CLASSIFICAÇÃO

MANOELA KOHLER

Prof.manoela@ica.ele.puc-rio.br

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

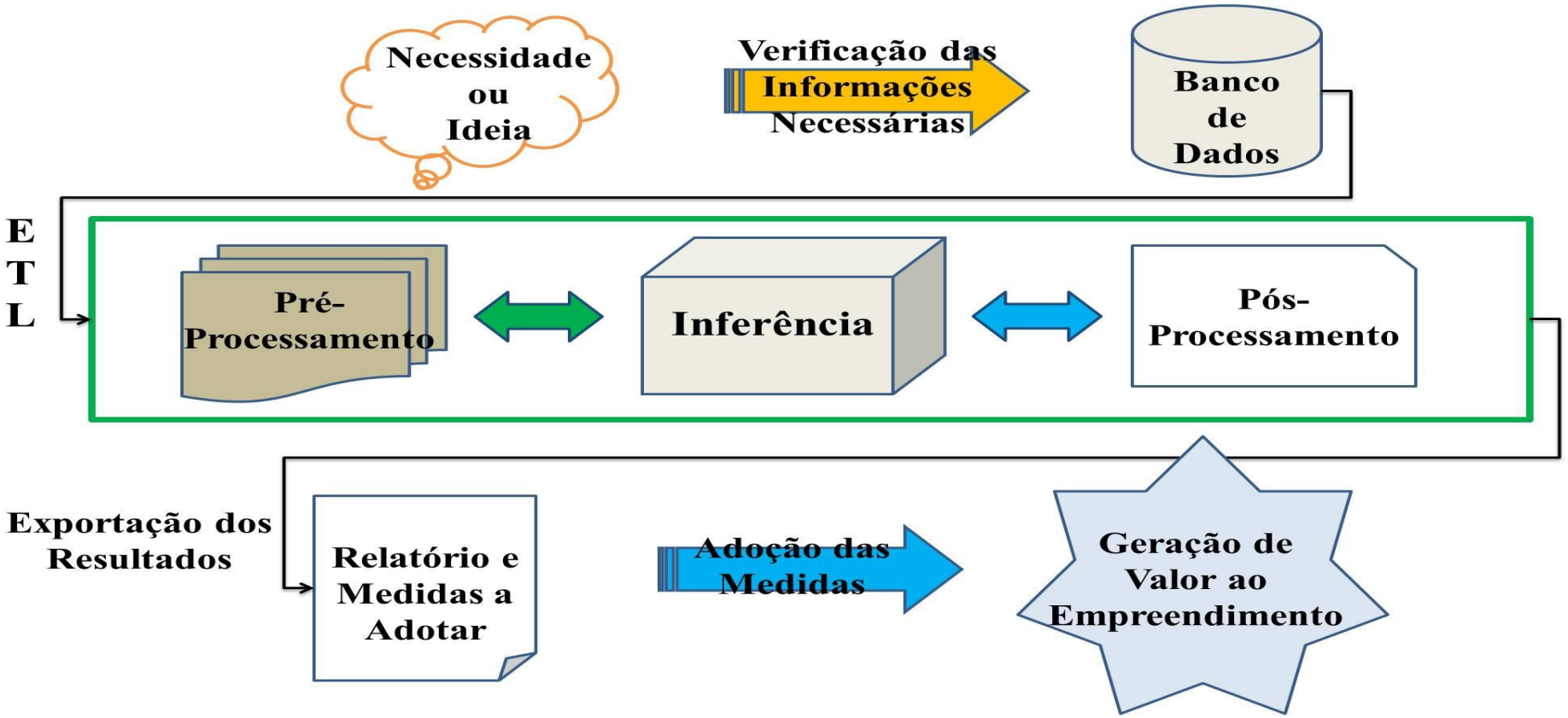
- Naive
- Média Móvel
- Amortecimento exponencial
- Auto-regressivo integrados de média móvel
- Auto regressivo não linear

BI Master

Recapitulação

ETAPAS DE UM PROJETO DE DATA MINING

ESQUEMA BÁSICO DE UM PROJETO DE DM



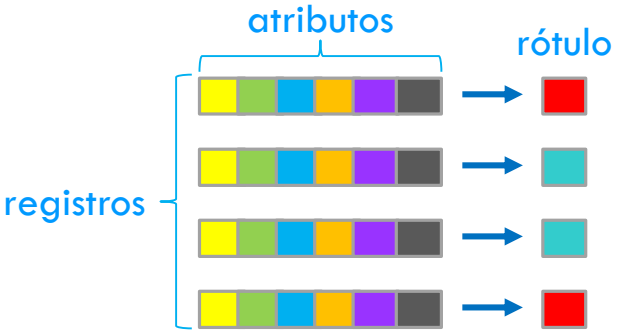
CLASSIFICAÇÃO

5 CLASSES DE PROBLEMAS DE DM

Machine Learning

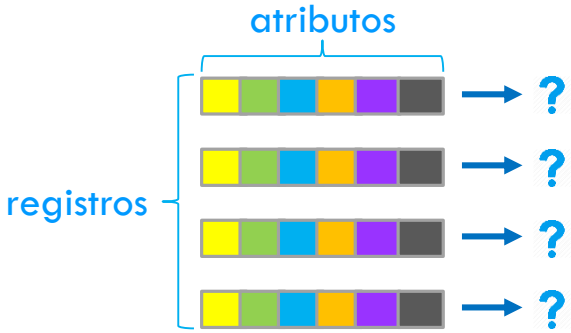
Supervisionado

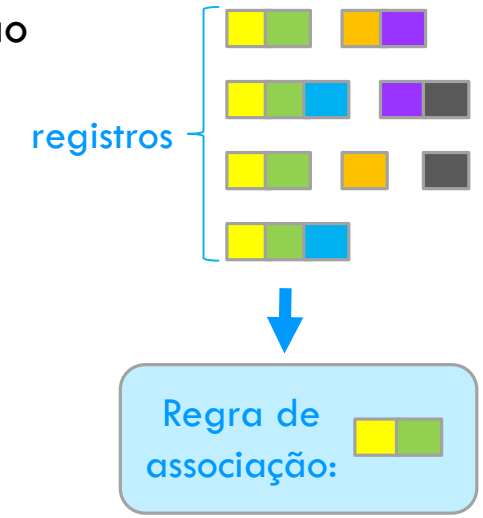
Classificação
Regressão
Previsão de séries temporais



Não Supervisionado

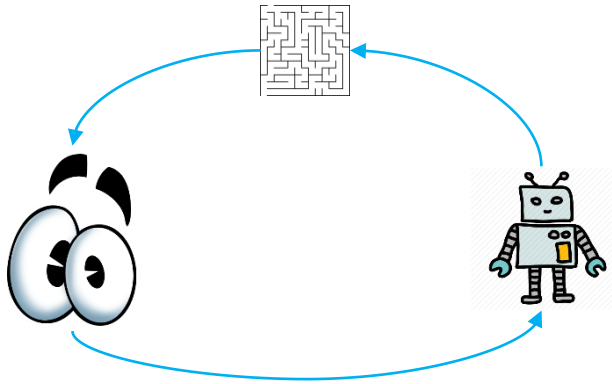
Agrupamento
Associação





Reforço

Aprendizado através da interação de agentes com um ambiente

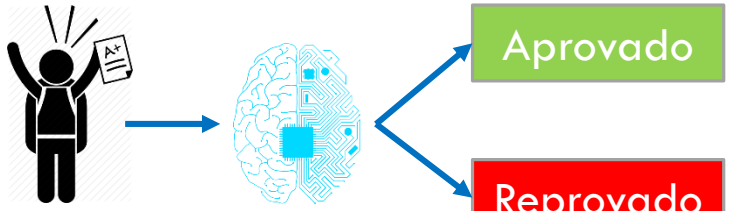


SUPERVISIONADO

- Aproximador: função mapeia entradas e saída.

Classificação

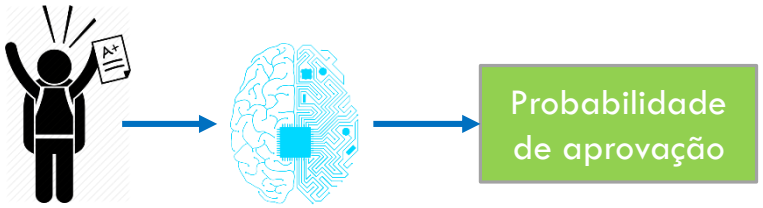
Rótulo é categórico.



Dados do estudante

Regressão

Rótulo é contínuo.



Dados do estudante

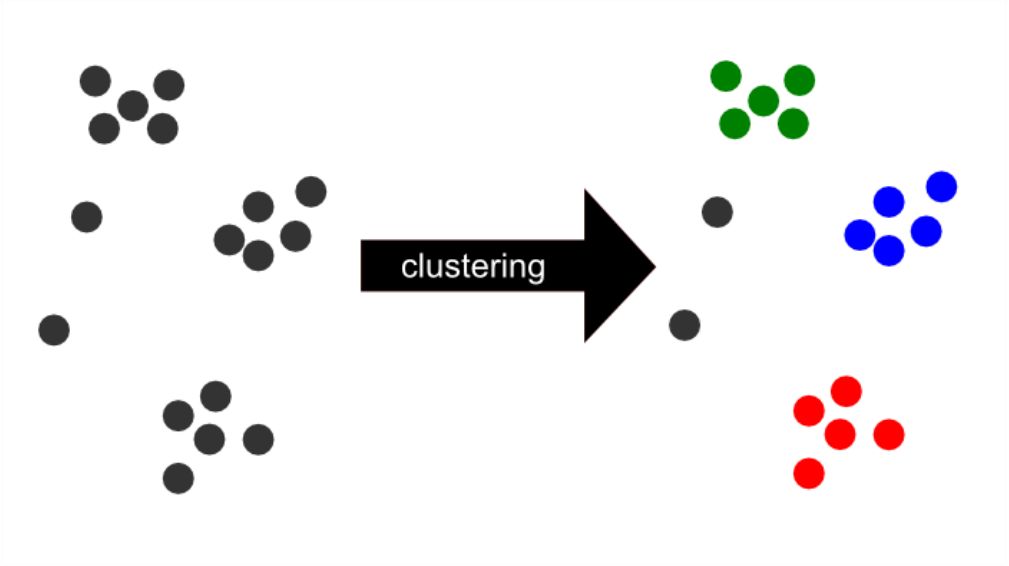
Previsão de Séries

Rótulo é contínuo e dependente do tempo.

NÃO SUPERVISIONADO

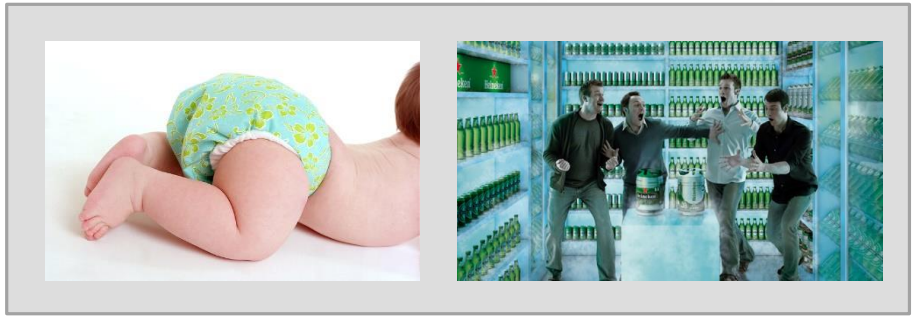
Agrupamento

Descoberta de semelhanças e grupos entre registros.



Associação

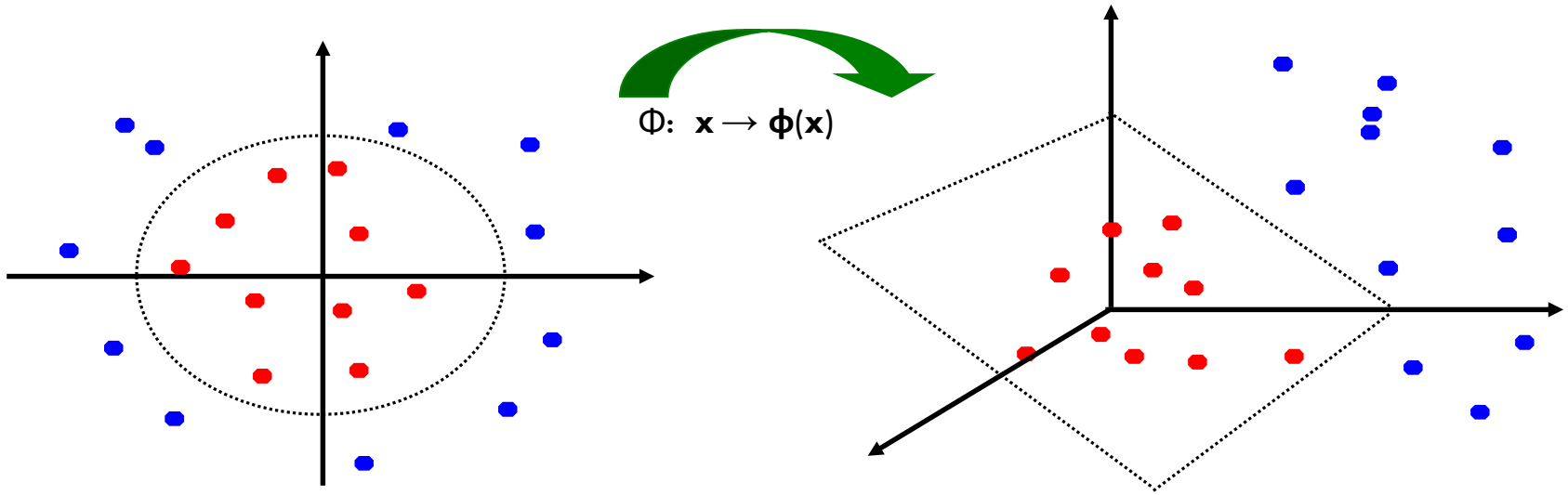
Descoberta de relações entre variáveis.



SVM - Support Vector Machine

SVM NÃO LINEAR

Ideia geral: o espaço de atributos original pode – com alta probabilidade (Teorema de Cover) – ser mapeado para um espaço de atributos maior, onde os dados podem ser separados:



Árvores de Decisão

ÁRVORES DE DECISÃO – ID3

Predictors | Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Regras de decisão

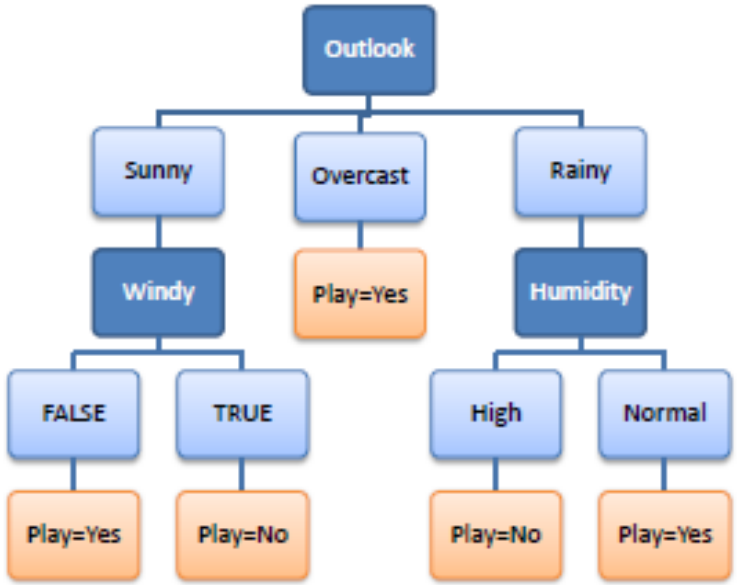
R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

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

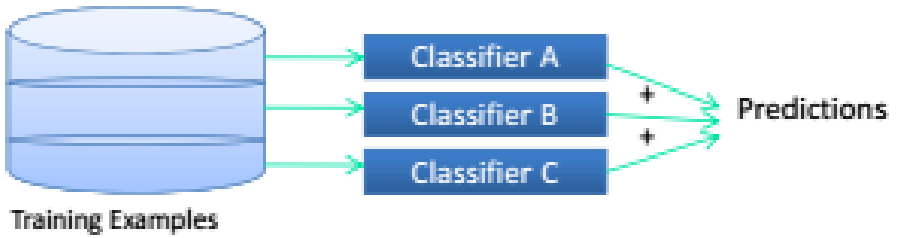


COMITÊS

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:

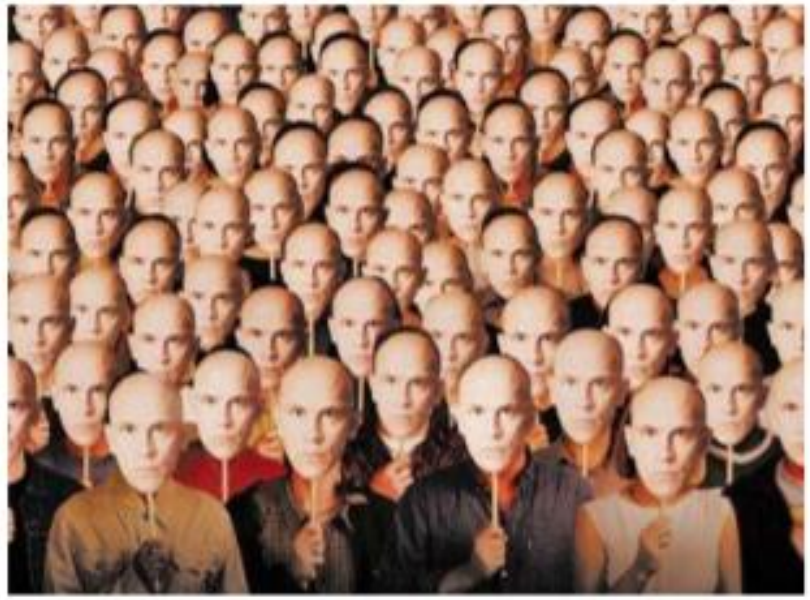
- Maioria do votos
- Maioria ponderada dos votos
- Borda count
- Média
- Média ponderada
- Soma
- Soma ponderada
- Produto
- Máximo
- Mínimo
- Mediana

APRENDIZADO DO COMITÊ

Aprendizado de Comitês

Ás 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

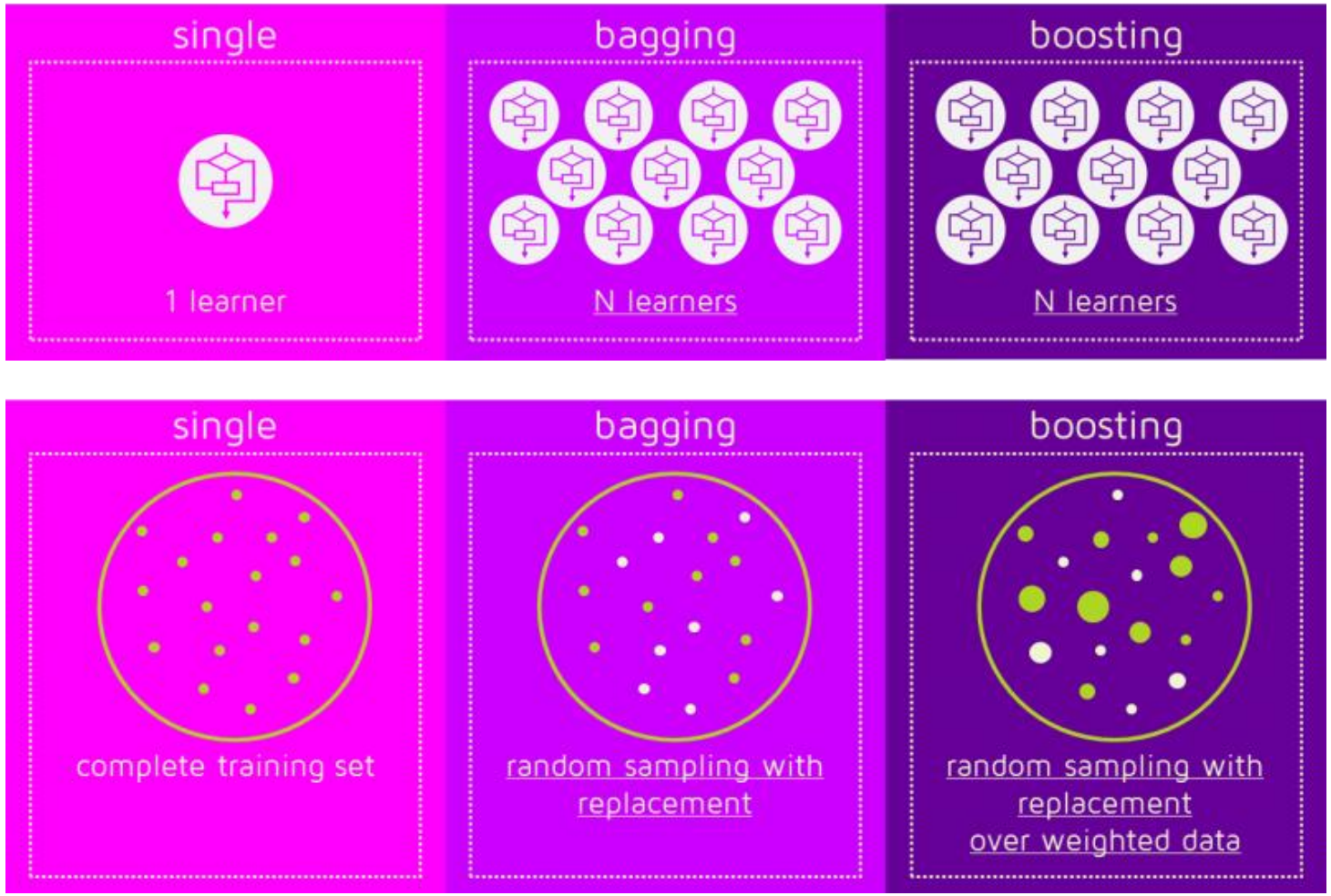
- A combinação de modelos pode apresentar melhor desempenho que um modelo só;
- Neutraliza ou minimiza fortemente a instabilidade inerente aos algoritmos de aprendizagem;

Desvantagens

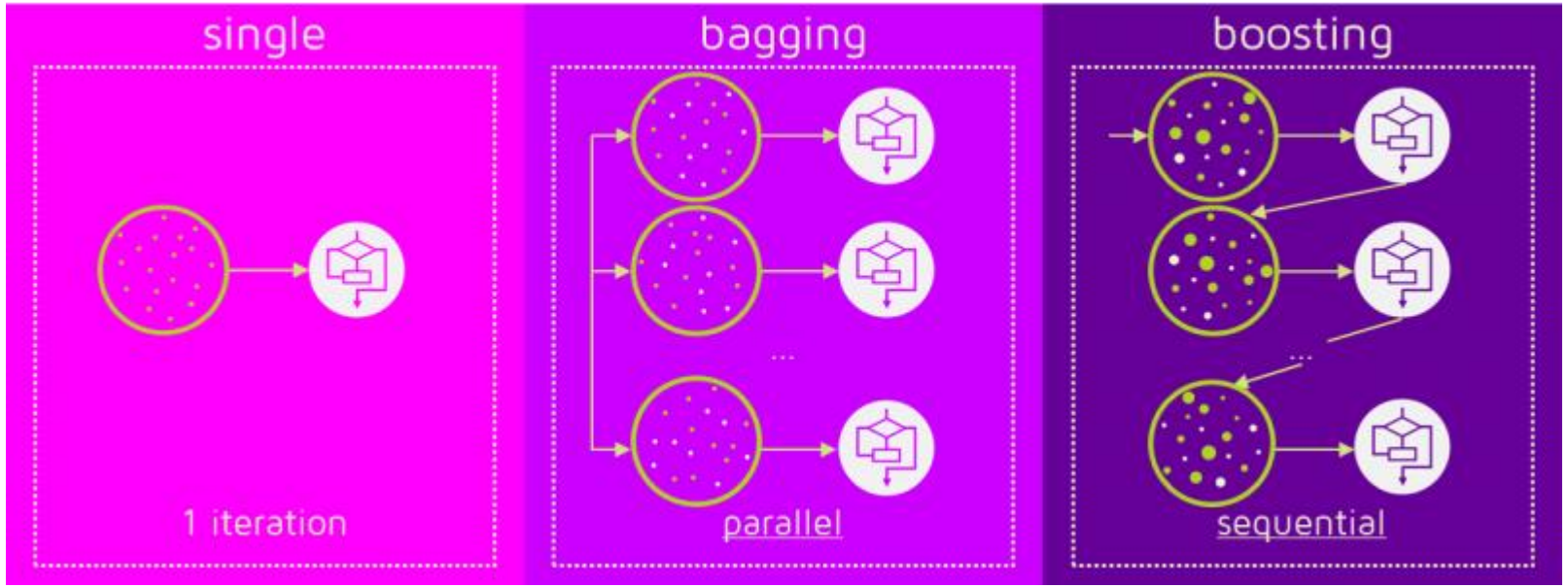
- Não há garantia de que as estruturas modulares apresentem os melhores resultados;
- Modelos combinados são mais difíceis de analisar;
- O custo é alto.

COMITÊS - TÉCNICAS

COMITÊS: BAGGING X BOOSTING

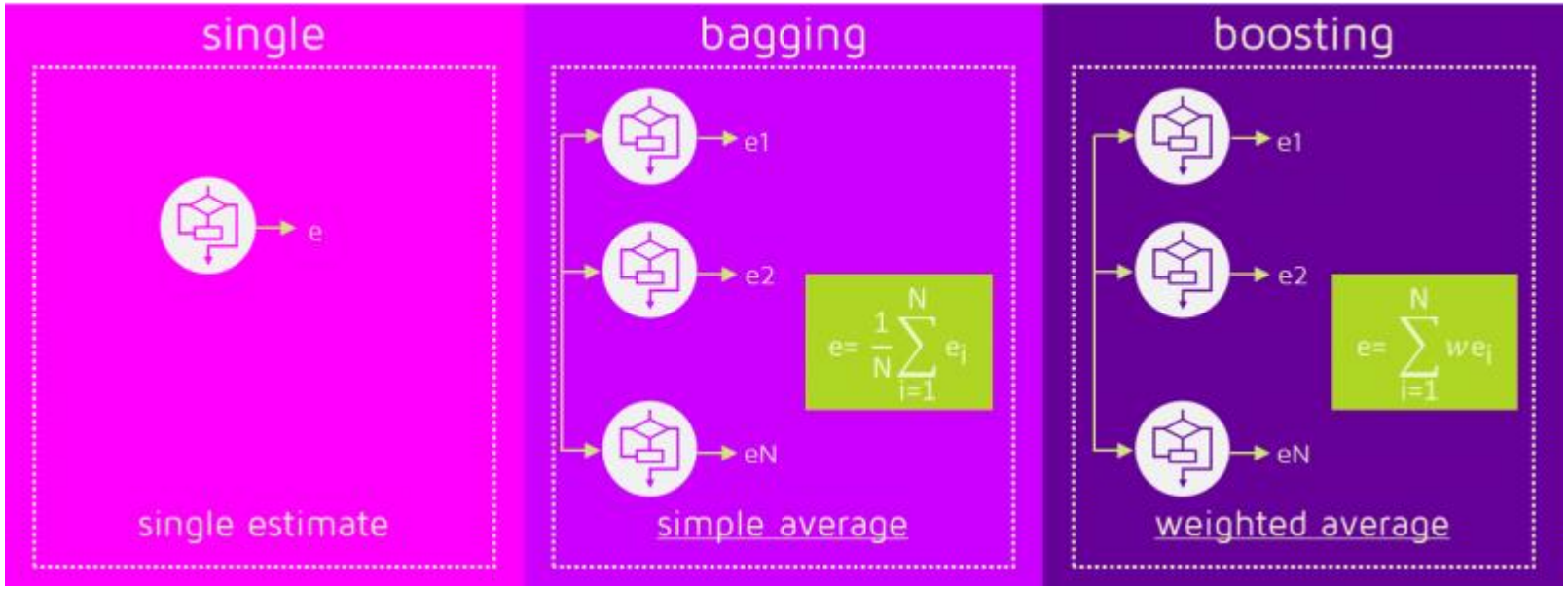


COMITÊS: BAGGING X BOOSTING



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).

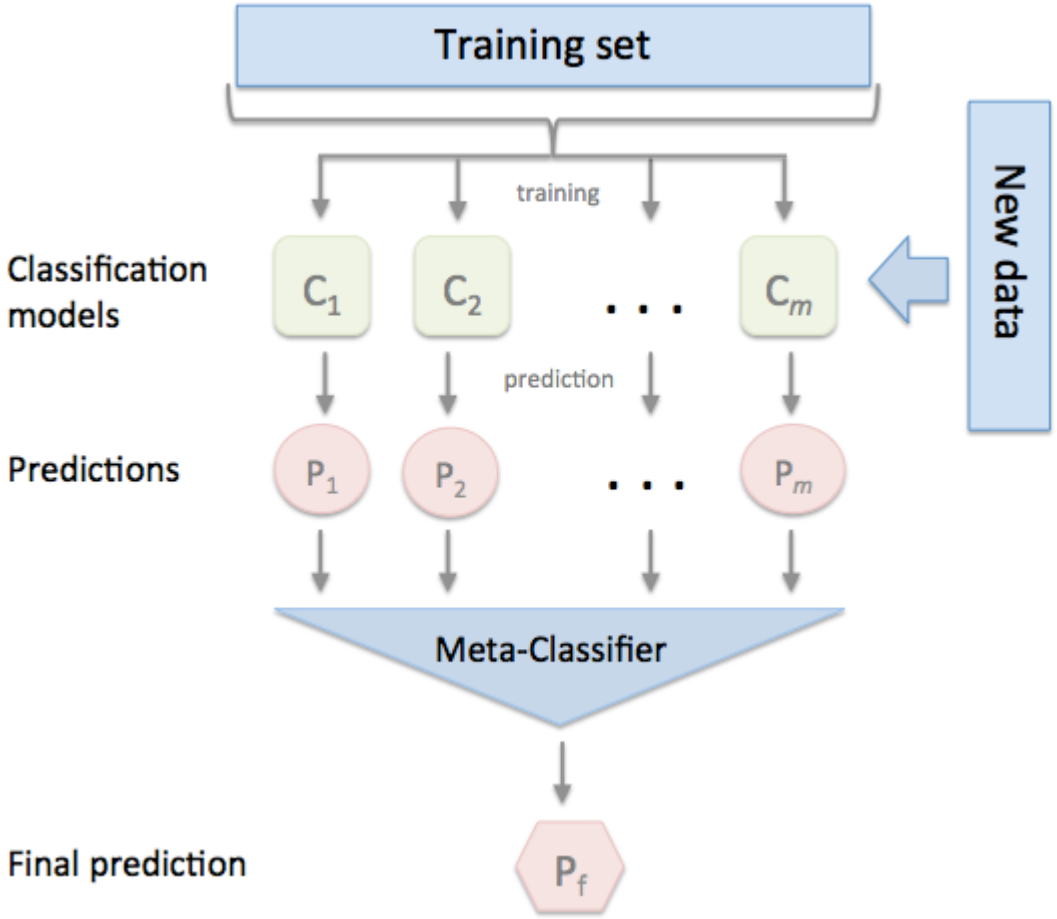
COMITÊS: BAGGING X BOOSTING

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



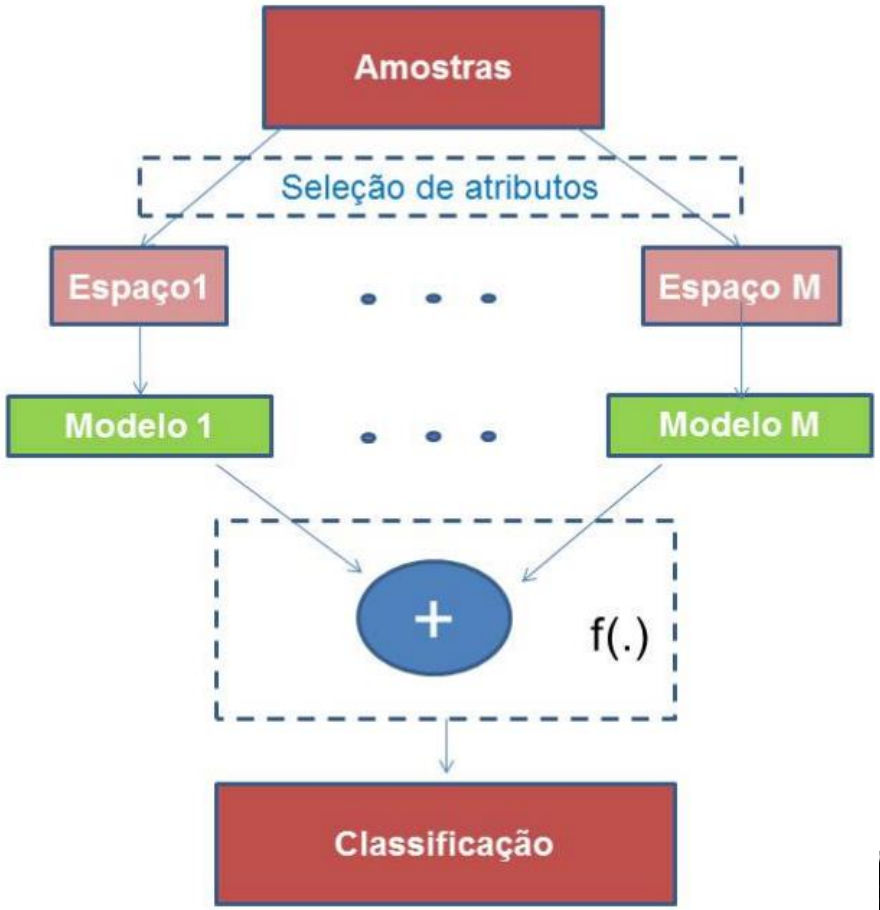
COMITÊS: RMS

Random Subspace Method (RMS)

Similar ao Bagging, mas com aleatorização sobre os atributos.

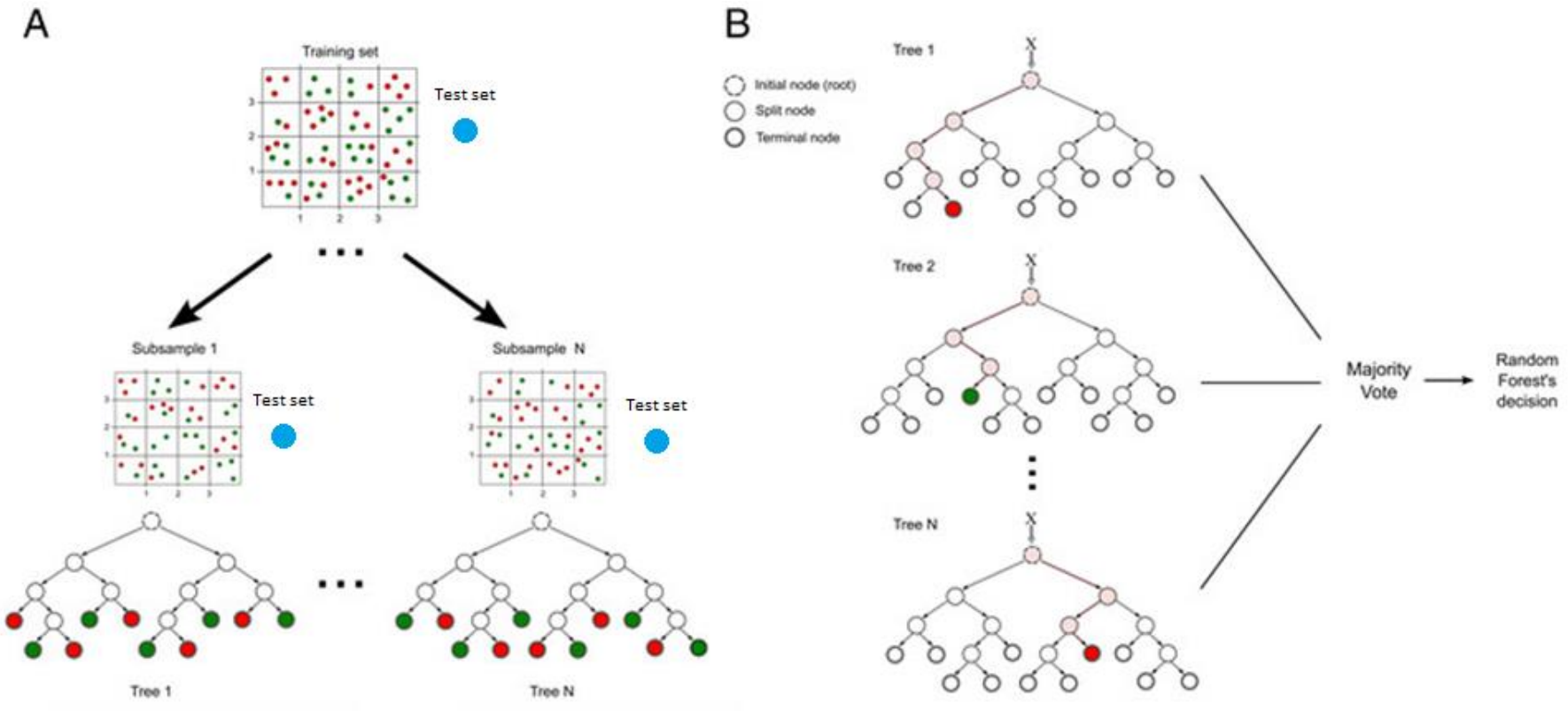
Classificadores-base aprendem nos subespaços *S* de mesma dimensão.

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**

  Bagging (Bootstrap) + Random subspace method

- **Recall**

  Maioria de votos

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

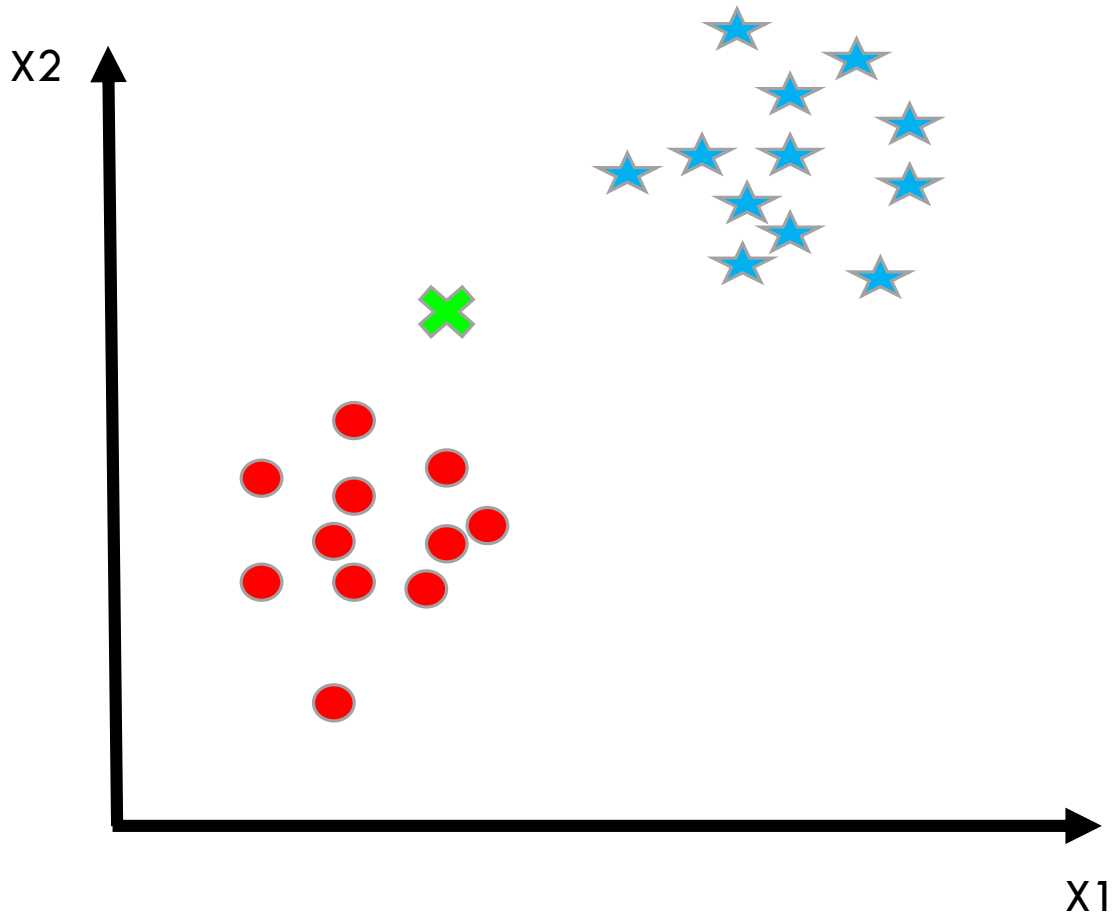
Tarefa de aprendizado supervisionado

- 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.

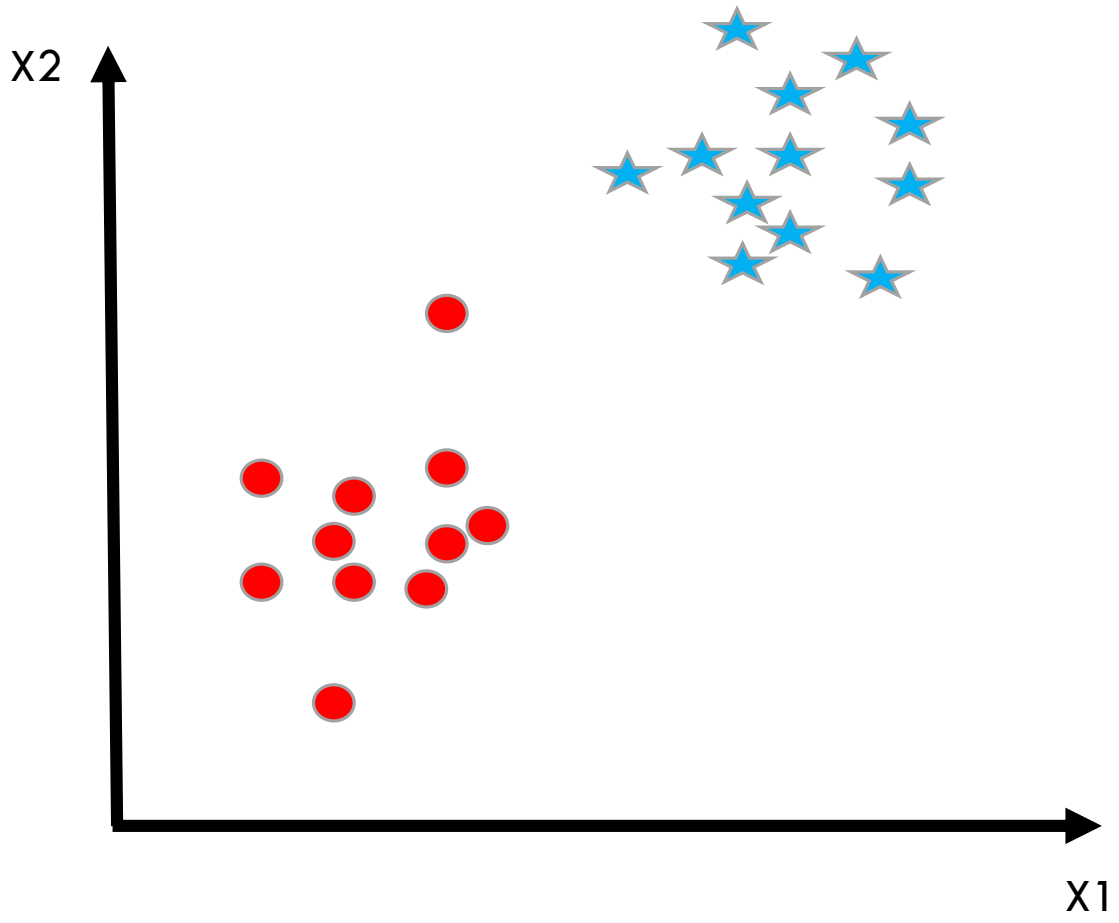
ESTUDOS DE CASO

KNN

KNN



KNN



KNN

1. Determinar o valor de K, ou número de vizinhos
2. Calcular a distância entre cada par de registros
3. Determinar quais são os K registros (vizinhos) mais próximos do novo registro
4. Dentre esses K vizinhos, contar o número de vizinhos em cada classe
5. O novo registro vai ser da classe majoritária entre os vizinhos mais próximos

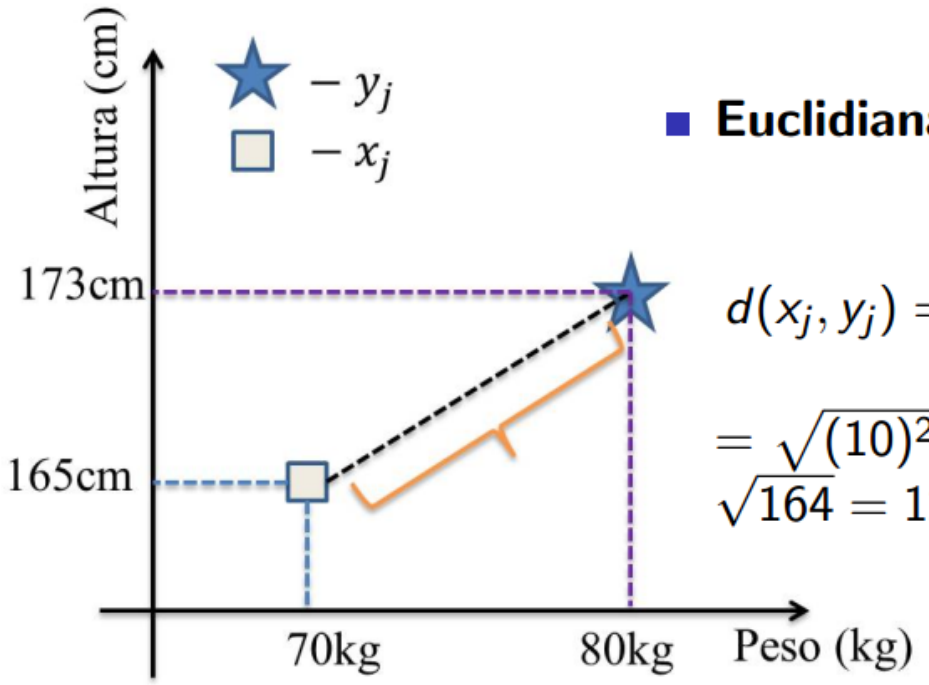
KNN

A classe do novo padrão é igual ao da K maioria mais próxima.

Pendências:

- Qual tipo de distância usar?
- Qual valor de K?
- Como Desempatar?

DISTÂNCIA



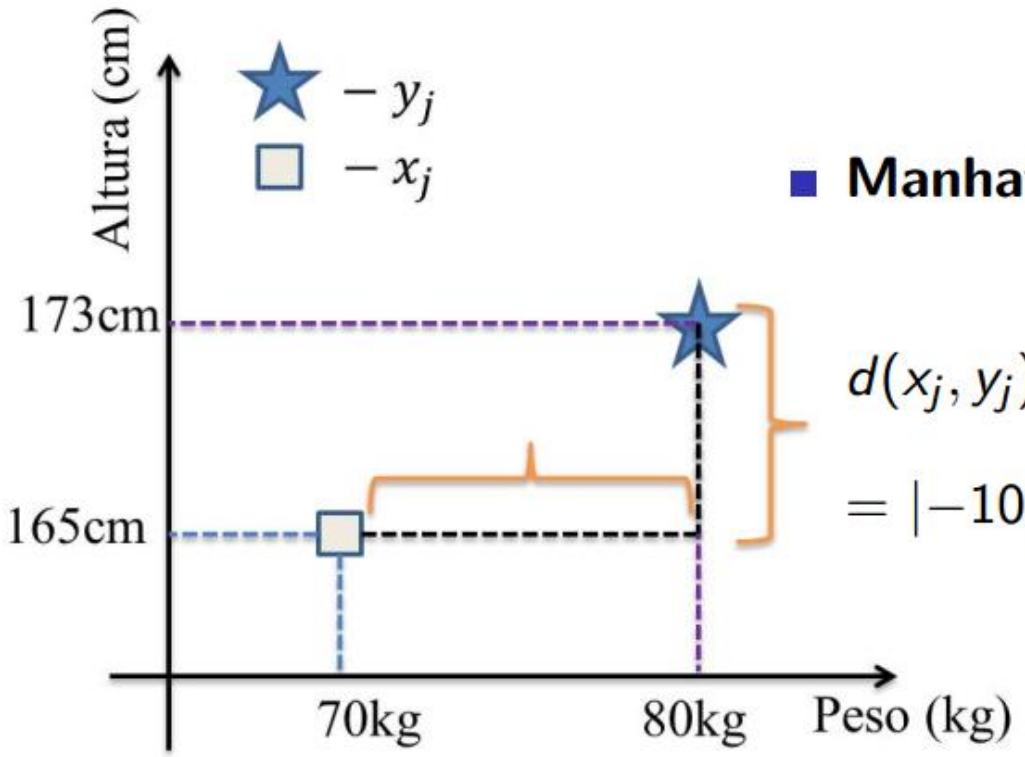
- **Euclidiana:**

$$d(x_j, y_j) = \sqrt{(70-80)^2 + (165-173)^2} =$$

$$= \sqrt{(10)^2 + (8)^2} = \sqrt{100+64} =$$

$$\sqrt{164} = 12.81$$

DISTÂNCIA

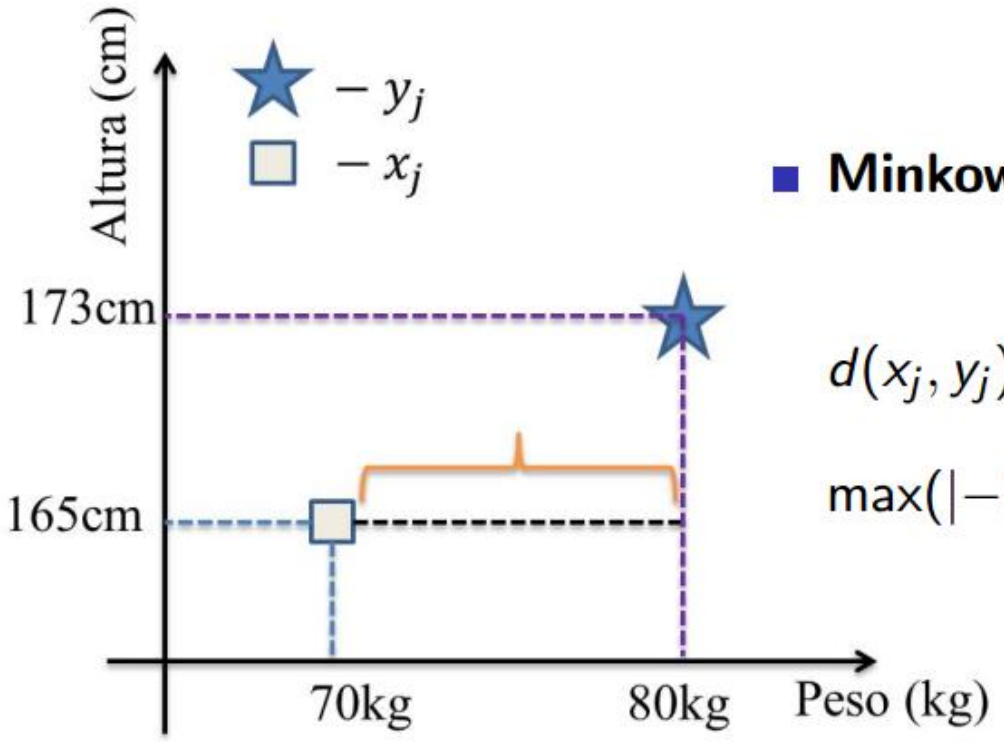


- **Manhattan:**

$$d(x_j, y_j) = |70 - 80| + |165 - 173| =$$

$$= |-10| + |-8| = 10 + 8 = 18$$

DISTÂNCIA



- **Minkowski:**

$$d(x_j, y_j) = \max(|70-80|, |165-173|) =$$

$$\max(|-10|, |-8|) = \max(10, 8) = 10$$

KNN

A classe do novo padrão é igual ao da K maioria mais Próxima.

Pendências:

- Qual tipo de distância usar?
- Qual valor de K?
- Como Desempatar?

Escolha experimental

KNN

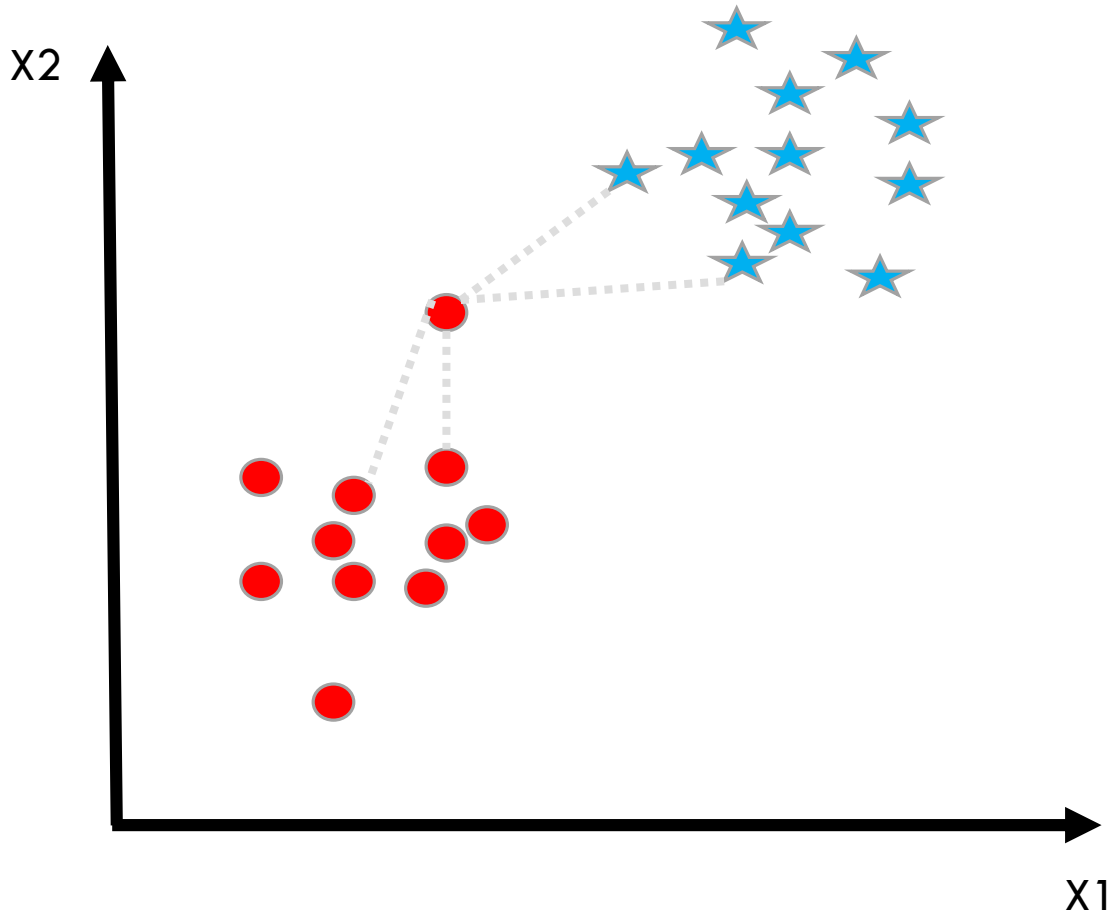
A classe do novo padrão é igual ao da K maioria mais Próxima.

Pendências:

- Qual tipo de distância usar?
- Qual valor de K?
- Como Desempatar?

Escolha aletória
Escolha aletória ponderada
Classe mais próxima

DESEMPATE

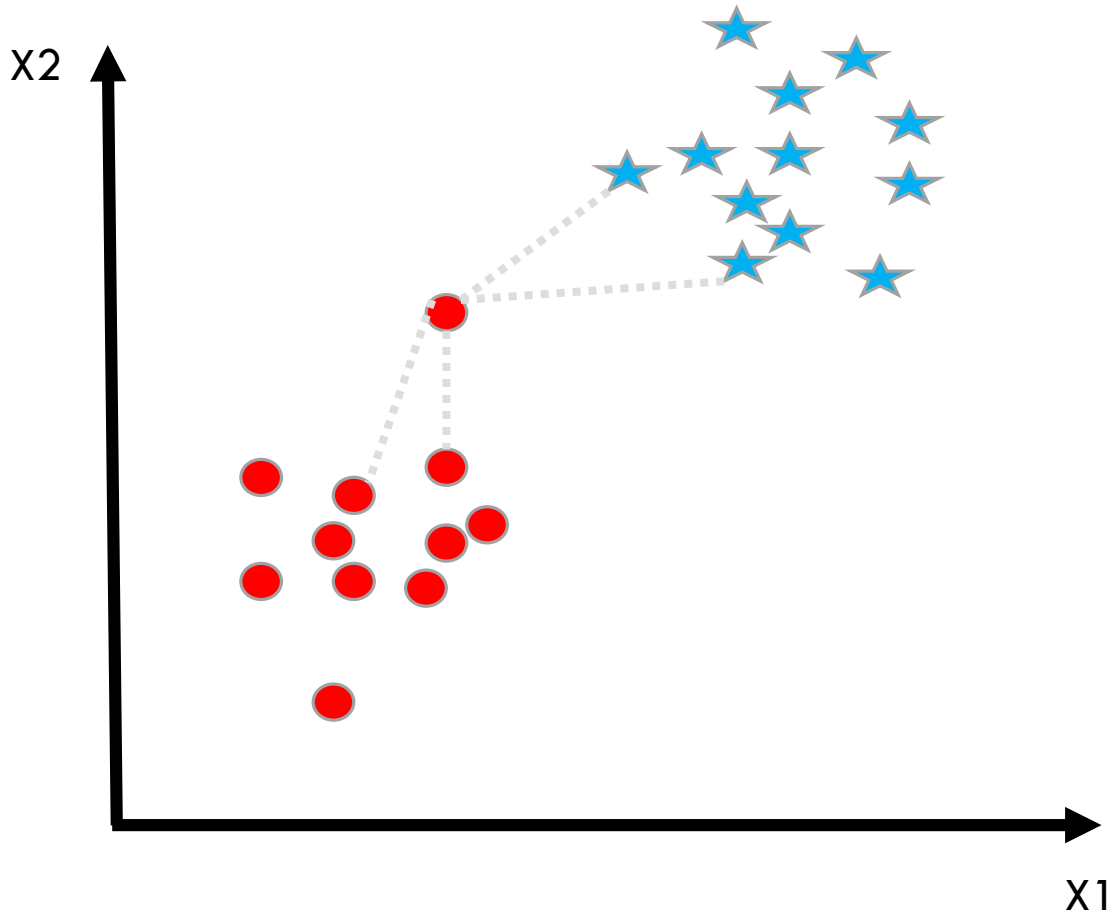


Supondo K = 4

Escolha Aleatória:

"Jogue uma moeda honesta": caso saia cara, escolha a classe vermelha, caso saia coroa, escolha a classe azul.

DESEMPATE

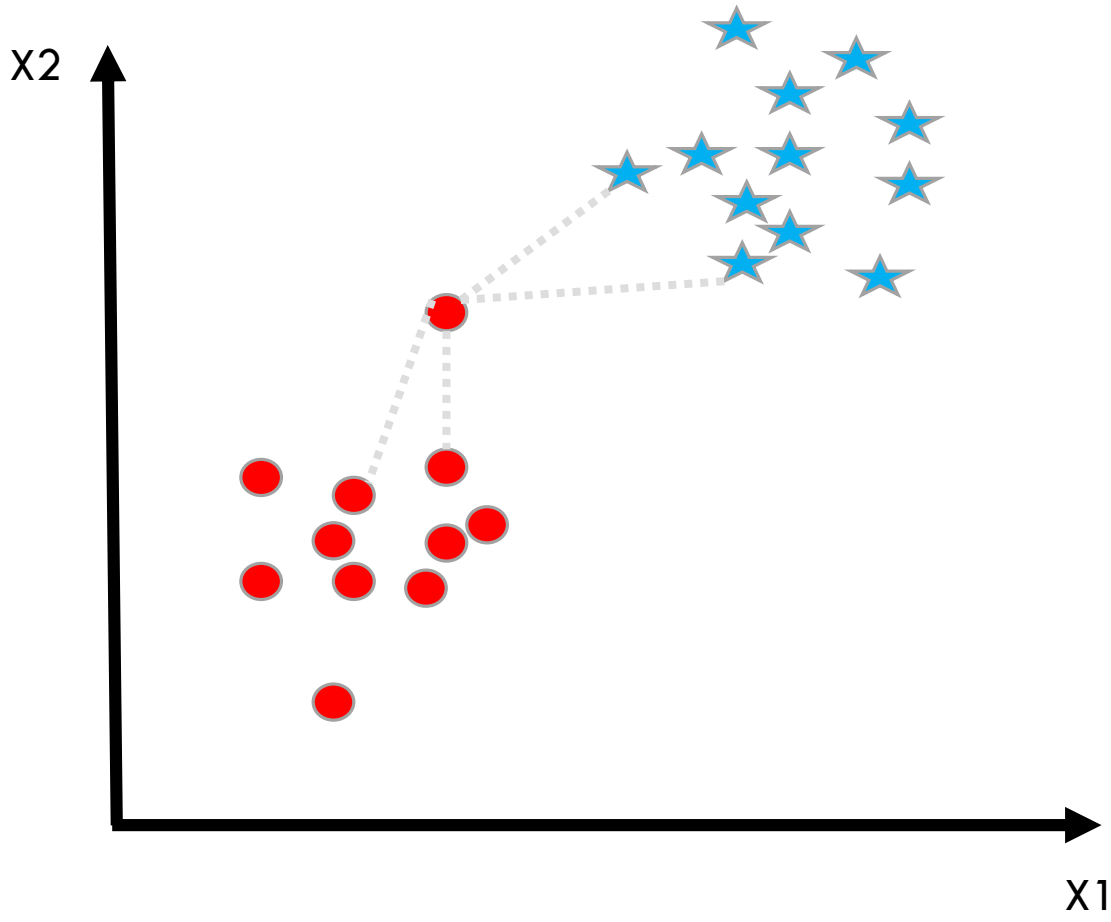


Supondo K = 4

Escolha Aleatória Ponderada:

"Jogue uma moeda desonesta": Dê mais chance à classe que está relacionada a classe que possua mais padrões.
Caso saia cara, escolha a classe vermelha, caso saia coroa, escolha a classe azul.

DESEMPATE



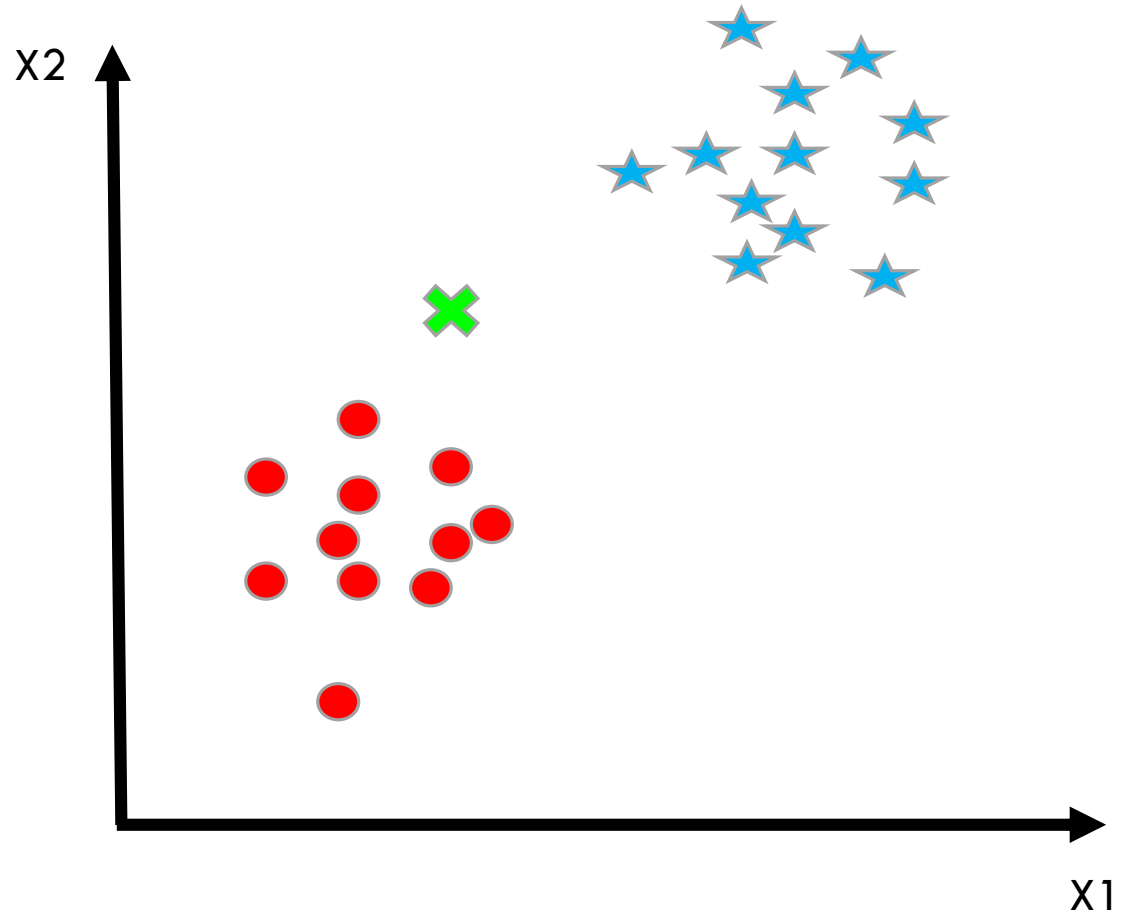
Supondo K = 4

Classe mais próxima:

Selecione a classe cuja distância é menor.

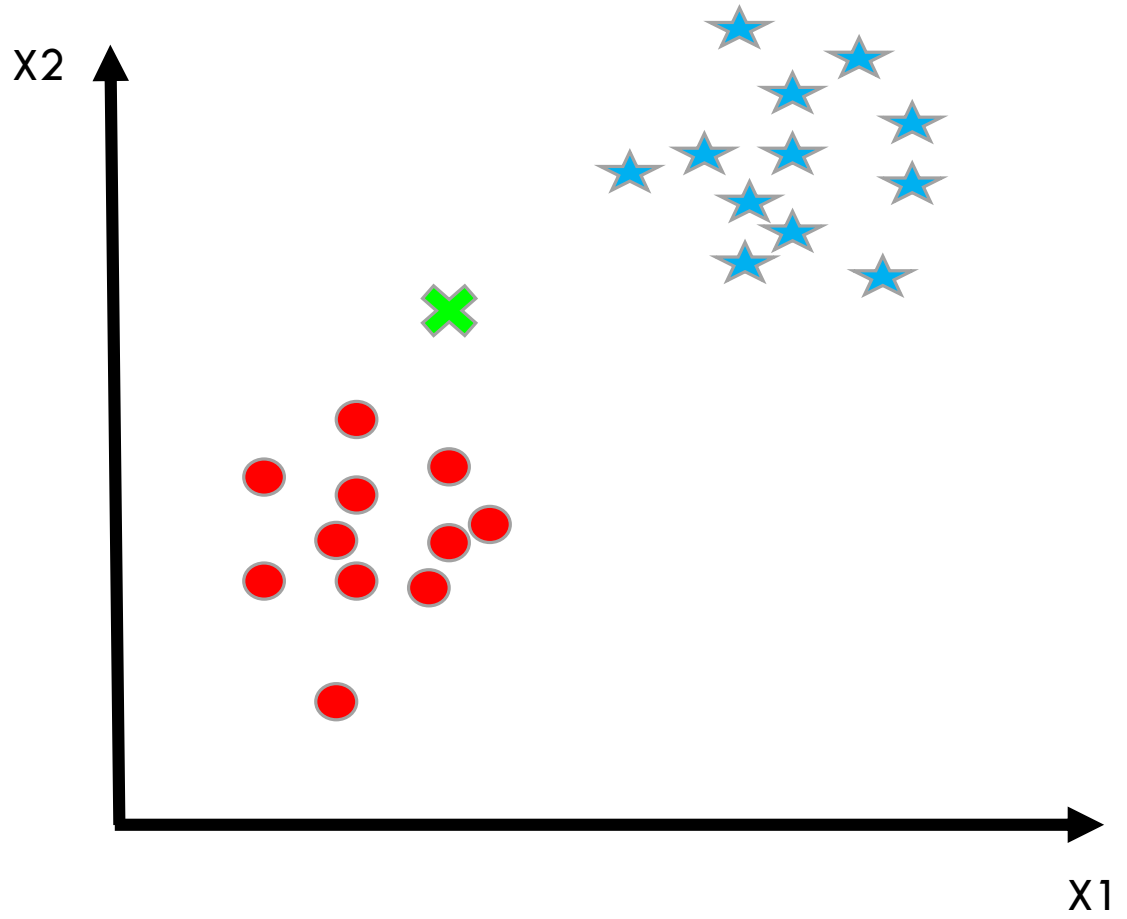
KNN

Passo 1: Determinar o valor de K, ou número de vizinhos **K = 5**



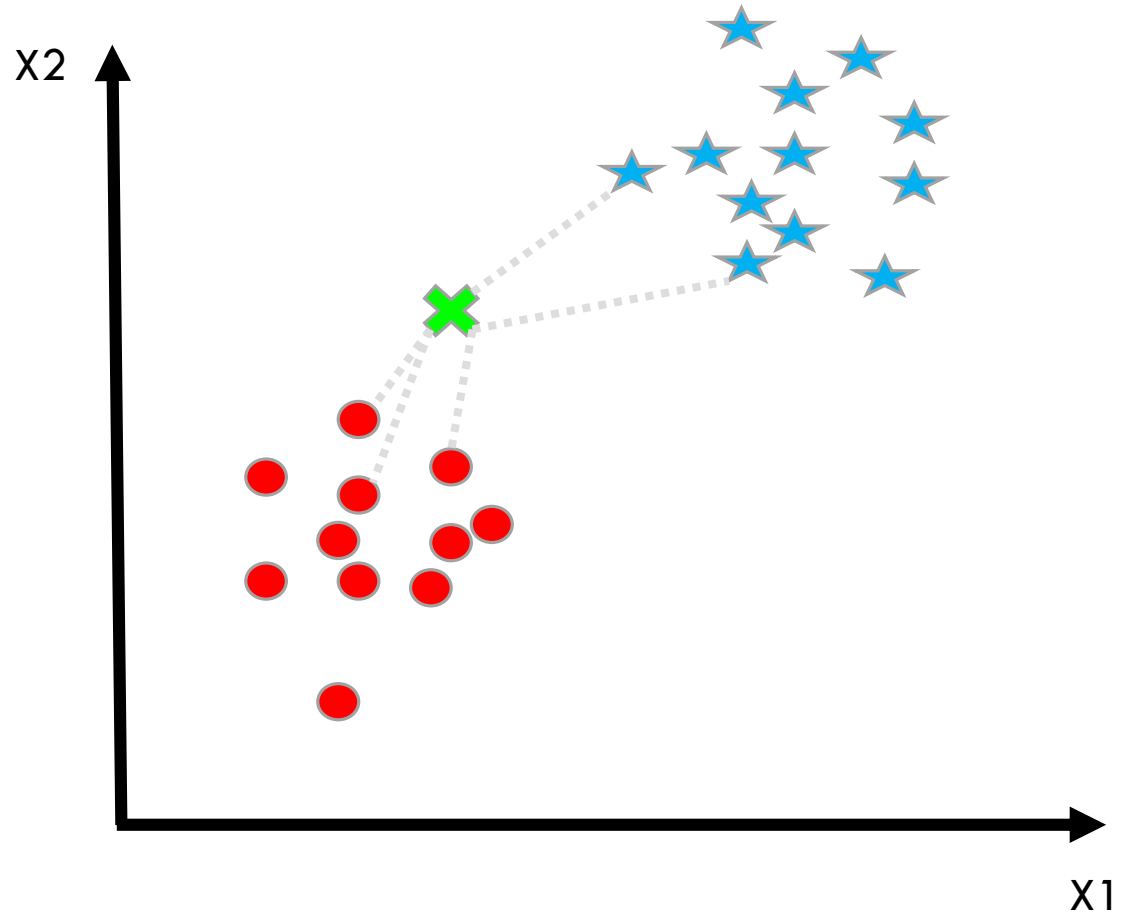
KNN

Passo 2: Calcular a distância entre cada par de registros



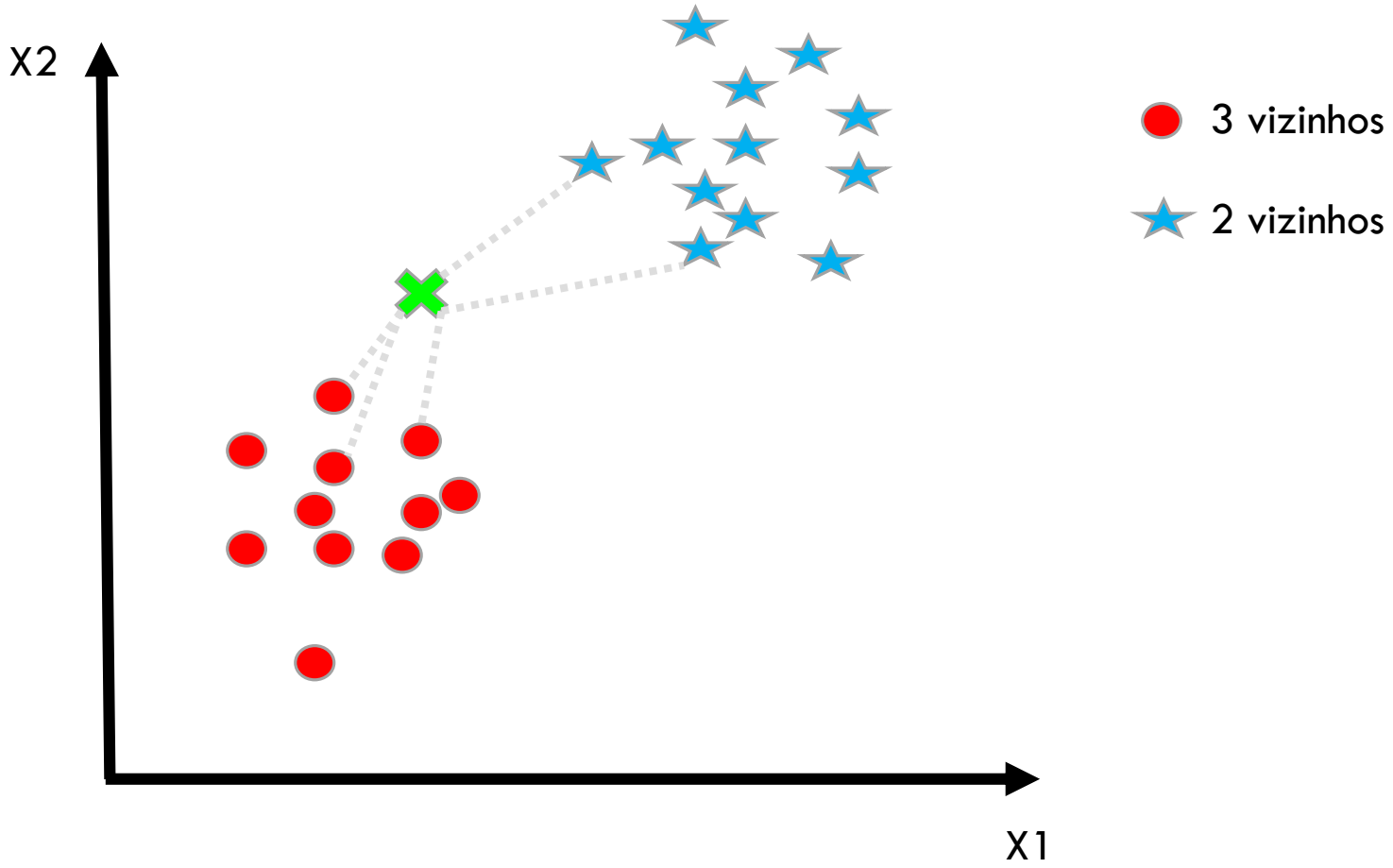
KNN

Passo 3: Determinar quais são os 5 vizinhos mais próximos (distância euclidiana) do novo registro



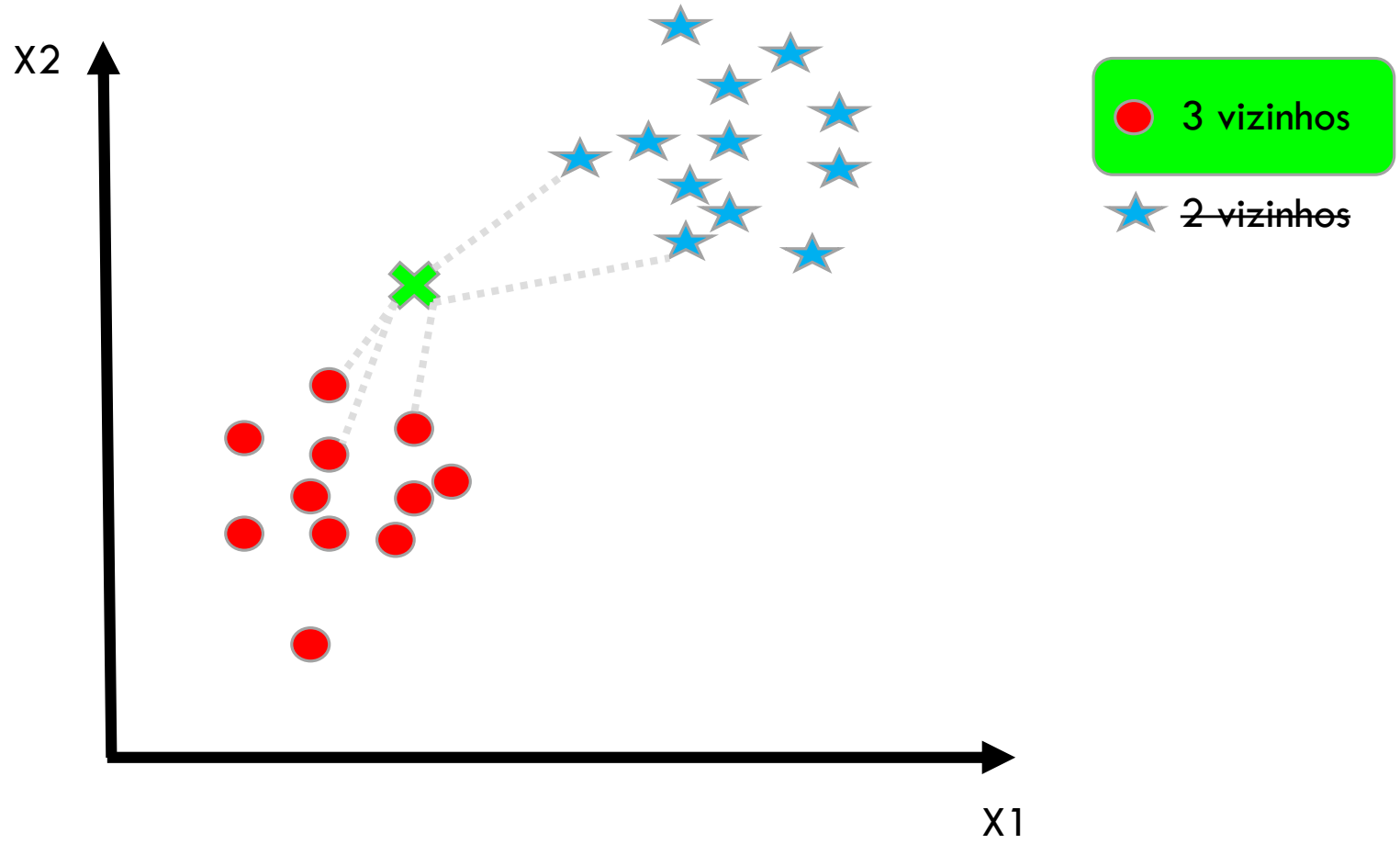
KNN

Passo 4: Dentre os 5 vizinhos, contar o número de vizinhos em cada classe



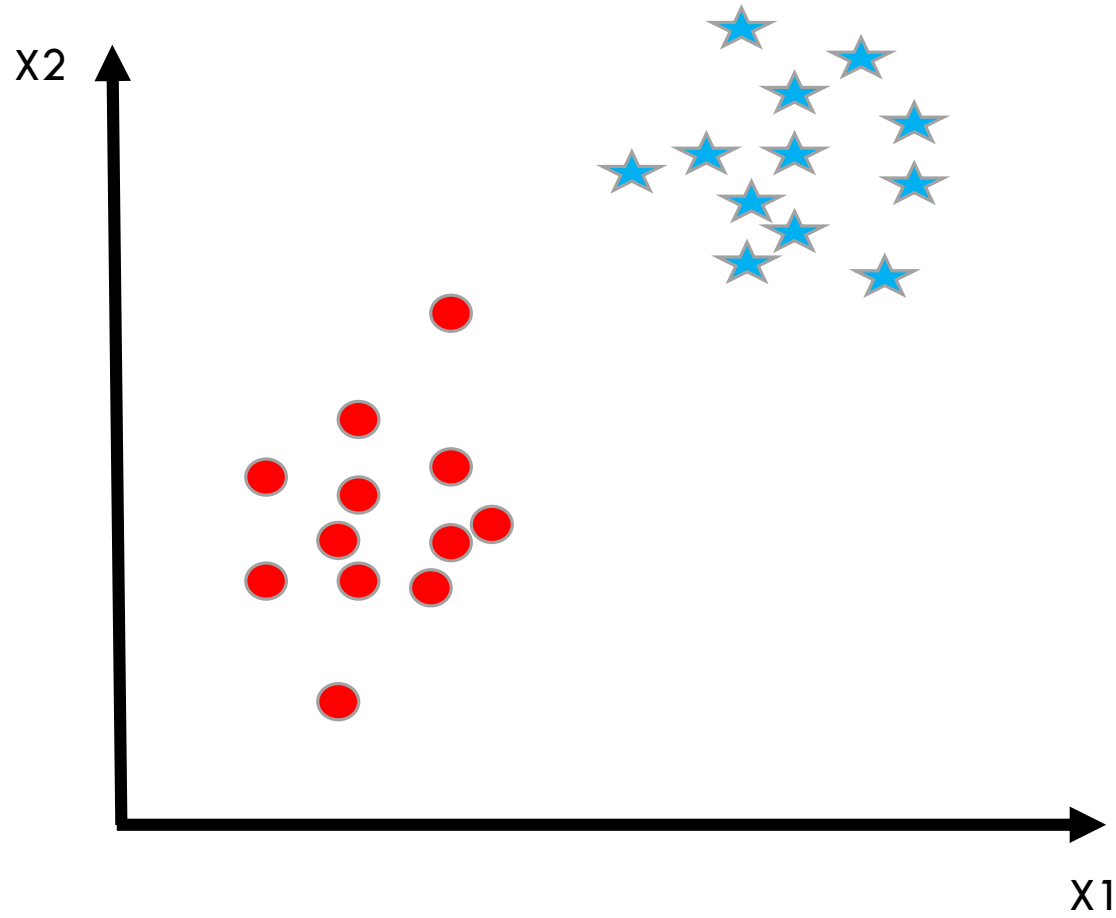
KNN

Passo 5: O novo registro vai ser da classe majoritária entre os vizinhos mais próximos



KNN

Passo 5: O novo registro vai ser da classe majoritária entre os vizinhos mais próximos



ESTUDOS DE CASO

ANÁLISE DE CRÉDITO BANCÁRIO

A base de dados contém **2077 exemplos** de créditos concedidos ou não.

Possui **11 atributos** de entrada e **2 classes de saída**.

A saída indica se o **cliente pagou** o empréstimo (=1) ou **se não pagou** (=0).

Master

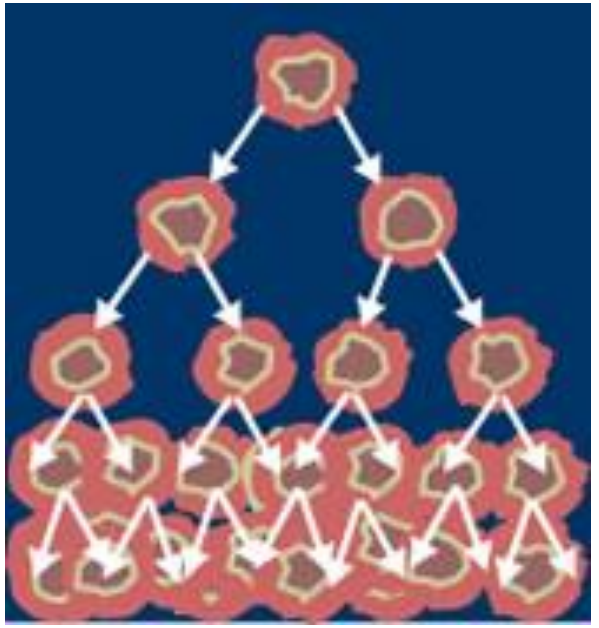
ANÁLISE DE CRÉDITO BANCÁRIO

	Nome das Variáveis	Descrição	Tipo	Valores possíveis
1	ESTC	**Est**ado **c**ivil	Categórica	0,1,2,3
2	NDEP	**N**úmero de **dep**endentes	Categórica	0,1,2,3,4,5,6,7
3	RENDA	**Renda** Familiar	Numérica	300-9675
4	TIPOR	**Tipo** de **r**esidência	Categórica	0,1
5	VBEM	**V**alor do **bem** a ser adquirido	Numérica	300-6000
6	NPARC	**N**úmero de **parc**elas	Numérica	1-24
7	VPARC	**V**alor da **parc**ela	Numérica	50-719
8	TEL	Se o cliente possui **tel**efone	Categórica	0,1
9	IDADE	**Idade** do cliente	Numérica	18-70
10	RESMS	Tempo de moradia (**Res**idência) (em **m**eses)	Numérica	0-420
11	ENTRADA	Valor da **entrada**	Numérica	0-1300
=	CLASSE	=1 se o cliente pagou a dívida	Categórica	0,1

Master

CÂNCER DE MAMA

University of Wisconsin, Clinical Sciences Center



30 atributos + classe + id:

- Raio: distância media do centro à pontos no perímetro do tumor;
- Textura: desvio padrão dos valores em escala de cinza;
- Perímetro;
- Área;
- Etc...

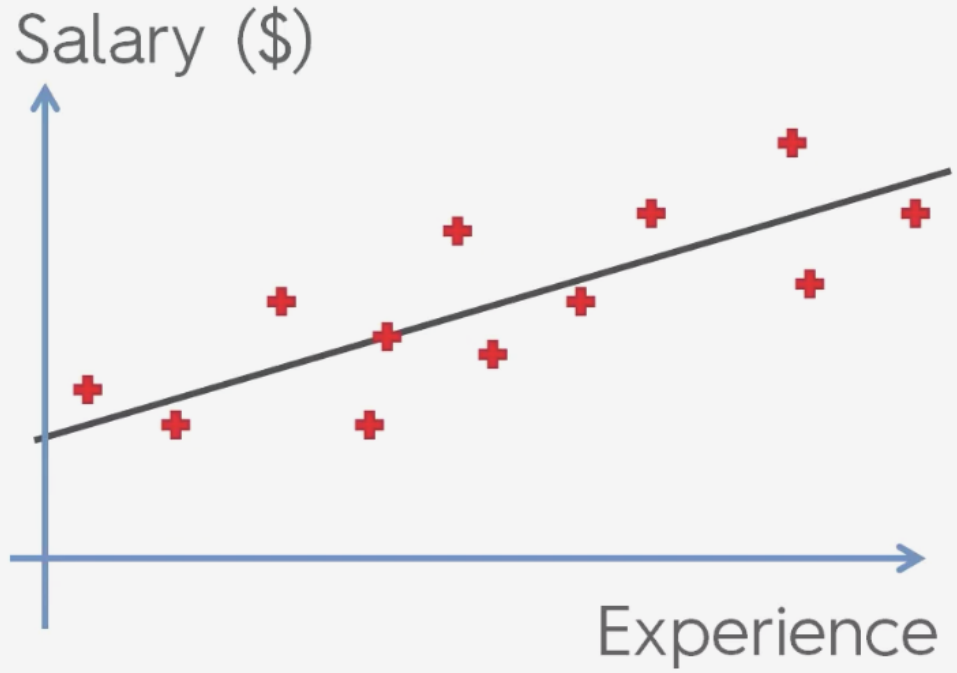
569 instâncias: 357 Benignas / 212 Malignas

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/home

REGRESSÃO LOGÍSTICA

REGRESSÃO LINEAR SIMPLES

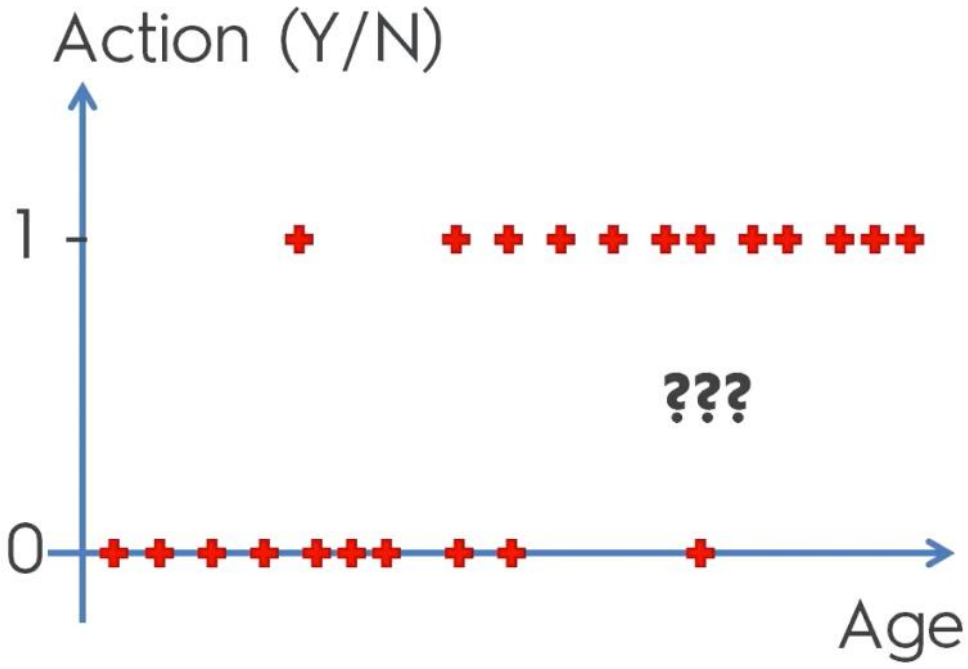
Simple Linear Regression:

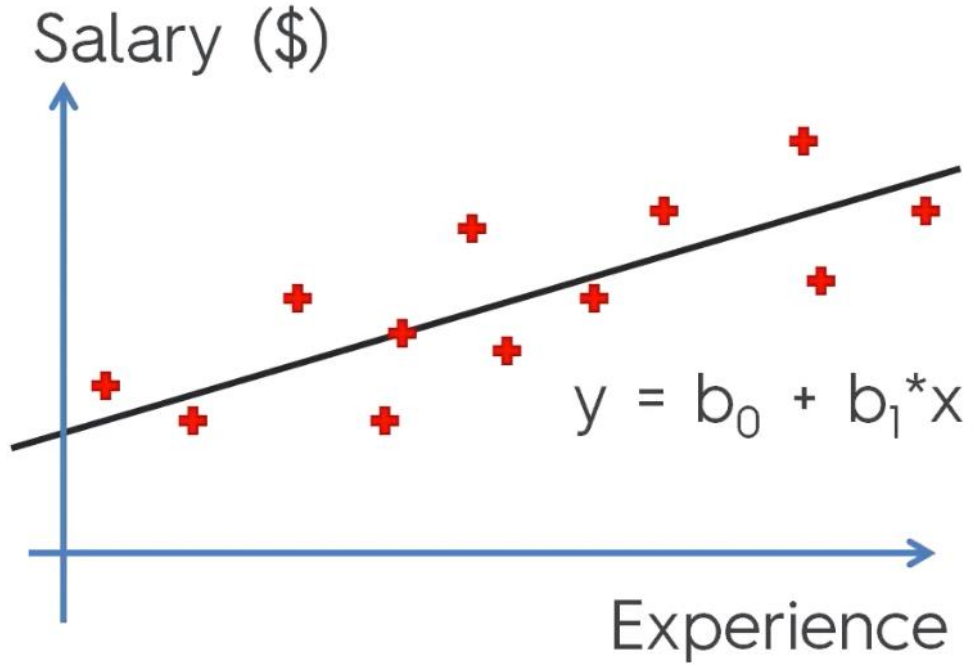


$y = b_0 + b_1 * x$

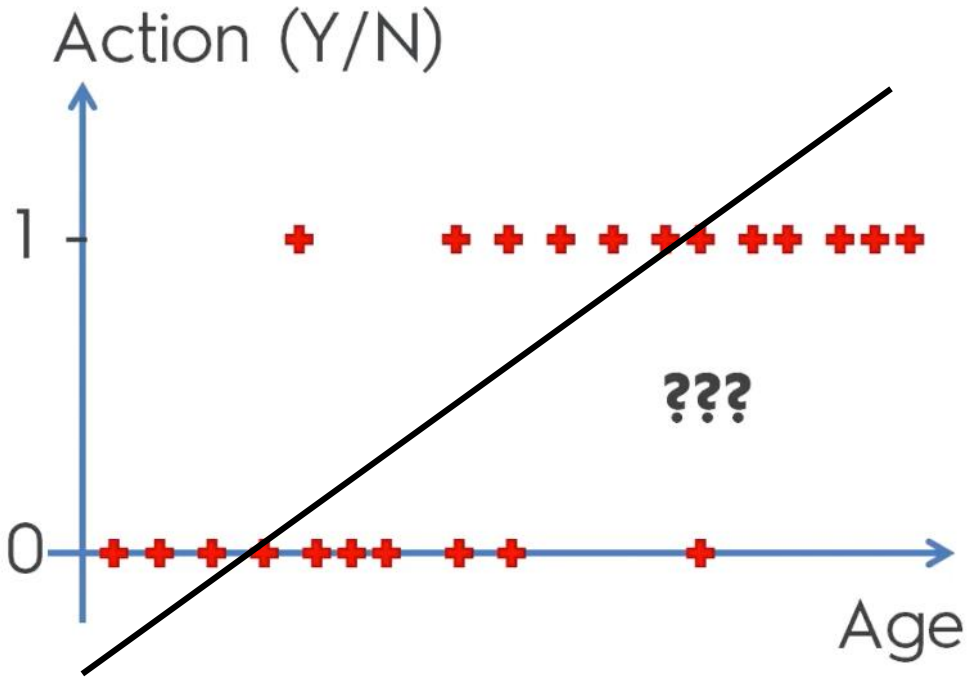
$\text{Salary} = b_0 + b_1 * \text{Experience}$

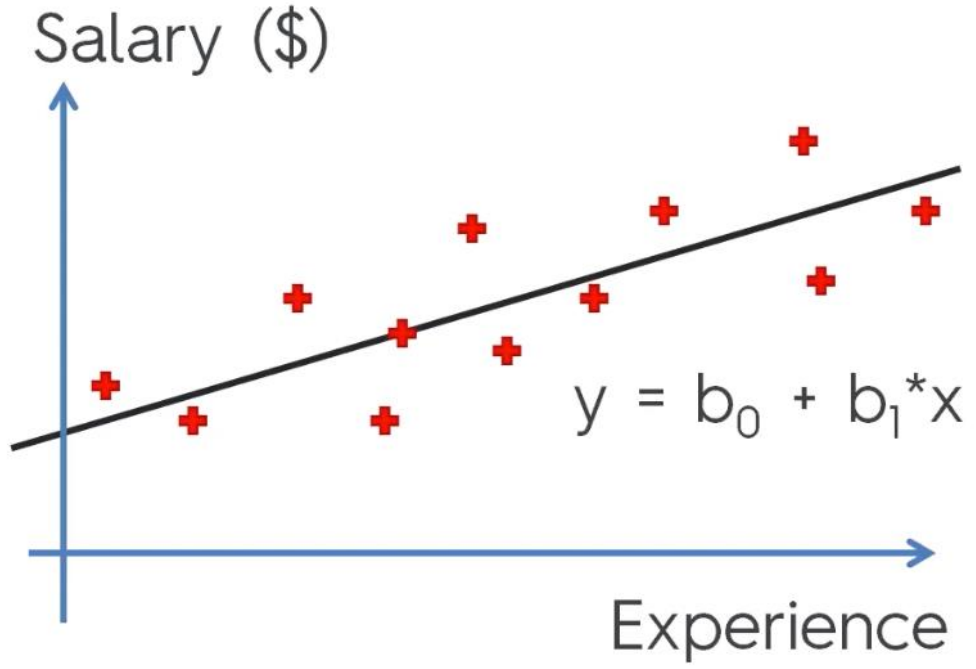
REGRESSÃO LOGÍSTICA



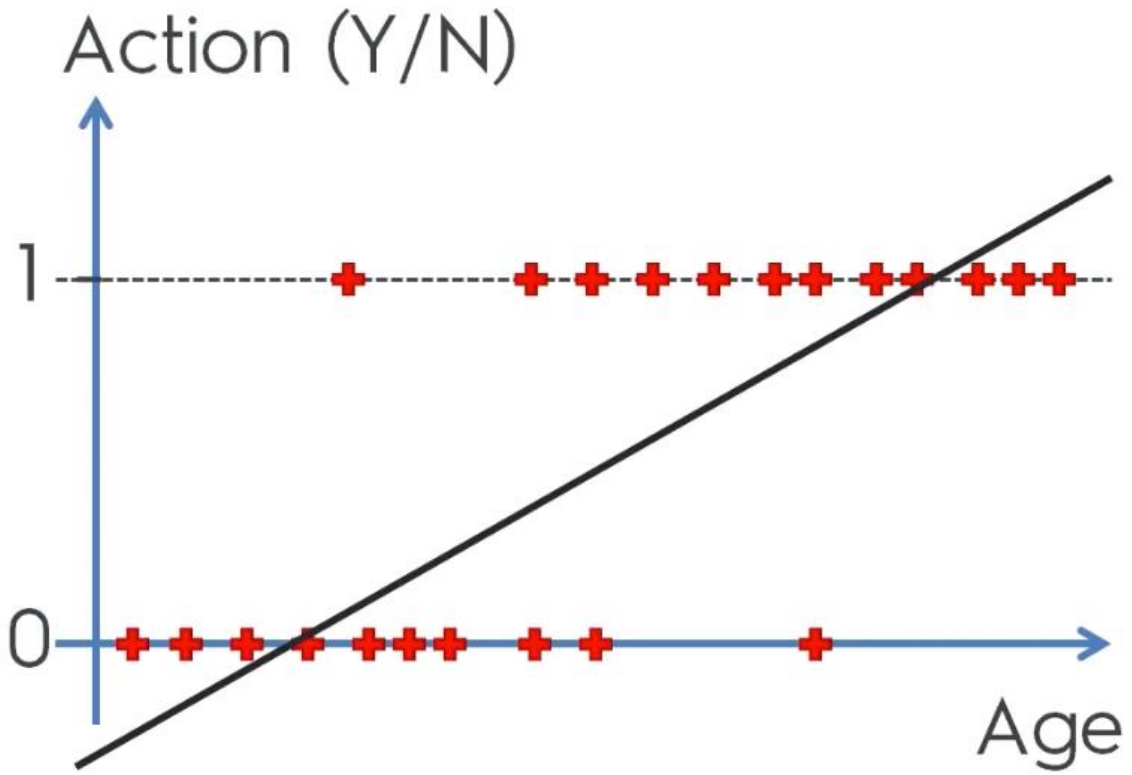


REGRESSÃO LOGÍSTICA

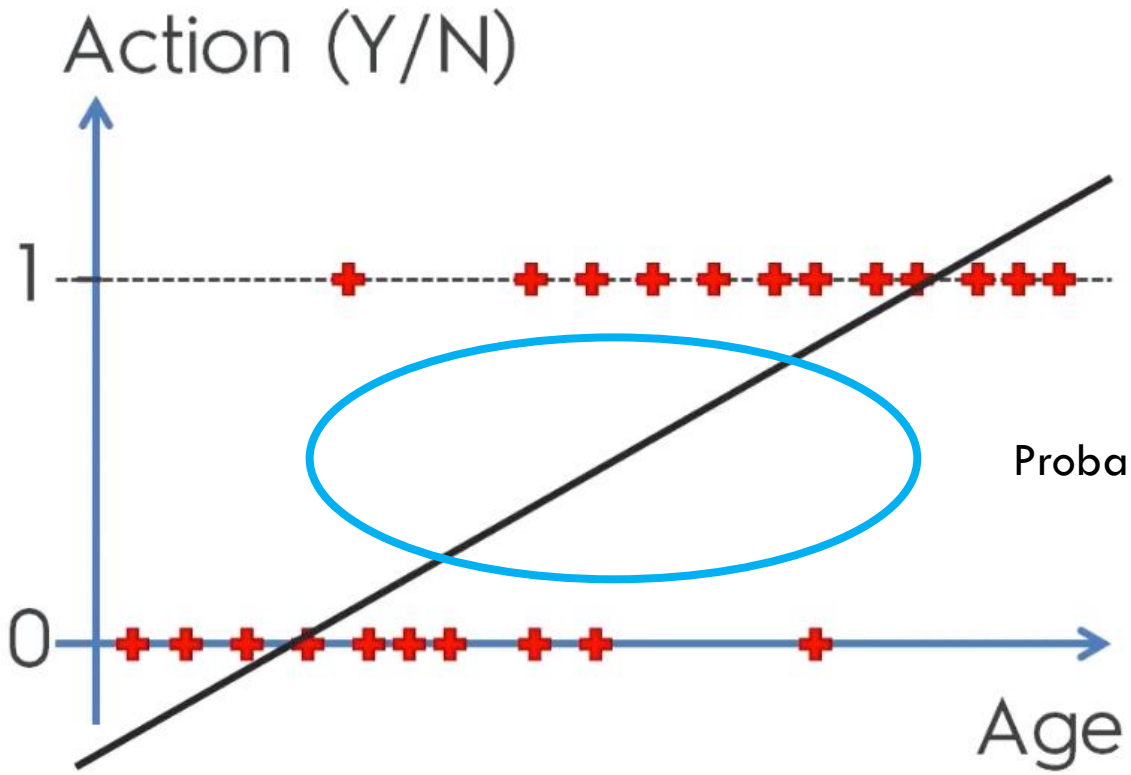




REGRESSÃO LOGÍSTICA



REGRESSÃO LOGÍSTICA

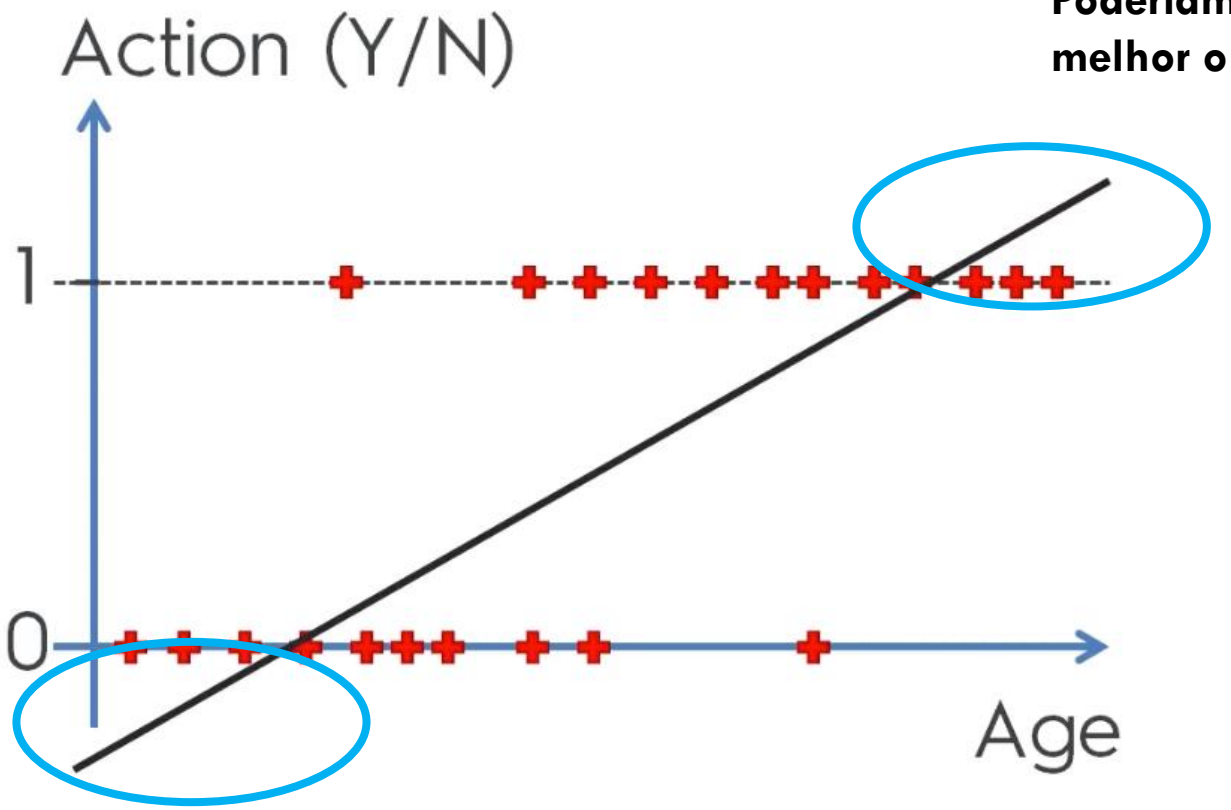


Probabilidades?

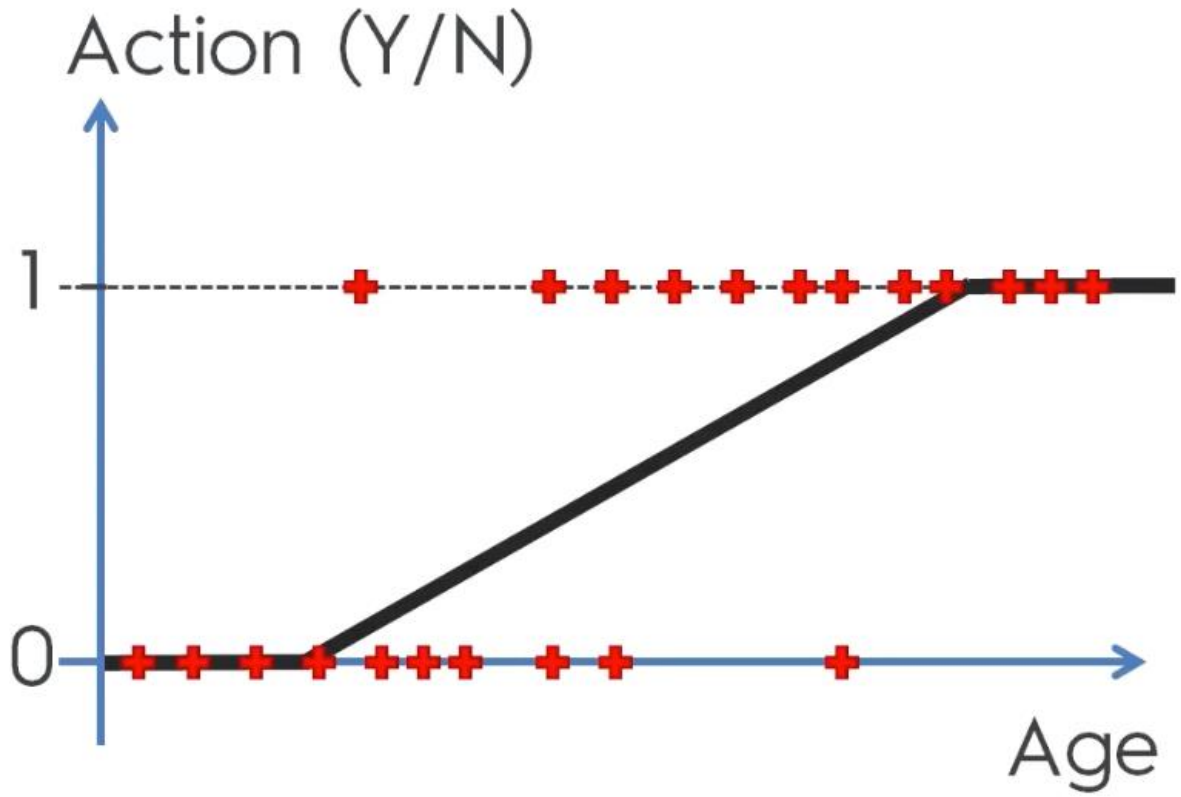
REGRESSÃO LOGÍSTICA

Pensando em probabilidades, esses 'pedaços' não fazem mais sentido.

Poderíamos alterar para modelar melhor o problema.



REGRESSÃO LOGÍSTICA



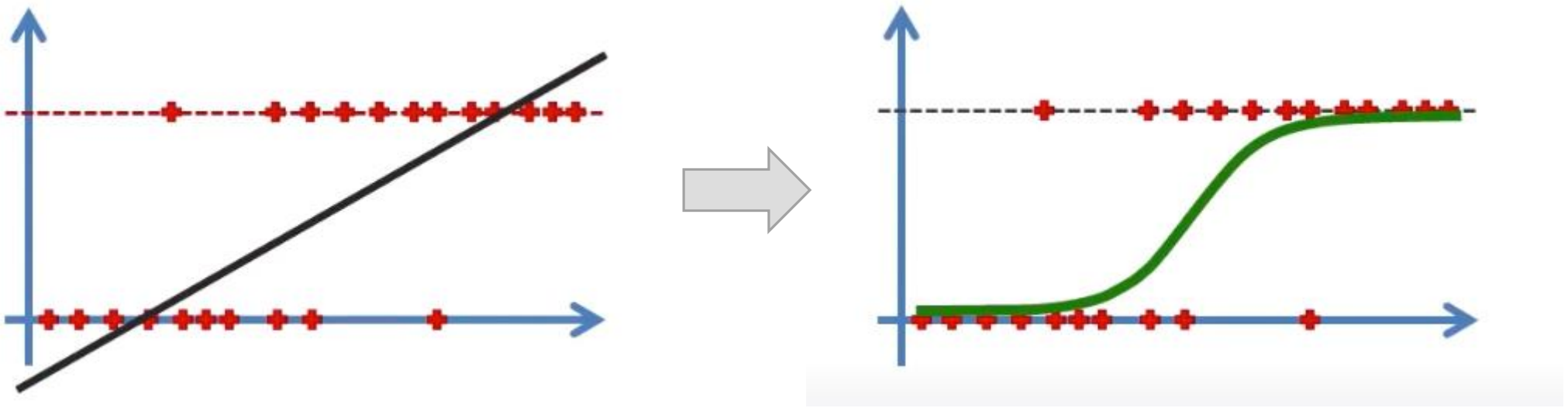
REGRESSÃO LOGÍSTICA

Regressão Linear ➔ Aplica sigmoidal ➔ Regressão Logística

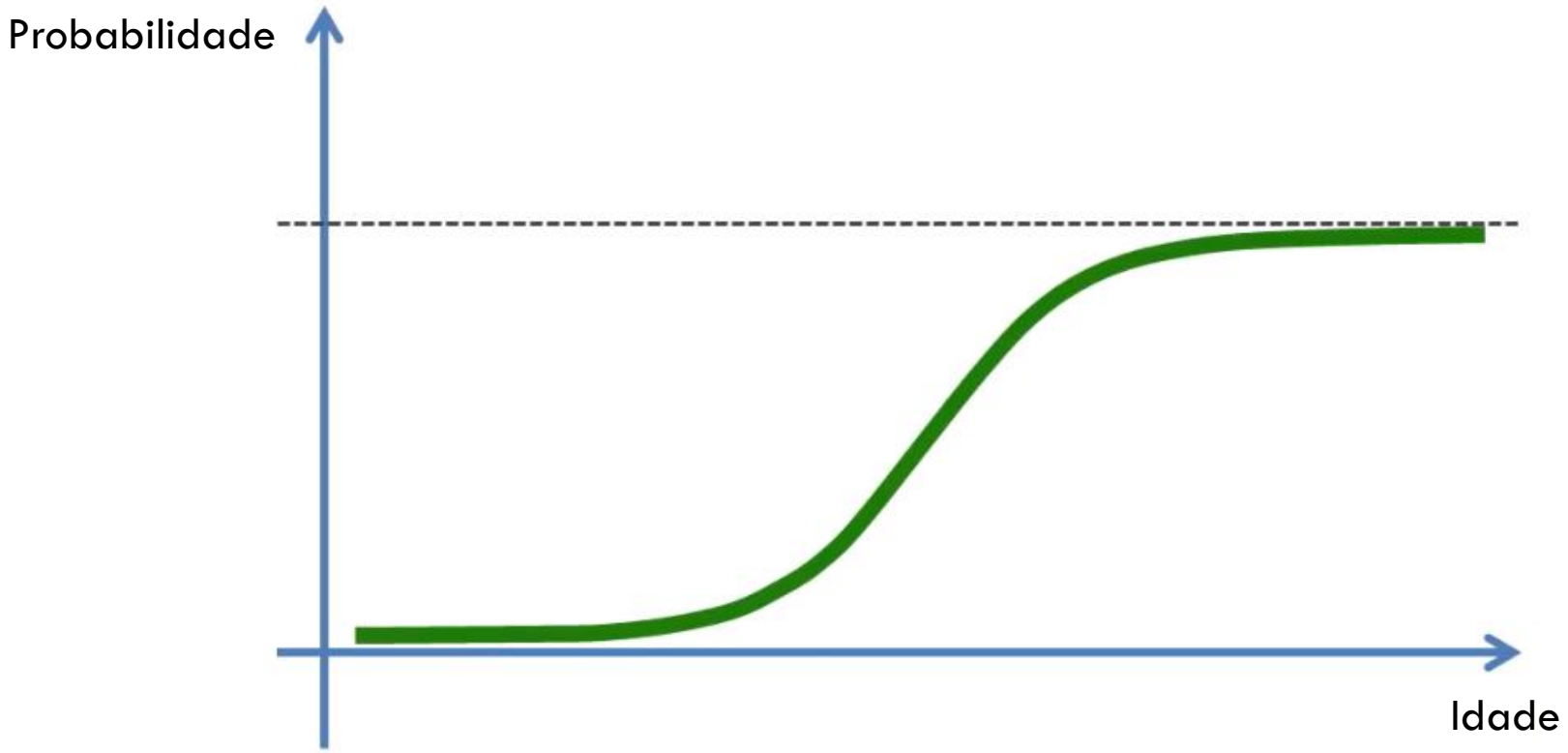
$$y = b_0 + b_1 x_1 \Rightarrow p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1$$

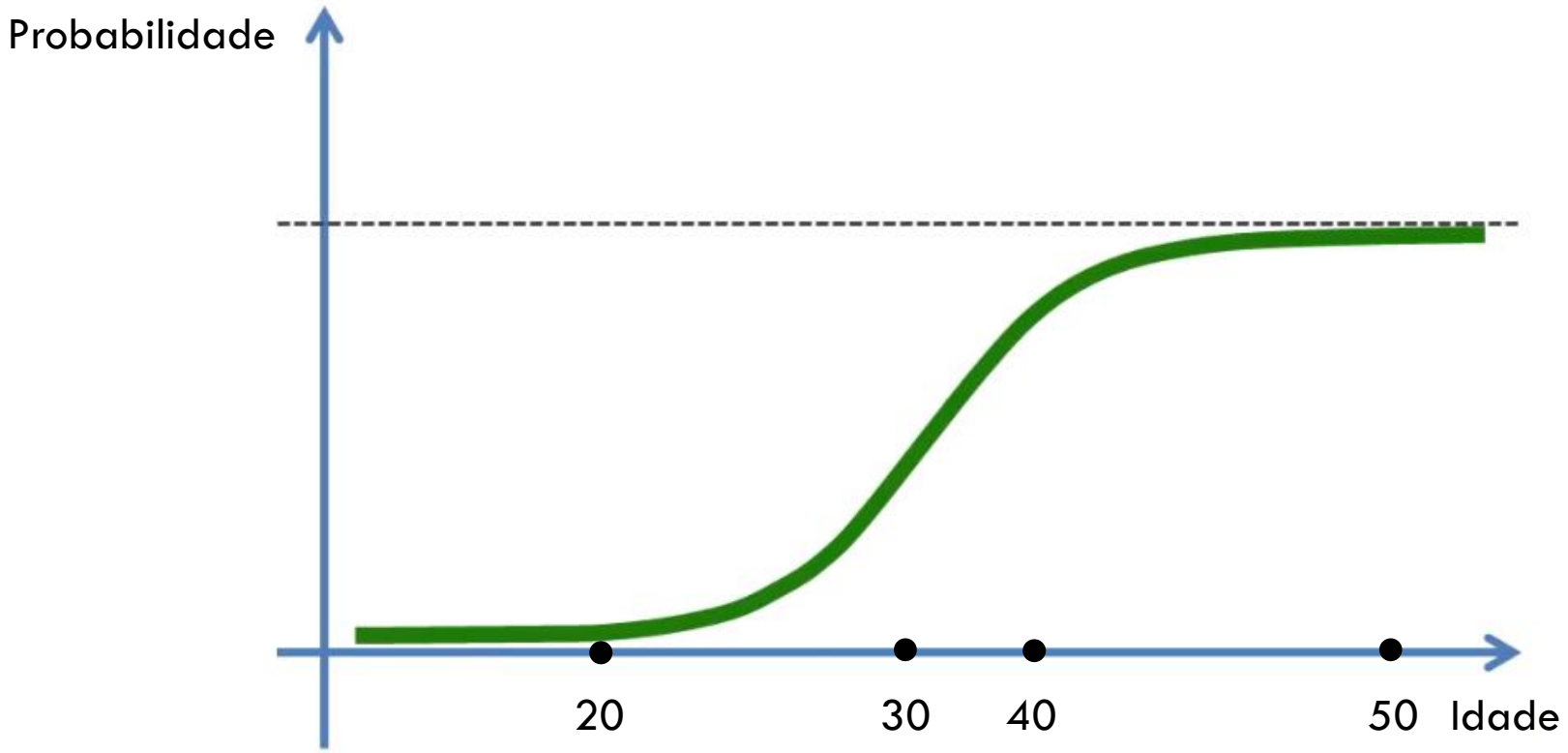
REGRESSÃO LOGÍSTICA



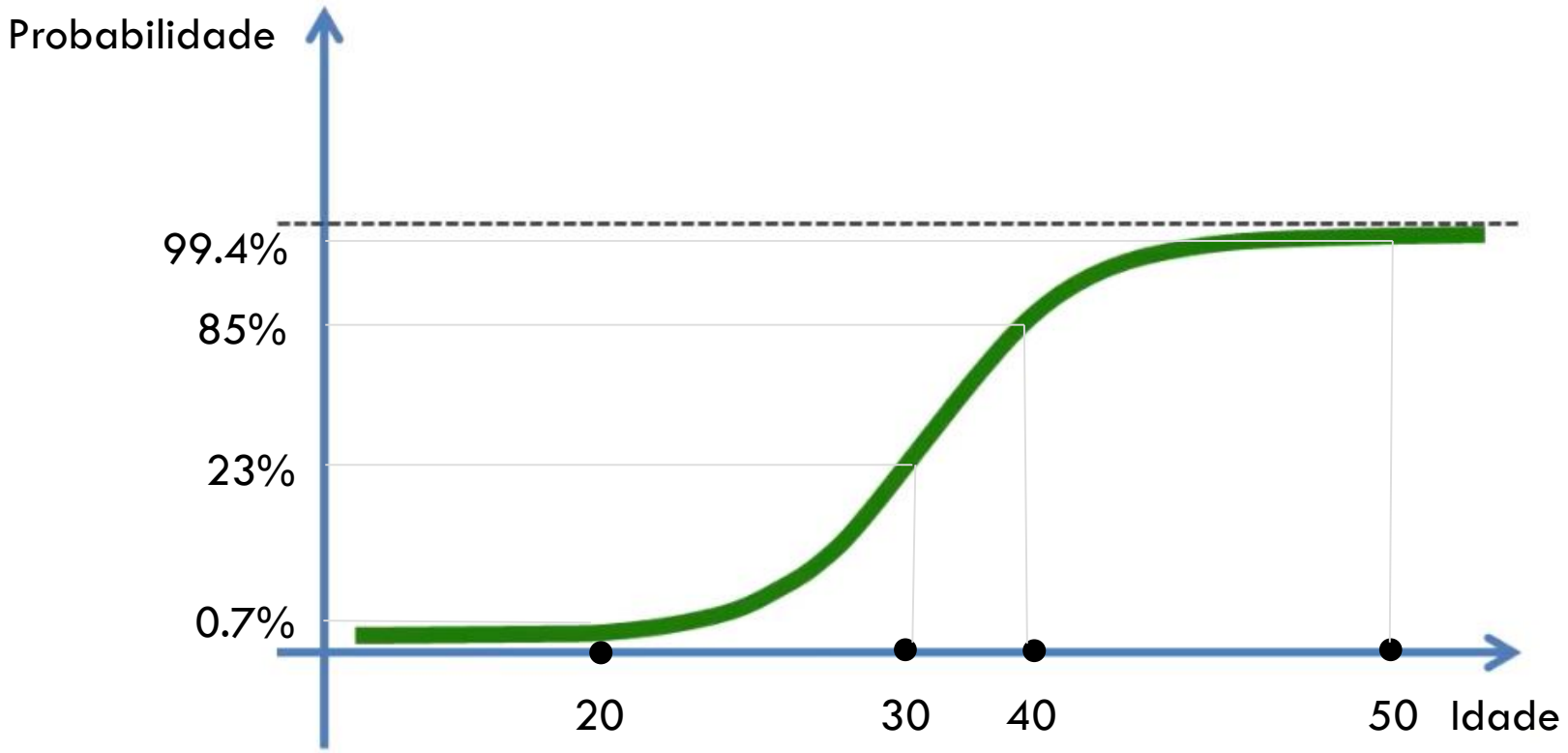
REGRESSÃO LOGÍSTICA



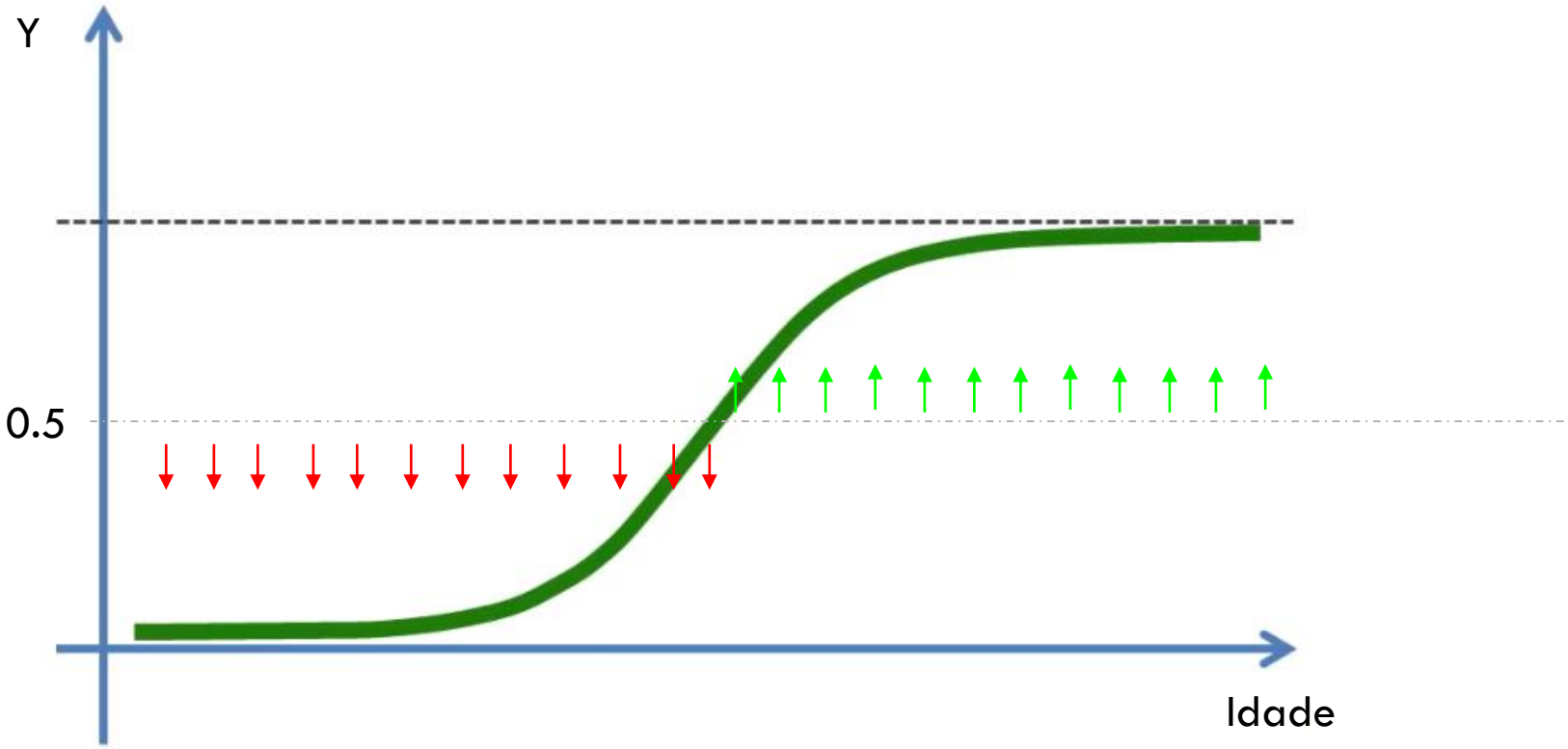
REGRESSÃO LOGÍSTICA



REGRESSÃO LOGÍSTICA



REGRESSÃO LOGÍSTICA - INFERÊNCIA



ESTUDOS DE CASO

Análise de Crédito Bancário

ANÚNCIOS EM REDES SOCIAIS

400 registros para inferir se determinado cliente vai ou não comprar o produto anunciado.

- Gênero
- Idade
- Salário Estimado

Trabalho