Introdução à Data Science

Termos em destaque

- Machine Learning
- Data science
- Data mining
- Data analysis
- Statistical learning
- Knowledge discovery
- Pattern recognition

Dados em todo lugar

- Google: processa 24 peta bytes de dados por dia.
- Facebook: 10 milhões de uploads de fotos por hora.
- Youtube: upload de 1 hora de vídeo a cada segundo.
- Twitter: 400 milhões de tweets por dia.
- …

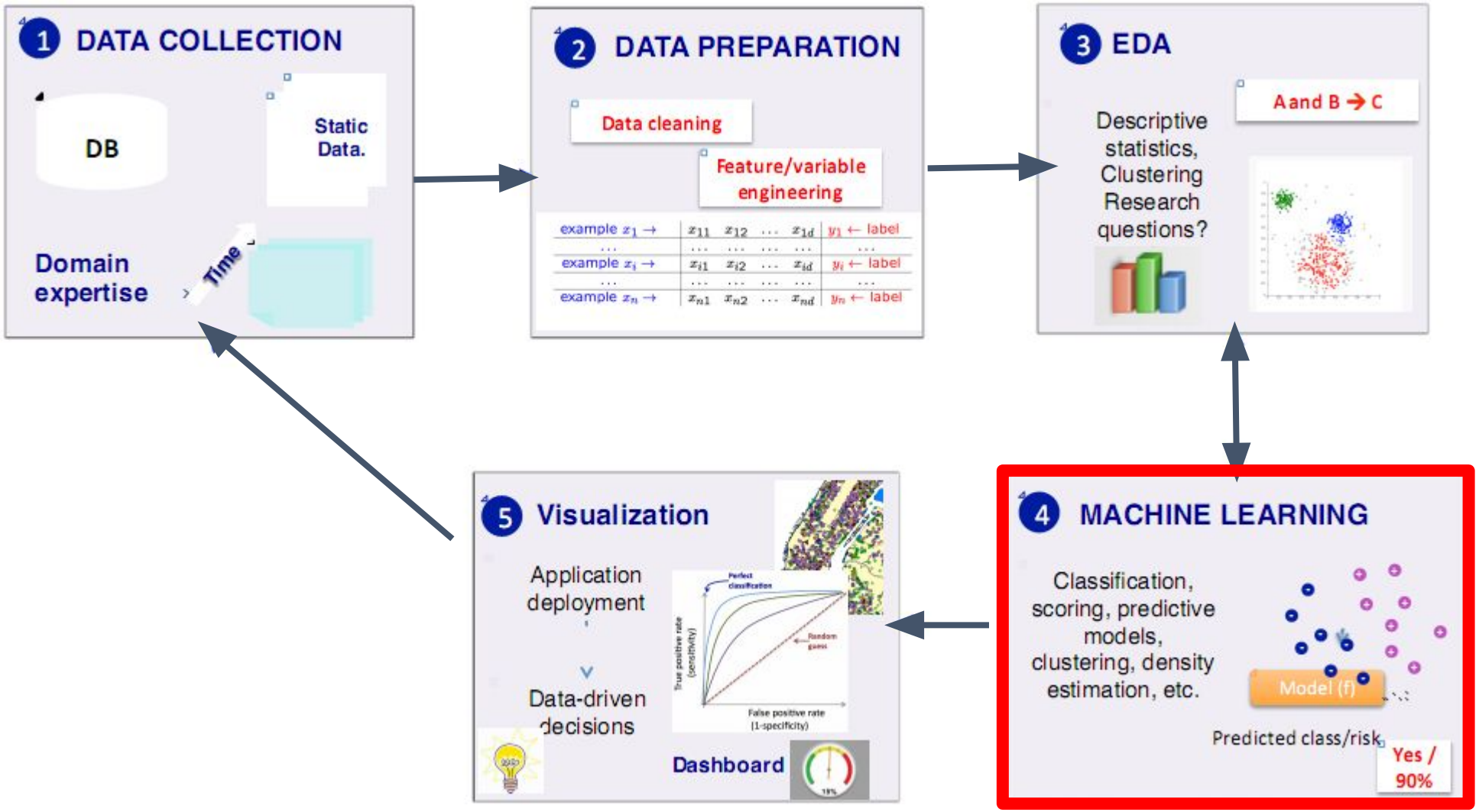
"Por volta de 2020 o universo digital alcançará 44 zettabytes (44 trilhões de gigabytes)"

The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, April 2014.

Tipos de dados

- Texto
- Números
- Clickstreams
- Grafos
- Tabelas
- Imagens
- Transações
- Vídeos
- ...

Ciência dos Dados

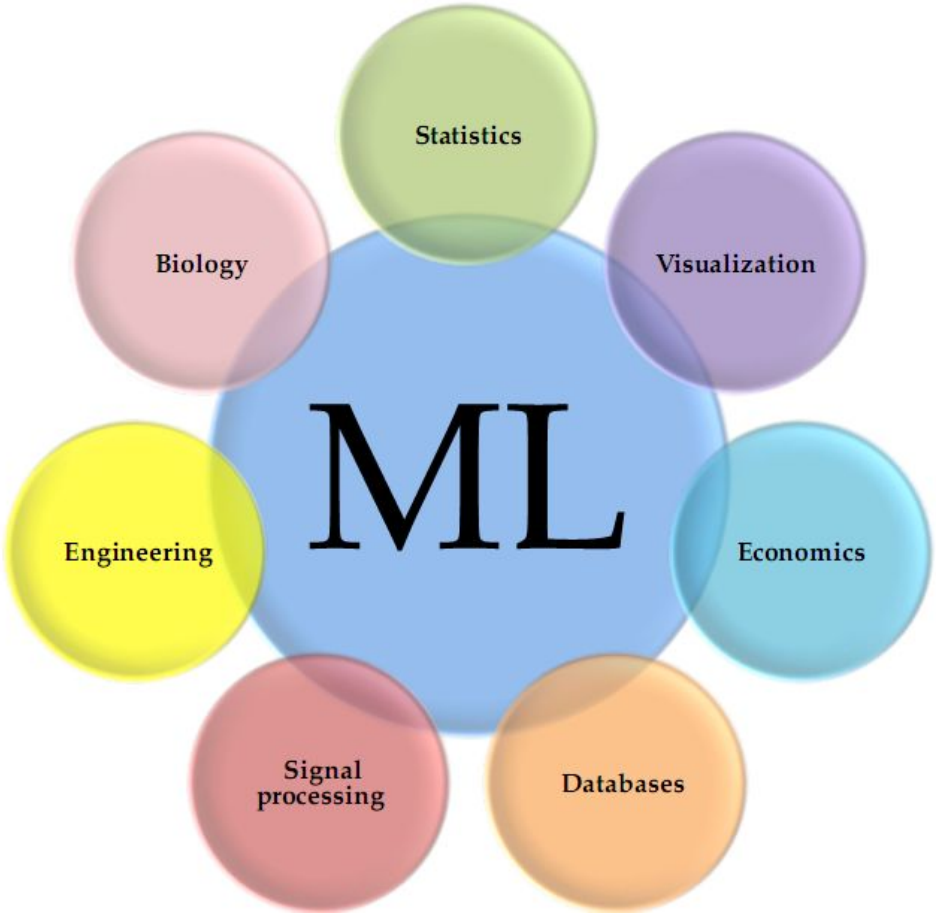


Aplicações de ML

- Filtragem de Spam.
- Detecção de fraude em cartão de crédito.
- Diagnóstico médico.
- Detecção de faces em imagens.
- Sistema de recomendação.
- Engenhos de busca.
- Reconhecimento manuscrito.
- Classificação de cena.
- ...

Aplicações de ML

- Filtragem de Spam.
- Detecção de fraude em cartão de crédito.
- Diagnóstico médico.
- Detecção de faces em imagens.
- Sistema de recomendação.
- Engenhos de busca.
- Reconhecimento manuscrito.
- Classificação de cena.
- ...



Tipos de aprendizagem

- Supervisionada
- Não-supervisionada

Aprendizagem supervisionada

Revisão

Se diz que um computador **aprende** a partir de uma experiência E em relação à uma tarefa T considerando-se uma medida de performance M; se a performance na tarefa T, medida por M, melhora com a experiência E.

Tom Mitchell. Machine Learning 1997.

Aprendizagem supervisionada

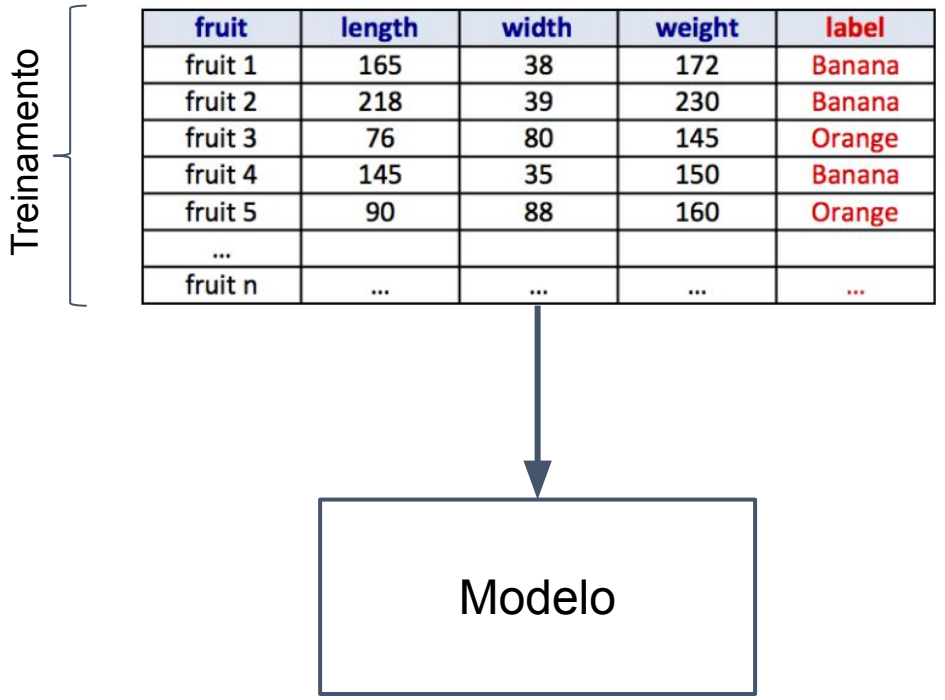
Aprender um modelo a partir de **dados rotulados.**

Aprendizagem supervisionada

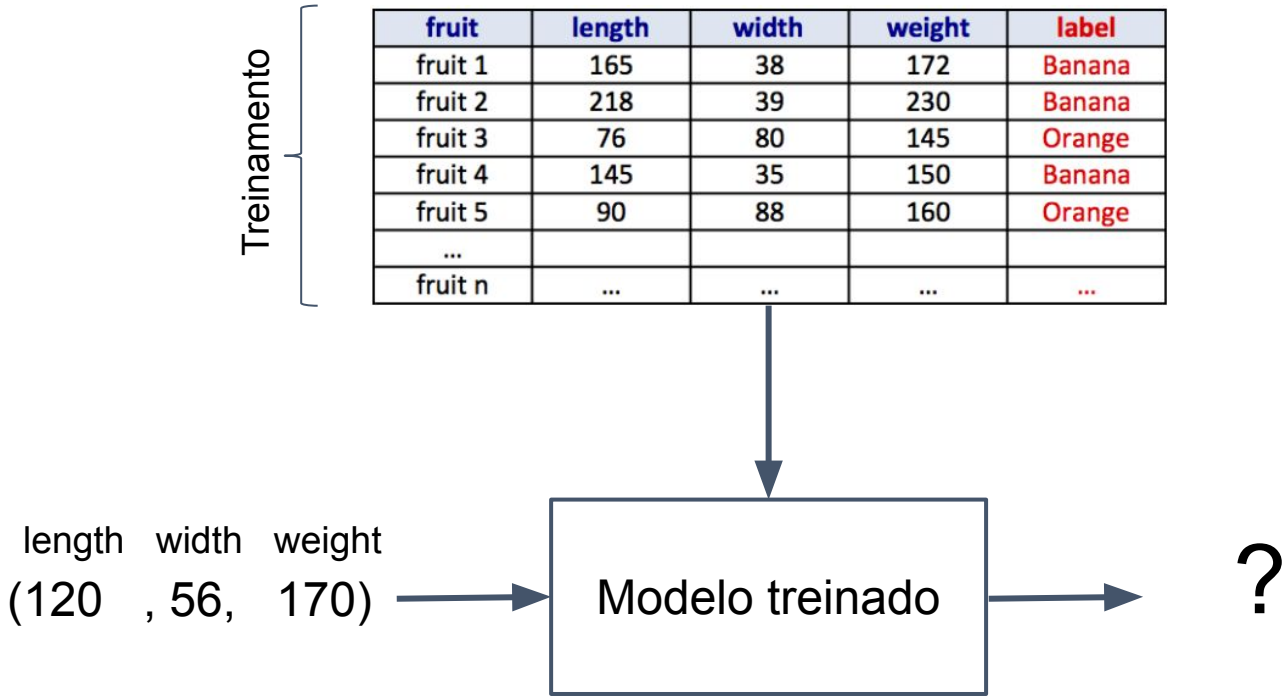
Aprender um modelo a partir de **dados de entrada e saída.**

fruit	length	width	weight	label
fruit 1	165	38	172	Banana
fruit 2	218	39	230	Banana
fruit 3	76	80	145	Orange
fruit 4	145	35	150	Banana
fruit 5	90	88	160	Orange
...				
fruit n	...	...	...	...

Aprendizagem supervisionada



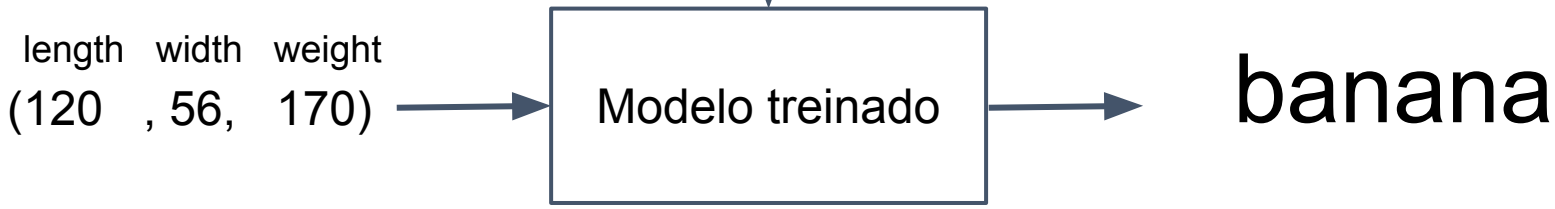
Aprendizagem supervisionada



Aprendizagem supervisionada

Treinamento

fruit	length	width	weight	label
fruit 1	165	38	172	Banana
fruit 2	218	39	230	Banana
fruit 3	76	80	145	Orange
fruit 4	145	35	150	Banana
fruit 5	90	88	160	Orange
...				
fruit n	...	...	...	...



Problemas supervisionados

- Classificação
- Regressão

Problema supervisionado: Classificação

Problema de classificação

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	S	S	Peq	S	Doente
Pedro	N	N	Grd	N	Saudável
Maria	S	S	Peq	N	Saudável
José	S	N	Grd	S	Doente
Ana	S	N	Peq	S	Saudável
Leila	N	N	Grd	S	Doente

Nome	Febre	Enjôo	Manchas	Dores
Luis	N	N	Peq	S
Laura	S	S	Grd	S

Problema de classificação

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	S	S	Peq	S	Doente
Pedro	N	N	Grd	N	Saudável
Maria	S	S	Peq	N	Saudável
José	S	N	Grd	S	Doente
Ana	S	N	Peq	S	Saudável
Leila	N	N	Grd	S	Doente

Nome	Febre	Enjôo	Manchas	Dores
Luis	N	N	Peq	S
Laura	S	S	Grd	S

Definição formal: Encontrar uma função aproximada que mapeia as variáveis de entrada à saída.

Problema de classificação

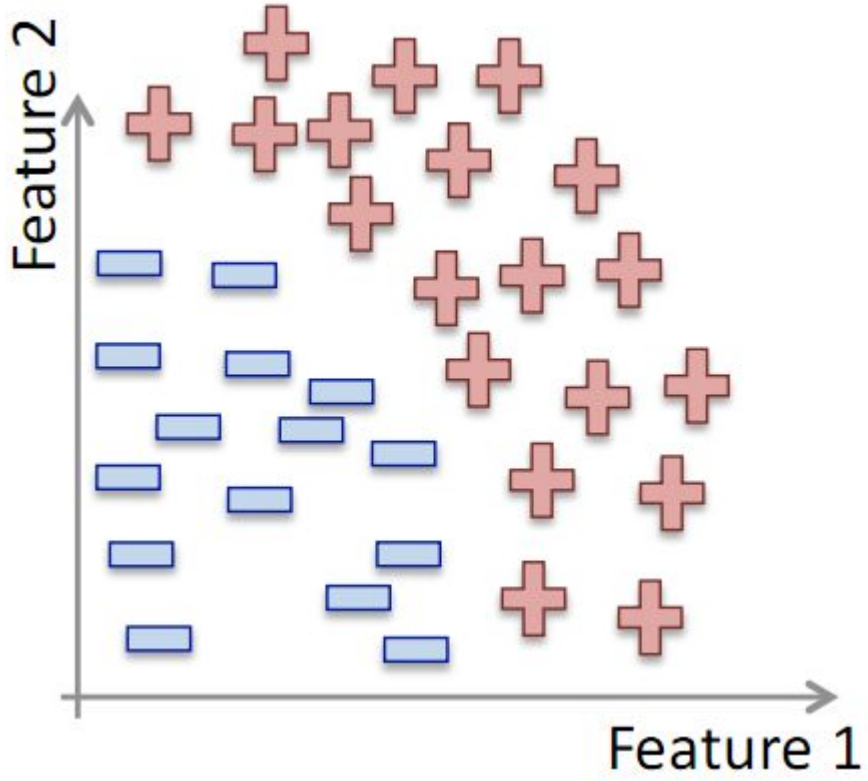
Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	S	S	Peq	S	Doente
Pedro	N	N	Grd	N	Saudável
Maria	S	S	Peq	N	Saudável
José	S	N	Grd	S	Doente
Ana	S	N	Peq	S	Saudável
Leila	N	N	Grd	S	Doente

Nome	Febre	Enjôo	Manchas	Dores
Luis	N	N	Peq	S
Laura	S	S	Grd	S

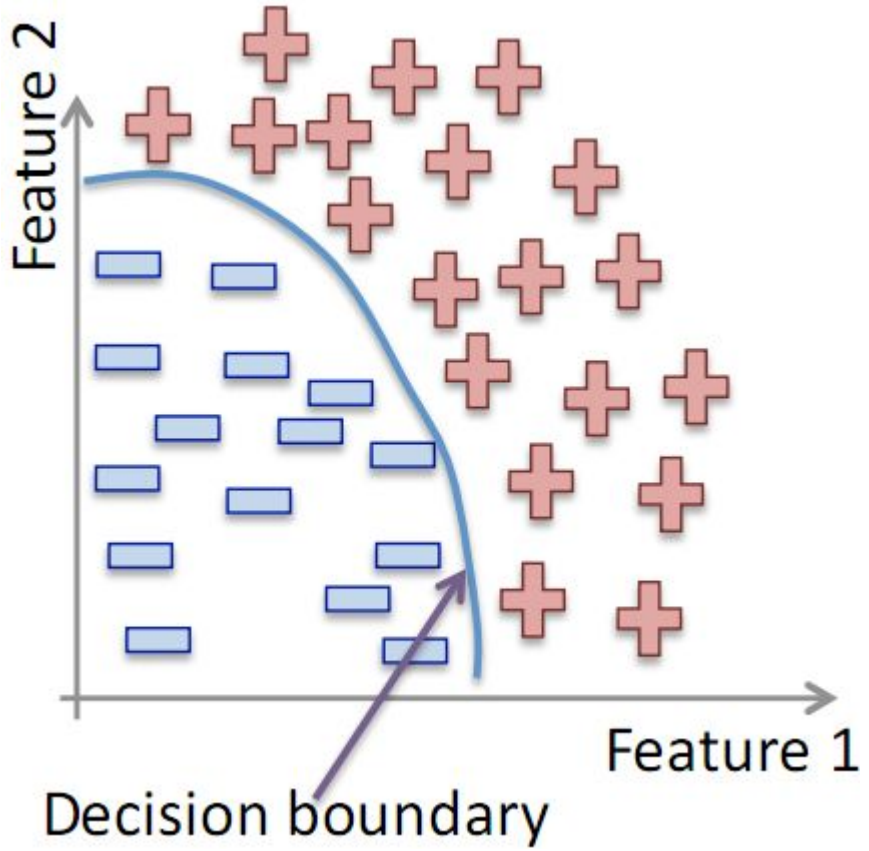
Definição formal: Encontrar uma função aproximada que mapeia as variáveis de entrada à saída.

Deve apresentar um número discreto de rótulos/classes!

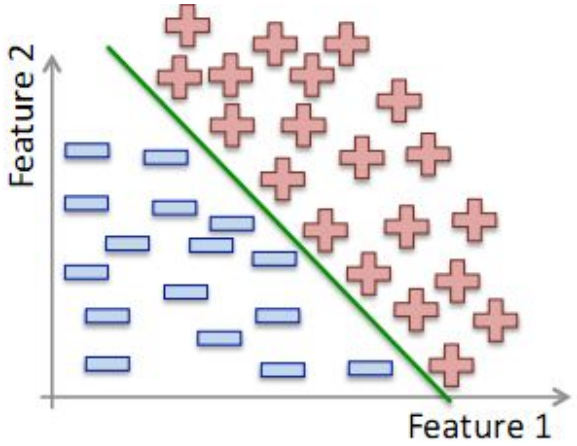
Problema de classificação

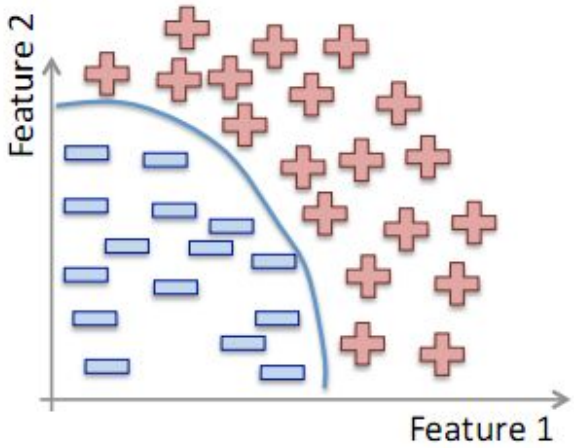


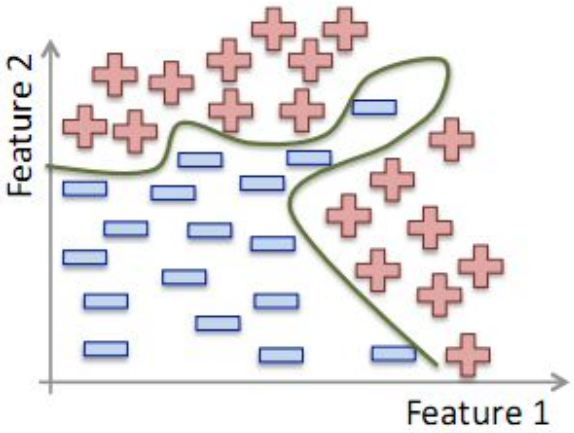
Problema de classificação

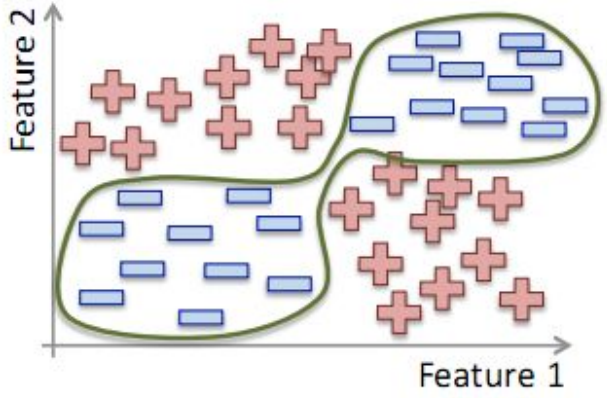


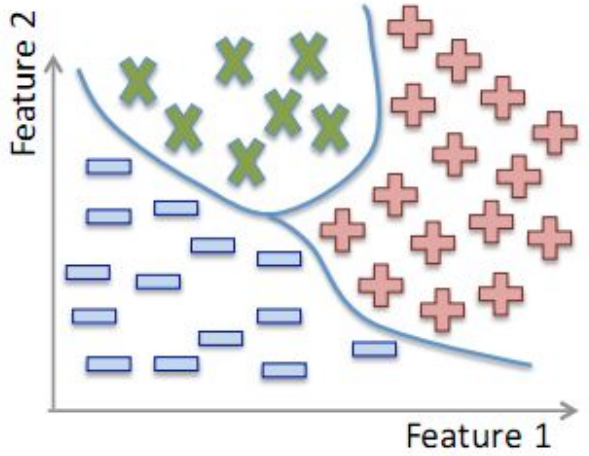
Problema de classificação



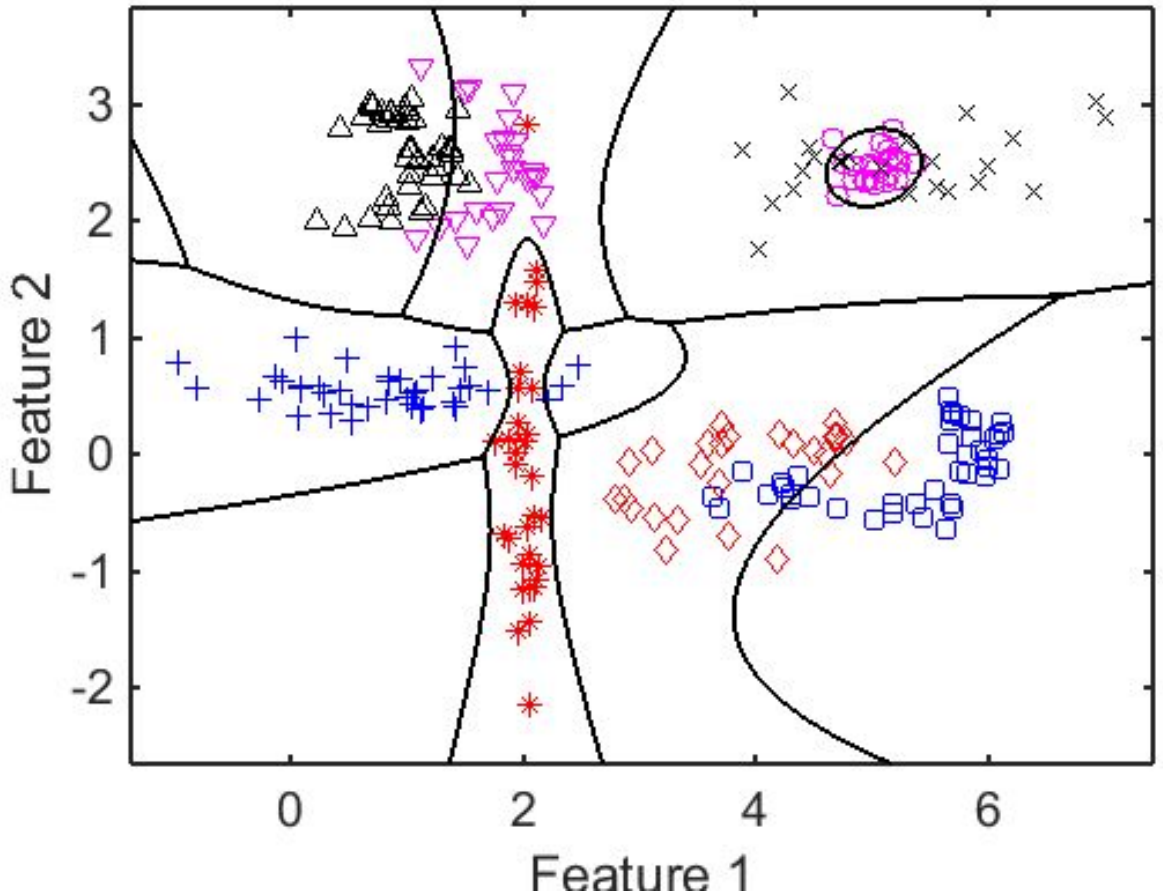




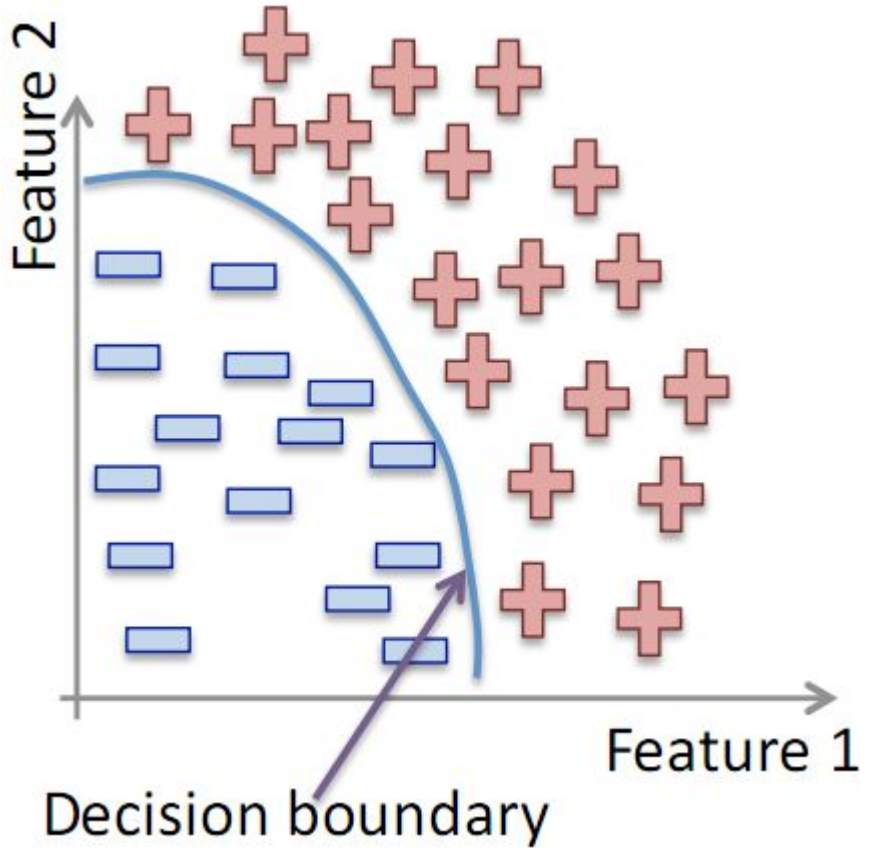




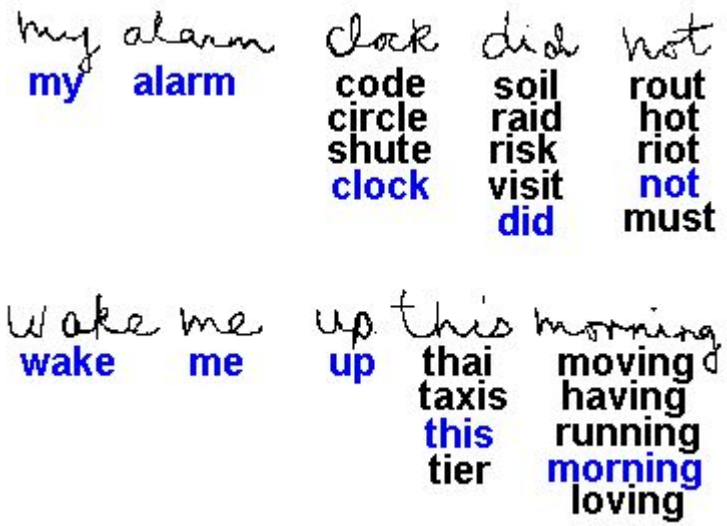
Problema de classificação



Problema de classificação



Exemplos de problemas



Vamos praticar!

E como seria com Regressão?

Aprendizagem não-supervisionada

Aprendizagem não-supervisionada

Aprender um modelo a partir de **dados não-rotulados.**

sepal		petal	
length	width	length	width
6.3	2.3	4.4	1.3
6.2	3.4	5.4	2.3
5.2	3.4	1.4	0.2
6.9	3.1	5.4	2.1
5.7	4.4	1.5	0.4
5.4	3.7	1.5	0.2
5	3.3	1.4	0.2
6.4	2.8	5.6	2.1
6	3	4.8	1.8
5.5	2.5	4	1.3

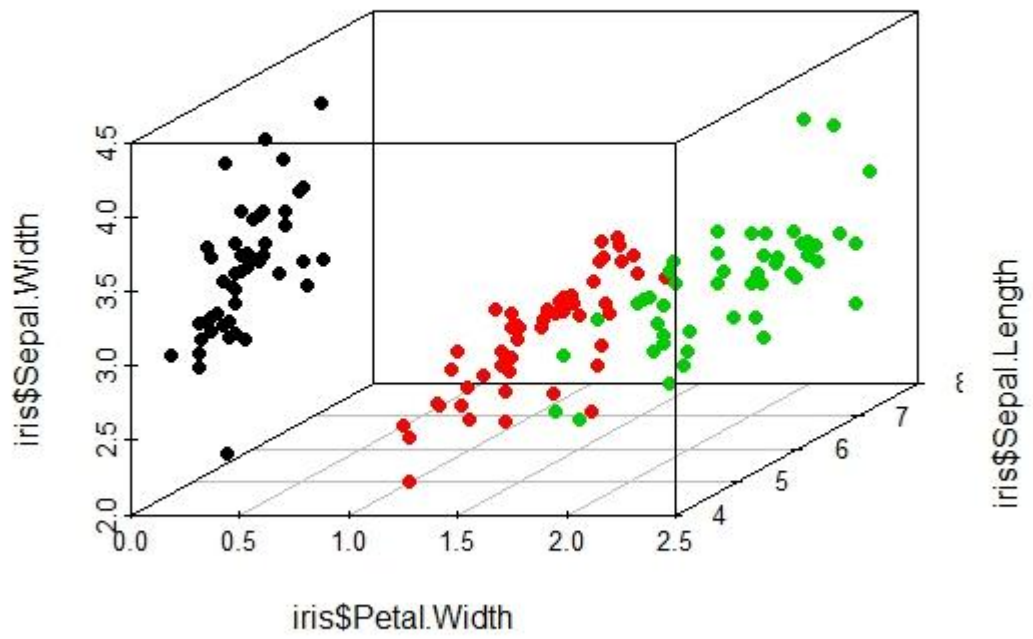
.

.

.

Aprendizagem não-supervisionada

A tarefa mais importante da aprendizagem não-supervisionada é o **agrupamento**.



Aprendizagem não-supervisionada

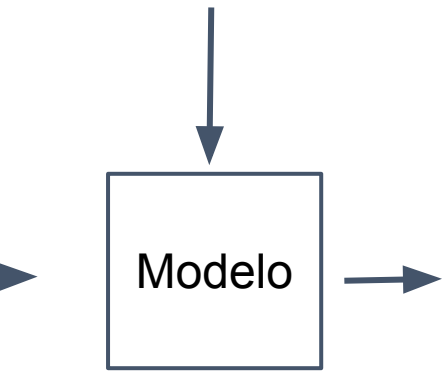
Treinamento

sepal		petal	
length	width	length	width
6.3	2.3	4.4	1.3
6.2	3.4	5.4	2.3
5.2	3.4	1.4	0.2
6.9	3.1	5.4	2.1
5.7	4.4	1.5	0.4
5.4	3.7	1.5	0.2
5	3.3	1.4	0.2
6.4	2.8	5.6	2.1
6	3	4.8	1.8
5.5	2.5	4	1.3

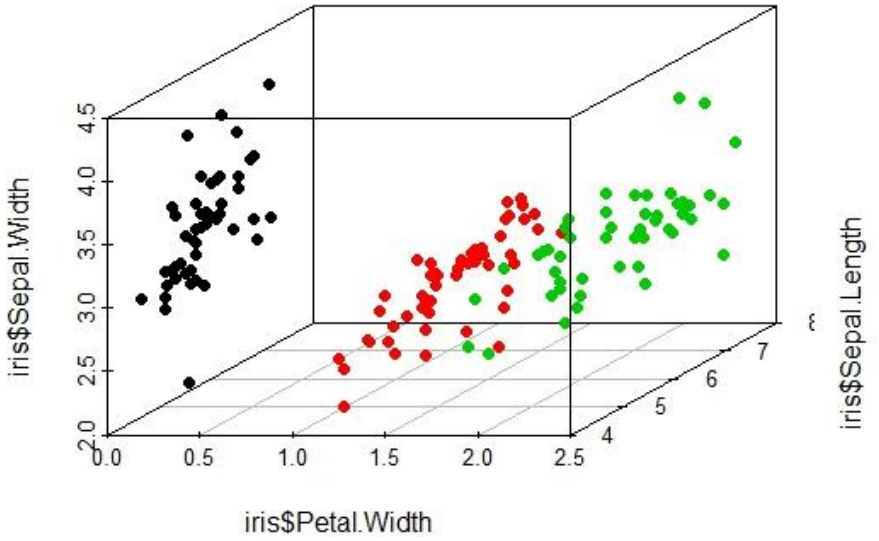
.
.
.

s.length s.width p.length p.width

(6.1, 1.3, 2.7, 0.8)



Modelo



?

Aprendizagem não-supervisionada

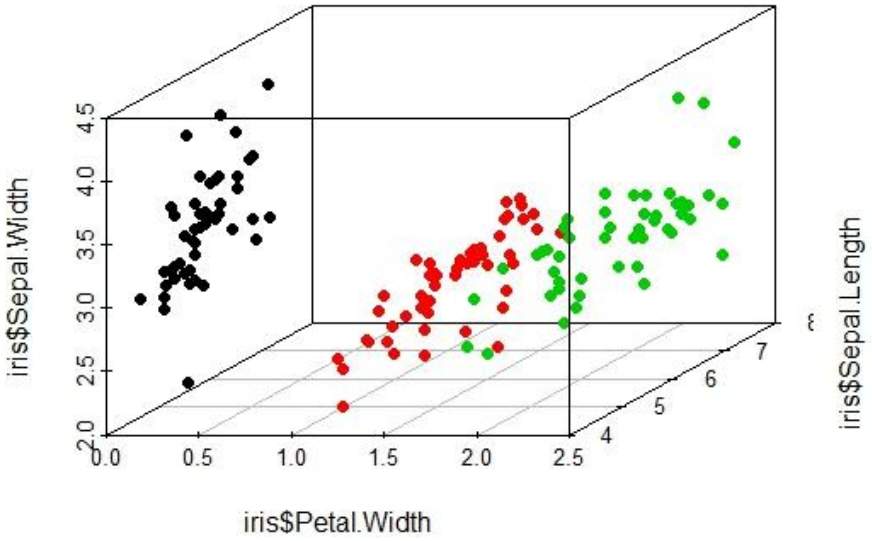
Treinamento

sepal		petal	
length	width	length	width
6.3	2.3	4.4	1.3
6.2	3.4	5.4	2.3
5.2	3.4	1.4	0.2
6.9	3.1	5.4	2.1
5.7	4.4	1.5	0.4
5.4	3.7	1.5	0.2
5	3.3	1.4	0.2
6.4	2.8	5.6	2.1
6	3	4.8	1.8
5.5	2.5	4	1.3

.
.
.

s.length s.width p.length p.width

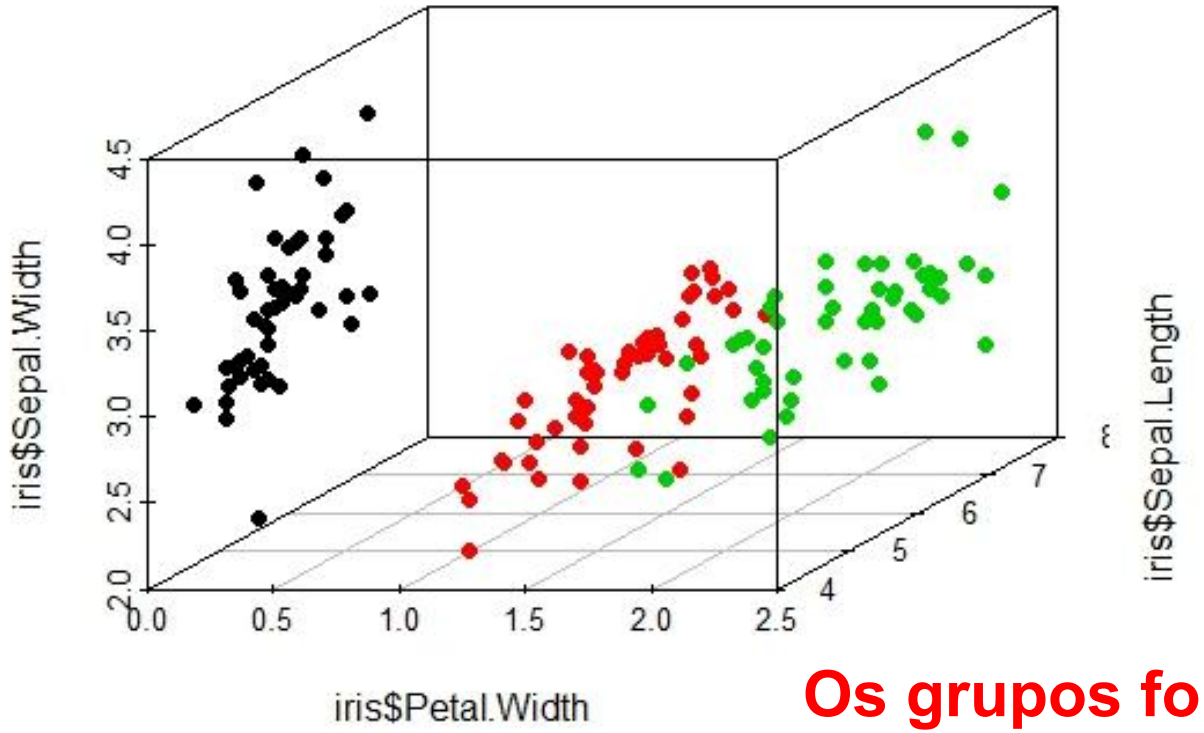
(6.1, 1.3, 2.7, 0.8)



Modelo

grupo preto

Aprendizagem não-supervisionada



Os grupos foram determinados. Ok!
Mas o que significa cada grupo?