



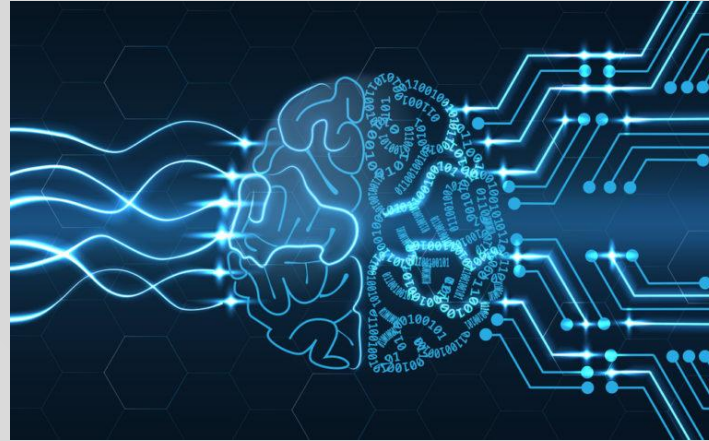
Curso de Extensão



Ciência dos Dados em Administração

Por Gustavo Alexandre

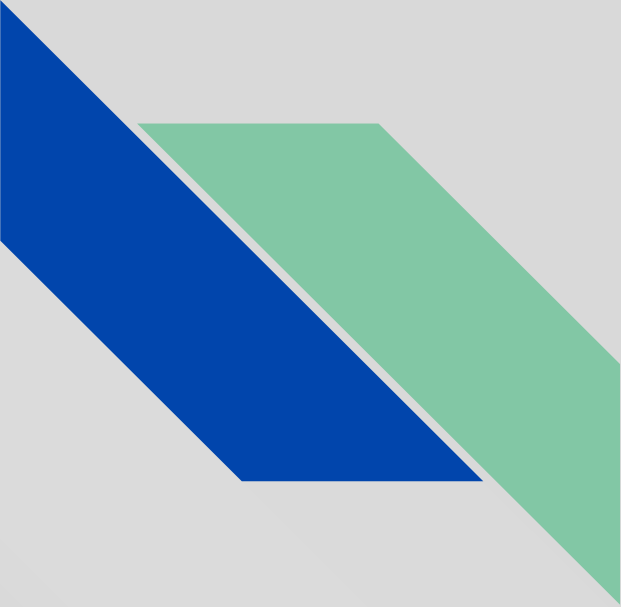
Aprendizado de Máquina





Agenda

- 1. Mercado**
- 2. Conceitos**
- 3. Dados**
- 4. Classificação**
- 5. Avaliação**
- 6. Técnicas**
- 7. Considerações**



Mercado

“Os erros causados por dados inadequados são muito menores do que aqueles devido à falta total de dados” *Charles Babbage*

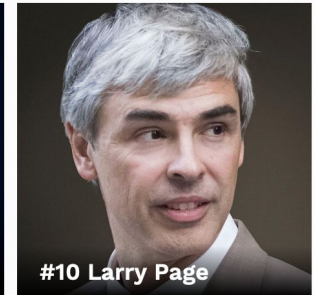
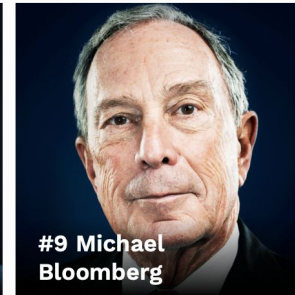
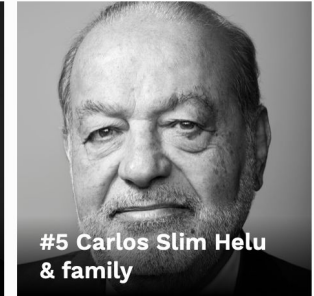
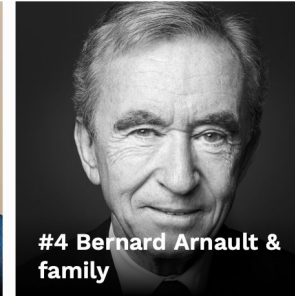


As Maiores Empresas do Mundo em 2019

Segundo a Forbes, ***Das trinta maiores empresas do mundo, nove são empresas de tecnologia:*** [Apple](#), [AT&T](#), [Samsung](#), [Microsoft](#), [Alphabet](#), [Verizon](#), [China Mobile](#) e [Amazon](#).

Fonte: Forbes Global 2000 (2019)

Os homens mais ricos do Mundo em 2019



Fonte: Forbes Global 2000 (2019)

Os homens mais ricos do Mundo em 2015

Top 100^[1] [\[edit \]](#)

No. ↕	Name ↕	Citizenship ↕	Age ↕	Net Worth USD billion ↕	Source(s) of wealth ↕
1	Bill Gates	United States	59	79.20	Microsoft
2	Carlos Slim Helu	Mexico	75	77.10	Telecom
3	Warren Buffett	United States	84	72.70	Berkshire Hathaway
4	Amancio Ortega	Spain	78	64.50	Zara
5	Larry Ellison	United States	70	54.30	Oracle
6	Charles Koch	United States	79	42.90	Diversified
6	David Koch	United States	74	42.90	Diversified
8	Christy Walton	United States	60	41.70	Wal-Mart
9	Jim Walton	United States	67	40.60	Wal-Mart
10	Liliane Bettencourt	France	92	40.10	L'Oréal
11	Alice Walton	United States	65	39.40	Wal-Mart
12	S. Robson Walton	United States	71	39.10	Wal-Mart
13	Bernard Arnault	France	66	37.20	LVMH
14	Michael Bloomberg	United States	73	35.50	Bloomberg LP
15	Jeff Bezos	United States	51	34.80	Amazon.com
16	Mark Zuckerberg	United States	30	33.40	Facebook
17	Li Ka-shing	Hong Kong	86	33.30	Diversified
18	Sheldon Adelson	United States	81	31.40	Casinos
19	Larry Page	United States	41	29.70	Google
20	Sergey Brin	United States	41	29.20	Google

Fonte: Forbes Global 2000 (2015)

Apple





Microsoft



Microsoft



America Movil



Samsung



Facebook





Google



Biggest Tech Company Ever

Alphabet Inc.



is for Google



O Case Amazon





Transformação Digital

É o uso da tecnologia para resolver problemas tradicionais baseado-se em soluções digitais promovendo eficiência e automação aos procedimentos e atividades que suportam os processos de negócio [[Christian Matt *et al.*, 2014](#)]



Transformação de Negócio

É o processo de "reestruturação fundamental" dos sistemas, dos processos, das pessoas e da tecnologia em toda uma empresa ou unidade de negócio, para alcançar melhorias mensuráveis em eficiência, eficácia e satisfação das partes interessadas [[Cruise, 2017](#)]



Leitura Complementar

- ❑ Estratégia Digital do Governo Federal 2017
- ❑ Information Economy Report 2015
- ❑ Digital Transformation
- ❑ Digital Transformation Strategies
- ❑ Industry 4.0
- ❑ Business Data Mining - A Machine Learning Perspective
- ❑ Business Intelligence and Analytics: From Big Data to Big Impact



Aprendizado de Máquina

"Há três tipos de mentiras: as mentiras, as mentiras descabeladas, e as estatísticas"
Benjamin Disraeli



Conceitos

O **Aprendizado de Máquina (AM)** explora o estudo e a construção de algoritmos que podem aprender sobre dados e fazer previsões

Conceitos

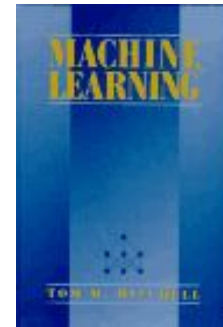
- Segundo Arthur Samuel (1959): “área de estudo que concede aos computadores a habilidade de aprender sem serem programados explicitamente.” [[Samuel, 1959](#)]



Conceitos

Tom Mitchell (1998): “Um programa de computador aprende com a experiência E em relação a tarefa T e alguma medida de desempenho P , se seu desempenho em T , medido por P , melhora com a experiência E .”

[[Mitchell, 1998](#)]





Aprendizado de Máquina

- Supervisionado
- Não-Supervisionado
- Semi-Supervisionado
- Profundo
- Por Reforço



Aprendizado de Máquina

- **Supervisionado** (*supervised learning*) visa construir um modelo estatístico a partir de um conjunto de dados que contém as entradas e as saídas desejadas (rotuladas):
 - Classificação
 - Regressão



Aprendizado de Máquina

- **Não-Supervisionado** (*unsupervised learning*) visa construir um modelo estatístico a partir de um conjunto de dados que contém apenas as informações de entradas e nenhuma identificação de saída (não rotulada):
 - Agrupamento
 - Sistemas de Recomendação
 - Filtragem



Aprendizado de Máquina

- **Semi-Supervisionado** (***semi-supervised learning***) visa construir um modelo estatístico a partir de um conjunto de dados de treinamento incompletos, em que uma parte da amostra possui rótulos e a outra não (rotulados e não-rotulados)



Aprendizado de Máquina

- **Profundo** (*deep learning*) compreende o uso das redes neurais artificiais em grandes volumes de dados (*big data*), ampliando continuamente sua capacidade de aprendizado, à medida que mais dados são processados



Aprendizado de Máquina

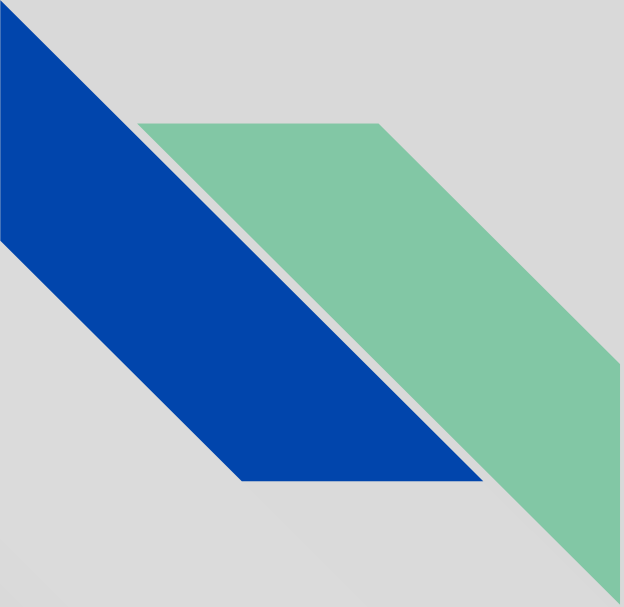
- Por Reforço (*reinforcement learning*) compreende a técnica de aprendizado interativo sobre a forma como agentes inteligentes (*multi-agent systems*) aprendem a agir em determinados ambientes, de modo a maximizar a noção de recompensa perante a execução das tarefas



Leitura Complementar

❏ Estratégia Digital do Governo Federal 2017





Problemas de AM



Identifique o problema de AM

- ❖ Dado um exame, como identificar se um tumor é **benigno ou maligno com base no seu tamanho e idade do paciente?**



Identifique o problema de AM

- ❖ Dado um conjunto de dados sobre o tamanho de casas no mercado imobiliário, como vamos **prever o preço de casas**, já que algumas instâncias foram atribuídas como padrão A, B e C?



Identifique o problema de AM

- ❖ Dada uma imagem de homem ou mulher, como podemos **prever a sua idade com base em dados da imagem?**



Identifique o problema de AM

- ❖ Dada uma coleção de milhares de pesquisas em uma universidade, como podemos encontrar uma maneira automática de **agrupar estas pesquisas que são de alguma forma semelhantes** por algumas variáveis, tais como a frequência das palavras, frases e contagem de páginas?



Leitura Complementar

- ❏ Estratégia Digital do Governo Federal 2017





Aprendizado Supervisionado

"Há três tipos de mentiras: as mentiras, as mentiras descabeladas, e as estatísticas"
Benjamin Disraeli



Aprendizado Supervisionado

- A máquina recebe as saídas identificadas
- Dois tipos (tarefas):
 - **Classificação**: prediz valor discreto
 - **Regressão**: prediz valor contínuo

Exemplos

Detecção de spam

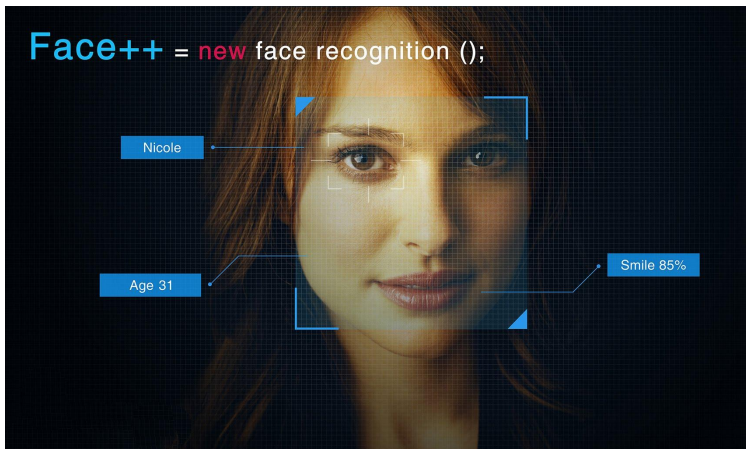


Reconhecimento de Voz

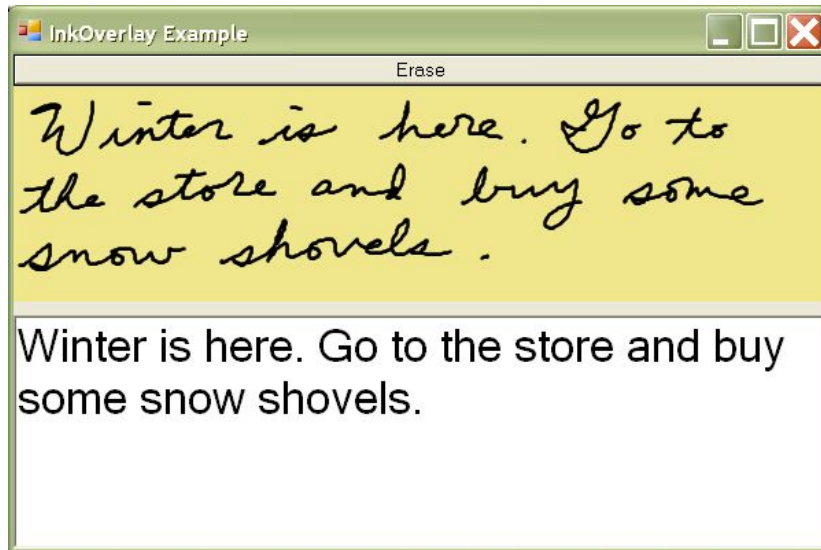


Exemplos

Reconhecimento de imagens



Reconhecimento de caracteres

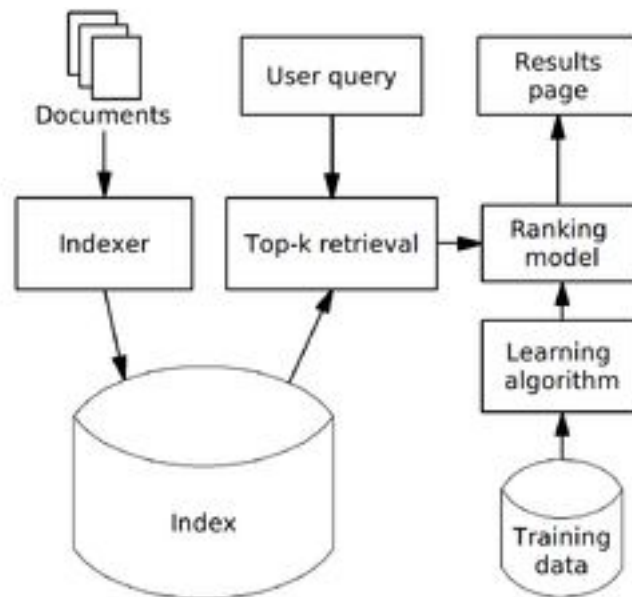


Exemplos

Tradução Automática



Learning to rank





Leitura Complementar

- ❏ Estratégia Digital do Governo Federal 2017





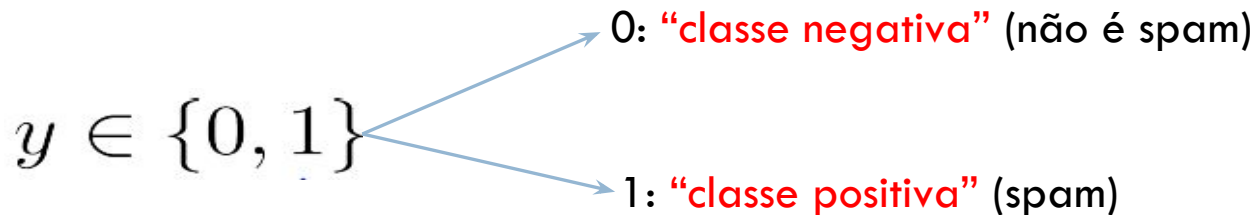
Classificação

"No futuro, o pensamento estatístico será tão necessário para a cidadania eficiente como saber ler e escrever" *H.G.Wells*

Classificação binária

- **Exemplos**

- **Email:** spam/ham (not spam)?
- **Transações financeiras:** fraudulenta/legítima?
- **Tumor:** maligno/benigno?

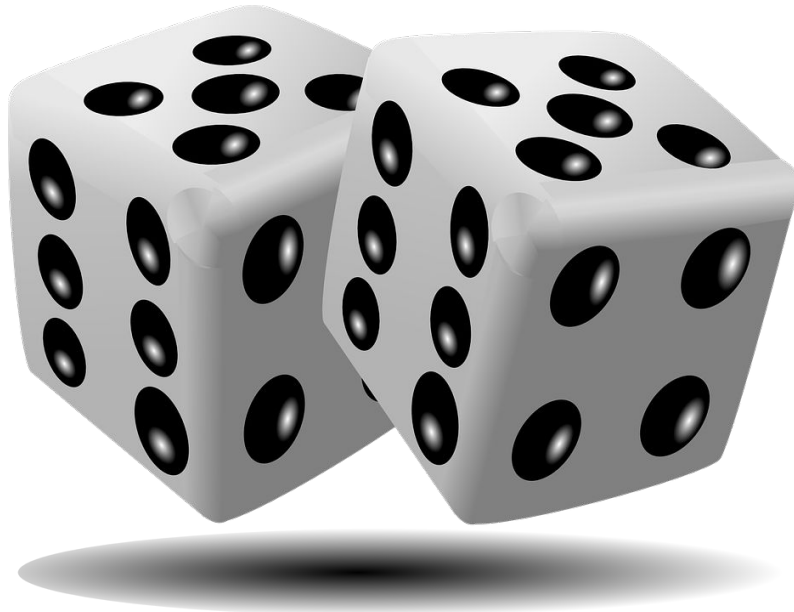




Terminologia

- ❖ Conjunto de Treinamento
- ❖ Conjunto de Teste
- ❖ Conjunto de Validação
- ❖ Função alvo
- ❖ Atributos
- ❖ Modelo
- ❖ Algoritmo de Aprendizado

Dados





Principais Pacotes para AM em **R**

- `library(dplyr)`
- `library(Hmisc)`
- `library(e1071)`
- `library(rtree.part)`
- `library(caret)`
- `library(ROCR)`



Conjunto de Dados em *R*

- Obtendo amostra de dados:
 - `data(iris)`
 - `data(Titanic)`
 - `data(economics)`
 - `data(mtcars)`
 - `data(diamonds)`



Manipulação de Dados em *R*

- Comandos para processar os dados:
 - **head**(iris)
 - **nrow**(Titanic)
 - **summary**(economics)
 - **unique**(mtcars)
 - **missing**(diamonds)



Pré-Processamento de Dados em *R*

- Conjuntos de dados:
 - Treinamento (***train***) e Teste (***test***)
 - **train**<-sample_frac(data, 0.8)
 - treino<-as.numeric(rownames(**train**))
 - **test**<data[-treino,]

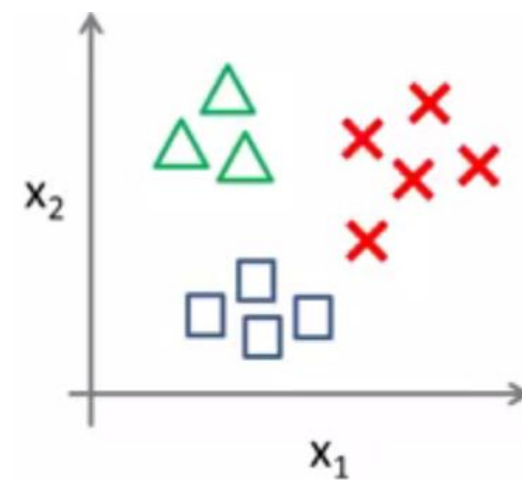
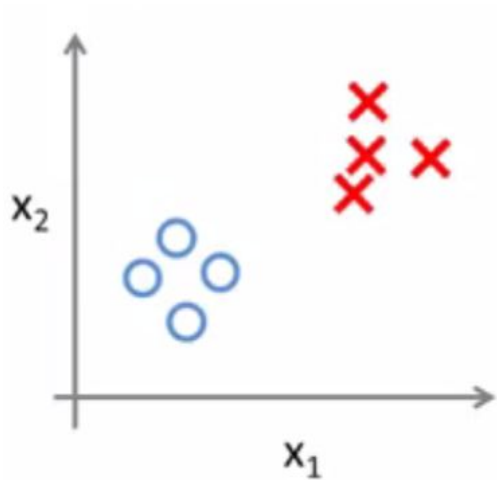




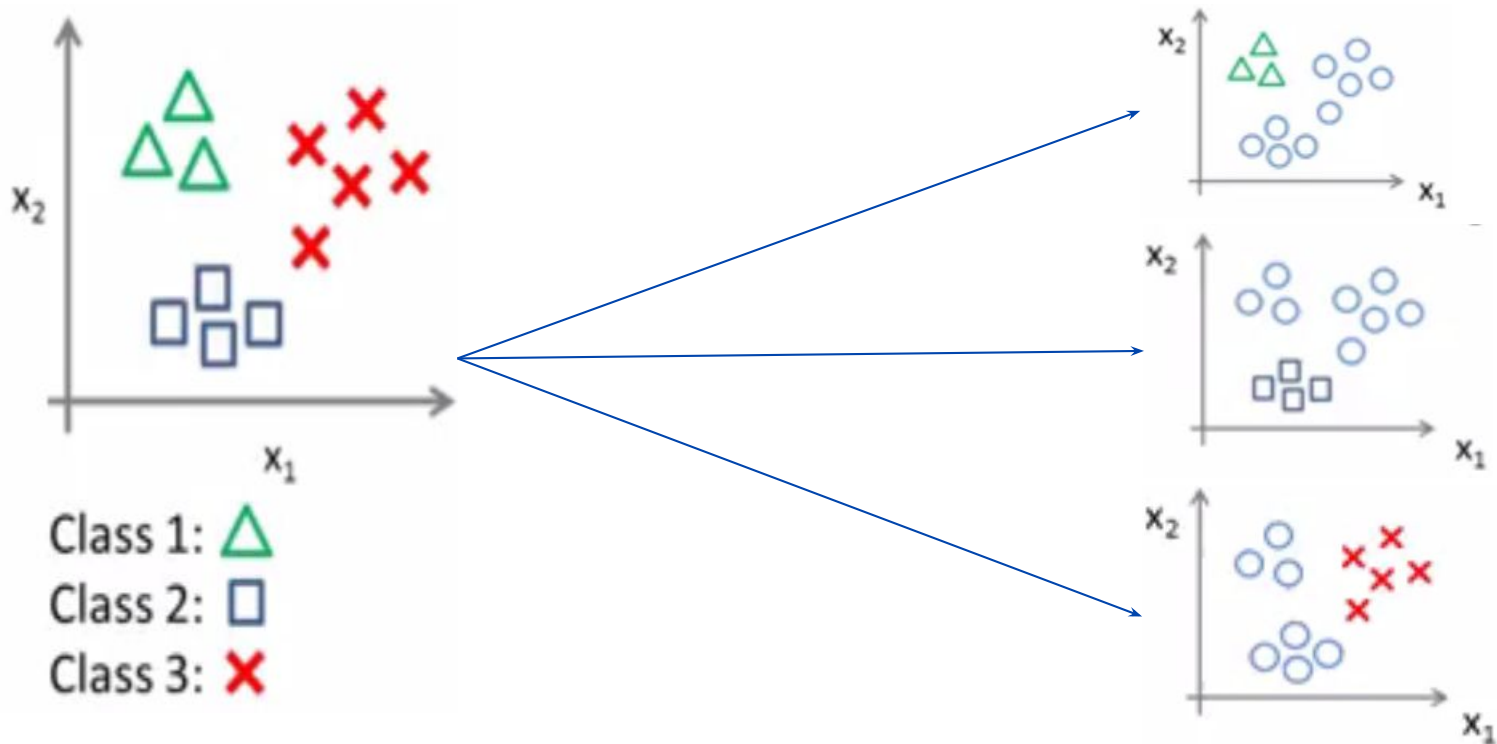
Abordagem multiclasse

- Organização de um portal de notícias: esportes, humor, política
- Diagnose médica: gripado, resfriado, dengue
- Condição do tempo: ensolarado, nublado, chuvoso

Abordagem multiclasse



Abordagem multiclasse





Procedimento multiclasse

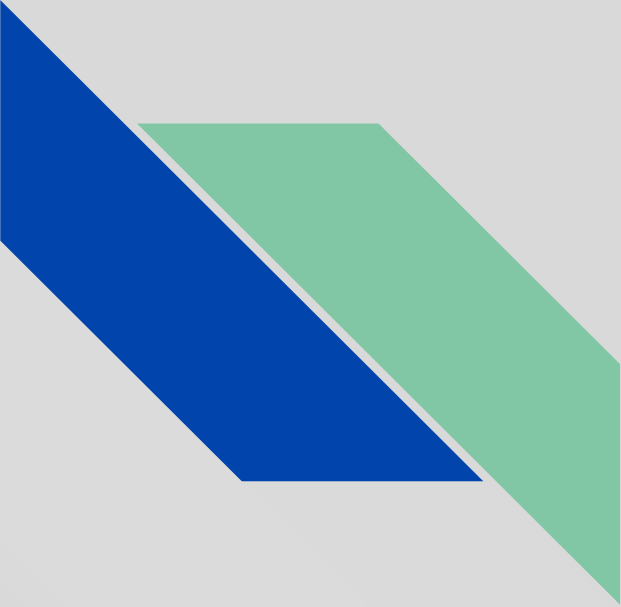
- Em geral para um problema de classificação com n classes, os passos são:
 - Treinar um classificador para cada uma das n classes
 - Para fazer a classificação de um novo exemplo x , seleccionar a classe que maximiza a hipótese correspondente.



Leitura Complementar

- ❏ Estratégia Digital do Governo Federal 2017





Avaliação

"No futuro, o pensamento estatístico será tão necessário para a cidadania eficiente como saber ler e escrever" *H.G.Wells*



Avaliação de Modelo

- A avaliação dos classificadores binários compara dois métodos de atribuição de um atributo binário, um dos quais é geralmente um método padrão e o outro está sendo investigado.



Avaliação de Modelo

- Existem muitas métricas que podem ser usadas para medir o desempenho de um classificador ou preditor e campos diferentes têm preferências diferentes para métricas específicas devido a objetivos diferentes.



Avaliação de Modelo

- Na **medicina**, sensibilidade e especificidade são frequentemente usadas
- Na **ciência da computação**, a precisão e a recordação são preferidas.



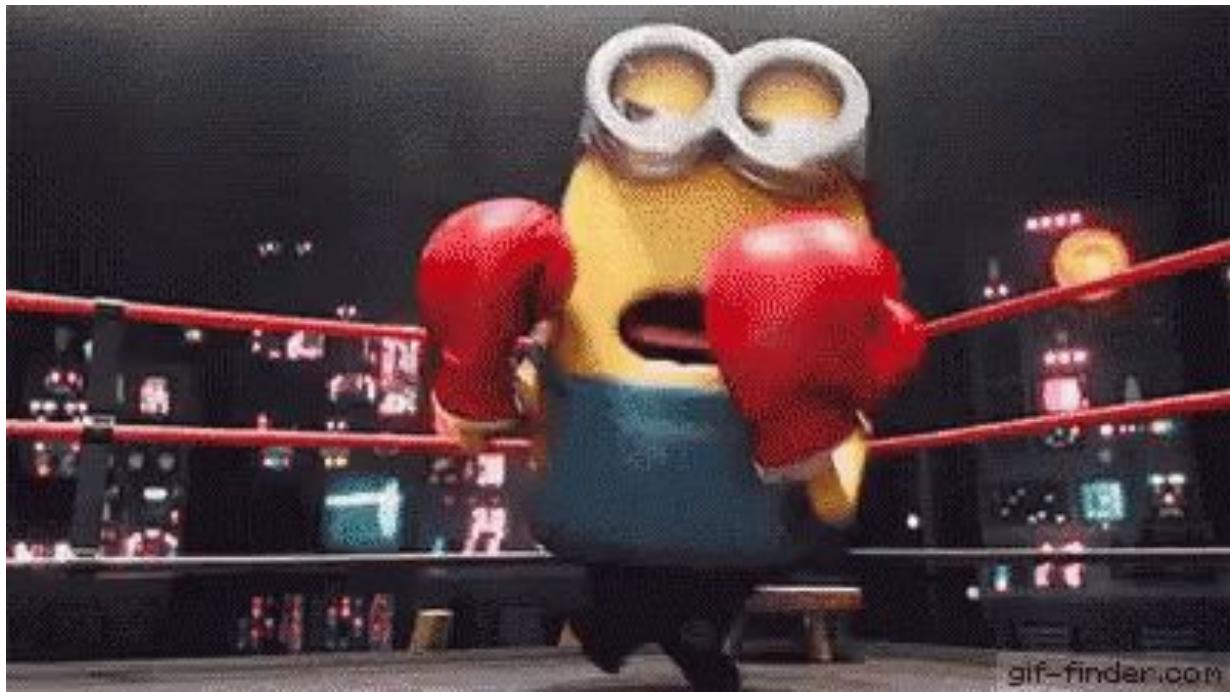
Definições

- **TP** = Verdadeiro Positivo (***True Positive***)
- **TN** = Verdadeiro Negativo (***True Negative***)
- **FP** = Falso Positivo (***False Positive***)
- **FN** = Falso Negativo (***False Negative***)



Métricas de Avaliação (Classificação)

- **Acurácia**
- **Medida F1 (*F-Score*)**
- **Matriz de Confusão**
- **Curva ROC**





Acurácia

- Segundo a Física, é a exatidão de uma medição ou de um instrumento de medição.
- Para o AM, é uma métrica para avaliar modelos de classificação.
- Pode-se dizer, informalmente, que acurácia é a fração de predições que nosso modelo acertou. Formalmente, a acurácia tem a seguinte definição:
- $Acuracia = (TP + TN)/(TP+TN+FP+FN)$



Acurácia



Medida F1 (*F-Score*)

- Na avaliação estatística de Classificação Binária, a medida F1 (também *F-score* ou *F-measure*) é uma medida utilizada na precisão de um teste.
- F1 tem sido amplamente utilizado na literatura de processamento de linguagem natural, como a avaliação do reconhecimento de entidade nomeada e segmentação de palavras.
- F1 é a média harmônica da **precisão** e da **recordação**, em que a pontuação de F1 atinge seu melhor valor em **1** (precisão e recordação perfeitas) e a pior em **0**. Formalmente, F1 tem a seguinte definição:
- $$F1 = 2 * (\text{Precisao} * \text{Recordacao}) / (\text{Precisao} + \text{Recordacao})$$



Medida F1 (*F-Score*)



Coeficiente de Correlação de Matthews (MCC)

- O MCC é usado em AM como uma medida da qualidade das classificações binárias.
- Foi um método introduzido pelo bioquímico Brian W. Matthews em 1975.
- É um coeficiente de correlação entre as classificações binárias observadas e preditas, retornando um valor entre **-1** (discordância total) e **+1** (previsão perfeita). Formalmente, a acurácia tem a seguinte definição:
- $$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$



Coeficiente de Correlação de Matthews (MCC)



Matriz de Confusão

- Para problemas de classificação em AM, a matriz de confusão é também conhecida como **matriz de erro**, com um *layout* de tabela específico que permite a visualização do desempenho de um algoritmo, tipicamente em Aprendizado Supervisionado.
- Em Aprendizado Não-Supervisionado, é geralmente chamada de **matriz de correspondência**.
- É uma tabela com duas dimensões e conjuntos classes em ambas as dimensões.



Matriz de Confusão

		Valor Previsto	
		Positivo	Negativo
Valor Verdadeiro	Negativo	Verdadeiros Positivos	Falsos Negativos
	Positivo	Falsos Positivos	Verdadeiros Negativos



Matriz de Confusão

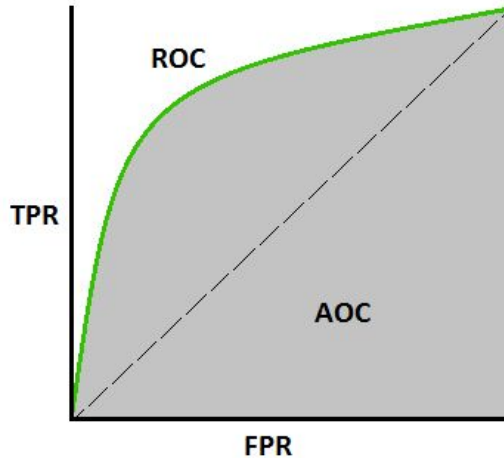


Curva ROC

- É uma representação gráfica que ilustra o desempenho de um classificador binário e como o seu limiar de discriminação é variado.
- A curva ROC foi desenvolvida por engenheiros durante a Segunda Guerra Mundial para detecção de objetos inimigos nas batalhas e já foi implementado na psicologia, medicina, radiologia, aprendizado de máquina e mineração de dados.
- Esta métrica está relacionada com a análise de custo/benefício do diagnóstico da Tomada de Decisão.

Curva ROC

- É obtido pela representação da fração de Positivos Verdadeiros dos Positivos Totais ($TPR = PV/P$) versus a fração de Positivos Falsos dos Negativos Totais ($FPR = PF/N$), em várias configurações do limite. Formalmente, a Curva ROC pode ser expressa da seguinte forma:





Curva ROC



Leitura Complementar

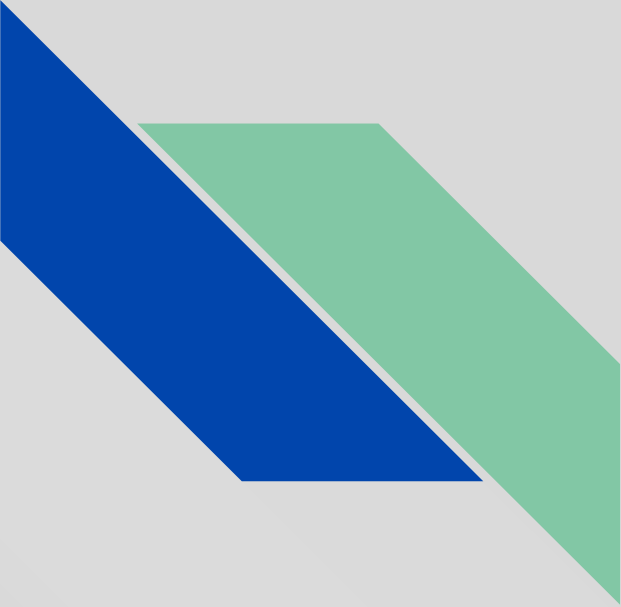
- ❏ Estratégia Digital do Governo Federal 2017



**Que tal uma
parada?**







Técnicas

"Não são tanto as coisas que não sabemos que nos metem em confusões. São as coisas que pensamos que sabemos." *Artemus Ward*



Aprendizado de Máquina Supervisionado

- Aprendizado de máquina (AM) é o estudo científico de algoritmos e modelos estatísticos que os sistemas de computador usam para realizar uma tarefa específica sem usar instruções explícitas, confiando em padrões e inferência



Técnicas de Classificação

- Constróem um modelo matemático com base nos dados de amostra e seus rótulos, para fazer previsões ou conduzir decisões sem ser explicitamente programado para executar as tarefas



Técnicas de Classificação

- Naive Bayes
- Árvore de Decisão
- Regressão Logística
- Máquina de Vetor de Suporte
- Florestas Aleatórias
- Aprendizado baseado em instância



Naive Bayes

Por Gustavo Alexandre



Naive Bayes

O **Naive Bayes** é um algoritmo probabilístico simples baseado no teorema de Bayes.

Este utiliza dados de treino para formar um modelo probabilístico baseado na evidência das features nos dados.

O algoritmo supõe que há uma independência entre as features do modelo.

Isso significa que a presença de uma determinada feature não tem nenhuma relação com as outras.



Naive Bayes





Árvore de Decisão

Por Gustavo Alexandre



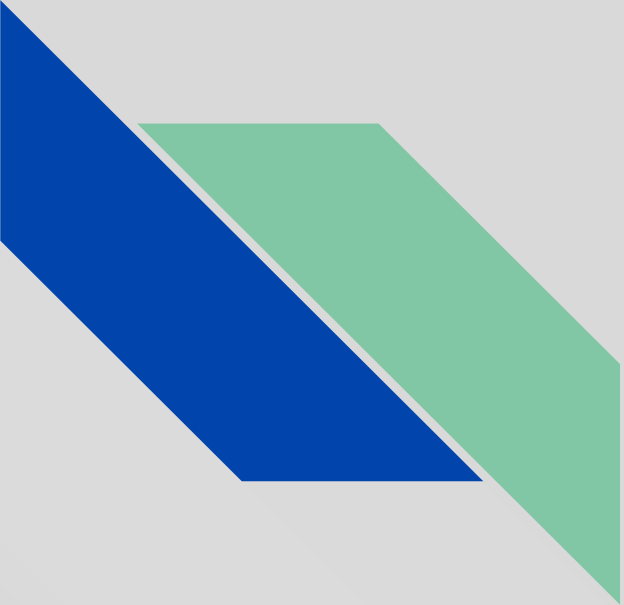
Árvore de Decisão

Uma árvore de decisão é uma ferramenta de **suporte à tomada de decisão** que usa um gráfico no formato de árvore e demonstra visualmente as condições e as probabilidades para se chegar a resultados. O algoritmo utilizado para chegar na representação visual da árvore pertence ao grupo de **aprendizado de máquina supervisionado**, e funciona tanto para **regressão** quanto para **classificação**.



Árvore de Decisão





Regressão Logística (Logit)

Por Gustavo Alexandre



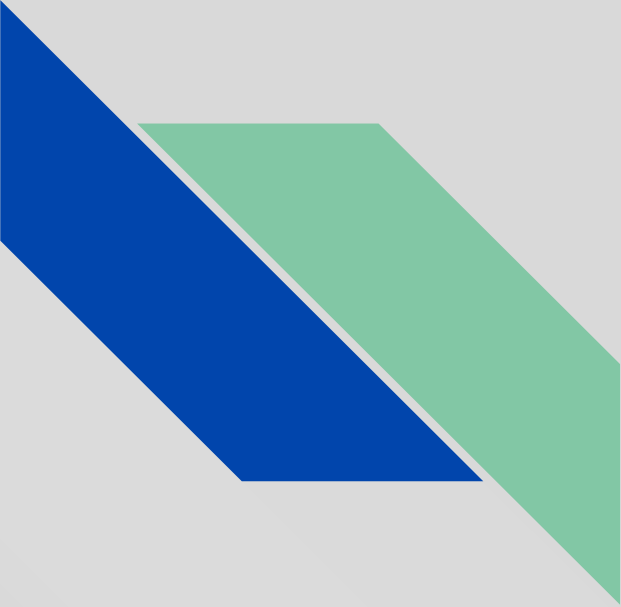
Regressão Logística

- **Regressão logística** é uma técnica estatística muito poderosa, utilizada para modelagem de saídas binárias (sim ou não). Quando se quer medir a relação de uma variável dependente binária com uma ou mais variáveis independentes, é comum utilizar esta técnica.
- Pense, por exemplo, numa empresa que empresta dinheiro para um cliente. Com base nas informações deste cliente (idade, profissão, etc.), é interessante a empresa tentar prever se o cliente vai pagar a dívida ou não. Uma forma de tentar prever isso é utilizando a regressão logística.



Regressão Logística





Máquina de Vetor de Suporte (SVM)

Por Gustavo Alexandre

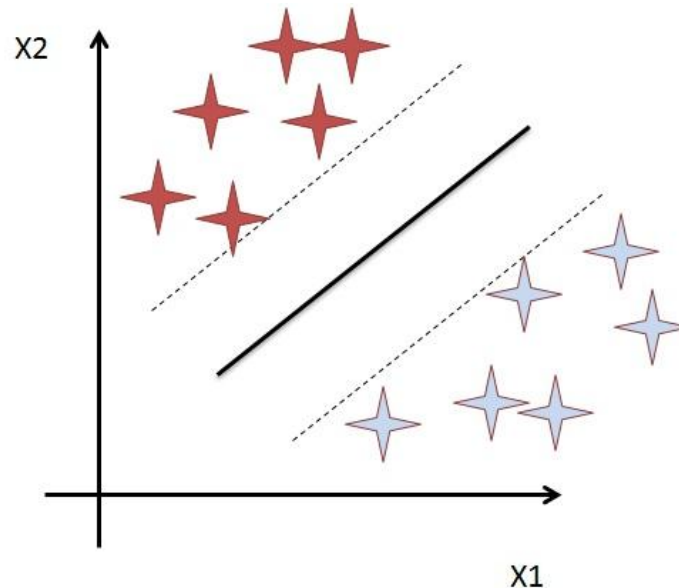


Máquina de Vetor de Suporte (SVM)

- Uma **máquina de vetores de suporte** (SVM, do inglês *support vector machine*) é um conceito na ciência da computação para um conjunto de métodos do aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão.
- O SVM padrão toma como entrada um conjunto de dados e prediz, para cada uma entrada, qual das possíveis classes essa entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico.
- Dado um conjunto de exemplos de treinamento, cada um marcado como pertencente a uma de duas categorias, um algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos a uma categoria ou outra.

Máquina de Vetor de Suporte (SVM)

- SVM busca um *hiperplano* entre os dados a serem classificados e visa maximizar a distância entre os pontos, separando cada uma das classes:





Máquina de Vetor de Suporte (SVM)





Seleção de Atributos

(Feature Selection)



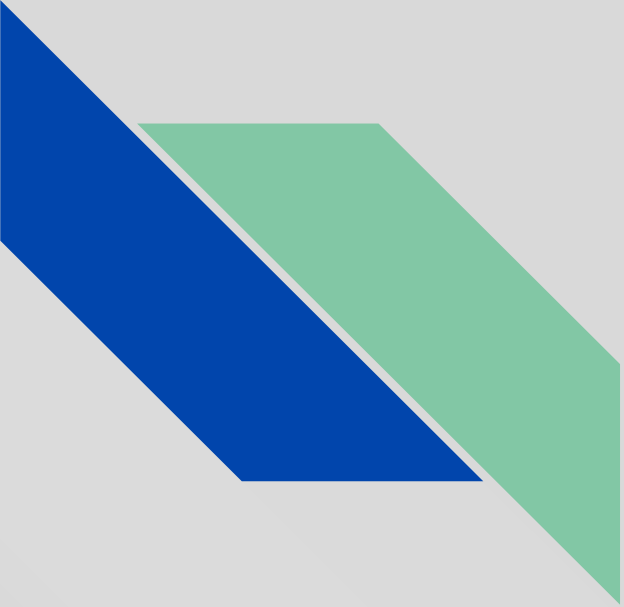
Seleção de Atributos

- A premissa central ao aplicar **seleção de atributos** é que os dados podem conter atributos redundantes ou irrelevantes e, portanto, são possíveis de remoção sem incorrer perdas significativas de informações.
- Redundantes e irrelevantes são duas noções distintas, uma vez que uma característica relevante pode ser redundante na presença de outra característica relevante com a qual ela é fortemente correlacionada.



Seleção de Atributos





Florestas Aleatórias

(Random Forests)



Florestas Aleatórias (*Random Forest*)

- Florestas aleatórias (*Random Forest*) são um método de aprendizado conjunto para classificação, regressão e outras tarefas que operam construindo uma multiplicidade de árvores de decisão no momento do treinamento e gerando a classe que é o modo das classes (classificação) ou previsão média (regressão) das árvores individuais.
- Florestas de decisão aleatória corrigem o hábito de **overfitting** de árvores de decisão em seu conjunto de treinamento
- O primeiro algoritmo para florestas de decisão aleatória foi criado por [Tin Kam Ho](#) usando o método de subespaço aleatório, que, na formulação de Ho, é uma maneira de implementar a abordagem de discriminação estocástica para classificação proposta por Eugene Kleinberg.



Florestas Aleatórias (*Random Forest*)





Aprendizado baseado em Instância

Por Gustavo Alexandre



Aprendizagem Baseada em Instâncias

- Uma vantagem que o aprendizado baseado em instância tem sobre outros métodos de aprendizado de máquina é sua capacidade de adaptar seu modelo a dados nunca vistos antes.
- Ou seja, pode simplesmente armazenar uma nova instância ou descartar uma instância antiga.
- Exemplos de algoritmo de aprendizado baseado em instâncias são o algoritmo KNN, máquinas kernel e redes RBF.



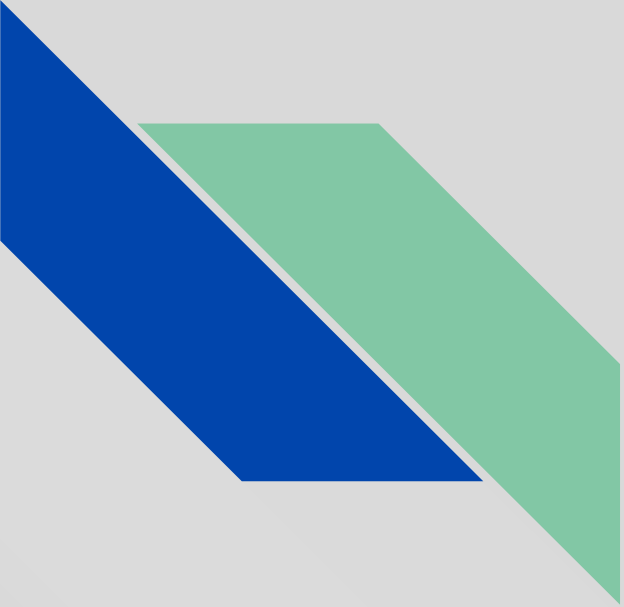
Aprendizagem Baseada em Instâncias

- Os algoritmos (KNN, RBF e máquinas de Kernel) armazenam seu conjunto de treinamento ao prever um valor/classe para uma nova instância.
- Esses algoritmos calculam distâncias ou semelhanças, dada uma instância a ser avaliada e as instâncias de treinamento já identificadas.
- Para combater a complexidade da memória de armazenar todas as instâncias de treinamento, bem como o risco de *overfitting* ao ruído no conjunto de treinamento, algoritmos de redução de instâncias foram propostos.



Leitura Recomendada

- ❏ Estratégia Digital do Governo Federal 2017
- ❏



Considerações

"Alguns usam a estatística como os bêbados usam postes: mais para apoio do que para iluminação" *Andrew Lang*

