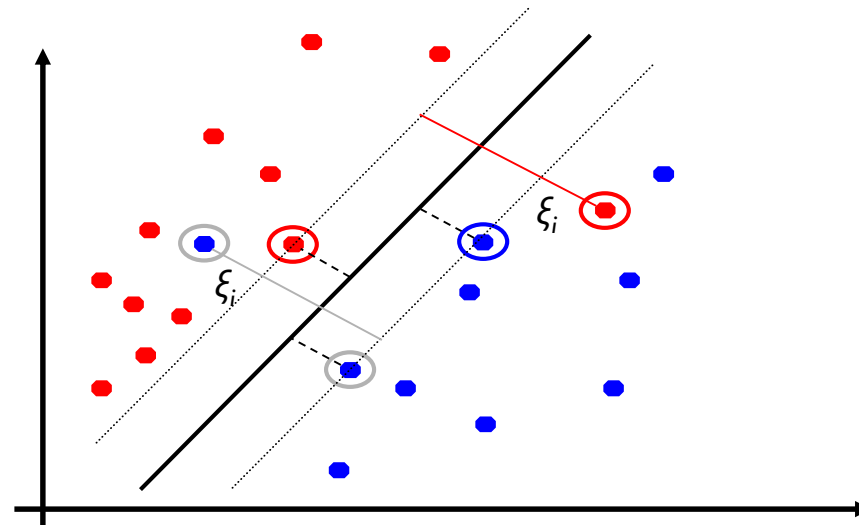
X_1



SUPPORT VECTOR MACHINE

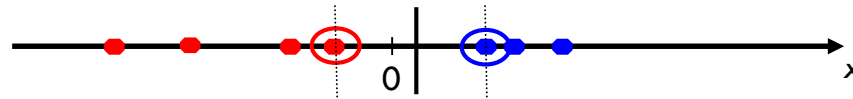
E se os dados não forem linearmente separáveis?

*Variáveis de folga ξ_i podem ser incluídas para permitir erro de classificação de exemplos difíceis ou com ruído: *soft margin*.*

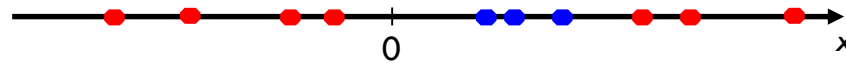


SVM NÃO LINEAR

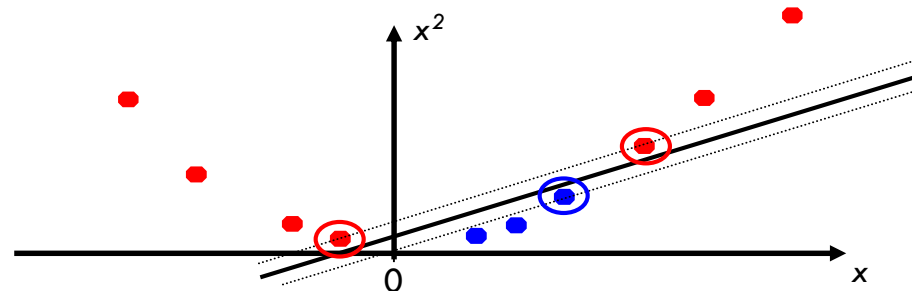
Dados que são linearmente separáveis (mesmo com algum ruído) funcionam bem:



Mas e se a base de dados for mais complexa que isso?

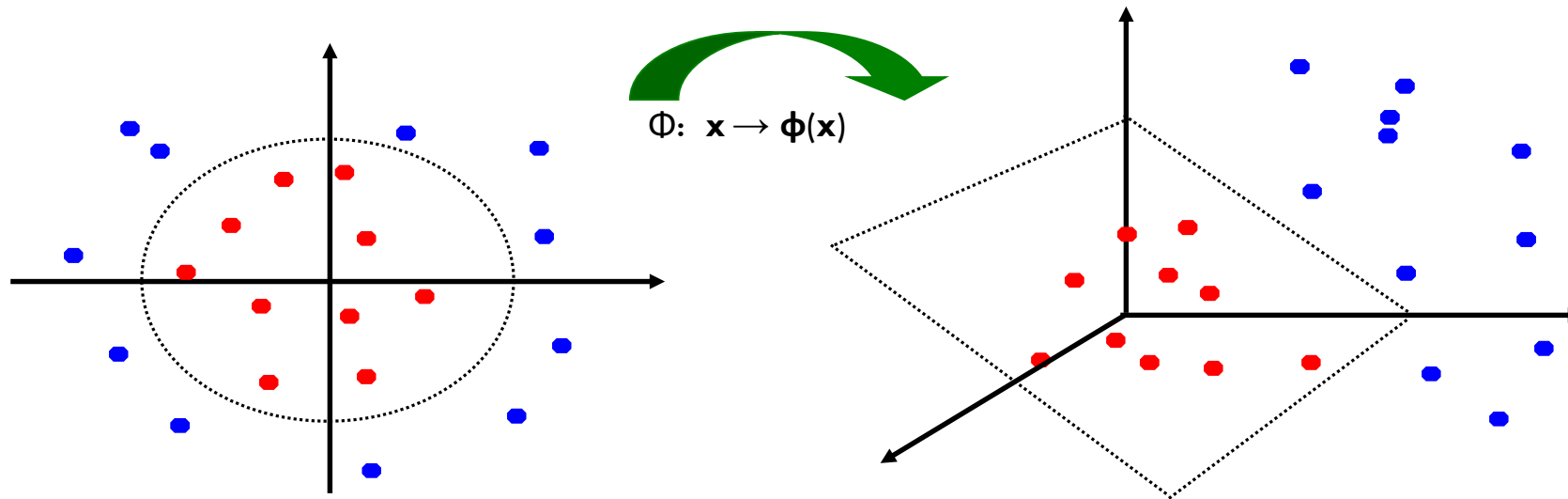


E se mapearmos os dados para uma dimensão maior?



SVM NÃO LINEAR

Ideia geral: o espaço de atributos original sempre pode ser mapeado para um espaço de atributos maior, onde os dados podem ser separados:



THE “KERNEL TRICK”

A forma mais simples de separar grupos de dados é com uma reta ou hiperplano;

Mas existem situações em que não é possível separar dados com um separador linear, ou que uma região não linear é capaz de separar os grupos de forma mais eficiente;

SVM trata esse problema usando uma função kernel (não linear) para mapear os dados em um espaço de atributos diferente (maior);

Isso significa que uma função não linear é aprendida por uma máquina de aprendizado linear em um espaço de atributos de dimensão maior;

Essa técnica é chamada de **truque do kernel**.

SVM

Vantagens:

- Consegue lidar bem com grandes conjuntos de exemplos.
- Trata bem dados de alta dimensão.
- O processo de classificação é rápido.

ESTUDO DE CASO

BASE MUSHROOM

Mushroom: <https://www.kaggle.com/uciml/mushroom-classification>

Este conjunto de dados está desfrutando de novos **picos de popularidade**. Aprenda quais **características soletram a morte certa e quais são mais palatáveis** neste conjunto de dados de características de cogumelo.

Este conjunto de dados inclui descrições de amostras hipotéticas correspondentes a **23 espécies de cogumelos**. Cada espécie é identificada como **definitivamente comestível, definitivamente venenosa, ou de consumo desconhecido e não recomendado**. Esta última classe foi combinada com a venenosa.

O guia (de onde foram retiradas as características dos cogumelos) afirma claramente que **não existe uma regra simples para determinar a comestibilidade de um cogumelo**.

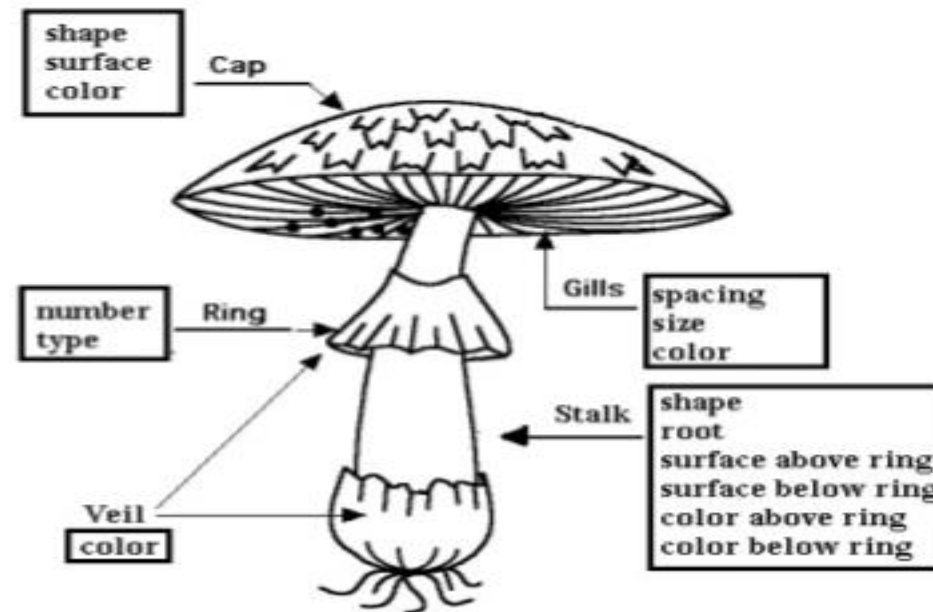


BASE MUSHROOM

Mushroom characteristics

5 Sections:

1. Cap
2. Ring
3. Veil
4. Gills
5. Stalk



Exercício

ESTUDO DE CASO

CRÉDITO BANCÁRIO

Árvores de Decisão

ÁRVORES DE DECISÃO

Árvores de decisão criam modelos de **classificação** na forma de estruturas;

Quebra-se um conjunto de dados em menores e menores **subconjuntos** enquanto simultaneamente são criadas árvores de decisão associadas;

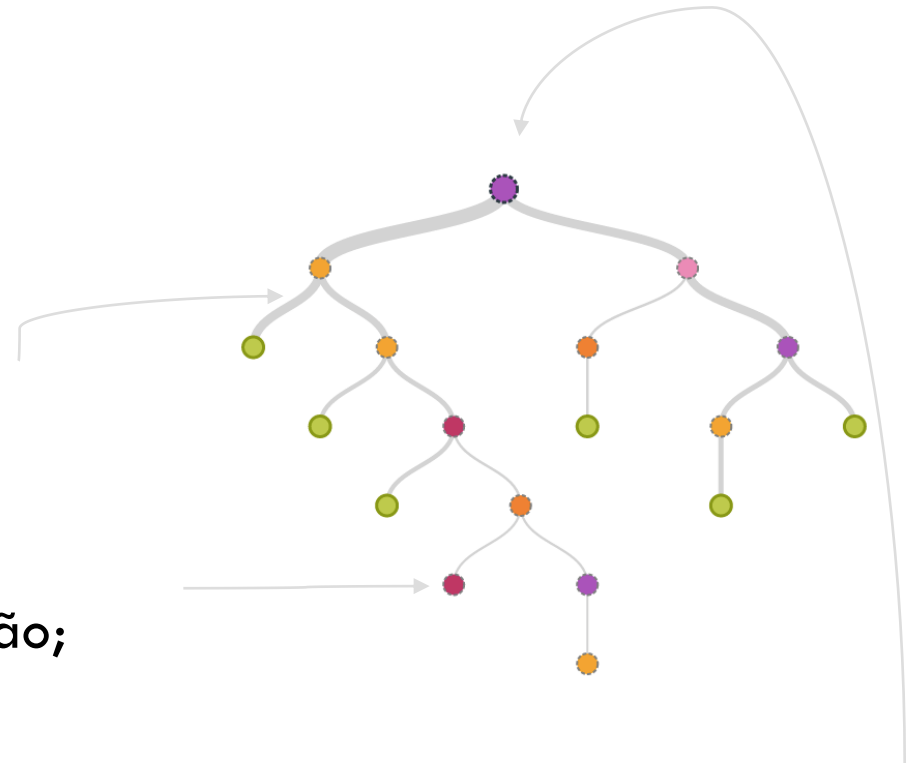
O resultado final é uma **árvore** com nós de decisão e nós folhas;

Uma árvore de decisão possui um ou mais ramos;

Nós folhas representam uma classificação ou decisão;

O nó mais alto da árvore corresponde ao melhor classificador, chamado de nó raiz;

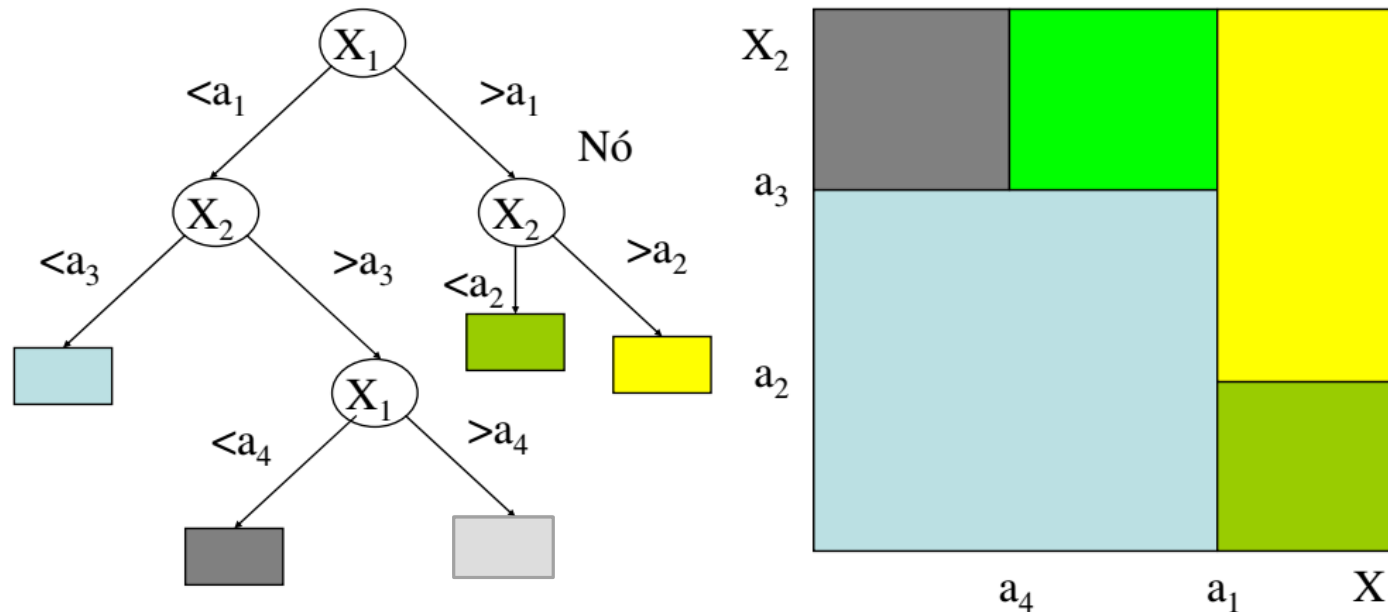
Árvores de decisão podem lidar tanto com dados numéricos quanto categóricos.



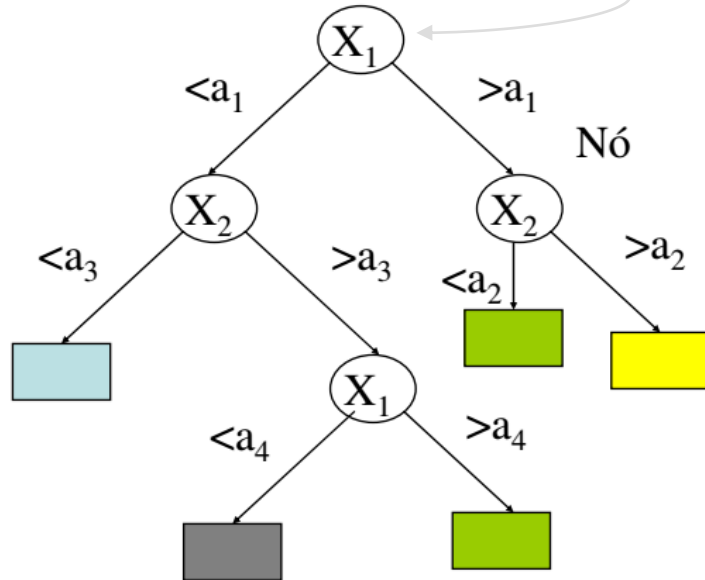
ÁRVORES DE DECISÃO

Fáceis de serem implementadas e interpretadas;

Dividir para conquistar;



ÁRVORES DE DECISÃO



Cada nó de decisão contém um teste de um atributo;

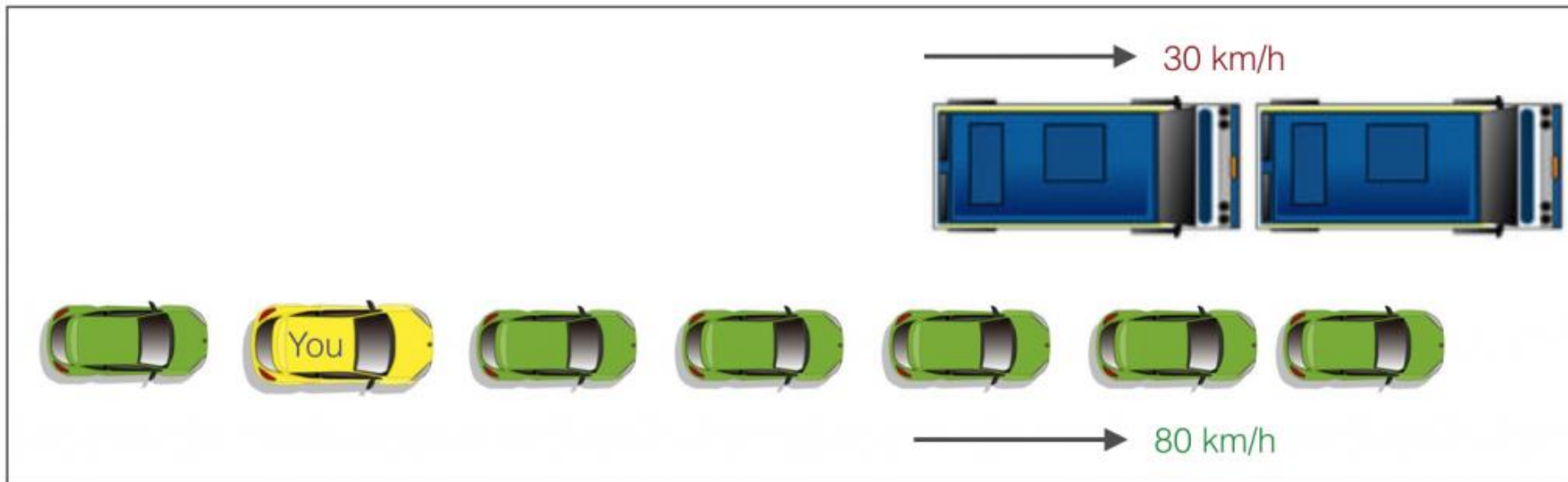
Cada folha está associada a uma classe;

Cada percurso na árvore (raiz à folha) corresponde a uma regra de classificação.

ÁRVORES DE DECISÃO – ID3

Algoritmo ID3 (J. R. Quinlan)

Busca gulosa de cima para baixo pelo espaço de possíveis ramos;



ÁRVORES DE DECISÃO – ID3

Algoritmo ID3 (J. R. Quinlan)

Busca gulosa de cima para baixo pelo espaço de possíveis ramos;

Sem backtracking;

Pode ficar preso em um ótimo local;

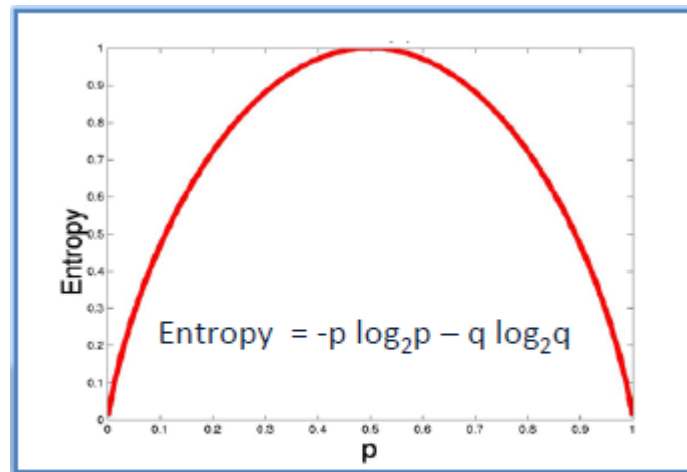
Difícil de se usar em variáveis contínuas;



ÁRVORES DE DECISÃO – ID3

Algoritmo ID3 (J. R. Quinlan)

Utiliza a entropia para calcular a homogeneidade dos dados. Se os dados são completamente homogêneos, a entropia é zero. Se os dados estão divididos igualmente, a entropia é 1.



$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

ÁRVORES DE DECISÃO – ID3



1. Calcula-se a entropia para classe (0 = dados homogêneos; 1 = dados igualmente distribuídos).

Play Golf	
Yes	No
9	5

$$\begin{aligned}
 \text{Entropy(PlayGolf)} &= \text{Entropy}(5,9) \\
 &= \text{Entropy}(0.36, 0.64) \\
 &= -(0.36 \log_2 0.36) - (0.64 \log_2 0.64) \\
 &= 0.94
 \end{aligned}$$

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

ÁRVORES DE DECISÃO – ID3

2. Divide-se a base nos diferentes atributos e calcula-se a entropia para cada um deles. Calcula-se o ganho de informação (entropia para classe – entropia para atributo).

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$$\begin{aligned} E(\text{PlayGolf}, \text{Outlook}) &= P(\text{Sunny}) * E(3,2) + P(\text{Overcast}) * E(4,0) + P(\text{Rainy}) * E(2,3) \\ &= (5/14) * 0.971 + (4/14) * 0.0 + (5/14) * 0.971 \\ &= 0.693 \end{aligned}$$

$$\begin{aligned} G(\text{PlayGolf}, \text{Outlook}) &= E(\text{PlayGolf}) - E(\text{PlayGolf}, \text{Outlook}) \\ &= 0.940 - 0.693 = 0.247 \end{aligned}$$

O ganho de informação é baseado na diminuição da entropia depois que uma base de dados é subdividida em um atributo.

ÁRVORES DE DECISÃO – ID3

2. Divide-se a base nos diferentes atributos e calcula-se a entropia para cada um deles. Calcula-se o ganho de informação (entropia para classe – entropia para atributo).

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			


		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

O ganho de informação é baseado na diminuição da entropia depois que uma base de dados é subdividida em um atributo.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$


ÁRVORES DE DECISÃO – ID3

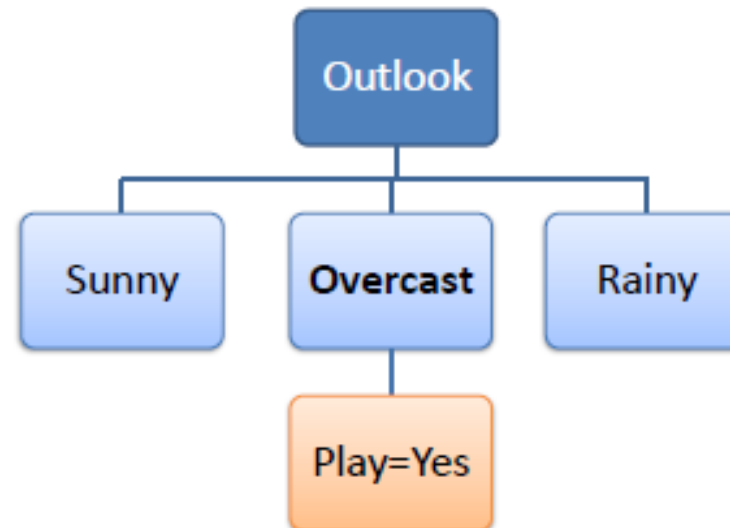
3. Escolhe-se atributo com maior ganho de informação como nó de decisão.

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

ÁRVORES DE DECISÃO – ID3

3. Escolhe-se atributo com maior ganho de informação como nó de decisão.
 - a. Um ramo que tenha entropia 0 é uma folha.

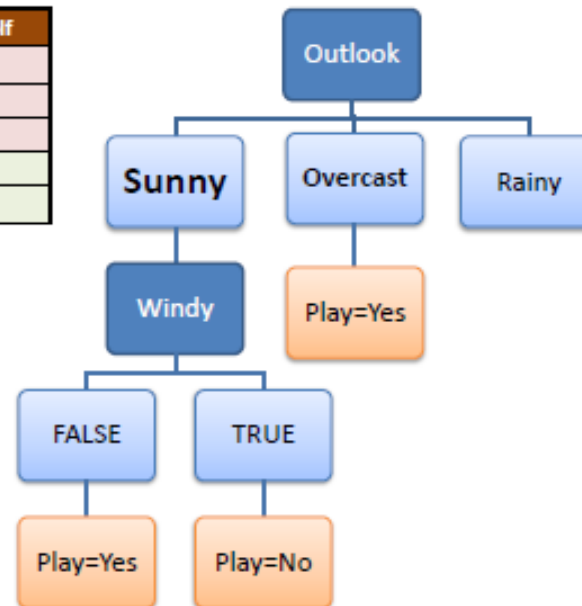
		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			



ÁRVORES DE DECISÃO – ID3

3. Escolhe-se atributo com maior ganho de informação como nó de decisão.
- a. Um ramo que tenha entropia 0 é uma folha.
 - b. Um ramo com entropia maior que 0 precisa ser subdividido.

Temp.	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



★		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

ÁRVORES DE DECISÃO – ID3

4. Algoritmo é rodado recursivamente para todos os ramos sem folha até que todos os dados sejam classificados.

Regras de
decisão

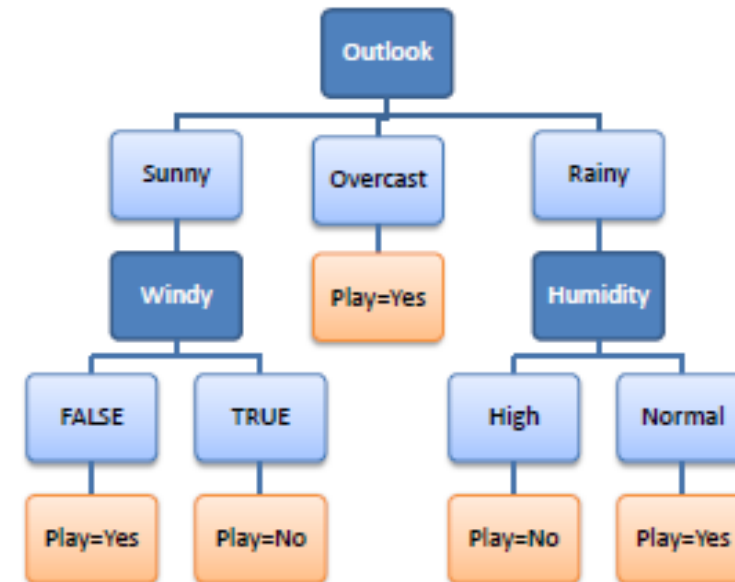
R_1 : IF (Outlook=Sunny) AND
(Windy=FALSE) THEN Play=Yes

R_2 : IF (Outlook=Sunny) AND
(Windy=TRUE) THEN Play=No

R_3 : IF (Outlook=Overcast) THEN
Play=Yes

R_4 : IF (Outlook=Rainy) AND
(Humidity=High) THEN Play=No

R_5 : IF (Outlook=Rain) AND
(Humidity=Normal) THEN
Play=Yes

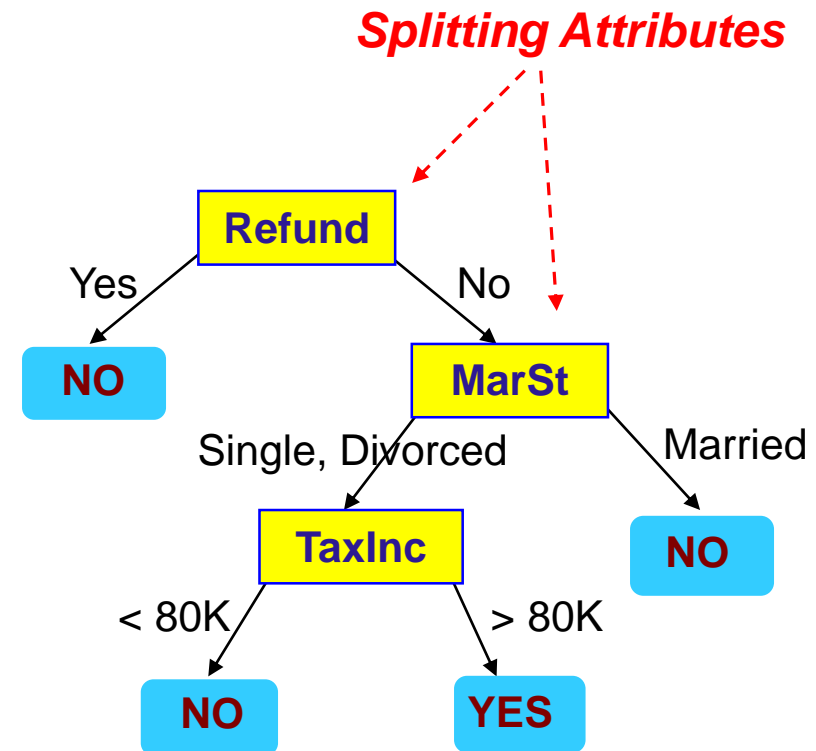


EXEMPLO

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

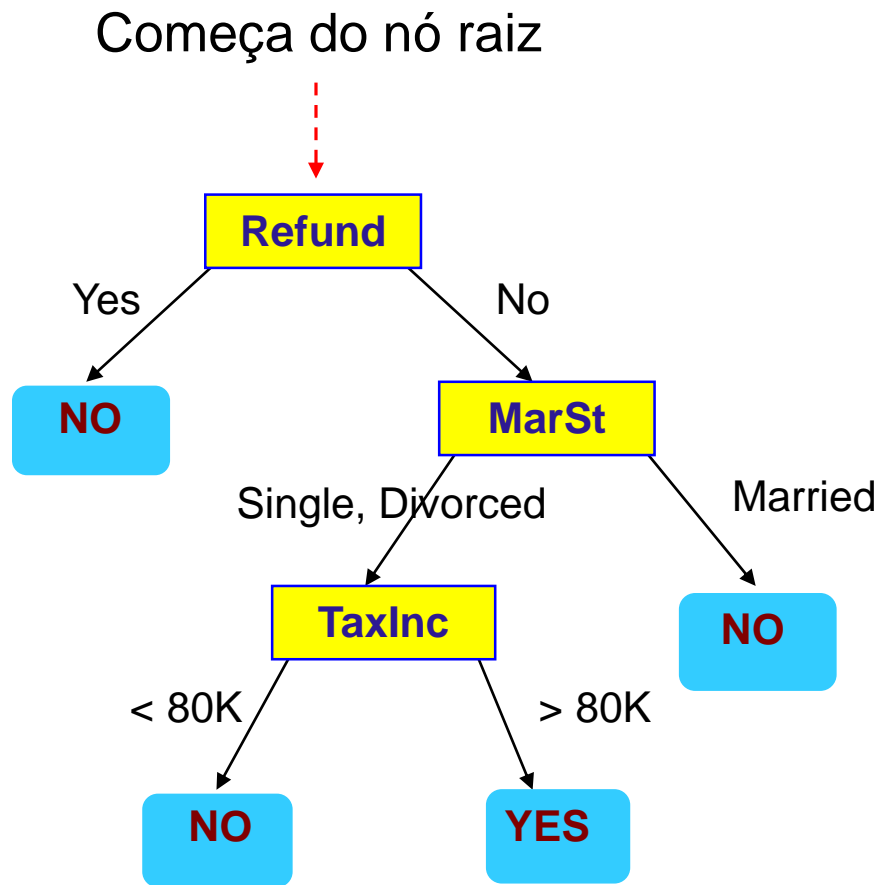
categorical
categorical
continuous
class

Training Data



Model: Decision Tree

APLICANDO O MODELO PARA INFERIR UM DADO



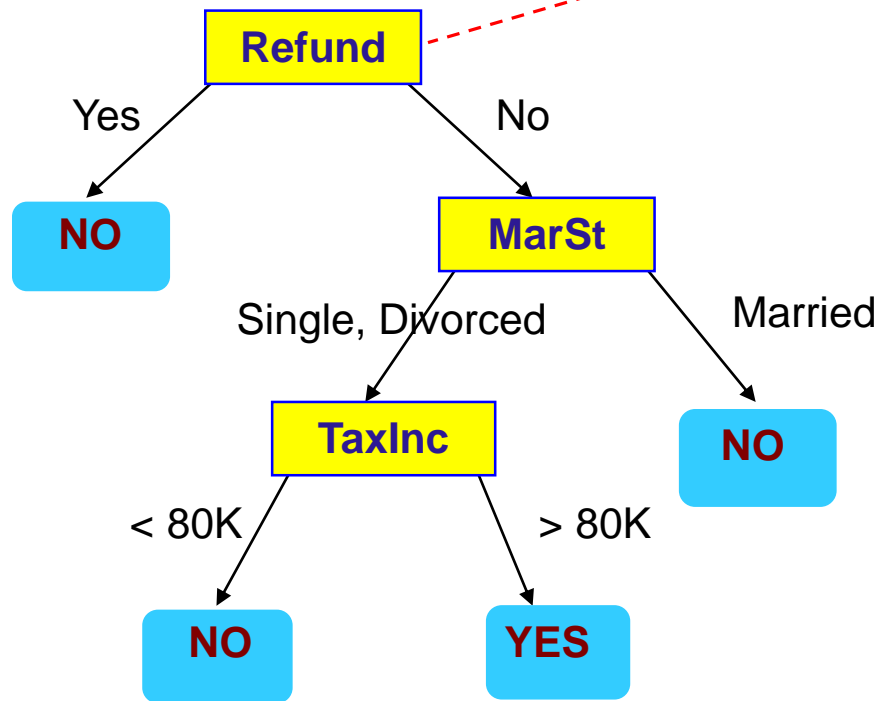
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

APLICANDO O MODELO PARA INFERIR UM DADO

Test Data

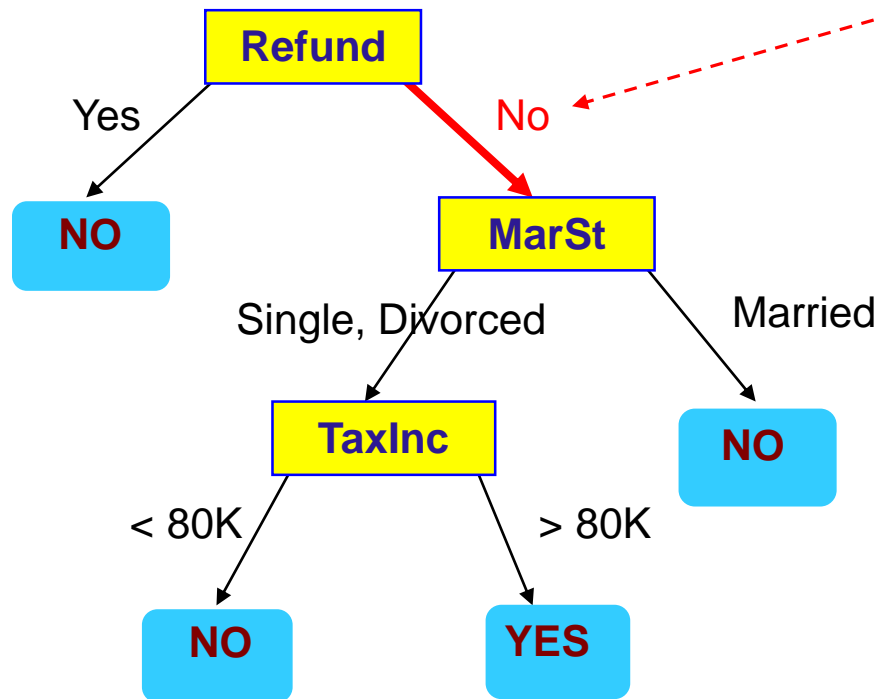
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APLICANDO O MODELO PARA INFERIR UM DADO

Test Data

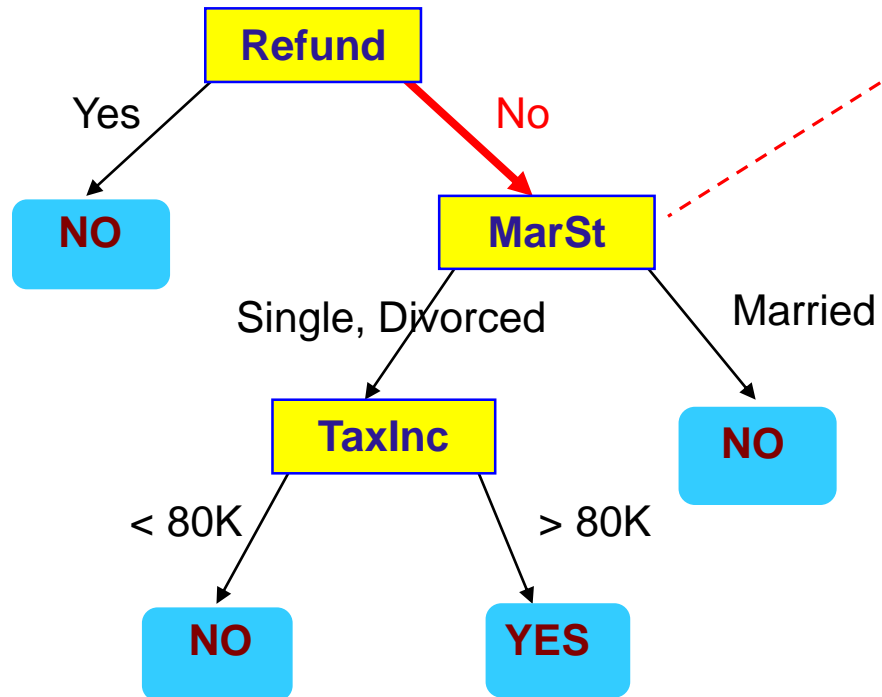
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APLICANDO O MODELO PARA INFERIR UM DADO

Test Data

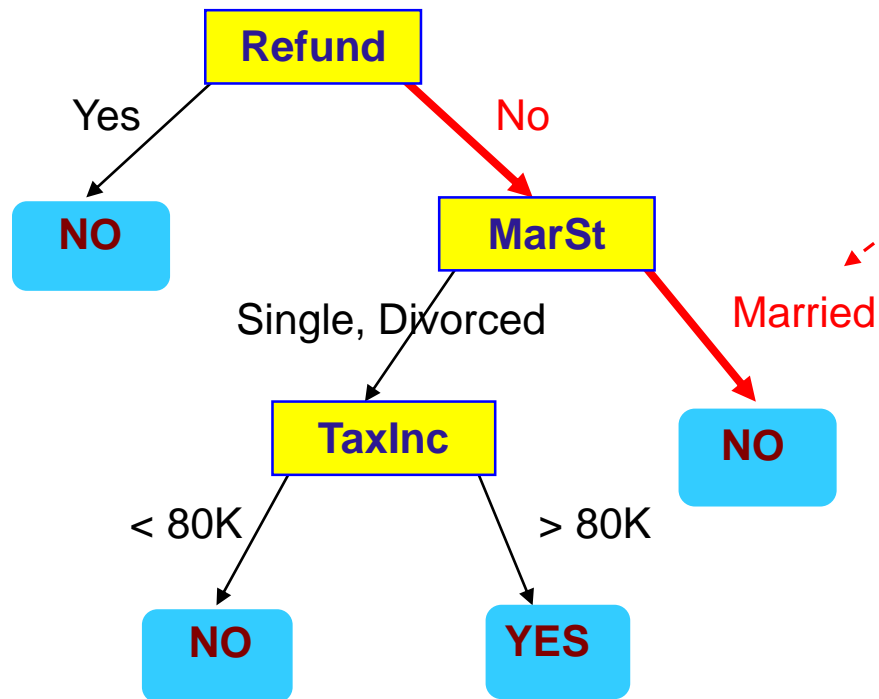
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APLICANDO O MODELO PARA INFERIR UM DADO

Test Data

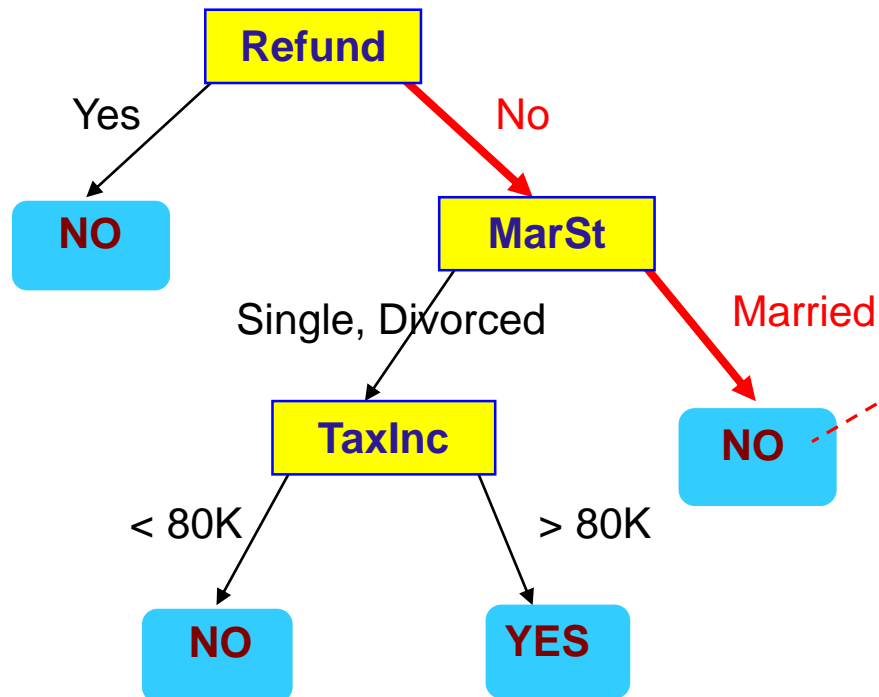
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APLICANDO O MODELO PARA INFERIR UM DADO

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to "No"

ÁRVORES DE DECISÃO

Árvores de decisão são fáceis de se entender;

Elas funcionam mais eficientemente com atributos discretos;

Extremamente rápidas em classificar dados novos;

ESTUDO DE CASO

Exercício

ESTUDO DE CASO

CRÉDITO BANCÁRIO

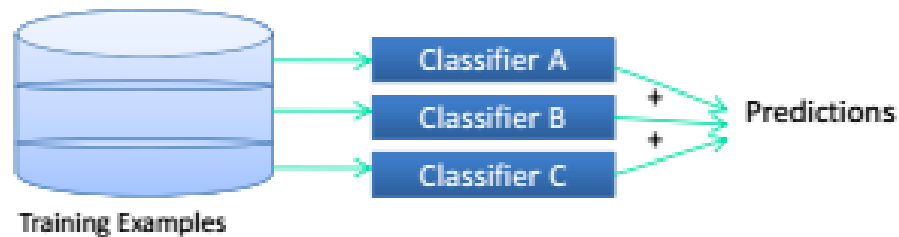
COMITÊS



COMITÊS

Agregar múltiplos modelos treinados com o objetivo de melhorar a acurácia do modelo conjunto.

Intuição: simula o que fazemos quando combinamos conhecimento de especialistas em um processo de tomada de decisão.



COMITÊS

Conhecidos também por:

- Comitês especialistas;
- Sistemas múltiplos de classificação;
- Comitê de classificadores;
- Máquina de Comitê;
- Mistura de especialistas;
- Aprendizado em conjunto.

Diversos estudos demonstram sua utilização com sucesso em problemas onde um único especialista não funciona bem.

COMITÊS

Votação:

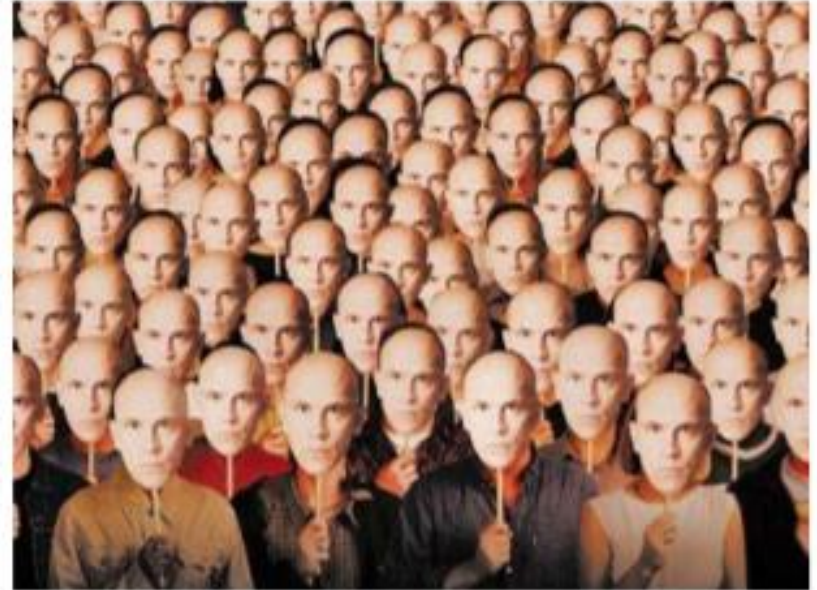
- Maioria do votos
- Maioria ponderada dos votos
- Borda count
- Média
- Média ponderada
- Soma
- Soma ponderada
- Produto
- Máximo
- Mínimo
- Mediana

APRENDIZADO DO COMITÊ

Aprendizado de Comitês

Às vezes cada técnica de aprendizado retorna diferentes 'hipóteses' (funções), mas nenhuma hipótese perfeita.

Poderíamos combinar várias hipóteses imperfeitas para se ter uma hipótese melhor?



MOTIVAÇÃO

Analogias:

Eleições combinam votos de eleitores para escolher um candidato bom;

Comitês combinam opiniões de especialistas para tomar decisões melhores;

Estudantes trabalhando em conjunto em um projeto.

Intuição:

Indivíduos cometem erros, mas a maioria é menos propensa a erros;

Indivíduos em geral têm conhecimento parcial. Um comitê pode juntar conhecimento para tomar decisões melhores.

COMITÊS

Quando usar?

- Temos um conjunto muito grande de dados;
- A região de domínio do problema é muito complexa;
- Queremos melhorar os resultados de classificadores individuais.

COMITÊS

Vantagem

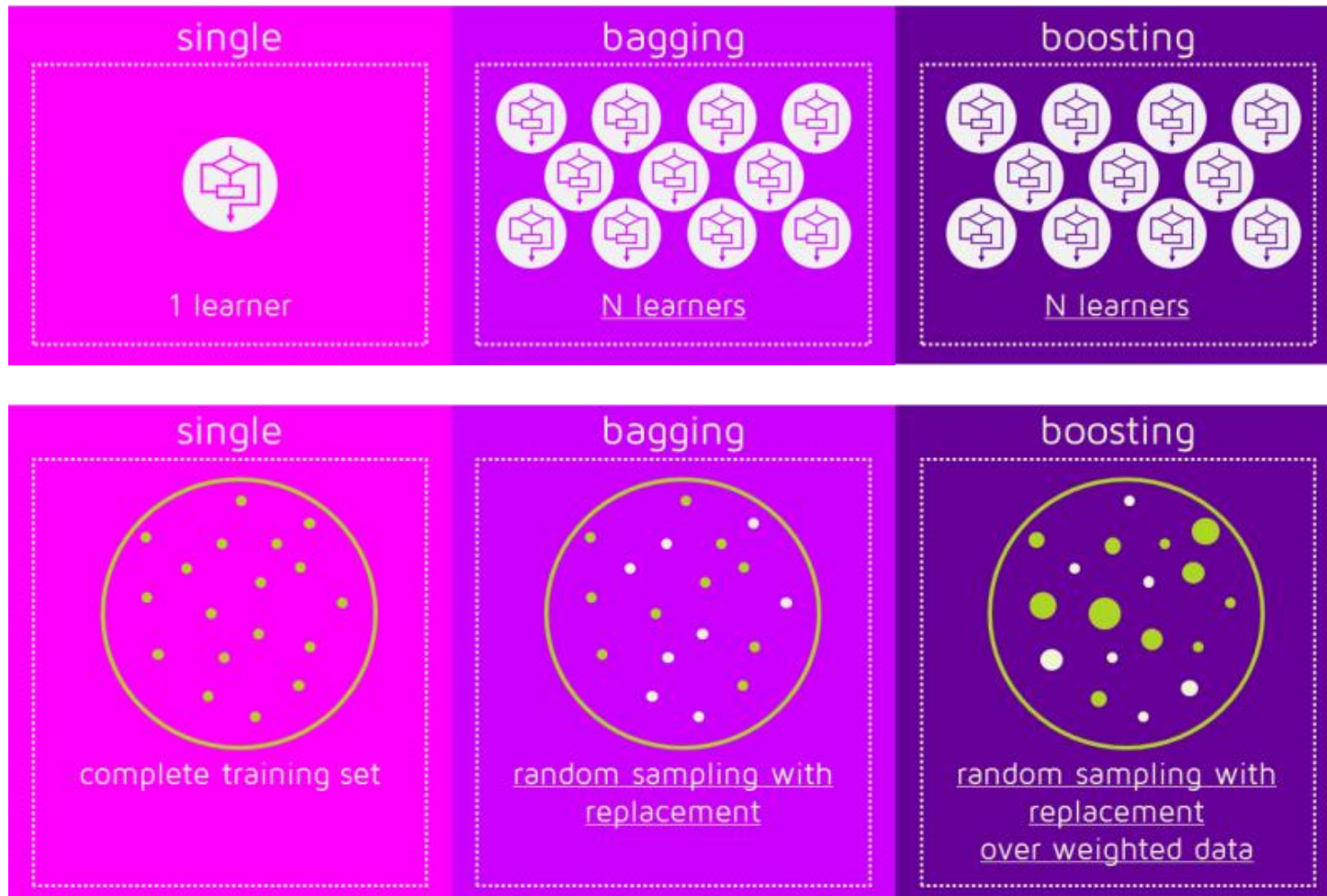
- A combinação de modelos pode apresentar melhor desempenho que um modelo só;
- Neutraliza ou minimiza fortemente a instabilidade inerente aos algoritmos de aprendizagem;

Desvantagens

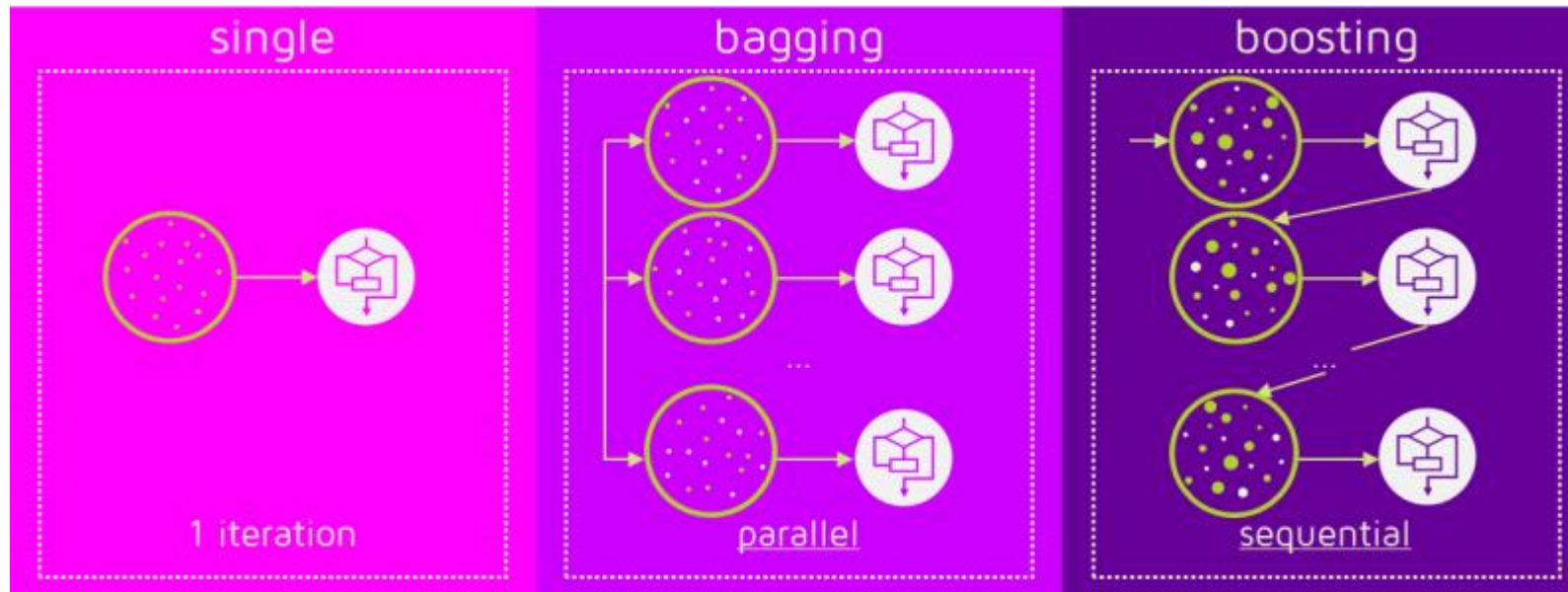
- Não há garantia de que as estruturas modulares apresentem os melhores resultados;
- Modelos combinados são mais difíceis de analisar;
- O custo é alto.

COMITÊS - TÉCNICAS

COMITÊS: BAGGING X BOOSTING

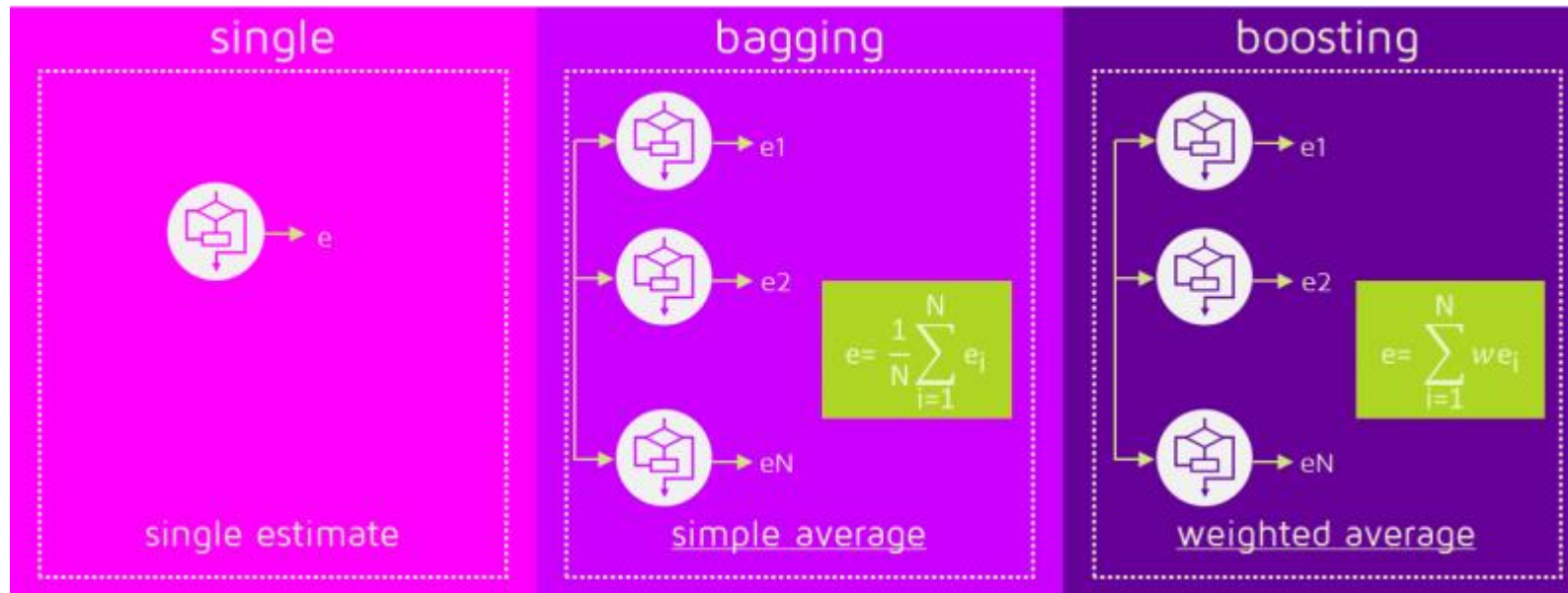


COMITÊS: BAGGING X BOOSTING



Boosting: Observações classificadas incorretamente são atribuídas maiores pesos!

COMITÊS: BAGGING X BOOSTING



- Bagging: média ou maioria de votos.
- Boosting: média ponderada (classificadores com resultados melhores, têm pesos maiores).

COMITÊS: BAGGING X BOOSTING

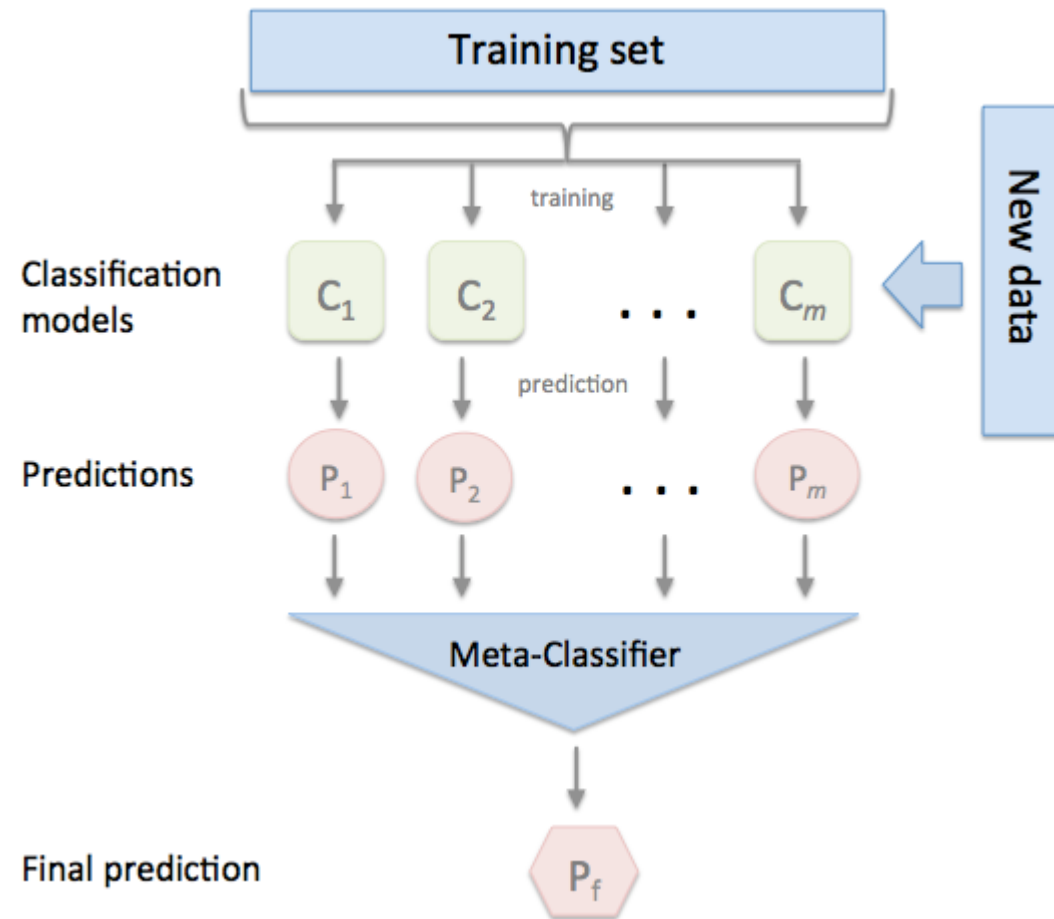


Boosting:

- Durante etapa de treinamento, erro de treinamento é guardado.
- Existe uma condição para determinar se um modelo será utilizado ou será descartado.

▪ AdaBoost, LPBoost, XGBoost, GradientBoost, BrownBoost

COMITÊS: STACKING



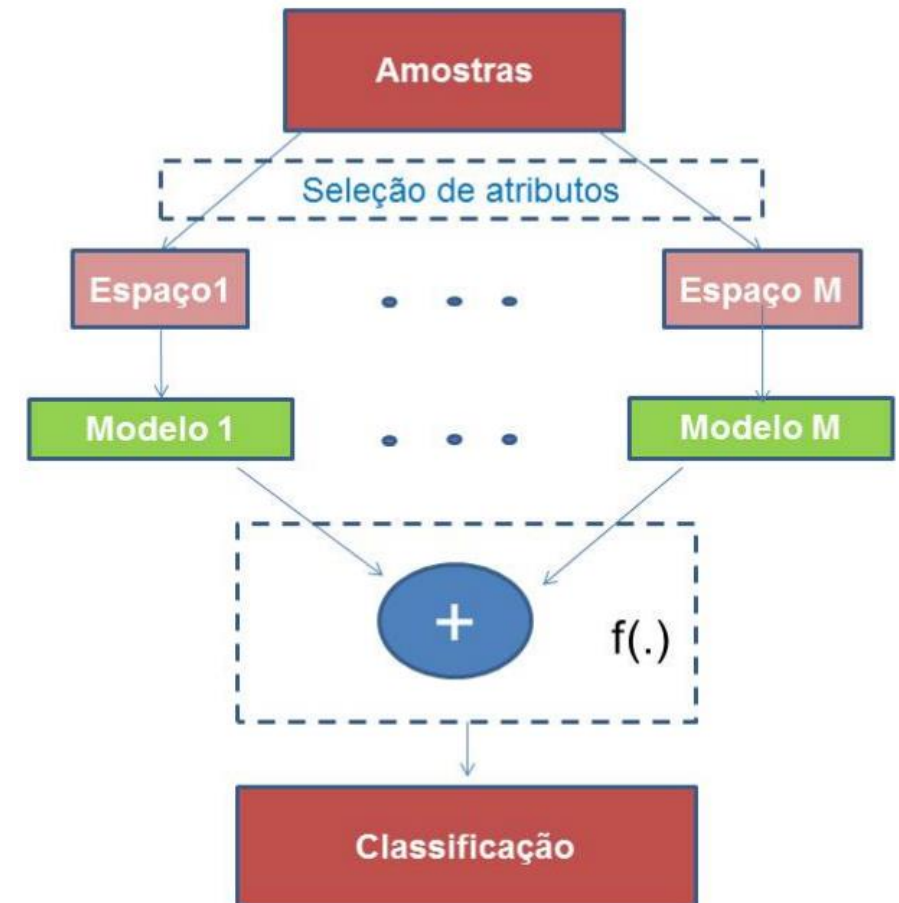
COMITÊS: RMS

Random Subspace Method (RMS)

Similar ao Bagging, mas com aleatorização sobre os atributos.

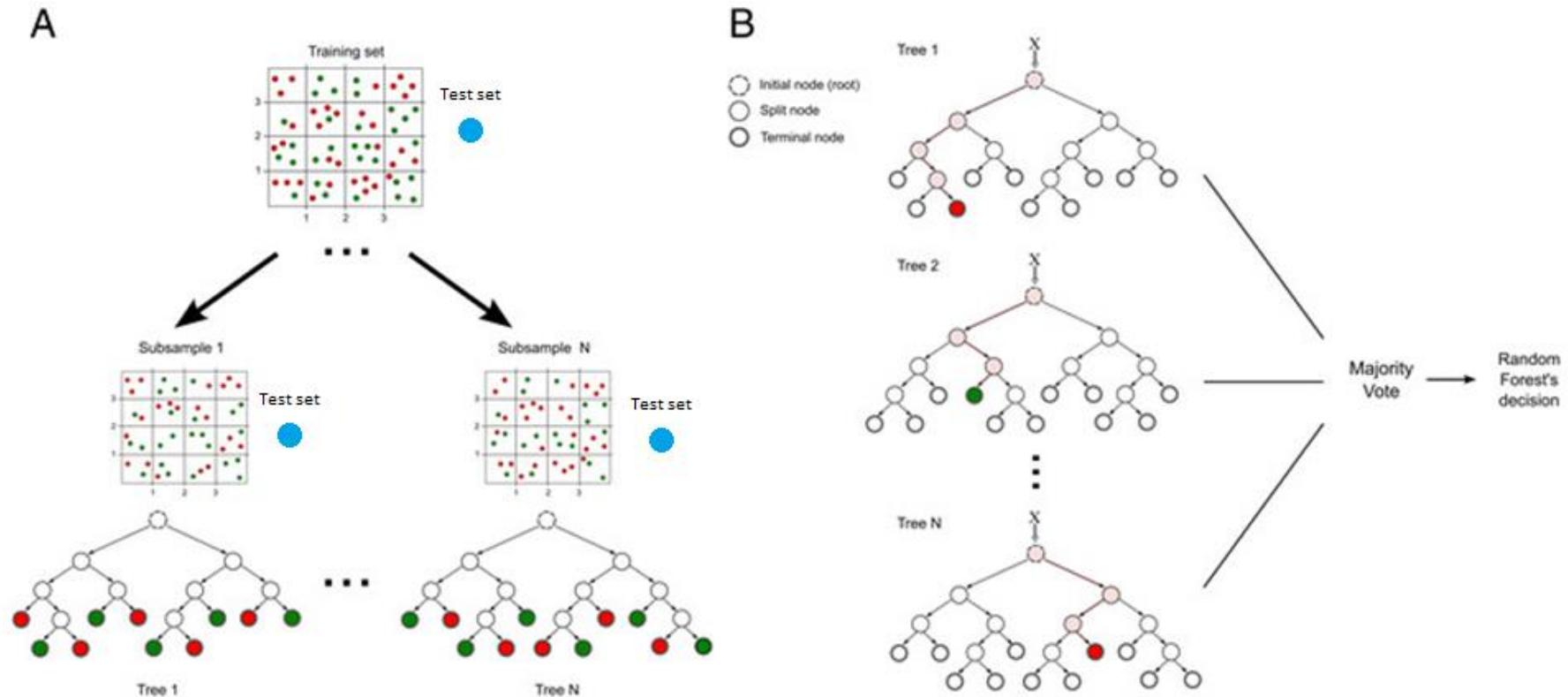
Classificadores-base aprendem nos subespaços S de mesma dimensão.

Decisão final é por votação.



RANDOM FOREST

RANDOM FOREST: WORKFLOW



A proporção de votos diferentes da classe target em relação ao total de votos é o erro OOB (Out-Of-Bag estimate)

RANDOM FOREST - CLASSIFICAÇÃO

Criada através de árvores de decisão individuais cujos parâmetros podem variar aleatoriamente.

- **Treinamento**



- **Recall**



ESTUDOS DE CASO



Prêmio de 1 milhão de dólares

- Melhora na acurácia do sistema de recomendação de filmes da Netflix em 10%.
- Os melhores times combinaram diversos modelos e algoritmos em um comitê.

http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf

NETFLIX PRIZE

Tarefa de aprendizado supervisionado

- Dados de treinamento são formados por um conjunto de usuários e as avaliações dos filmes (1,2,3,4,5 estrelas) feitas por esses usuários;
- Construir um classificador que dado um usuário e um filme não avaliado, classifique corretamente aquele filme como 1, 2, 3, 4, ou 5 estrelas;
- Prêmio de \$1 milhão para 10% em melhora na acurácia em relação ao modelo atual.

ESTUDO DE CASO

Exercício



Master

POS-GRADUAÇÃO EM CIÊNCIA DE DADOS

ESTUDO DE CASO

CRÉDITO BANCÁRIO