



**Instituto de Informática**



**Especialização**

Big Data Analítica

**Disciplina**

Mineração de Dados

**Classificação de dados**

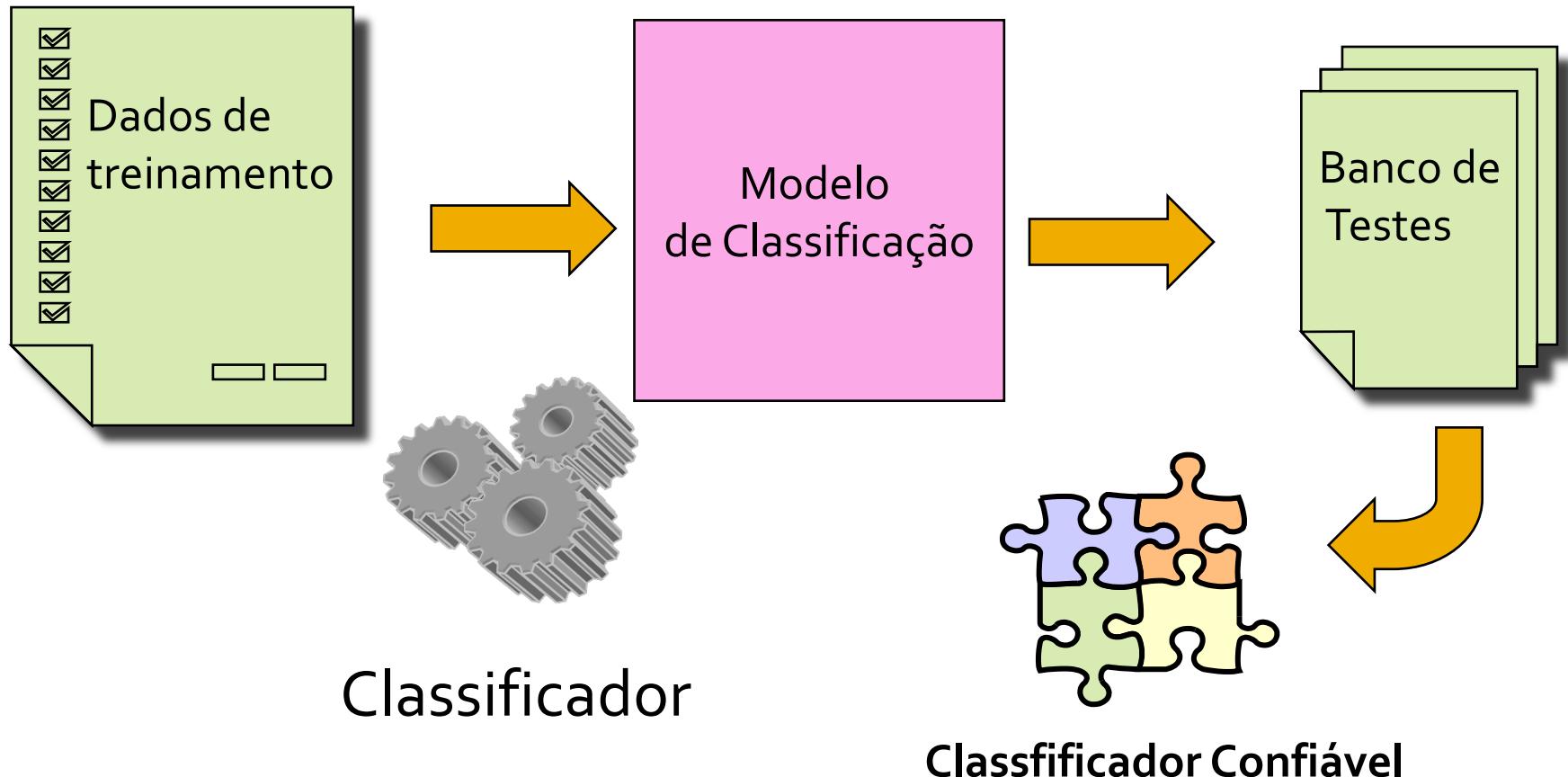
Nádia Félix Felipe da Silva, Dra.

# Classificação

Nome	Idade	Renda	Profissão	Classe
Daniel	$\leq 30$	Média	Estudante	Sim
João	31..50	Média-Alta	Professor	Sim
Carlos	31..50	Média-Alta	Engenheiro	Sim
Maria	31..50	Baixa	Vendedora	Não
Paulo	$\leq 30$	Baixa	Porteiro	Não
Otavio	$> 60$	Média-Alta	Aposentado	Não

**SE Idade  $\leq 30$  E Renda é Média ENTÃO Compra-Video-Game = SIM.**

# Etapas do Processo



# Métodos de Classificação

## ■ **Classificadores eager (espertos)**

A partir da amostragem, constroem um modelo de classificação capaz de classificar novas tuplas. Uma vez pronto o modelo, as amostras não são mais utilizadas na classificação de novos objetos (tuplas)

- Arvores de Decisão
- Redes Bayseanas
- Máquinas de Suporte Vetorial

## ■ **Classificadores lazy (preguiçosos)**

Cada nova tupla é comparada com todas as amostras e é classificada segundo a classe da amostra à qual é mais similar.

- Método kNN (k-nearest-neighbor)

## ■ **Outros Métodos**

- Algoritmos Genéticos
- Conjuntos Difusos

# Critérios de Comparação dos Métodos

- **Acurácia** – capacidade de classificar corretamente novas tuplas
- **Rapidez** – tempo gasto na classificacao
- **Robustez** – habilidade de classificar corretamente em presença de ruidos e valores desconhecidos
- **Escalabilidade** – eficiência do classificador em grandes volumes de dados
- **Interpretabilidade** – facilidade de um usuário entender as regras produzidas pelo classificador

# Acurácia – Taxa de erros

- $\text{Acc}(M)$  = porcentagem das tuplas dos dados de teste que são corretamente classificadas.
- $\text{Err}(M) = 1 - \text{Acc}(M)$
- Matriz de Confusão

		Classes Preditas	
		C1	C2
Classes Reais	C1	Positivos verdadeiros	Falsos Negativos
	C2	Falsos Positivos	Negativos verdadeiros

# Outras medidas mais precisas

- Exemplo :  $\text{acc}(M) = 90\%$

$C_1$  = tem-câncer (4 pacientes)

$C_2$  = não-tem-câncer (500 pacientes)

Classificou corretamente 454 pacientes que não tem câncer

Não acertou nenhum dos que tem câncer

Pode ser classificado como “bom classificador”  
mesmo com acurácia alta ?

- **Sensitividade** =  $\frac{\text{true-pos}}{\text{pos}}$

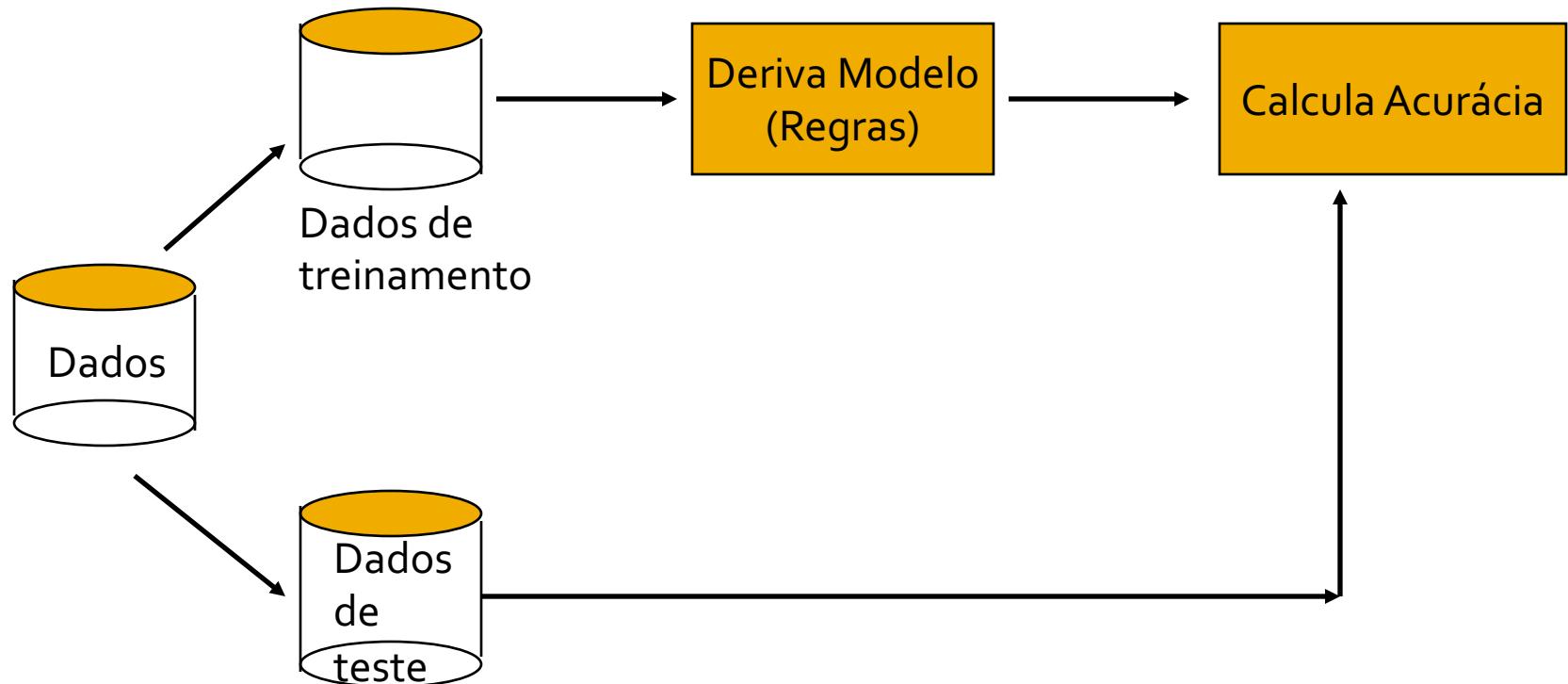
% pacientes classificados **corretamente com câncer** dentre todos os que **realmente tem câncer**

- **Especificidade** =  $\frac{\text{true-neg}}{\text{neg}}$

% pacientes classificados **corretamente com câncer** dentre todos os que foram classificados **com câncer**

- **Precisão** =  $\frac{\text{true-pos}}{\text{true-pos} + \text{falso-pos}}$

# Processo de Classificação



# Preparação dos Dados

- Limpeza dos dados : remove ruidos e resolve problemas de dados incompletos
- Análise de Relevância : elimina atributos irrevelantes para a classificação
- Transformação dos dados
  - Categorização
  - Generalização
    - Ex: Rua pode ser substituído por Cidade
  - Normalização : todos os valores dos atributos em  $[0,1]$

# Classificadores *Lazy*

*Aprendem a partir de seus vizinhos*

***Não constrói um modelo de classificação.***

*Para cada nova tupla que se quer classificar, o banco de dados de treinamento é analisado.*

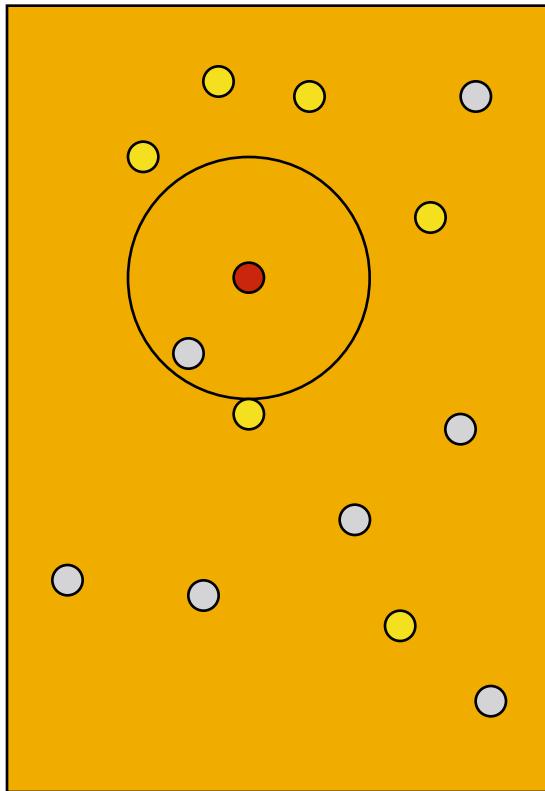
# História

- Método introduzido nos anos 50.
- Muito dispendioso computacionalmente.
- Só ganhou popularidade a partir dos anos 60, como o aumento da capacidade computacional dos computadores.
- Muito usado na área de *Reconhecimento de Padrões*.

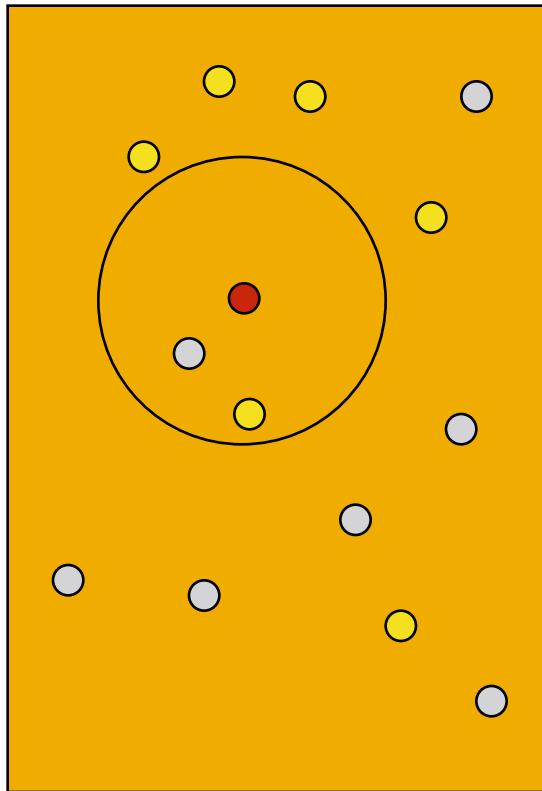
# Descrição do Método KNN

- Dados: Banco de Dados de m tuplas classificadas ( $a_1, \dots, a_n, C$ )
- Uma tupla  $X = (x_1, \dots, x_n)$  não classificada
- Os valores dos atributos são **normalizados**.  
Valor normalizado =  $(v.\text{real} - \text{MinA}) / (\text{MaxA} - \text{MinA})$
- Calcula-se a distância de  $X$  a cada uma das tuplas do banco de dados.
- Pega-se as  $k$  tuplas do banco de dados mais próximas de  $X$ .
- A classe de  $X$  é a classe que aparece com mais frequência entre as  $k$  tuplas mais próximas de  $X$ .

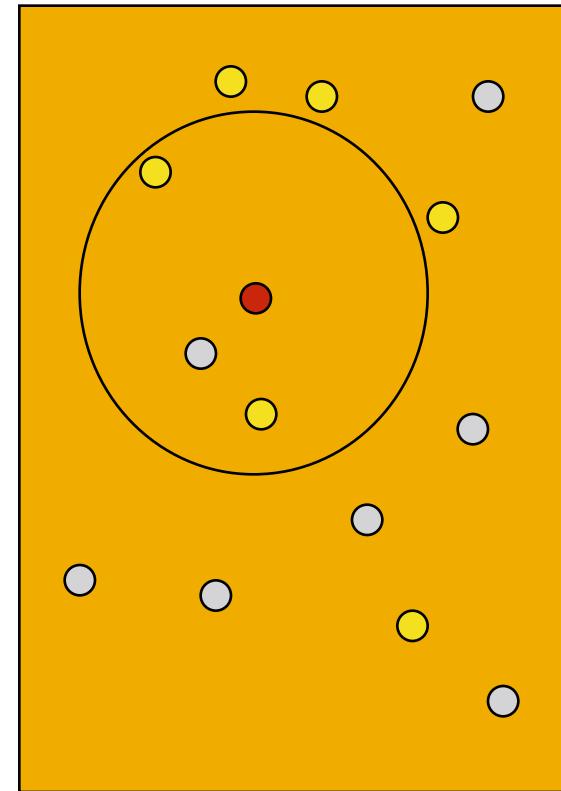
# Diferentes valores de K



$K = 1$



$K = 2$



$K = 3$

# Algumas questões

- Como calcular a distância entre duas tuplas ?
  - Para atributos contínuos : distância Euclidiana  
$$d(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$
  - Para atributos categóricos
    - Se  $x_i = y_i$  então  $x_i - y_i = 0$
    - Se  $x_i$  e  $y_i$  são distintos:  $x_i - y_i = 1$
- Como lidar com valores incompletos (ausentes) ao calcular a distância entre duas tuplas X e Y ?
  - Se  $x_i$  e  $y_i$  são ausentes:  $x_i - y_i = 1$
  - Se  $x_i$  é ausente e  $y_i$  não:  $x_i - y_i = \max \{ |1 - y_i|, |0 - y_i| \}$
- Como determinar o melhor valor de K (=número de vizinhos) ?  
Obtido repetindo-se os experimentos.

# Vantagens e Desvantagens

## ■ **Performance**

- Não constrói um modelo de classificação.
- Processo de classificação de uma tupla é lento.
- Classificadores *Eager* gastam tempo para construir o modelo. O processo de classificação de uma tupla é rápido.

## ■ **Sensível a ruídos**

- KNN faz predição baseando-se em informações locais à tupla sendo classificada.
- Arvores de decisão, redes neurais e bayesianas encontram modelo global que se leva em conta todo o banco de dados de treinamento.

# Exemplo

ID	IDADE	RENDA	ESTUDANTE	CREDITO	CLASSE
1	$\leq 30$	Alta	Não	Bom	Não
2	$\leq 30$	Alta	Sim	Bom	Não
3	31...40	Alta	Não	Bom	Sim
4	$> 40$	Média	Não	Bom	Sim
5	$> 40$	Baixa	Sim	Bom	Sim
6	$> 40$	Baixa	Sim	Excelente	Não
7	31...40	Baixa	Sim	Excelente	Sim
8	$\leq 30$	Média	Não	Bom	Não
9	$\leq 30$	Baixa	Sim	Bom	Sim
10	$> 40$	Média	Sim	Bom	Sim
11	$\leq 30$	Média	Sim	Excelente	Sim
12	31...40	Média	Não	Excelente	Sim
13	31...40	Alta	Sim	Bom	Sim
14	$> 40$	Média	Não	Excelente	Não

$$X = (\leq 30, \text{Média}, \text{Sim}, \text{Bom})$$

# Exemplo

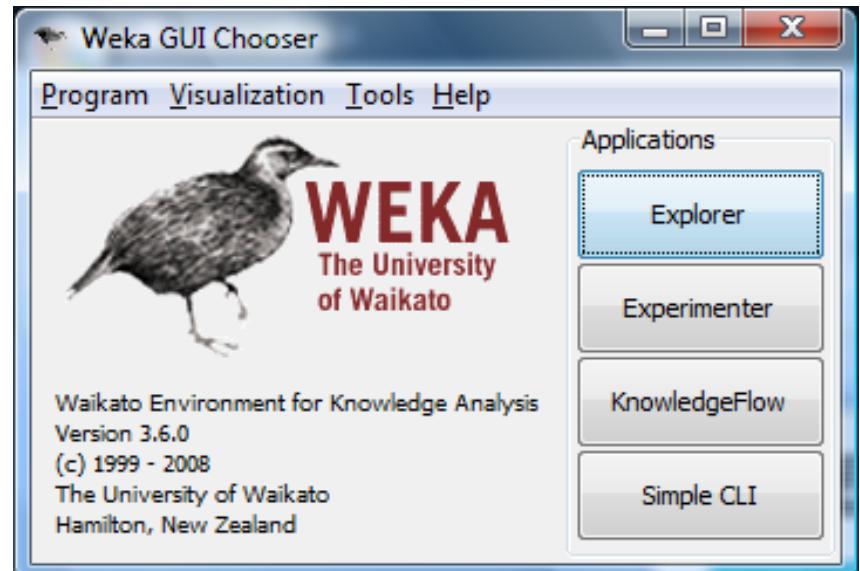
Distância	VALOR
$d(X,1)$	1,41
$d(X,2)$	1
$d(X,3)$	1,73
$d(X,4)$	1,41
$d(X,5)$	1,41
$d(X,6)$	1,73
$d(X,7)$	1,73
$d(X,8)$	1
$d(X,9)$	1
$d(X,10)$	1
$d(X,11)$	1
$d(X,12)$	1,73
$d(X,13)$	1,41
$d(X,14)$	1,73

# Exemplo

- $K = 5$
- Os 5 vizinhos mais próximos são
  - $X_1 = (\leq 30 \quad \text{Alta} \quad \text{Sim} \quad \text{Bom}) \quad \text{Classe} = \text{Não}$
  - $X_2 = (\leq 30 \quad \text{Média} \quad \text{Não} \quad \text{Bom}) \quad \text{Classe} = \text{Não}$
  - $X_3 = (\leq 30 \quad \text{Baixa} \quad \text{Sim} \quad \text{Bom}) \quad \text{Classe} = \text{Sim}$
  - $X_4 = (> 40 \quad \text{Média} \quad \text{Sim} \quad \text{Bom}) \quad \text{Classe} = \text{Sim}$
  - $X_5 = (\leq 30 \quad \text{Média} \quad \text{Sim} \quad \text{Exc.}) \quad \text{Clase} = \text{Sim}$
- Logo, X é classificada na classe = Sim

# WEKA

- ferramenta para data mining com muitos algoritmos implementados.
- Desenvolvida pela Universidade de Waikato, NZ
- Muito usada nos meios acadêmicos, free
- Site: <http://www.cs.waikato.ac.nz/ml/weka/>
- Desenvolvido em Java



# Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

## Filter

Choose **None**

Apply

## Current relation

Relation: None  
Instances: None

Attributes: None

## Selected attribute

Name: None  
Missing: None  
Distinct: None  
Type: None  
Unique: None

## Attributes

All

None

Invert

Pattern

Visualize All

Remove

## Status

Welcome to the Weka Explorer

Log



# Após abrir arquivo:

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose None Save the world Apply

Current relation

Relation: weather Instances: 14 Attributes: 5

Attributes

All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> tempo
2	<input checked="" type="checkbox"/> temperatura
3	<input checked="" type="checkbox"/> umidade
4	<input checked="" type="checkbox"/> vento
5	<input checked="" type="checkbox"/> jogo

Remove

Selected attribute

Name: tempo Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	sol	5
2	nublado	4
3	chuva	5

Class: jogo (Nom) Visualize All

The figure consists of three vertical bars, each divided into two horizontal sections: red on top and blue on bottom. The left bar has a value of 5 above it. The middle bar has a value of 4 above it. The right bar has a value of 5 above it. These values likely correspond to the counts of 'sol' (5), 'nublado' (4), and 'chuva' (5) respectively, as shown in the 'Selected attribute' table.

Status

OK Log x 0

# Aba de classificação

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose ZeroR

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) jogo

Start Stop

Result list (right-click for options)

Status

OK

Log x 0

# Exemplo de saída

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose J48 -C 0.25 -M 2

Test options

Use training set  
 Supplied test set Set...  
 Cross-validation Folds 10  
 Percentage split % 66  
More options...

(Nom) jogo

Start Stop

Result list (right-click for options)

- 10:27:40 - trees.J48
- 10:32:37 - trees.J48
- 10:32:48 - trees.J48
- 10:32:54 - trees.J48

Classifier output

Correctly Classified Instances	9	64.2857 %
Incorrectly Classified Instances	5	35.7143 %
Kappa statistic	0.186	
Mean absolute error	0.2857	
Root mean squared error	0.4818	
Relative absolute error	60 %	
Root relative squared error	97.6586 %	
Total Number of Instances	14	

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
sim	0.778	0.6	0.7	0.778	0.737	0.789	sim
nao	0.4	0.222	0.5	0.4	0.444	0.789	nao
Weighted Avg.	0.643	0.465	0.629	0.643	0.632	0.789	

==== Confusion Matrix ====

a	b	<- classified as
7	2	a = sim
3	2	b = nao

Status

OK

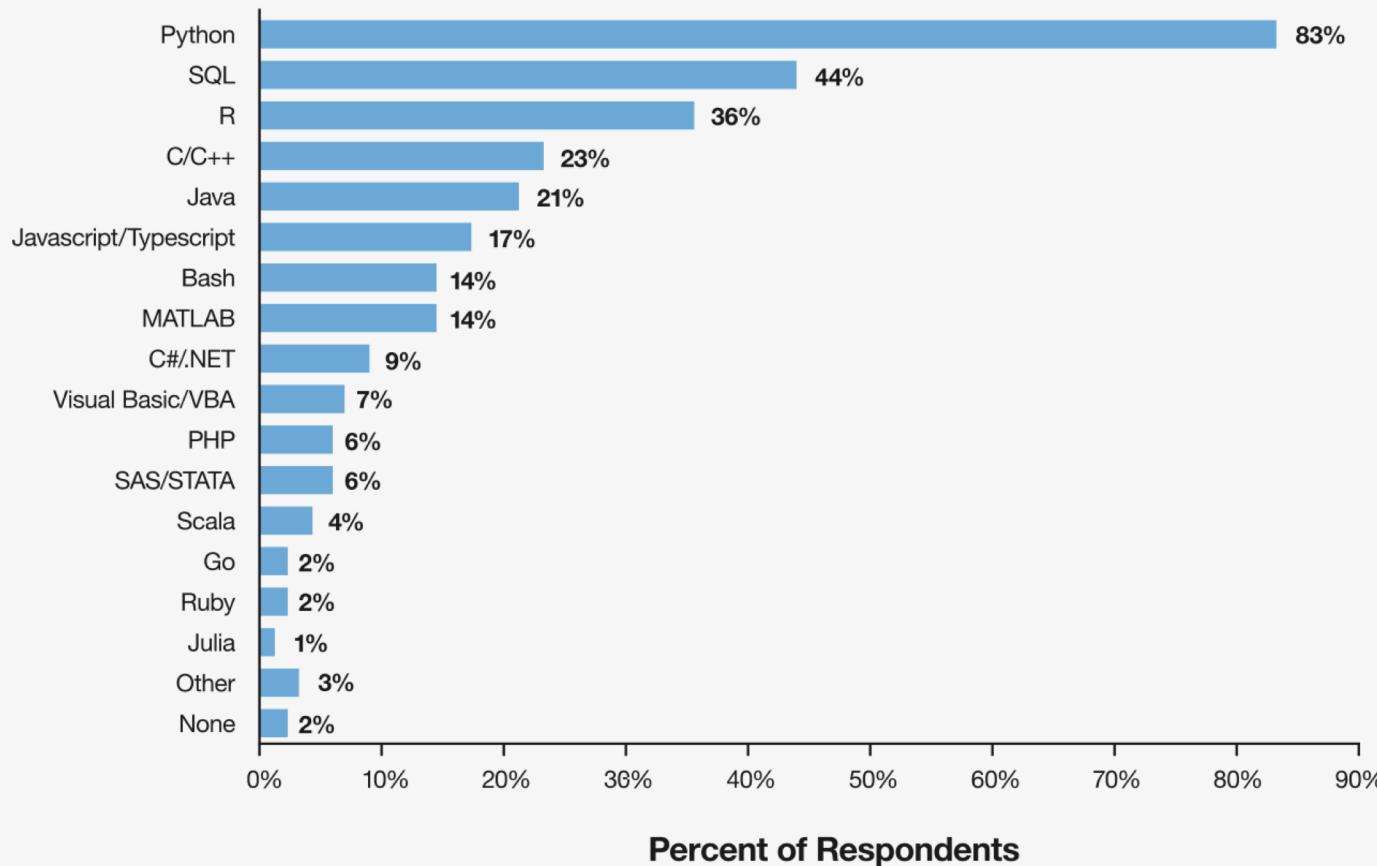
Log x 0



# Weka

- Exemplos de uso

## What Programming Language Do You Use on a Regular Basis?



**Figure 19** Data scientists have many options for programming languages to develop machine learning models. Python has become a popular choice.

*Figura obtida em <https://c3.ai/>*

# Python

- <https://www.continuum.io/downloads>

The screenshot shows the Anaconda website at https://www.continuum.io/downloads. The header includes the Anaconda logo and navigation links for Cloud, Documentation, Search, and Contact. Below the header are links for PRODUCTS, SUPPORT & SOLUTIONS, COMMUNITY, ABOUT, and RESOURCES. A large green circular graphic with a grid pattern and colored triangles (green, blue, and yellow) is on the right. The main content features a large green "DOWNLOAD ANACONDA NOW" button. Below it, there are download links for Windows, Mac, and Linux. A section titled "GET SUPERPOWERS WITH ANACONDA" describes Anaconda as an open data science platform. A sidebar on the right provides advice on choosing the correct Python version.

https://www.continuum.io/downloads

ANA CONDA  
Powered by Continuum Analytics®

Anaconda Cloud Documentation Search Contact

PRODUCTS SUPPORT & SOLUTIONS COMMUNITY ABOUT RESOURCES

## DOWNLOAD ANACONDA NOW

Download for

### GET SUPERPOWERS WITH ANACONDA

Anaconda is the leading open data science platform powered by Python. The open source version of Anaconda is a high performance distribution of Python and R and includes over 100 of the most popular Python, R and Scala packages

**Which version should I download and install?**

With Anaconda you can run multiple versions of Python in isolated environments, so choose the download with the Python version that you use more often, as that will be your default Python version.

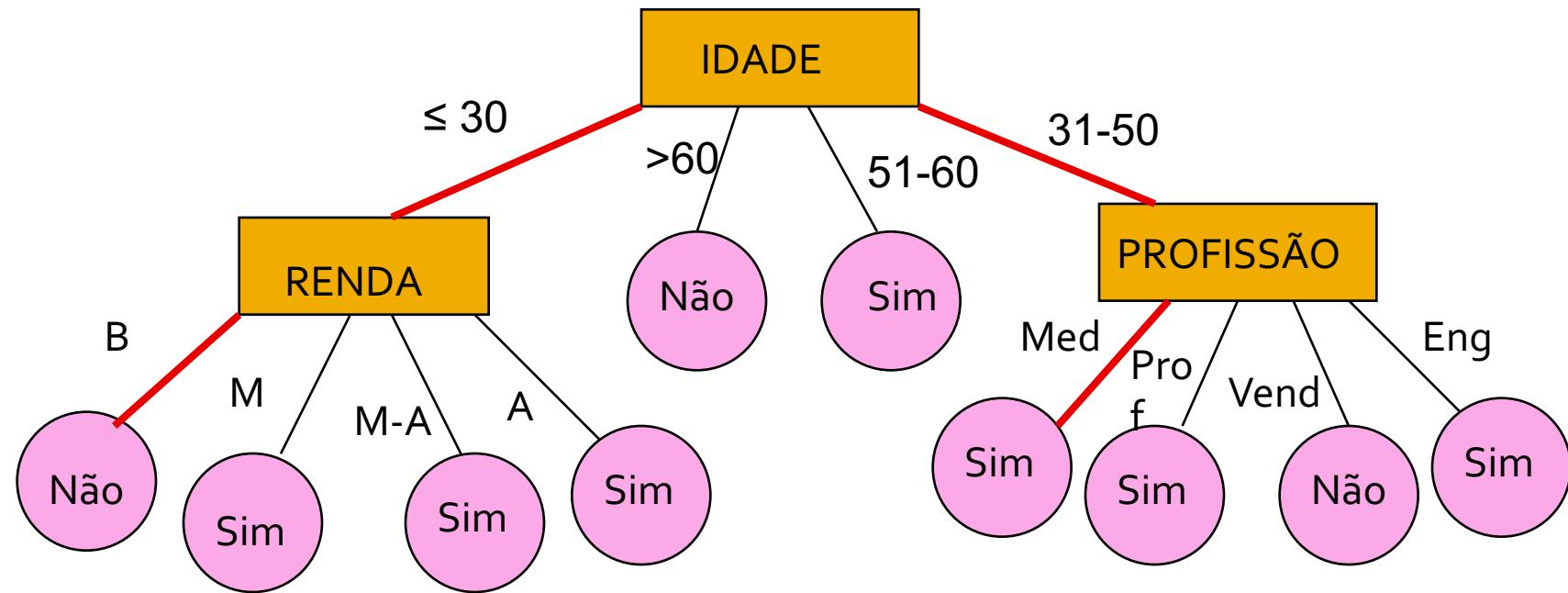
# Exemplo com python

- Usando o spyder

# Classificadores *Eager*

*Constróem um **modelo** de classificação.  
Modelo é utilizado para classificar nova tupla.*

# Modelo: Árvore de Decisão



Se Idade  $\leq 30$  e Renda é Baixa então **Compra Video Game**

Se Idade = 31-50 e Prof é Médico então **Compra Video Game**

# Como criar uma Árvore de Decisão – Algoritmo ID3

CASO 1

A	B	C	CLASSE
a1	b1	c1 g1	X
a1	b2	c1	X
a2	b1	c1	X
a2	b2	c2	X
a1	b2	c2	X

# Como criar uma Árvore de Decisão

CASO 2

A	B	C	CLASSE
a1	b1	c1	X
a1	b	A	
a2	b1	c1	Y
a2	b2	c2	X
a1	b2	c2	Y

Atributo-Teste =

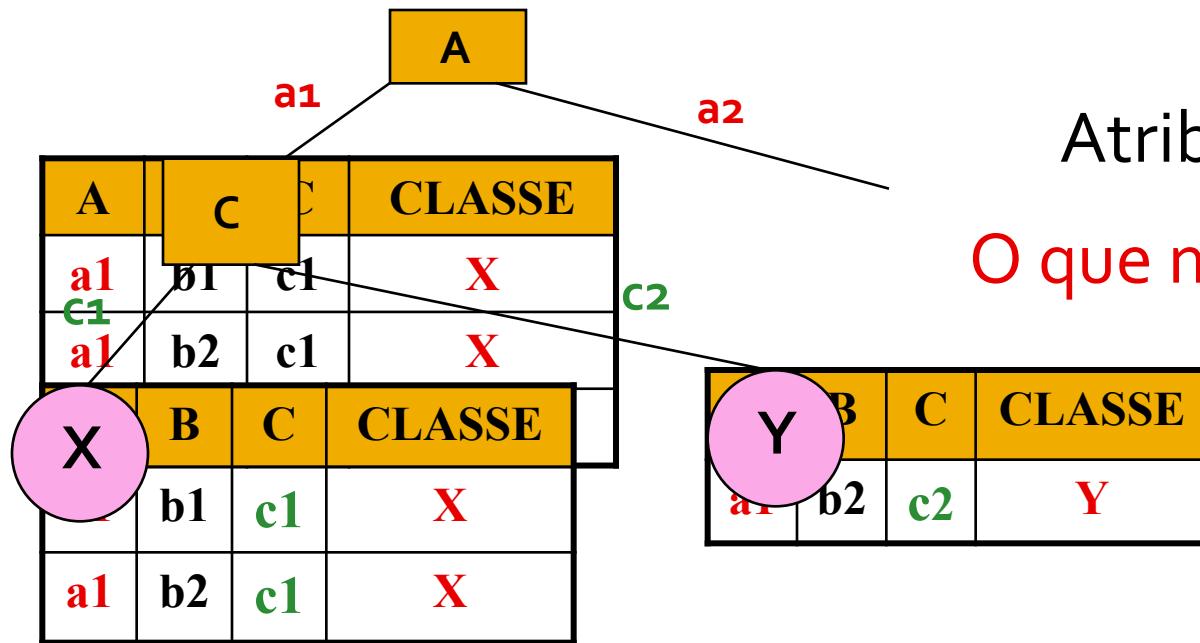
O que mais reduz  
a entropia

A	B	C	CLASSE
a1	b1	c1	X
a1	b2	c1	X

A	B	C	CLASSE
a2	b1	c1	Y
a2	b2	c2	X

LISTA-ATRIBUTOS = { A, B, C }

# Como criar uma Árvore de Decisão



Atributo-Teste =

O que mais reduz a entropia  
= C

LISTA-ATRIBUTOS = { B, C }

# Qual é o Atributo-Teste ?

- Divide-se o nó segundo cada atributo.
- Para cada divisão calcula-se a entropia produzida caso fosse escolhido este atributo.
- Considera-se o atributo cuja divisão resulta numa *maior redução da entropia*.

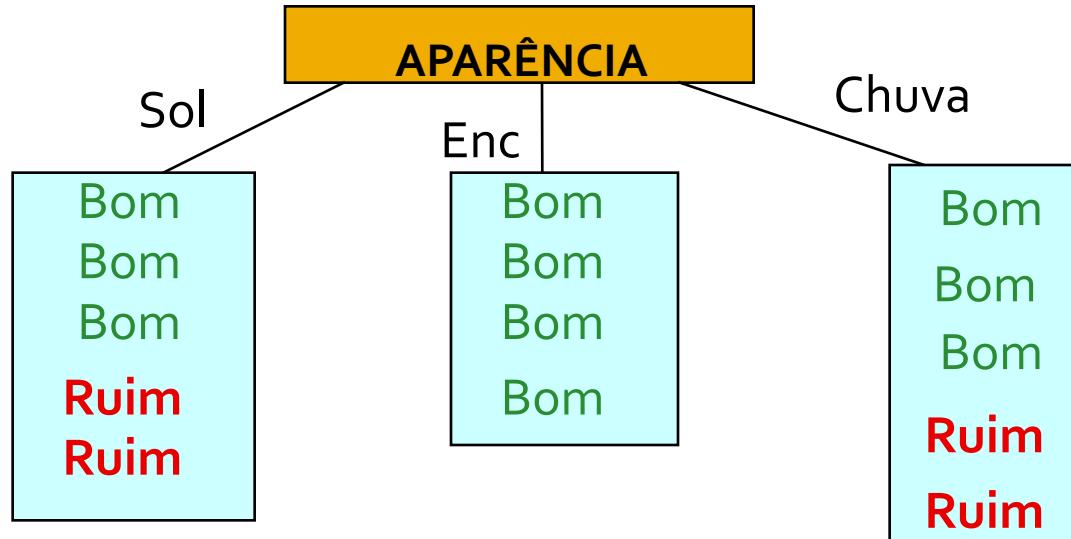
# Informação ganha na divisão

- $\text{Entrop}(A) = -\frac{\text{NC}_1}{\text{Tot}} \log_2 \frac{\text{NC}_1}{\text{Tot}} - \frac{\text{NC}_2}{\text{Tot}} \log_2 \frac{\text{NC}_2}{\text{Tot}}$
- $\text{Entrop}(D) = \frac{\text{NF}_1}{\text{Tot}} * \text{Entrop}(F_1) + \frac{\text{NF}_2}{\text{Tot}} * \text{Entrop}(F_2)$
- $\text{Info}(\text{Divisão}) = \text{Entrop}(A) - \text{Entrop}(D)$
- Maior Info(Divisão) → Atributo escolhido

# Um Exemplo

Aparência	Temperatura	Humidade	Vento	Classe
Sol	Quente	Alta	Não	Ruim
Sol	Quente	Alta	Sim	Ruim
Encoberto	Quente	Alta	Não	Bom
Chuvoso	Agradável	Alta	Não	Bom
Chuvoso	Frio	Normal	Não	Bom
Chuvoso	Frio	Normal	Sim	Ruim
Encoberto	Frio	Normal	Sim	Bom
Sol	Agradável	Alta	Não	Ruim
Sol	Frio	Normal	Não	Bom
Chuvoso	Agradável	Normal	Não	Bom
Sol	Agradável	Normal	Sim	Bom
Encoberto	Agradável	Alta	Sim	Bom
Encoberto	Quente	Normal	Não	Bom
Chuvoso	Agradável	Alta	Sim	Ruim

# 1a divisão possível : Aparência



$$\text{Entrop}(D) = \frac{5}{14} * \text{Entrop}(F_1) + \frac{4}{14} * \text{Entrop}(F_2) + \frac{5}{14} * \text{Entrop}(F_3) = 0.693$$

$$\text{Entrop}(F_1) = -\frac{3}{5} \log_2(\frac{3}{5}) - \frac{2}{5} \log_2(\frac{2}{5}) = 0.971$$

$$\text{Entrop}(F_2) = -\frac{4}{4} \log_2(\frac{4}{4}) = 0$$

$$\text{Entrop}(F_3) = -\frac{3}{5} \log_2(\frac{3}{5}) - \frac{2}{5} \log_2(\frac{2}{5}) = 0.971$$

# Redução da Entropia

$$\text{Entrop(A)} = - (9/14 * \log_2(9/14) + 5/14 * \log_2(5/14)) = \\ = \mathbf{0.940}$$

$$\text{INFO(APARÊNCIA)} = \text{Entrop(A)} - \text{Entrop(D)} = \\ = \mathbf{0.940 - 0.693 = 0.247}$$

# Comparando as 4 possibilidades

- $\text{Info}(\text{Aparência}) = 0.247$
- $\text{Info}(\text{Temperatura}) = 0.029$
- $\text{Info}(\text{Humidade}) = 0.152$
- $\text{Info}(\text{Vento}) = 0.020$

# Algoritmo ID3

- **Input:** Banco de dados de amostras A (com os valores dos atributos categorizados), lista de atributos Cand-List
- **Output :** Uma árvore de decisão

Begin

**Gera-árvore(A, Cand-List)**

End

# Algoritmo ID3

## Gera-árvore(A, Cand-List)

- Cria um nó N; Associa a este nó o banco de dados A
  - Se todas as tuplas de A são da mesma classe C: transforma N numa folha com label C e PÁRA
  - Caso contrário: Se Cand-List = vazio então transforma N numa folha com label igual a classe mais frequente de A
  - Caso contrário: X:= Ganco(Cand-List)  
*% esta função retorna o atributo X com maior ganho de informação (que causa maior redução de entropia)*
5. Etiqueta N com o atributo X
6. Para cada valor  $a$  do atributo X
1. Cria nó-filho F ligado a X por um ramo com label  $a$  e associa a este nó o conjunto A' de amostras que tem X =  $a$
  2. Se A' é vazio: transforma o nó F numa folha com label C onde C é a classe mais frequente de A
  3. Caso contrário: chama a rotina Gera(A', Cand-List-{X}) e associa ao nó F a árvore resultante deste cálculo.

# Implementações

- SKlearn

<https://scikit-learn.org/stable/modules/tree.html>

- Software Weka

Machine Learning Software in Java  
<http://www.cs.waikato.ac.nz/ml/weka/>

- Dados reais para testes

UCI Machine Learning Repository  
<http://archive.ics.uci.edu/ml/datasets.html>

# Weka

- Exemplos de uso

# Exemplo com python

- Usando o spyder

# Classificadores Bayesianos

## Introduction to Data Mining (Cap. 5-3)

### Tan, Steinbach, Kumar



---

Profa. Nádia Félix  
INF-UFG



# Introdução - Conceitos

- Classificadores Bayesianos são classificadores estatísticos que tem a função de classificar um objeto numa determinada classe, baseando-se na probabilidade deste objeto pertencer a esta classe.
- Deve-se atentar ao fato de que:
  - Em muitas aplicações, a relação entre o conjunto de atributos e a variável classe são não-determinísticos.



# Introdução - Exemplo

- Predizer quando uma pessoa tem doença no coração considerando os fatores **alimentação saudável e freqüência que pratica exercícios**.



# Introdução - Exemplo





# Teorema de Bayes

- Fornece o cálculo das probabilidades de que uma determinada amostra de dados pertença a cada uma das classes possíveis, predizendo para a amostra, a **classe mais provável**.
- Considerando X e Y variáveis aleatórias, uma probabilidade condicional  $P(Y|X)$  refere-se a probabilidade de Y assumir um valor determinado, observando-se o valor assumido por X.

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$



# Teorema de Bayes na Classificação

Exemplo:

ID	Idade	Renda	Estudante	Crédito	Compra_VideoGame
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

Classificar os seguintes valores:

X = (Idade <= 30, Renda = Media, Estudante = sim, Crédito = bom)

Y = Compra\_Computador?



# Teorema de Bayes na Classificação

Exemplo:

ID	Idade	Renda	Estudante	Crédito	Compra_VideoGame
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

$P(Y=\text{sim})$  e  $P(Y=\text{não})$

$$P(Y=\text{sim}) = 9/14 = 0,643$$

$$P(Y=\text{não}) = 5/14 = 0,357 = 1 - P(Y=\text{sim})$$



# Teorema de Bayes na Classificação

X = (**Idade <= 30**, Renda = Media, Estudante = sim, Crédito = bom)

ID	Idade	Renda	Estudante	Crédito	Compra_VideoGame
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

## Probabilidades:

$$P[\text{Idade} \leq 30 \mid Y = \text{sim}] = 2/9 = 0,222$$

$$P[\text{Idade} \leq 30 \mid Y = \text{não}] = 3/5 = 0,6$$



# Teorema de Bayes na Classificação

X = (Idade <= 30, Renda = Media, Estudante = sim, Crédito = bom)

ID	Idade	Renda	Estudante	Crédito	Compra_VideoGame
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

Probabilidades:

$$P[Renda = Media \mid Y = sim] = 4/9 = 0,444$$

$$P[Renda = Media \mid Y = não] = 2/5 = 0,4$$



# Teorema de Bayes na Classificação

X = (Idade <= 30, Renda = Media, Estudante = sim, Crédito = bom)

ID	Idade	Renda	Estudante	Crédito	Compra_videogame
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

## Probabilidades:

$$P[\text{Estudante} = \text{sim} | Y = \text{sim}] = 6/9 = 0,667$$

$$P[\text{Estudante} = \text{sim} | Y = \text{não}] = 1/5 = 0,2$$



# Teorema de Bayes na Classificação

X = (Idade <= 30, Renda = Media, Estudante = sim, Crédito = bom)

ID	Idade	Renda	Estudante	Crédito	Compra_VideoGame
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

## Probabilidades:

$$P[\text{Credito} = \text{bom} \mid Y = \text{sim}] = 6/9 = 0,667$$

$$P[\text{Credito} = \text{bom} \mid Y = \text{não}] = 2/5 = 0,4$$



# Teorema de Bayes na Classificação

- Calculamos isoladamente o valor da probabilidade condicional de cada atributo, mas para que eles sejam calculado de forma intersecccionada, temos:

$$P[x_1, x_2, \dots, x_d | C] = P(x_1 | C) * P(x_2 | C) * \dots * P(x_d | C)$$

- Com isso, é possível chegar a uma forma mais geral do Teorema de Bayes:

$$P(Y | X) = \frac{P(Y) \prod_{i=1}^d P(X_i | Y)}{P(X)}$$

# Teorema de Bayes na Classificação



Temos:

$$P(X|Y=\text{sim}) = 0,222 * 0,444 * 0,667 * 0,667 = 0,044$$

$$P(X|Y=\text{não}) = 0,6 * 0,4 * 0,2 * 0,4 = 0,019$$

Pela lei da probabilidade total:

$$P(X|Y=\text{sim}) * P(Y=\text{sim}) / P(X) = 0,044 * 0,643 = 0,028 / 0,035 = 0,8$$

$$P(X|Y=\text{não}) * P(Y=\text{não}) / P(X) = 0,019 * 0,357 = 0,007 / 0,035 = 0,2$$

- Ou seja,  $P(X|Y=\text{sim}) > P(X|Y=\text{não})$
- O classificador Bayesiano prediz que a tupla  $X$  é classificada na classe **Compra-VideoGame = sim**



# Classificador Naive Bayes

- Um classificador Naive Bayes estima a probabilidade de classe condicional  $P(X|Y)$ .
- Pré-considerações:
  - Assume-se que os atributos são condicionalmente independentes (Naive Bayes ingênuo ou simples);
  - As probabilidades condicionais são estimadas para os atributos de acordo com a sua classificação:
    - Categórico;
    - Contínuo.

# Atributos Condisionalmente Independentes



- São atributos que apresentam independência estatística entre si:
  - Dois eventos são estatisticamente independentes se a probabilidade da ocorrência de um evento não é afetada pela ocorrência do outro evento.

## Exemplo:

- Tamanho do braço x Habilidades de Leitura
  - Considerando a Idade, a dependência não ocorre.



# Atributos Categóricos

- É aquele atributo para o qual é possível estabelecer um conjunto de valores finito.
- Exemplo:
  - Sexo: {Masculino, Feminino}
  - Cor da Pele: {Branca, Marrom, Amarela, Preta}



# Atributos Categóricos

- Para uso no algoritmo Naive Bayes:
  - Estima-se a fração das instâncias de treinamento de acordo com cada valor da classe.

## Exemplo:

Casa Própria	Estado Civil	Inadimplente
Sim	Casado	Sim
Sim	Solteiro	Não
Não	Casado	Não
Sim	Divorciado	Não

- $P(\text{Casa Própria}=\text{Sim}|\text{Não})$



# Atributos Categóricos

Exemplo:

Casa Própria	Estado Civil	Inadimplente
Sim	Casado	Sim
Sim	Solteiro	Não
Não	Casado	Não
Sim	Divorciado	Não

- $P(\text{Casa Própria}=\text{Sim}|\text{Não}) = 2/3$



# Atributos Contínuos

- São considerados contínuos os atributos que possuem muitos ou infinitos valores possíveis
- Exemplo:
  - Idade:  $\in \mathbb{N} \geq 0$
  - Peso:  $\in \mathbb{R} \geq 0$



# Atributos Contínuos

- Existem duas formas de estimar a probabilidade de classe condicional para atributos contínuos:
  - Discretização dos atributos;
  - Distribuição Gaussiana.



# Atributos Contínuos

- Discretização de atributos contínuos:
  - Os atributos contínuos são divididos em intervalos discretos, que substituem os valores desses atributos.
  - Esta abordagem transforma os atributos contínuos em atributos ordinais.
- A transformação dos atributos contínuos em atributos discretos permite que sejam tratados como atributos categóricos.



# Atributos Contínuos

## ■ Distribuição Gaussiana:

- Assume uma certa forma de distribuição de probabilidade para variáveis contínuas, e estima os parâmetros da distribuição usando os dados de treinamento.
- Caracterizada por dois parâmetros:
  - Média ( $\mu$ )
  - Variância ( $\sigma^2$ ) da amostra



# Atributos Contínuos

- Para cada valor de classe  $y$ , a probabilidade da classe condicional para o atributo  $X$  é:

$$P(X = x | Y = y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

$$\mu = \frac{\sum y}{n}$$

$$\sigma^2 = \frac{\sum(x - \mu)^2}{n-1}$$

Obs.:  $\exp(x)$  é equivalente a  $e^x$ , e o "e" é uma constante universal  
 $e = 2.718281828\dots$

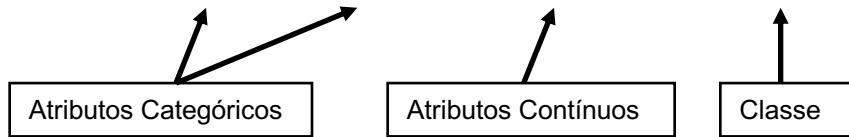


# Classificador Naive Bayes

## ■ Exemplo:

- Dado o seguinte conjunto de treinamento:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim





# Classificador Naive Bayes

## ■ Cálculo dos atributos categóricos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

$$P(\text{Casa Própria=Sim}|\text{Não}) = 3/7$$



# Classificador Naive Bayes

## ■ Cálculo dos atributos categóricos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

$$P(\text{Casa Própria}=\text{Sim}|\text{Não}) = 3/7$$

$$P(\text{Casa Própria}=\text{Não}|\text{Não}) = 4/7$$



# Classificador Naive Bayes

## ■ Cálculo dos atributos categóricos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

$$P(\text{Casa Própria=Sim}|\text{Não}) = 3/7$$

$$P(\text{Casa Própria=Não}|\text{Não}) = 4/7$$

$$P(\text{Casa Própria=Sim}|Sim) = 0$$



# Classificador Naive Bayes

## ■ Cálculo dos atributos categóricos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

$P(\text{Casa Própria=Sim}|\text{Não}) = 3/7$

$P(\text{Casa Própria=Não}|\text{Não}) = 4/7$

$P(\text{Casa Própria=Sim}|\text{Sim}) = 0$

$P(\text{Casa Própria=Não}|\text{Sim}) = 1$



# Classificador Naive Bayes

## ■ Cálculo dos atributos categóricos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

$P(\text{Casa Própria=Sim|Não}) = 3/7$

$P(\text{Casa Própria=Não|Não}) = 4/7$

$P(\text{Casa Própria=Sim|Sim}) = 0$

$P(\text{Casa Própria=Não|Sim}) = 1$

$P(\text{Estado Civil=Solteiro|Não}) = 2/7$

$P(\text{Estado Civil=Divorciado|Não}) = 1/7$

$P(\text{Estado Civil=Casado|Não}) = 4/7$

$P(\text{Estado Civil=Solteiro|Sim}) = 2/3$

$P(\text{Estado Civil=Divorciado|Sim}) = 1/3$

$P(\text{Estado Civil=Casado|Sim}) = 0$



# Classificador Naive Bayes

## ■ Cálculo dos atributos contínuos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Média:

$$\mu = (125 + 100 + 70 + 120 + 60 + 220 + 75) / 7 = 110$$

Variância:

$$\sigma^2 = (125-110)^2 + (100-110)^2 + (70-110)^2 + (120-110)^2 + (60-110)^2 + (220-110)^2 + (75-110)^2 / 6 = 2975$$

■ Para a classe Não



# Classificador Naive Bayes

## ■ Cálculo dos atributos contínuos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Média:

$$\mu = (95 + 85 + 90) / 3 = \mathbf{90}$$

Variância:

$$\sigma^2 = (95-90)^2 + (85-90)^2 + (90-90)^2 / 2 = \mathbf{25}$$

■ Para a classe **Sim**



# Classificador Naive Bayes

## ■ Resultado dos cálculos básicos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

$$P(\text{Casa Própria=Sim}|\text{Não}) = 3/7$$

$$P(\text{Casa Própria=Não}|\text{Não}) = 4/7$$

$$P(\text{Casa Própria=Sim}|Sim) = 0$$

$$P(\text{Casa Própria=Não}|Sim) = 1$$

$$P(\text{Estado Civil=Solteiro}|\text{Não}) = 2/7$$

$$P(\text{Estado Civil=Divorciado}|\text{Não}) = 1/7$$

$$P(\text{Estado Civil=Casado}|\text{Não}) = 4/7$$

$$P(\text{Estado Civil=Solteiro}|Sim) = 2/3$$

$$P(\text{Estado Civil=Divorciado}|Sim) = 1/3$$

$$P(\text{Estado Civil=Casado}|Sim) = 0$$

Para o cálculo da Renda Anual:

Classe Não:

Média: 110

Variância: 2975

Classe Sim:

Média: 90

Variância: 25



# Classificador Naive Bayes

- Dado o conjunto de treinamento anterior, qual a classe do seguinte registro de teste:

X = (Casa Própria=Não, Estado Civil=Casado, Renda Anual=120K)

- Avaliar qual a maior probabilidade entre as probabilidades posteriores:
  - $P(\text{Inadimplente}=\text{Não}|X)$  e  $P(\text{Inadimplente}=\text{Sim}|X)$



# Classificador Naive Bayes

- Para calcular as probabilidades posteriores  $P(\text{Não}|X)$  e  $P(\text{Sim}|X)$  necessitamos:
  - Calcular as classes condicionais  $P(X|\text{Nao})$  e  $P(X|\text{Sim})$
- $P(X|\text{Não})$
- $P(\text{Casa Própria}=\text{Não}|\text{Não}) * P(\text{Estado Civil}=\text{Casado}|\text{Não}) * P(\text{Renda Anual}=120\text{K}|\text{Não})$
- $4/7 * 4/7 * 0,0072$
- 0,0024
- $P(X|\text{Sim})$
- $P(\text{Casa Própria}=\text{Não}|\text{Sim}) * P(\text{Estado Civil}=\text{Casado}|\text{Sim}) * P(\text{Renda Anual}=120\text{K}|\text{Sim})$
- $1 * 0 * 1,2 \times 10^{-9}$
- 0



# Exemplo:Naive Bayes

Dia	Aspecto	Temperatura	Umidade	Vento	Decisão
1	Sol	Quente	Alta	Fraco	N
2	Sol	Quente	Alta	Forte	N
3	Nublado	Quente	Alta	Fraco	S
4	Chuva	Agradável	Alta	Fraco	S
5	Chuva	Fria	Normal	Fraco	S
6	Chuva	Fria	Normal	Forte	N
7	Nublado	Fria	Normal	Forte	S
8	Sol	Agradável	Alta	Fraco	N
9	Sol	Fria	Normal	Fraco	S
10	Chuva	Agradável	Normal	Fraco	S
11	Sol	Agradável	Normal	Forte	S
12	Nublado	Agradável	Alta	Forte	S
13	Nublado	Quente	Normal	Fraco	S
14	Chuva	Agradável	Alta	Forte	N

## Exemplo: Naive Bayes

Qual será a *decisão* (valor da classe), se o dia estiver com sol, a temperatura fria, a umidade alta e o vento forte ?

$P(\text{Jogar} = S \mid \text{Aspecto} = \text{Sol}, \text{Temperatura} = \text{Fria}, \text{Umidade} = \text{Alta} \text{ e } \text{Vento} = \text{Forte}) = ?$

$P(\text{Jogar} = N \mid \text{Aspecto} = \text{Sol}, \text{Temperatura} = \text{Fria}, \text{Umidade} = \text{Alta} \text{ e } \text{Vento} = \text{Forte}) = ?$

## Exemplo: Naive Bayes

$P(\text{Jogar} = S) = 9/14;$   $P(\text{Jogar} = N) = 5/14;$

$P(\text{Aspecto} = \text{Sol} \mid \text{Jogar} = S) = 2/9;$

$P(\text{Aspecto} = \text{Sol} \mid \text{Jogar} = N) = 3/5;$

$P(\text{Temperatura} = \text{Fria} \mid \text{Jogar} = S) = 3/9;$

$P(\text{Temperatura} = \text{Fria} \mid \text{Jogar} = N) = 1/5;$

$P(\text{Umidade} = \text{Alta} \mid \text{Jogar} = S) = 3/9;$

$P(\text{Umidade} = \text{Alta} \mid \text{Jogar} = N) = 4/5;$

$P(\text{Vento} = \text{Forte} \mid \text{Jogar} = S) = 3/9;$

$P(\text{Vento} = \text{Forte} \mid \text{Jogar} = N) = 3/5;$

## Exemplo: Naive Bayes

$P(\text{Aspecto} = \text{Sol}) = 5/14$

$P(\text{Temperatura} = \text{Fria}) = 4/14$

$P(\text{Umidade} = \text{Alta}) = 7/14$

$P(\text{Vento} = \text{Forte}) = 6/14$

## Exemplo: Naive Bayes

$P(\text{Jogar} = S \mid \text{Aspecto} = \text{Sol}, \text{Temperatura} = \text{Fria}, \text{Umidade} = \text{Alta} \text{ e } \text{Vento} = \text{Forte}) =$

$$\frac{P(\text{Sol}|S) * P(\text{Fria}|S) * P(\text{Alta}|S) * P(\text{Forte}|S) * P(S)}{P(\text{Sol}) * P(\text{Fria}) * P(\text{Alta}) * P(\text{Forte})}$$

$$= (2/9 * 3/9 * 3/9 * 3/9 * 9/14) / (5/14 * 4/14 * 7/14 * 6/14) =$$

$$= 0,0053 / 0,02186 = 0,242$$

## Exemplo: Naive Bayes

$P(\text{Jogar} = N \mid \text{Aspecto} = \text{Sol}, \text{Temperatura} = \text{Fria}, \text{Umidade} = \text{Alta} \text{ e } \text{Vento} = \text{Forte}) =$

$$\frac{P(\text{Sol}|N) * P(\text{Fria}|N) * P(\text{Alta}|N) * P(\text{Forte}|N) * P(N)}{P(\text{Sol}) * P(\text{Fria}) * P(\text{Alta}) * P(\text{Forte})} =$$

$$= (3/5 * 1/5 * 4/5 * 3/5 * 5/14) / (5/14 * 4/14 * 7/14 * 6/14)$$

=

$$= 0,0206 / 0,02186 = 0,942$$

Como  $(J=N) 0,942 > (J=S) 0,242$  Então **Jogar = Não**

# Naïve Bayes

- Vantagens:
  - rápido
  - Bons resultados em dados reais
- Desvantagens:
  - Resultados não tão bons em problemas complexos (milhares de atributos. Árvores Profundas...)
- Mozilla Thunderbird e Microsoft Outlook usam classificadores naive bayes para filtrar (marcar) emails que seriam spam

# Naive bayes no Weka

---



# Naive Bayes com Python

---



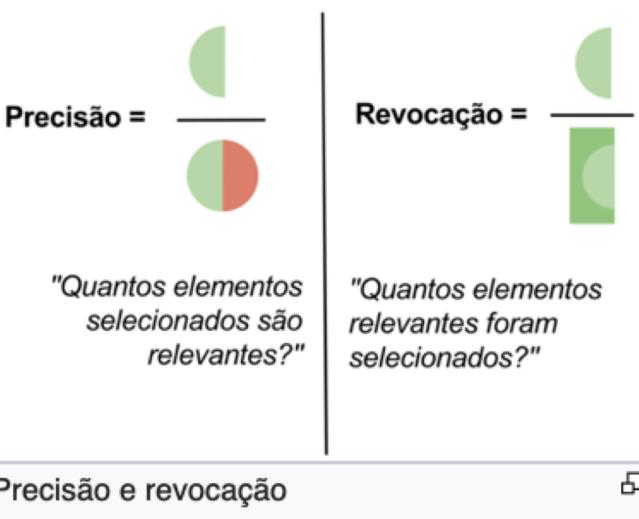
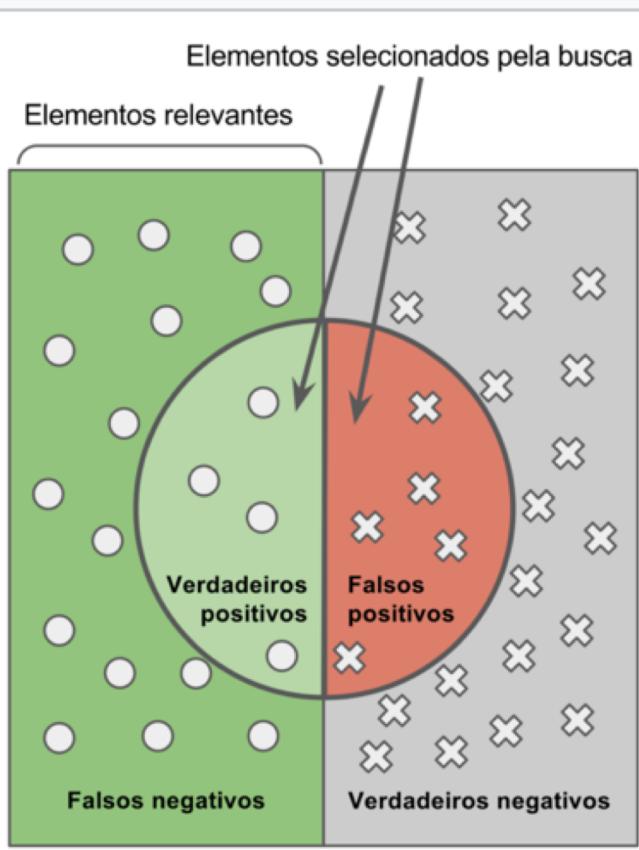
```
1 # -*- coding: utf-8 -*-
2 from pandas import read_csv
3 from sklearn.naive_bayes import GaussianNB
4
5 from sklearn.cross_validation import train_test_split
6
7 from sklearn.metrics import accuracy_score
8
9
10 data = read_csv('iris.csv')
11
12 Y = data['Species']
13 X = data[['Sepal.Length', 'Sepal.Width', 'Petal.Length', 'Petal.Width']]
14
15
16
17 X_train, X_test, y_train, y_test = train_test_split( X, Y)
18
19 classificador = GaussianNB()
20
21 classificador.fit(X_train, y_train)
22
23 y_pred = classificador.predict(X_test)
24
25 print(accuracy_score(y_test,y_pred)*100)|
```

# Método de Validação Cruzada

- Particionar aleatoriamente o conjunto de dados em três partes:
- Passo 1: Define pelo menos três conjuntos disjuntos:
  - 1. Conjunto de exemplos de treinamento
  - 2. Conjunto de exemplos de validação
  - 3. Conjunto de exemplos de teste
- Passo 2: Utiliza o Conjunto de Treinamento para fazer a aprendizagem do algoritmo. Utiliza o Conjunto de Validação para verificar a generalização do algoritmo (ajustar os parâmetros).

# Método de Validação Cruzada

- Passo 3: Depois do algoritmo treinado, avalia sua generalização sobre o Conjunto de Testes.
  - OBS 1: O Conjunto de Testes não pode ser utilizado para ajustar parâmetros!
  - OBS 2: Cuidar com a distribuição de exemplos por classe que compõem os conjuntos. Probabilidades a priori diferentes!
  - OBS 3: Método melhor adaptado a grandes conjuntos de dados.



## Medidas de Avaliação para classificadores:

$$\text{Precisão} = \frac{tp}{tp + fp}$$

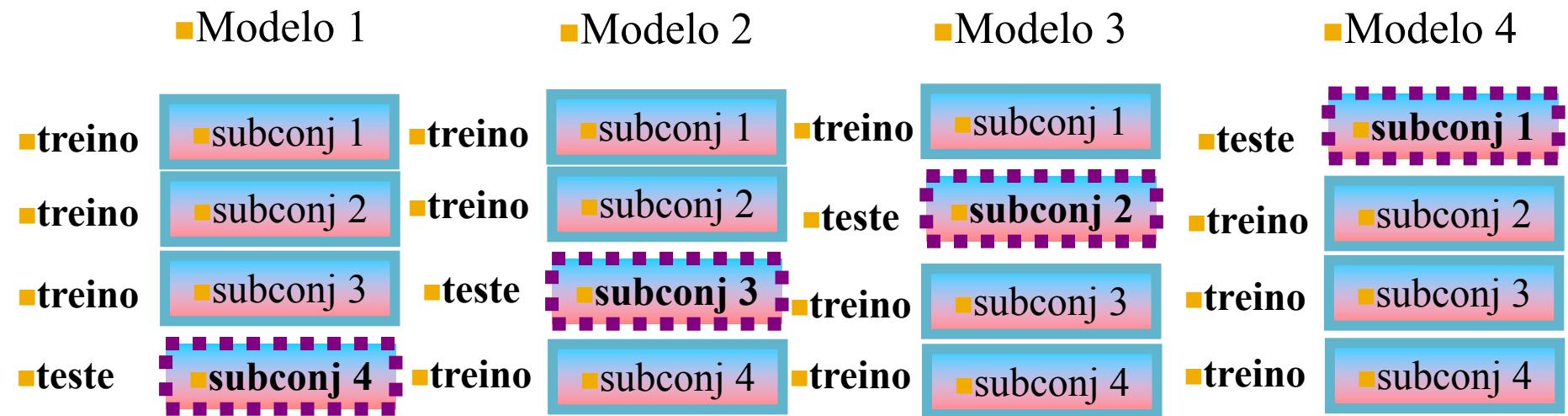
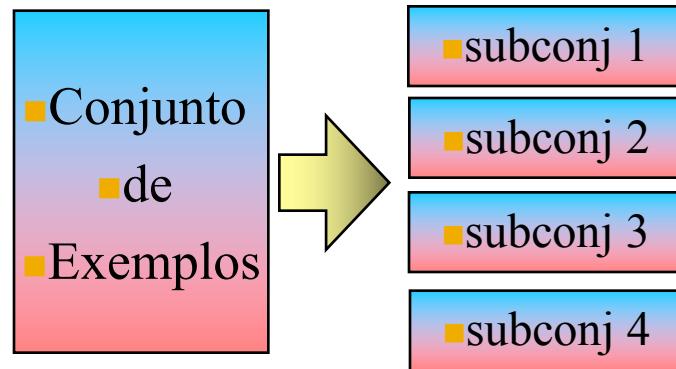
$$\text{Revocação} = \frac{tp}{tp + fn}$$

$$\text{Acurácia} = \frac{tp + tn}{tp + tn + fp + fn}$$

F-measure ou F1 ou F:

$$F = 2 \cdot \frac{\text{precis} \cdot \text{revoc}}{\text{precis} + \text{revoc}}$$

## ■ 4-fold-cross-validation



# Mais sobre validação cruzada

- Método padrão de avaliação: validação cruzada por dez vezes estratificada
- Por que dez? Experimentos demonstraram que esta é uma boa escolha para se obter uma estimativa precisa
- Estratificação reduz a variância da estimativa
- Melhor ainda: validação cruzada estratificada repetida
  - P. ex. se repete dez vezes a validação cruzada por dez vezes e se calcula a média (reduz variância)

## ■ Validação cruzada deixando um fora

- Validação cruzada deixando um fora (*leave-one-out c-v*):
  - O número de vezes é escolhido como o número de exemplos de treinamento
  - Isto é, deve-se construir  $n$  classificadores, onde  $n$  é o número de exemplos de treinamento
- Aproveita ao máximo os dados
- Computacionalmente muito custoso

# Navalha de Occam

- Dados dois modelos com mesma taxa de erros, o modelo mais simples é melhor (preferível)

# K-fold CrossValidation no Weka

---



# K-fold CrossValidation Python

---



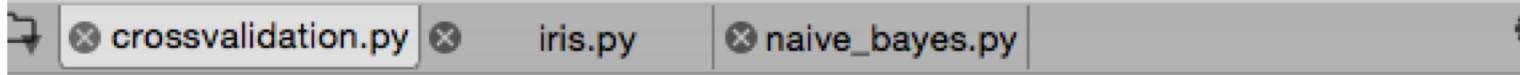
crossvalidation.py

iris.py

naive\_bayes.py

```
1 # -*- coding: utf-8 -*-
2 from pandas import read_csv
3 from sklearn.naive_bayes import GaussianNB
4
5 from sklearn.model_selection import cross_val_score
6
7
8 data = read_csv('iris.csv')
9
10 Y = data['Species']
11 X = data[['Sepal.Length', 'Sepal.Width', 'Petal.Length', 'Petal.Width']]
12
13
14
15 classificador = GaussianNB()
16
17 scores = cross_val_score(classificador, X, Y, cv=5)
18
19 print(scores)
20
```

# Fazendo a média das acuráncias dos folds:



```
1 # -*- coding: utf-8 -*-
2 from pandas import read_csv
3 from sklearn.naive_bayes import GaussianNB
4
5 from sklearn.model_selection import cross_val_score
6
7
8 data = read_csv('iris.csv')
9
10 Y = data['Species']
11 X = data[['Sepal.Length', 'Sepal.Width', 'Petal.Length', 'Petal.Width']]
12
13
14
15 classificador = GaussianNB()
16
17 scores = cross_val_score(classificador, X, Y, cv=5)
18
19 print(scores)
20
21 print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```