

Inteligência Artificial

Tópico 04 - Parte 01

Aprendizado de Máquina - Introdução e Conjuntos de Dados

Profa. Dra. Priscila Tiemi Maeda Saito
✉ priscilasaito@ufscar.br

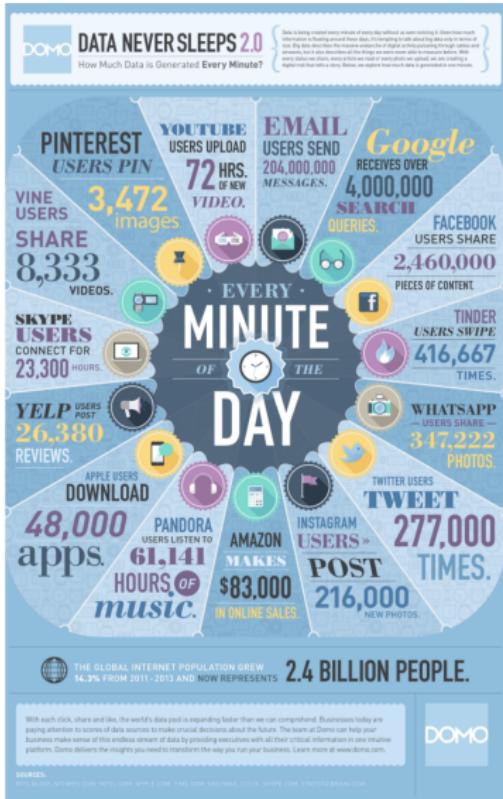
Roteiro

1 Aprendizado de Máquina

Introdução

- **Pessoas geram dados** a todo momento, sem perceber
 - ▶ Navegação pela internet
 - ▶ Consulta ao médico
 - ▶ Realização de cartão fidelidade
 - ★ empresa aérea, supermercado, lojas, entre outros
 - ▶ Compra com cartão de débito ou crédito
- **Máquinas estão constantemente **coletando e armazenando dados****

Dados Nunca Dormem¹



¹Dados 2014/2021. Origem: Domo business management platform. www.domo.com

Análise de Dados

- Análise de dados por **seres humanos**
 - ▶ falta de especialistas
 - ▶ custo elevado
 - ▶ subjetividade
 - ▶ grande volume
- **Técnicas tradicionais** para análise de dados
 - ▶ planilhas
 - ▶ sistemas de gerenciamento de bancos de dados

- **Técnicas tradicionais** de análise de dados permitem apenas consultas simples

- ▶ quantos itens de um produto específico foram vendidos em um dado dia?
- ▶ não conseguem responder consultas como:
 - ★ que filme novo eu gostaria de assistir?
 - ★ um dado tecido pode apresentar um tumor?
 - ★ qual a estrutura terciária de uma nova proteína?

técnicas mais sofisticadas → capazes de extrair conhecimento de grandes conjuntos de dados

Aprendizado de Máquina

- **Sub-campo da inteligência artificial**
 - ▶ dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador
“aprender” ou aperfeiçoar seu desempenho em alguma tarefa
- Essencial em áreas de reconhecimento de padrões e visão computacional



Aprendizado de Máquina

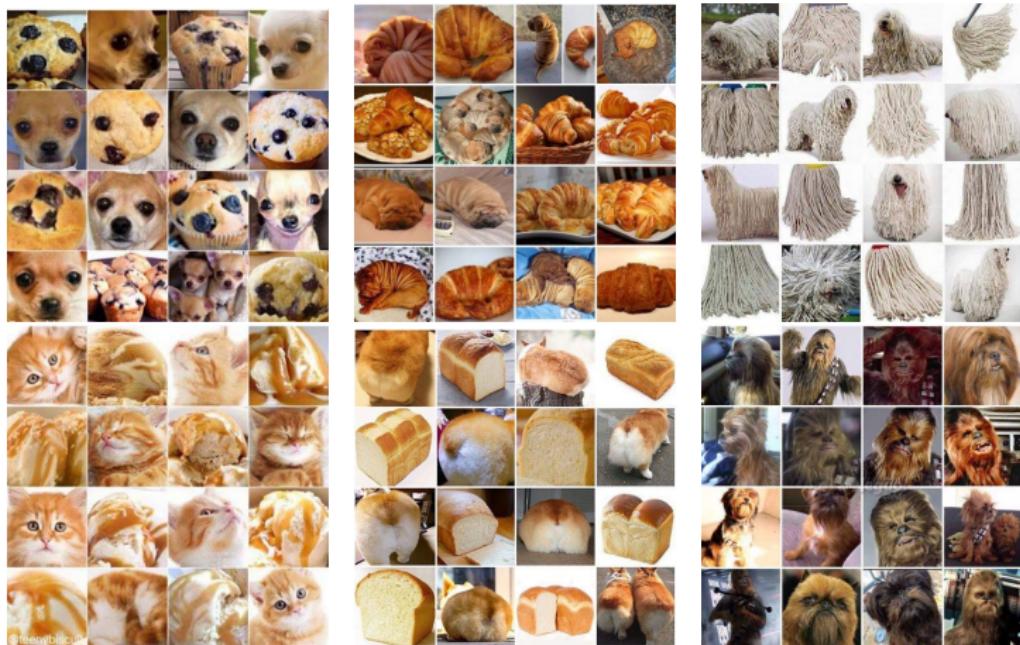
- **Sub-campo da inteligência artificial**
 - ▶ dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador
“aprender” ou aperfeiçoar seu desempenho em alguma tarefa
- Essencial em áreas de reconhecimento de padrões e visão computacional



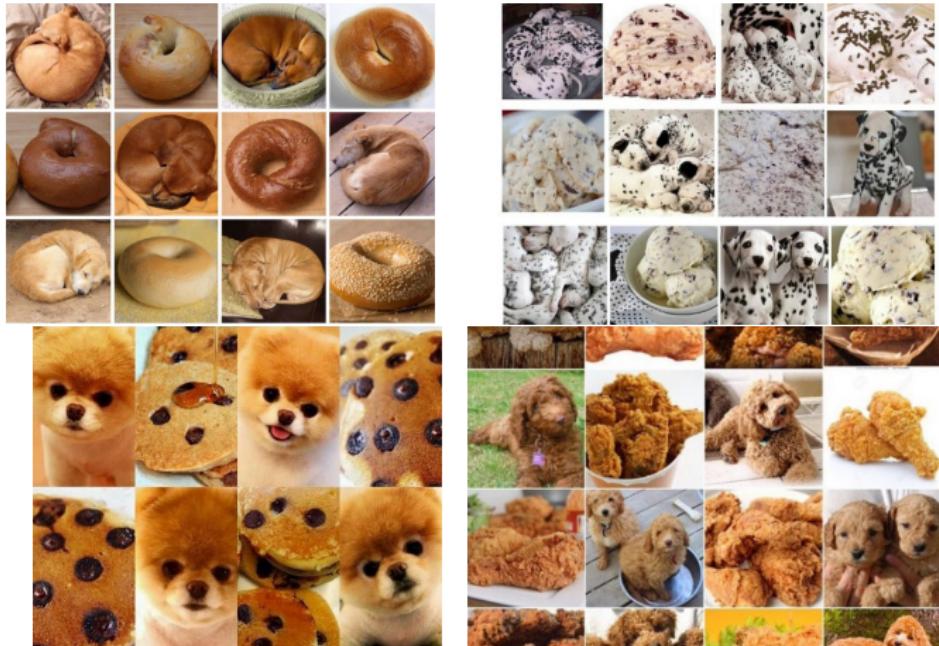
Aprendizado de Máquina



Aprendizado de Máquina



Aprendizado de Máquina



Aprendizado de Máquina



Realidade de fato x imagem da realidade

Aprendizado de Máquina

- Investiga técnicas computacionais capazes de adquirir automaticamente
 - ▶ novas habilidades, conhecimentos e formas de organizar o conhecimento existente

Definição

Área de pesquisa que dá aos computadores a habilidade de aprender sem ser explicitamente programado (Arthur Samuel, 1959)

Definição

Técnicas capazes de melhorar seu **desempenho** em uma dada **tarefa** utilizando **experiências** prévias (Mitchell, 1997)

Aplicações de Aprendizado de Máquina

- Programas baseados em AM têm sido bem sucedidos para:
 - ▶ análise de redes sociais
 - ▶ análise de dados biológicos
 - ▶ detecção de fraudes
 - ▶ diagnóstico médico
 - ▶ biometria
 - ▶ recomendação de filmes e séries

Aprendizado de Máquina

- Algoritmos de AM aprendem a partir de um conjunto de exemplos
 - ▶ indução de hipóteses ou modelos
 - ★ aplicados depois a novos dados
- Todo algoritmo de AM indutivo possui um **viés**
 - ▶ tendência a privilegiar uma dada hipótese ou conjunto de hipóteses

Viés Indutivo

- Algoritmos de AM precisam ter um **viés indutivo**
 - ▶ necessário para restringir o espaço de busca
 - ▶ se não houvesse viés, não haveria generalização
 - ★ regras/equações seriam especializados para os exemplos específicos

Algoritmos de AM

- Extraem conhecimento de um conjunto de dados
 - ▶ novo, útil e relevante
 - ▶ precisam ser tratados
 - ▶ precisam ser representativos
 - ★ cobrir situações que possam ocorrer
 - ▶ podem ser estruturados ou não

Conjuntos de Dados

- **Estruturados**

- ▶ mais facilmente analisados por técnicas de AM
- ▶ ex.: planilhas e tabelas atributo-valor

- **Não estruturados**

- ▶ mais facilmente analisados por seres humanos
 - ★ para AM, são geralmente convertidos em dados estruturados
- ▶ Ex.: sequência de DNA, textos, conteúdo de página web, emails

Conjuntos de Dados

Atributos de entrada (preditivos)

Exemplos
(objetos,
instâncias)

	Nome	Batim.	Temp.	Idade	Peso	Pressão	Diagn.
	João	70	37	70	94	12	Saudável
	Maria	38	39	30	40	14	Doente
	José	39	38.5	70	85	18	Doente
	Sílvia	38	37.5	15	60	13	Saudável
	Pedro	37	40	90	78	14	Doente

Atributo alvo

Conjuntos de Dados

Preparação dos dados

- Fase que antecede o processo de aprendizagem, para facilitar ou melhorar o processo
- Exemplos:
 - ▶ remover exemplos incorretos
 - ▶ transformar o formato dos exemplos para que possam ser usados com um determinado modelo
 - ▶ selecionar um subconjunto de atributos relevantes (FSS - Feature Subset Selection)

Conjuntos de Dados

- **Ruídos ou outliers**, exemplos imperfeitos que podem ser derivados do processo de aquisição, transformação ou rotulação das classes
- Ex.: exemplos com os mesmos atributos mas com classes diferentes

	x_1	x_2	x_3	x_4	y
	overcast	19	65	yes	dont_go
	rain	19	70	yes	dont_go
	rain	23	80	yes	dont_go
	sunny	23	95	no	dont_go
	sunny	28	91	yes	dont_go
	sunny	30	85	no	dont_go
	overcast	19	65	yes	go
	rain	21	80	no	go
	rain	22	95	no	go
	sunny	22	70	no	go
	overcast	23	90	yes	go
	rain	25	81	no	go
	sunny	25	72	yes	go
	overcast	26	75	no	go
	overcast	29	78	no	go

Conjuntos de Dados

- **Estatísticas** comuns no trato com dados multivariados
- Tais estatísticas se aplicam, de modo geral, a cada atributo do vetor de atributos

Conjuntos de Dados

• Amplitude Total

- ▶ trata-se da **dispersão** entre o maior e o menor valor de um determinado atributo

$$R = \max_j X_i(j) - \min_j X_i(j)$$

- ▶ Exemplo: para um atributo “idade”
 - ★ 20, 25, 27, 28, 40, 30, 31 e 19
 - ★ $R = 40 - 19 = 21$

Conjuntos de Dados

- **Média ou esperança**

- ▶ é o valor que aponta para onde mais se concentram os dados de uma distribuição
- ▶ pode também ser chamado de centróide
- ▶ a **média aritmética** é a forma mais simples de calcular uma média

Conjuntos de Dados

• Mediana

- ▶ dado um conjunto de dados organizados em ordem crescente, a **mediana** é o valor que ocupa a posição central do conjunto
 - ★ dado o conjunto {2, 2, 3, 5, 5, **6**, 7, 7, 9, 9, 10}
 - ★ mediana será igual a 6
- ▶ se a quantidade de valores é ímpar
 - ★ **mediana** será simplesmente o **valor central**
- ▶ se a quantidade de valores é par
 - ★ **mediana** será a **média** dos dois **valores centrais**
 - ★ dado o conjunto {0, 1, 1, 2, 3, **4**, **5**, 5, 6, 6, 7, 8}
 - ★ mediana será igual a $(4+5)/2 = 4.5$

Conjuntos de Dados

• Moda

- ▶ dado um conjunto de dados, a **moda** é o valor com maior frequência individual, ou seja, aquele que mais se repete dentro do conjunto de dados
 - ★ dado o conjunto {0, 1, 2, 2, 2, 3, 4, 5, 5, 6, 6, 7, 8}
 - ★ moda será igual a 2

Conjuntos de Dados

- **Normalizações**

- ▶ Min-Max
- ▶ Z-Score

- **Normalização Min-Max**

- ▶ valores do atributo são normalizados linearmente (entre [0,1]) com base nos valores máximo e mínimo

$$v' = \frac{(v - min1)}{max1 - min1}$$

- ▶ v' = novo valor do atributo 1
- ▶ v = valor original do atributo 1
- ▶ $min1$ = valor mínimo do atributo 1
- ▶ $max1$ = valor máximo do atributo 1

• Normalização Z-Score

- ▶ valores do atributo são normalizados com base na média e no desvio padrão do atributo

$$v' = \frac{(v - med1)}{desv - pad1}$$

- ▶ v' = novo valor do atributo 1
- ▶ v = valor original do atributo 1
- ▶ $med1$ = média do atributo 1
- ▶ $desv - pad1$ = desvio padrão do atributo 1

Abordagens de Aprendizado

- **Supervisionado**

- ▶ tarefa preditiva (mais comum) ou descritiva
- ▶ ensina ao modelo o que ele deve fazer
 - ★ fornece, para cada entrada, a saída desejada (correta)

- **Não supervisionado**

- ▶ tarefa descritiva (mais comum) ou preditiva
- ▶ algoritmo aprende por si só

- **Semi-supervisionado**

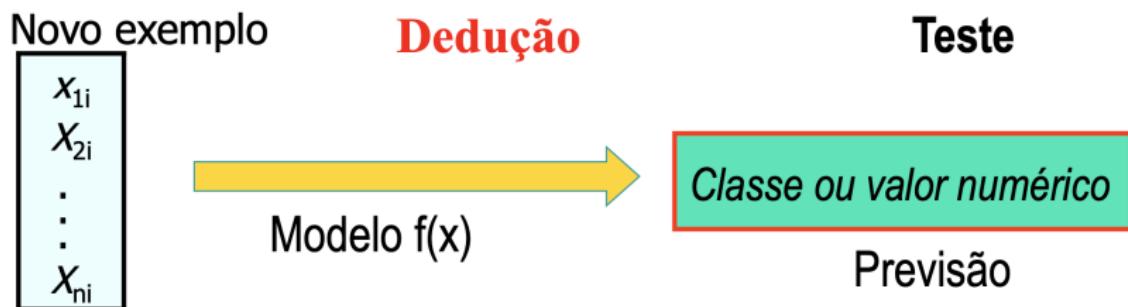
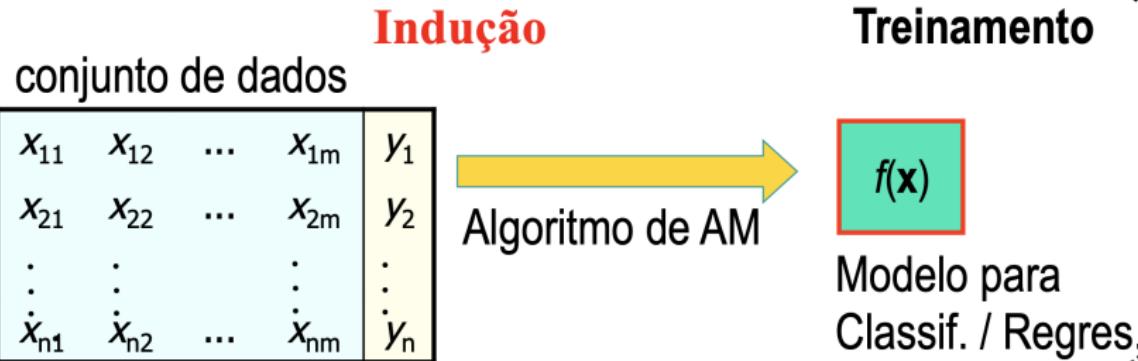
- **Aprendizado ativo**

- **Reforço**

Algoritmos de AM preditivos

- Induzem modelos (funções) preditivas
 - ▶ dados de **treinamento**
- Modelo pode ser aplicado a novos dados
 - ▶ dados de **teste**
 - ▶ predição
- **Classificação e regressão**

Algoritmos de AM preditivos



Regressão

- Objetivos: aprender uma função que mapeia um exemplo em um **valor real**
 - ▶ caso especial: análise de séries temporais

Exemplos

- prever valor de mercado de um imóvel
- prever o lucro de um empréstimo bancário
- prever tempo de internação de um paciente

Regressão

- Predição de preços de casa



\$70000

Regressão

- Predição de preços de casa



\$160000

Regressão

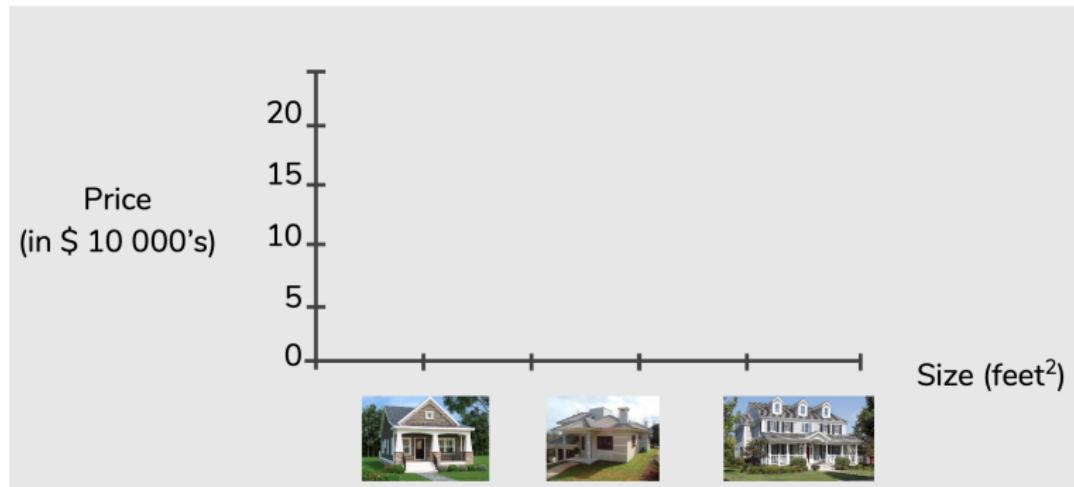
- Predição de preços de casa



???

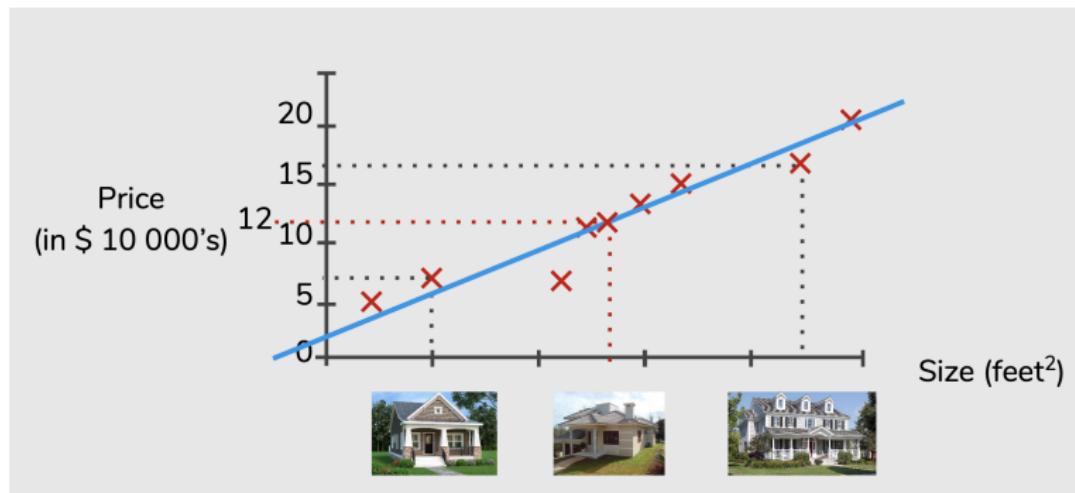
Regressão

- Predição de preços de casa



Regressão

- Predição de preços de casa
- Regressão Linear



House Sales in King County, USA

Predict house price using regression



harlfoxem • last updated a year ago

108

[Overview](#)[Kernels](#)[Discussion](#)[Activity](#)[Download \(778 KB\)](#)[New Kernel](#)[Tags](#)[finance](#)[home](#)[small](#)[featured](#)[Kernels](#)

[Feature Ranking w Random... 55](#)
run 2 days ago votes

[Step by Step House Price Pre... 41](#)
run 7 months ago votes

[House_Price_Prediction_Part_1 29](#)
run a year ago votes

[Discussion](#)

[Variable explanation 15](#)
6 days ago replies

[King County Geoclustering... 7](#)
9 days ago replies

[RF, RFE, linear models|特征... 3](#)
10 days ago replies

[Top Contributors](#)

harlfoxem

1st



Anisotropic

2nd



ArmanUygur

3rd

[Recent Activity](#)

[Jois Leonida Lobo](#) Ran version 10 of kernel [HousePricePrediction_SimpleLinearRegression](#) 13 hours ago

[Anisotropic](#) Ran version 41 of kernel [Feature Ranking w RandomForest, RFE, linear models](#) 2 days ago

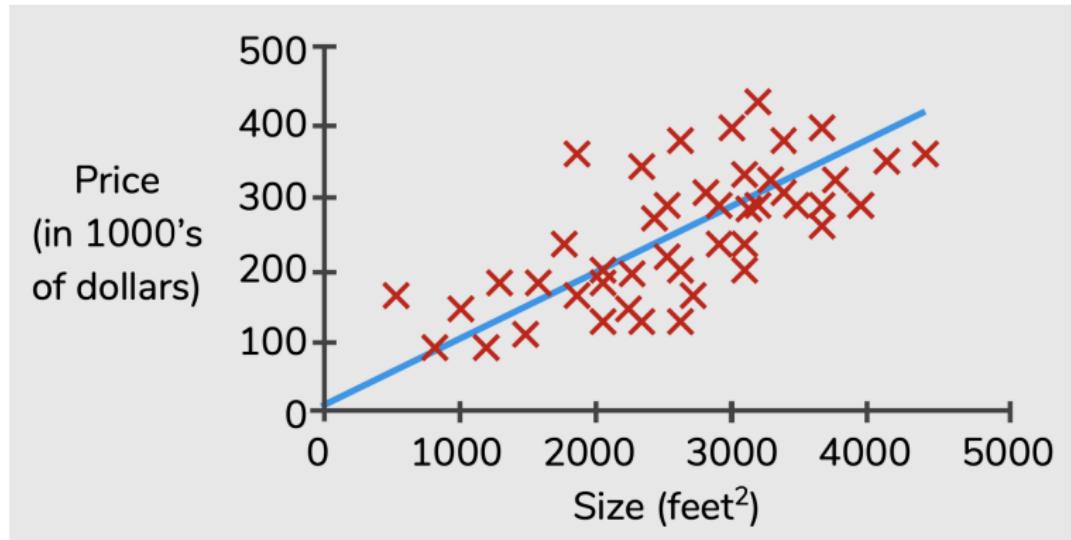
[DavidTan](#) Commented on dataset discussion [Variable explanation](#) 6 days ago

[Harsh Tyagi](#) Ran version 3 of kernel [King County's Housing Market, various techniques.](#) 8 days ago



Regressão

- Predição de preços de casa



Aprendizado Supervisionado

Fornece a “resposta correta” para cada exemplo nos dados

Regressão

Prediz saída com valor real

Regressão

- Treinamento a partir de conjunto de preços de casas

Tamanho em feet ² (x)	Preço (\$) em 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

Notação

m = número de exemplos de treinamento

x 's = variável de entrada / características

y 's = variável de saída / “target”

Classificação

- Objetivos: aprender uma função que associa descrição de um exemplo a uma **classe**

Exemplos

- definir a função de uma proteína
- distinguir emails entre spam ou não spam
- definir se um paciente tem ou não uma doença

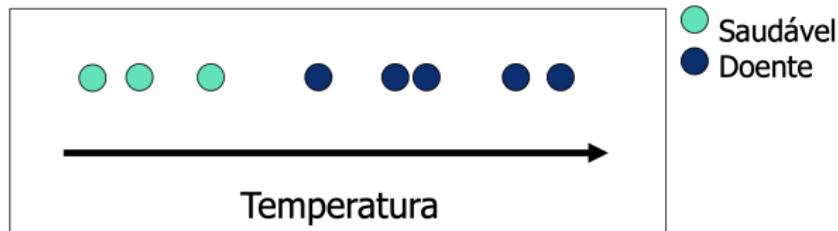
Classificação

- Posto médico X

- ▶ tem um histórico de vários atendimentos e diagnósticos
- ▶ João, ao sentir alguns sintomas, vai ao posto para uma consulta médica
- ▶ o único médico, faltou
 - ★ mas uma enfermeira pode anotar os sintomas
- ▶ é possível fazer um **pré-diagnóstico?**

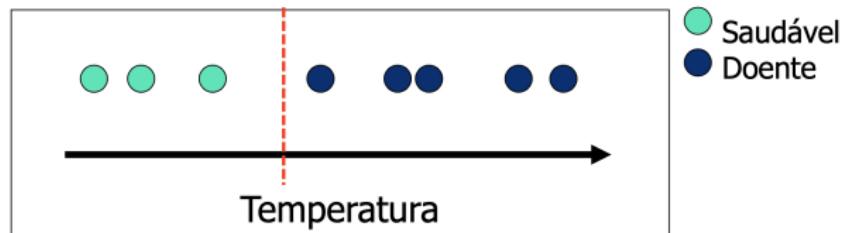
Classificação

- Sintomas coletados pela enfermeira
 - ▶ temperatura
- Forma mais simples



Classificação

- Forma mais simples



Função estimada: diagnóstico = $f(\text{temperatura})$

se temperatura > t
então doente
senão saudável

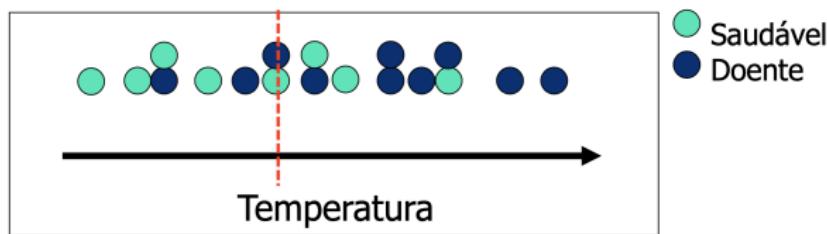
Classificação

- Basta encontrar um valor de temperatura que separa
 - ▶ doentes
 - ▶ saudáveis

Problema de classificação é simples assim?

Classificação

- Problema pode não ser tão simples

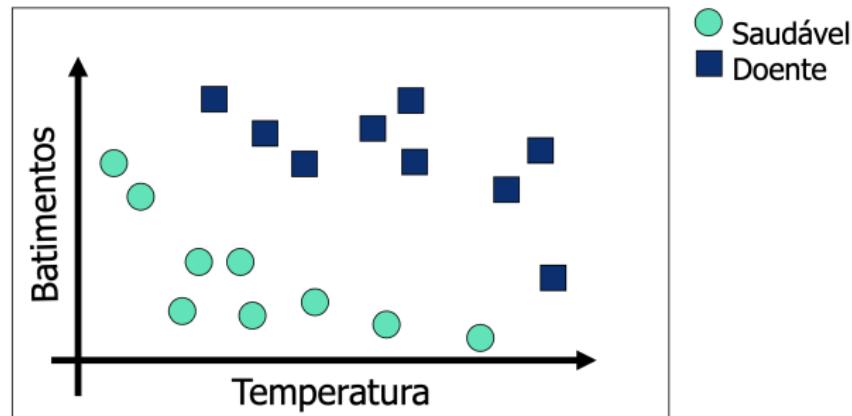


Alternativa

considerar outros sintomas para o diagnóstico

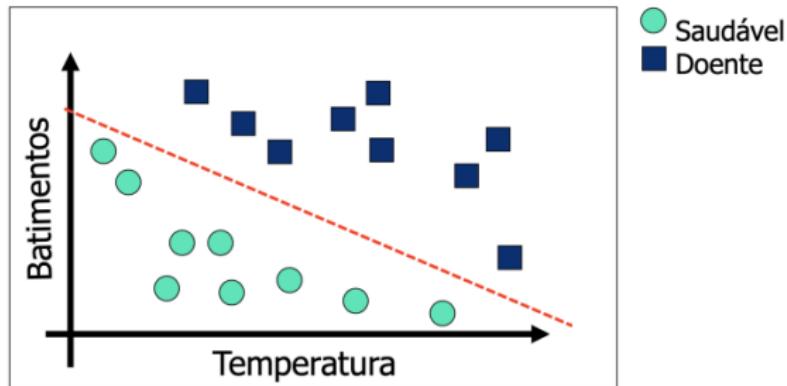
Classificação

- Sintomas coletados pela enfermeira:
 - ▶ batimentos cardíacos
 - ▶ temperatura
- Incluir número de batimentos



Classificação

- Função linear permite diagnóstico



Nova função:

se $a \cdot t + b > 0$

então doente

senão saudável

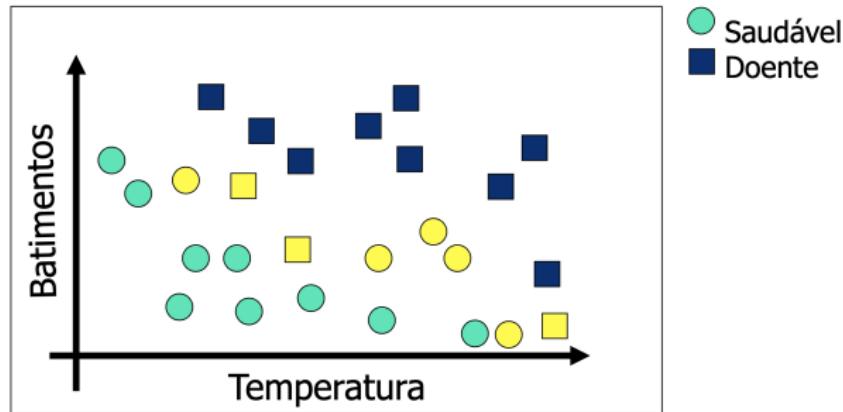
Classificação

- Basta encontrar uma função linear que separa doentes de saudáveis
 - ▶ inclinação da reta e ponto onde cruza o eixo da ordenada
- Espaço de pacientes
 - ▶ ordenada: número de batimentos
 - ▶ abscissa: temperatura

Toda tarefa de classificação é simples assim?

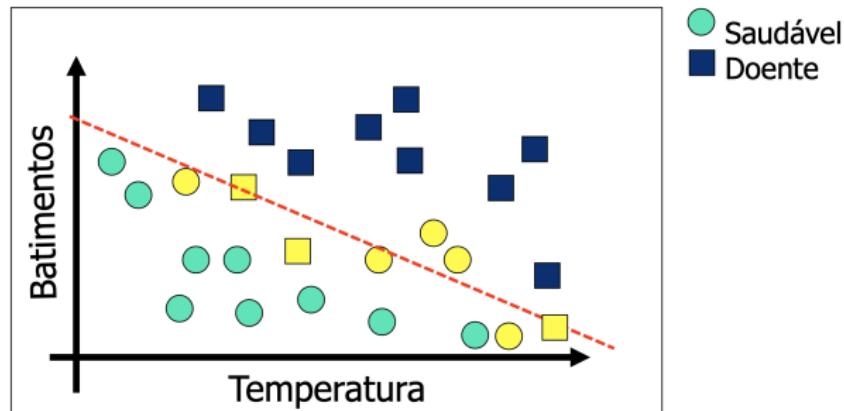
Classificação

- Supor inclusão de outros pacientes



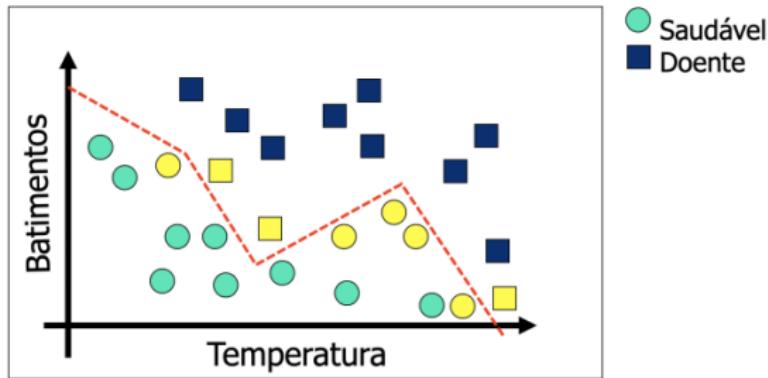
Classificação

- Supor inclusão de outros pacientes



Classificação

- Supor inclusão de outros pacientes

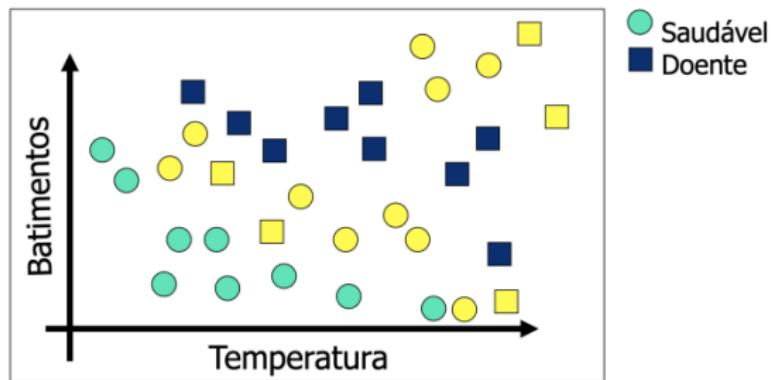


Nova função:

Muito complexa

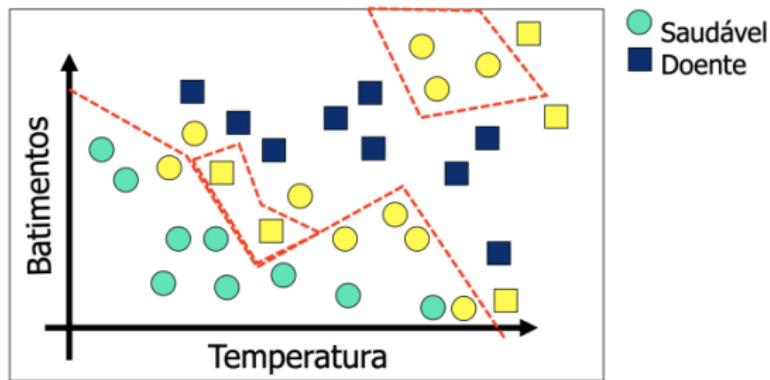
Classificação

- Supor inclusão de mais pacientes



Classificação

- Supor inclusão de outros pacientes



Nova função:

Muito complexa

Classificação

- Função para definir fronteira de decisão se torna **mais complexa**
 - ▶ difícil de obter por técnicas tradicionais
- Algoritmos de AM utilizam **heurísticas** para procurar essas funções
- **Conjuntos e atributos** utilizados podem
não representar bem a tarefa
 - ▶ dificultando a indução de bons modelos

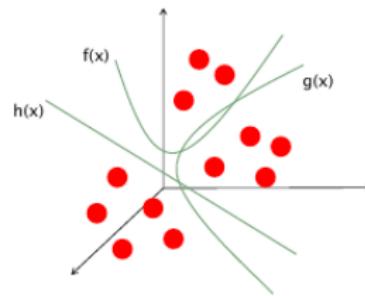
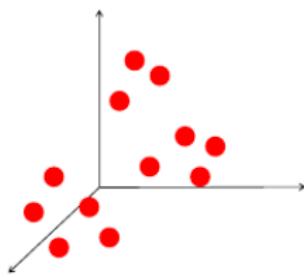
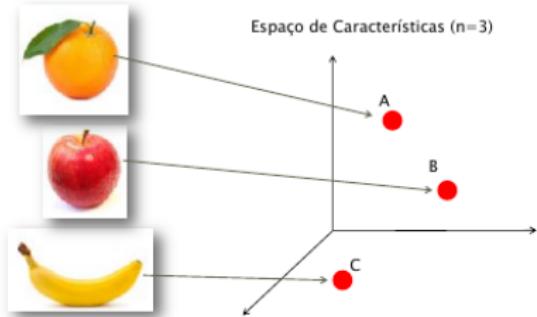
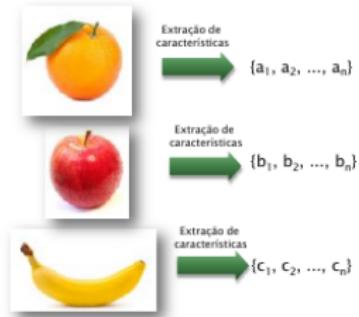
Classificação

- Sintomas (**atributos**) que poderiam permitir um melhor modelo para diagnóstico:
 - ▶ batimentos cardíacos
 - ▶ idade
 - ▶ peso
 - ▶ pressão
 - ▶ temperatura
 - ▶ taxas em uma amostra de sangue

Classificação

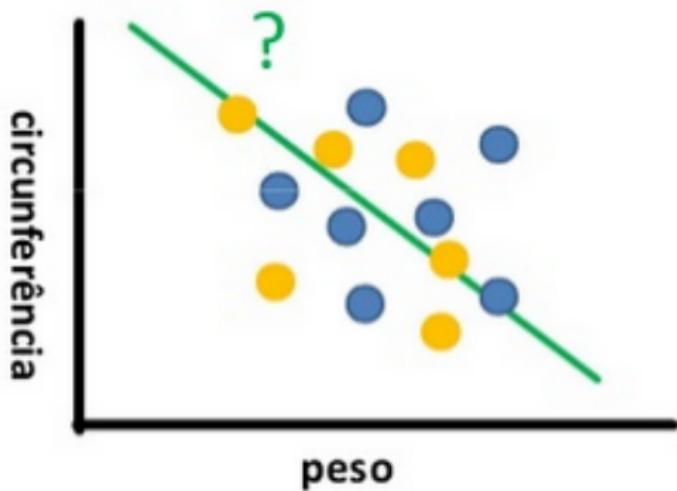
- Atributos preditivos procuram descrever a tarefa a ser resolvida
 - ▶ em geral, quanto **mais atributos** são extraídos, **melhor**
 - ★ facilitam a indução de bons modelos
 - ▶ no entanto
 - ★ dificultam visualizar a distribuição dos dados
 - ★ podem incluir **atributos irrelevantes, redundantes, ...**
 - ★ **maldição da dimensionalidade**

Classificação

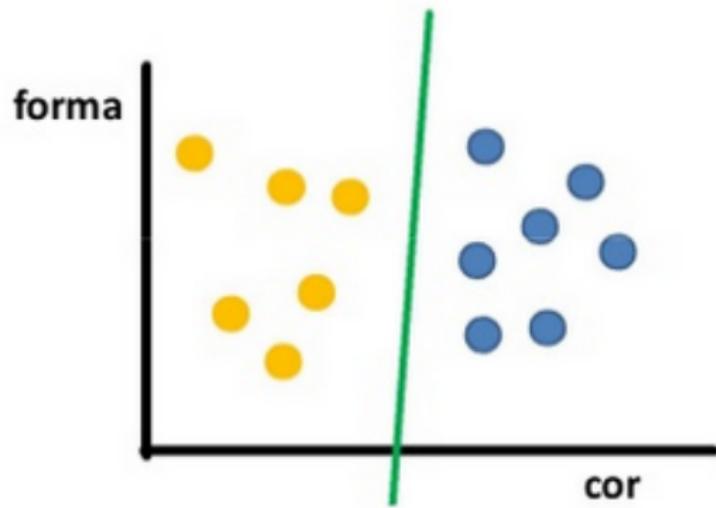


Objetivo: Encontrar modelos, funções ou regras que separem corretamente grupos de objetos

Classificação



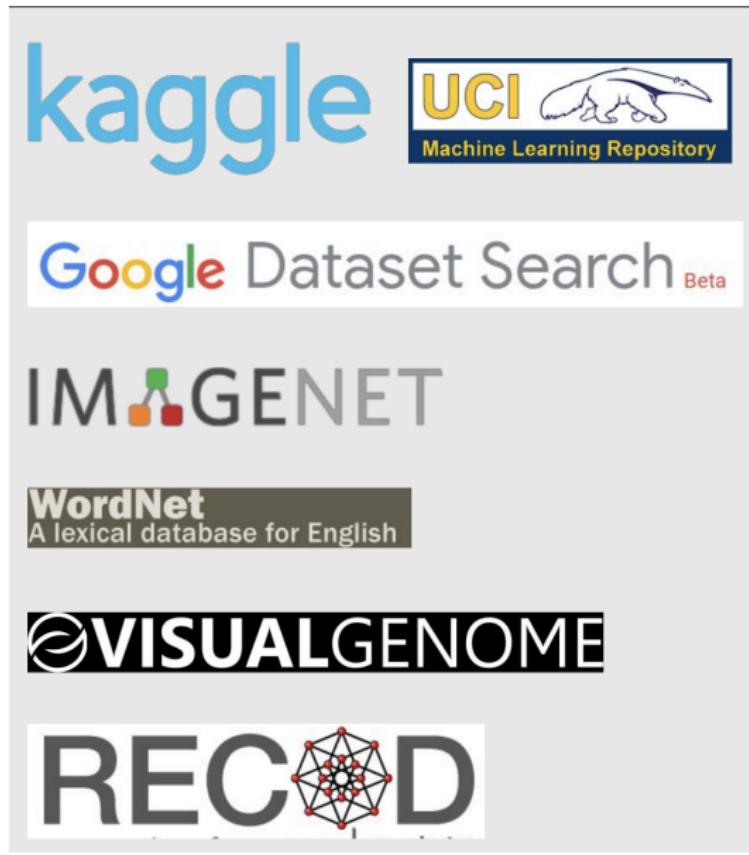
Classificação



Algoritmos de Classificação

- Indução de árvores de decisão
- Indução de conjuntos de regras
- Redes neurais
- Máquinas de vetores de suporte
- k-NN
- Regressão logística
- Redes bayesianas

Pacotes e Conjuntos de Dados



Pacotes e Conjuntos de Dados

- UCI Machine Learning Repository
 - ▶ <http://archive.ics.uci.edu/ml/>
- Weka
 - ▶ <http://www.cs.waikato.ac.nz/ml/weka/>
- Keel
 - ▶ <http://www.keel.es/>
- R Project
 - ▶ <http://www.r-project.org/>

Exercício

- Definição dos conjuntos de dados a serem utilizados nos trabalhos práticos!

Referências e Leituras Complementares

- Cap. 18 → livro Russel e Norvig
- Cap. 10 → livro Ben Coppin