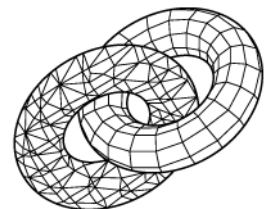
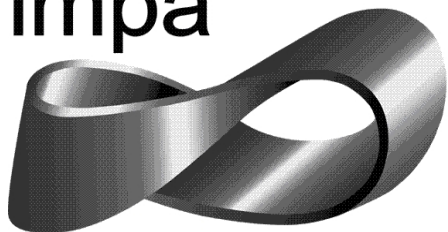


Ciência de Dados Aplicada

Aula 8: Introdução ao Aprendizado de Máquina

impa



VisgrafLab

O que é Aprendizado de Máquina?

O que é Aprendizado de Máquina?

Estudo de algoritmos que podem aprender pelo exemplo...

O que é Aprendizado de Máquina?

Estudo de algoritmos que podem aprender pelo exemplo...

... e em seguida generalizar com base nas experiências previamente observadas sobre uma determinada tarefa

O que é Aprendizado de Máquina?

Estudo de algoritmos que podem aprender pelo exemplo...

... e em seguida generalizar com base nas experiências previamente observadas sobre uma determinada tarefa

Principais exemplos:

- Reconhecimento de fala

- Carros autônomos

- Deteccção de Fraudes

- Reconhecimento de imagens (Medicina)

- Sistemas de Recomendação e pesquisa

Aprendizado de Máquina engloba estatística, ciência da computação e outras disciplinas

Métodos estatísticos

- Inferir conclusões sobre dados
- Estimar a confiabilidade das previsões

Computação

- Arquiteturas de computação em larga escala
- Algoritmos para captura, manipulação, indexação, combinação, recuperação e previsões sobre dados
- Pipelines de software que gerenciam a complexidade de várias subtarefas

Economia, biologia, psicologia

- Como um indivíduo ou sistema pode melhorar eficientemente seu desempenho em um determinado meio Ambiente?
- O que é aprendizado e como ele pode ser otimizado?

O que é Aprendizado de Máquina Aplicado?

Compreender os fundamentos e conceitos básicos, assim como o fluxo de trabalho (metodologia) do Machine Learning

Como aplicar corretamente o ML, isto é, como usar seus componentes e recursos

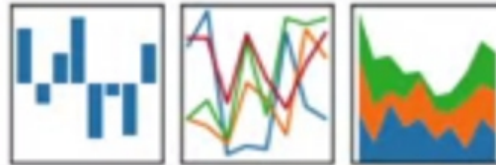
Aprendizado de máquina em Python usando o pacote scikit-learn

Python Tools

scikit-learn Homepage
<http://scikit-learn.org/>



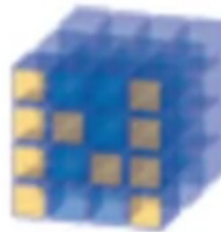
pandas
 $y_{it} = \beta^t x_{it} + \mu_i + \epsilon_{it}$



<http://pandas.pydata.org/>



<http://www.scipy.org/>



<http://www.numpy.org/>

matplotlib



<http://matplotlib.org/>

Tipos de problemas em Machine Learning

1. Aprendizado de máquina supervisionado

São apresentados ao computador exemplos de entradas e saídas desejadas

O objetivo é aprender uma regra geral que mapeia previsões futuras

Classificação (os valores-alvo são classes discretas)

Regressão (os valores-alvo são valores contínuos)

Aprendizado Supervisionado

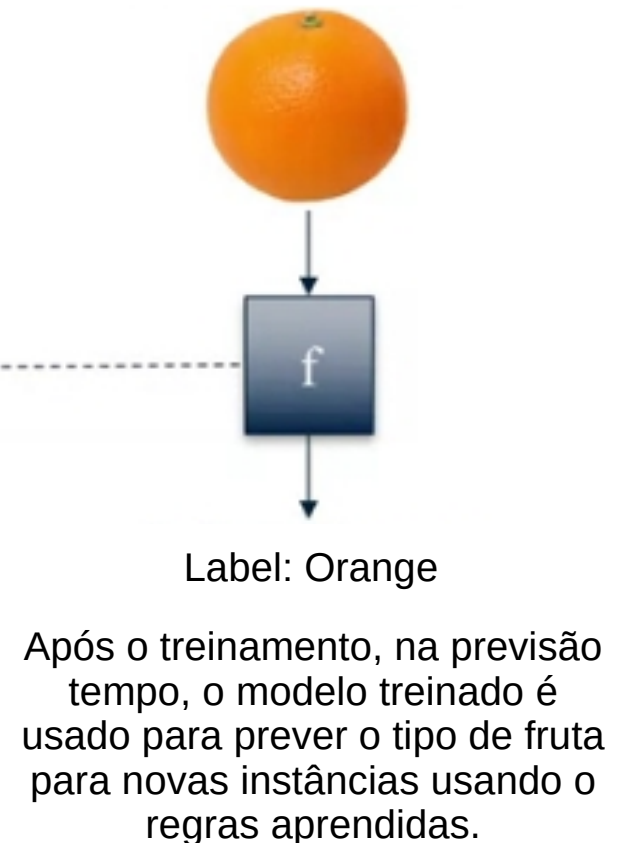
Classificação

Conjunto de Treinamento

X		Y	
Amostra		Valor-alvo (Label)	
	x_1	Apple	y_1
	x_2	Lemon	y_2
	x_3	Apple	y_3
	x_4	Orange	y_4



Amostra Futura



Tipos de problemas em Machine Learning

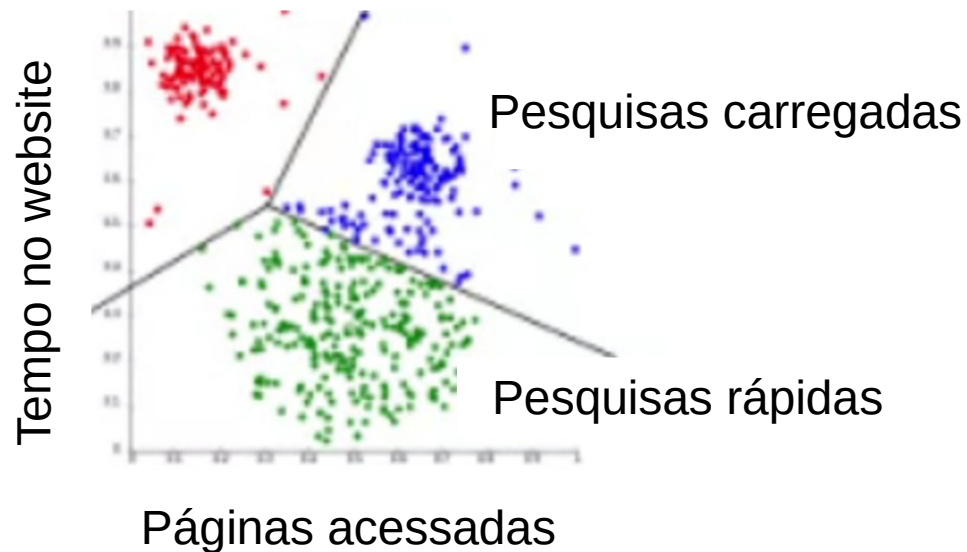
2. Aprendizado de máquina não supervisionado

Encontrar estrutura ou conhecimento útil nos dados quando não há ensinamento prévio ou rótulos disponíveis

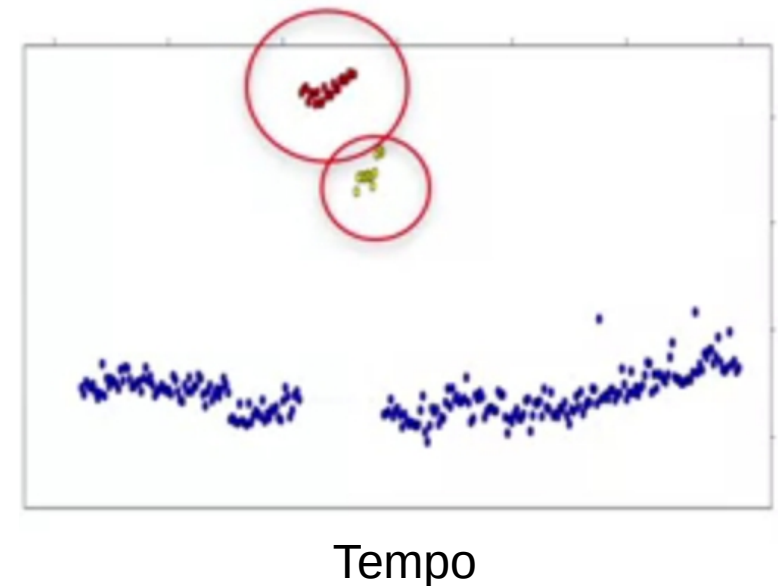
Grupos de instâncias semelhantes nos dados – cluster

Localizando padrões incomuns (detecção de outlier)

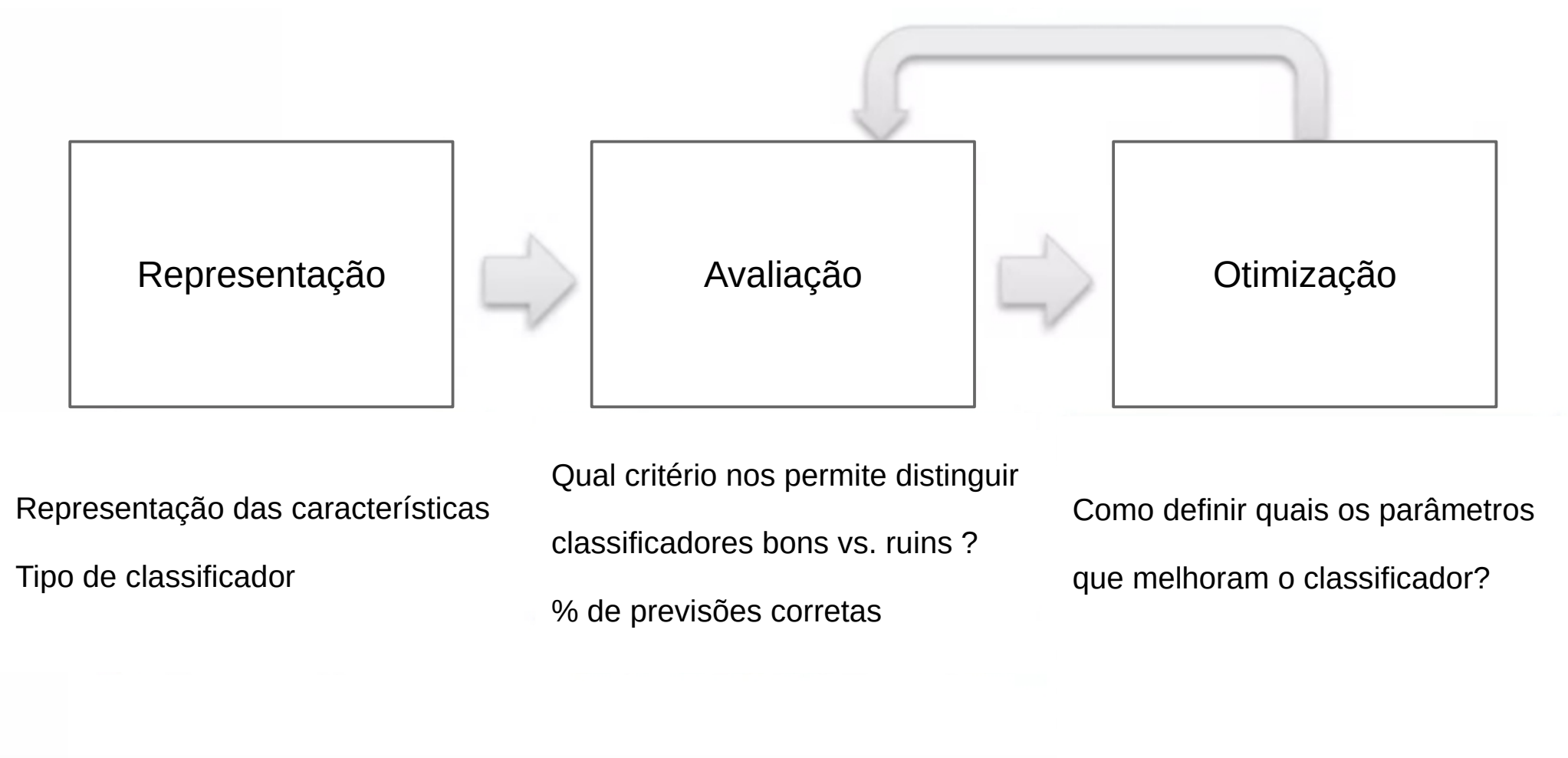
Pesquisas Cuidadosas



Acessos ao Servidor



Metodologia Básica para Aprendizado de Máquina



Representação de características

Email

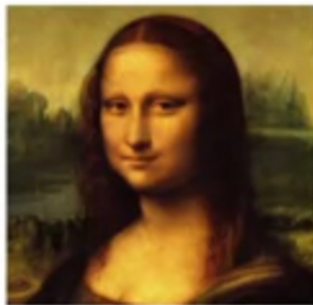
To: Chris Brooks
From: Daniel Romero
Subject: Next course offering
Hi Daniel,
Could you please send the outline for the
next course offering? Thanks! -- Chris

<u>Feature</u>	<u>Count</u>
to	1
chris	2
brooks	1
from	1
daniel	2
romero	1
the	2
...	

Representação da
Característica

Lista de palavras
com suas frequências

Figura



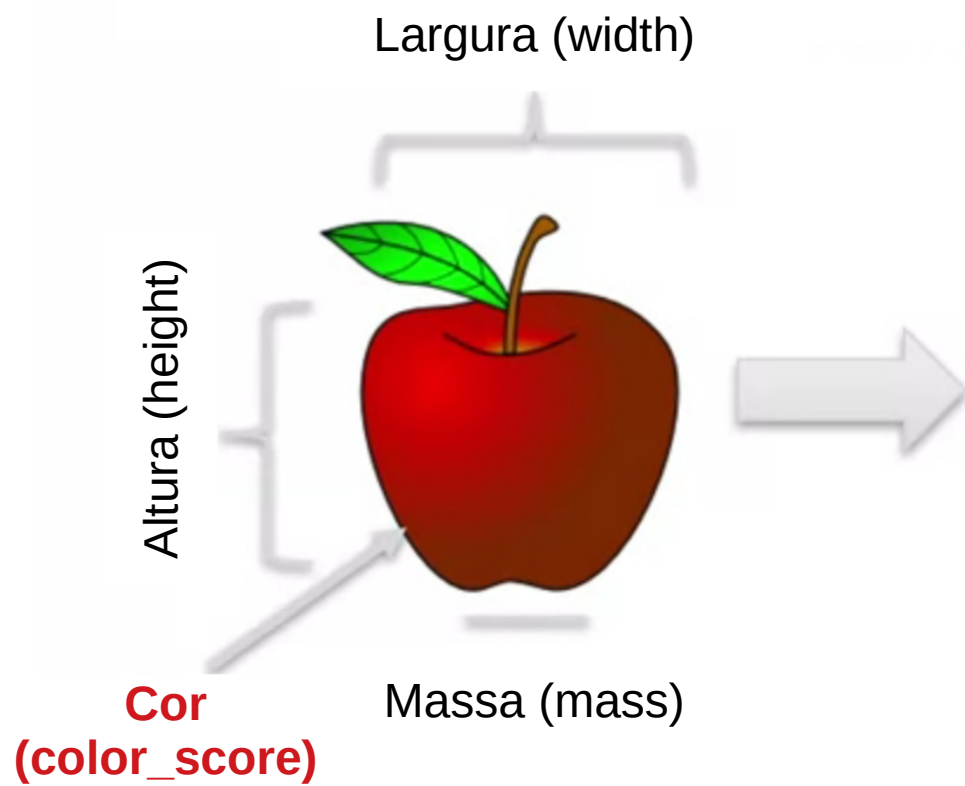
Matrix de valores
de cores (pixels)

Criaturas do Mar



<u>Feature</u>	<u>Value</u>
DorsalFin	Yes
MainColor	Orange
Stripes	Yes
StripeColor1	White
StripeColor2	Black
Length	4.3 cm

Conjunto de atributos



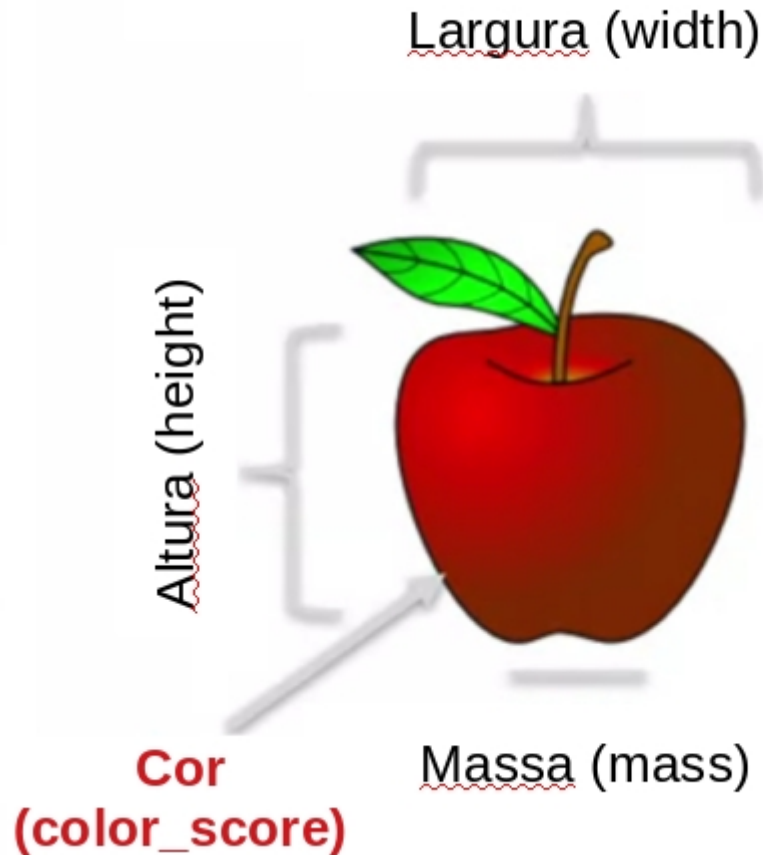
Representação da Característica

Label				Característica			
fruit_label	fruit_name	fruit_subtype		mass	width	height	color_score
18	1	apple	cripps_pink	162	7.5	7.1	0.83

Classificador

Prevê Label

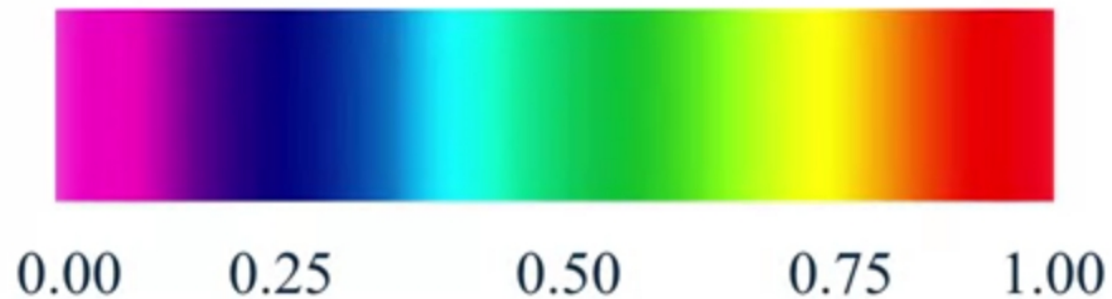
Base de dados - Frutas



	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	0.60
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	mandarin	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn	172	7.4	7.0	0.89
10	1	apple	braeburn	166	6.9	7.3	0.93
11	1	apple	braeburn	172	7.1	7.6	0.92
12	1	apple	braeburn	154	7.0	7.1	0.88
13	1	apple	golden_delicious	164	7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69
15	1	apple	golden_delicious	156	7.7	7.1	0.69
16	1	apple	golden_delicious	156	7.6	7.5	0.67

fruit_data_with_colors.txt

Representação da Característica Cor



cor	score
Vermelho	0.85 - 1.00
Laranja	0.75 - 0.85
Amarelo	0.65 - 0.75
Verde	0.45 - 0.65

Conjunto de Treinamento vs. Conjunto de Testes



O algoritmo K-NN (vizinho mais próximo) precisa de quatro coisas especificadas

- Uma métrica de distância
Euclidiana
- Quantos vizinhos 'mais próximos' para olhar?
- Função de ponderação opcional nos pontos vizinhos
- Como agregar as classes de pontos vizinhos
Voto por maioria simples
(Classe com mais representantes entre os vizinhos mais próximos)