

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. Some nodes are highlighted with blue circles or dots, while others are grey.

# **Ciência de dados**

# **Prática**

A decorative network diagram in the bottom-right corner, similar to the one in the top-left, with a web of interconnected nodes and lines, some highlighted in blue.

## Conteúdo

### **K-NN**

Algoritmo de classificação supervisionado.

### **Naive Bayes**

Algoritmo de classificação probabilístico, que usa como base o teorema de bayes.

### **K-Means**

Algoritmo de agrupamento não supervisionado.

### **Regressão**

Técnica estatística utilizada principalmente para estimação de valores futuros.



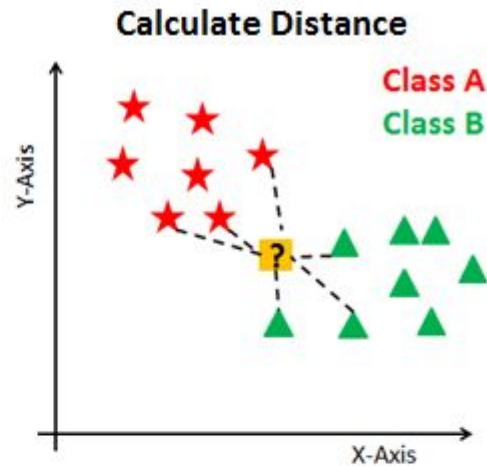
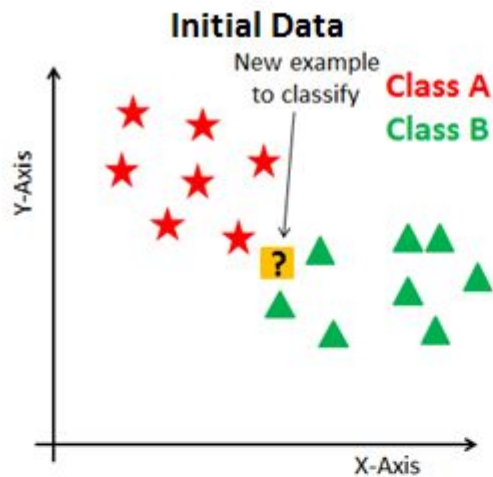
# K-NN

## Introdução

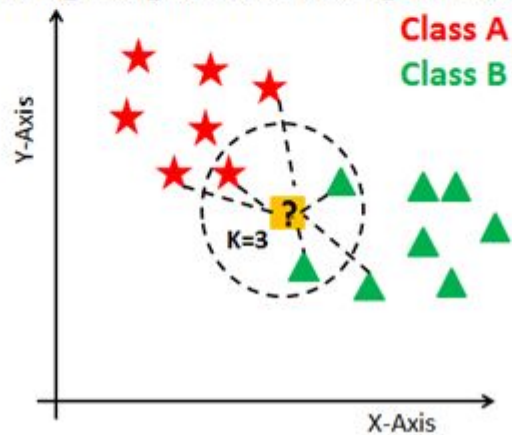
O algoritmo KNN (K Nearest Neighbor) é um dos algoritmos mais utilizados em Machine Learning e também um dos mais simplistas, analisando seu processo de cálculo. Este algoritmo pode ser aplicado em diversos segmentos de negócio, logo também se aplica em diversos problemas como finanças, saúde, ciência política, reconhecimento de imagem e reconhecimento de vídeos.

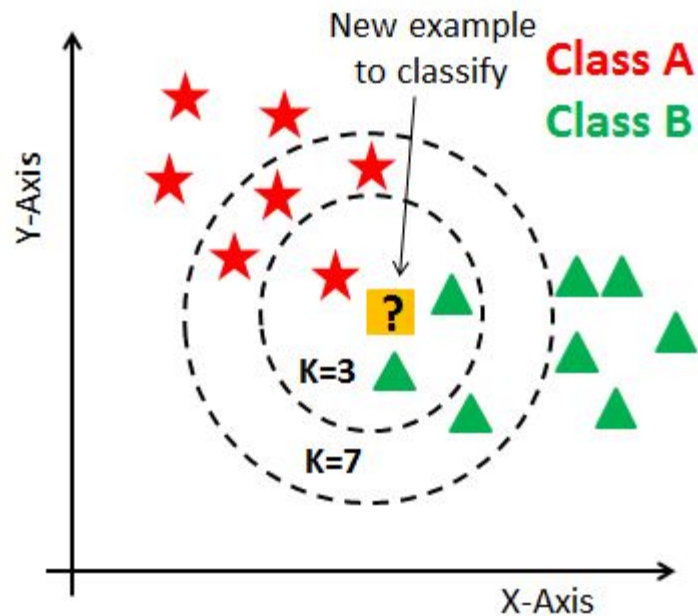
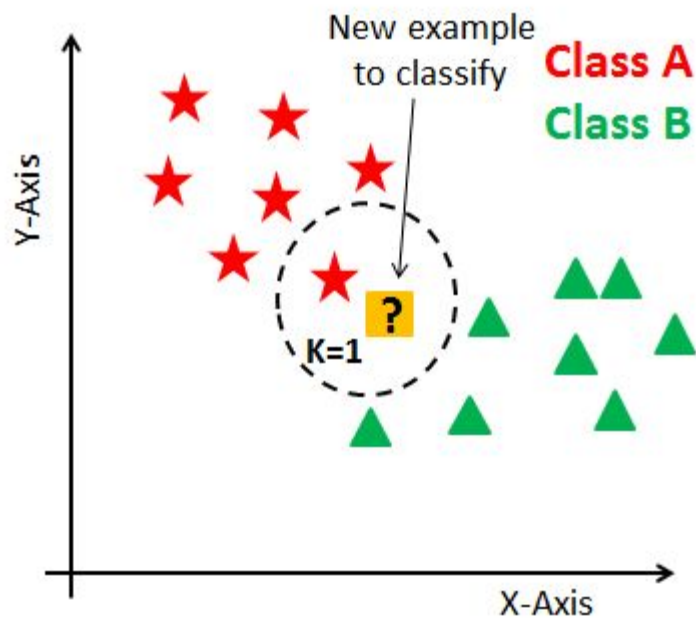
## Funcionamento

- ⦿ Calcular a distância do ponto a ser previsto para todos os outros pontos na base de dados;
- ⦿ Encontrar os vizinhos mais próximos;
- ⦿ Votar com base nos vizinhos mais próximos o valor para o ponto a ser previsto.

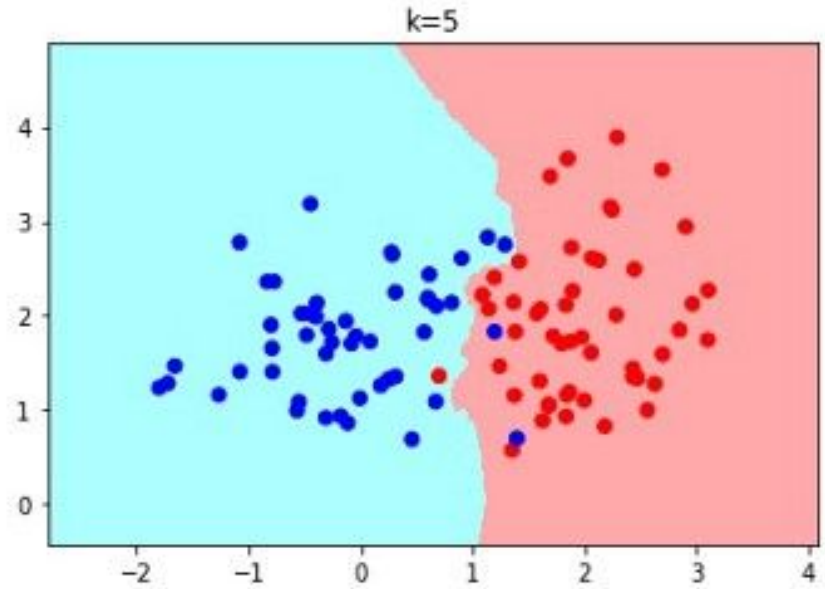
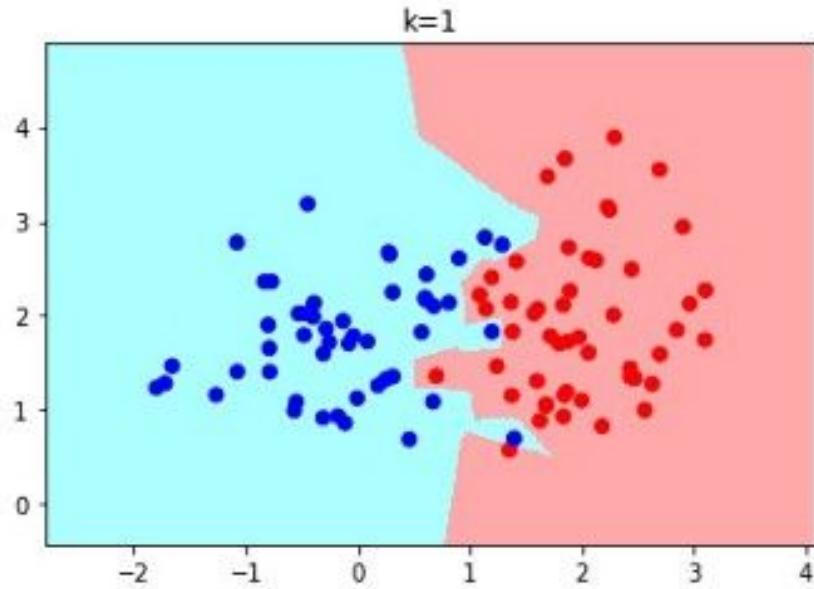


### Finding Neighbors & Voting for Labels





## Efeitos da alteração do valor K



## Prós

- ⊙ É um dos mais simples algoritmos a se implementar;
- ⊙ Boa precisão na maioria dos casos aplicados;
- ⊙ A facilidade para se efetuar tuning dos poucos parâmetros existentes (Valor de K e Medida de Distância);
- ⊙ Tempo para se efetuar treinamento dos dados também é diferenciado, sendo um dos mais rápidos para esta atividade;
- ⊙ O KNN pode ser utilizado em dados não lineares bem como para problemas de regressão.



## Contras

- © Uma demora excessiva na fase de teste e o alto consumo de memória para realizar esta atividade de teste, uma vez que o mesmo armazena todo dataset em memória;
- © KNN não é indicado a dados de grandes dimensões, imagina ter que calcular a distância de todos os pontos de dados entre si, quanto mais dados e mais dimensões desses registros, maior será o tempo para processar os cálculos;
- © Uma vez que o mesmo trabalha com medidas de distância necessita sempre que nos atentemos a escala dos valores utilizados, para que não gere resultados equivocados.

## Aplicações

- ◎ Sistemas de Recomendação;
- ◎ Pesquisar documentos semanticamente semelhantes;
- ◎ O KNN pode ser usado com eficácia na detecção de outliers. Um exemplo é a detecção de fraude de cartão de crédito.



# Naive Bayes

## Introdução

O Algoritmo Naive Bayes funciona como classificador e baseia-se na probabilidade de cada evento ocorrer, desconsiderando a correlação entre features. Por ter uma parte matemática relativamente simples, possui um bom desempenho e precisa de poucas observações para ter uma boa acurácia.

## Teorema de bayes

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

Diagram illustrating the components of Bayes' Theorem:

- $P(c | x)$ : Probabilidade posterior
- $P(x | c)$ : Probabilidade
- $P(c)$ : Probabilidade original da Classe
- $P(x)$ : Preditor da probabilidade posterior

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

## Funcionamento

- ◎ Converter o conjunto de dados em uma tabela de frequência;
- ◎ Use o teorema para calcular a probabilidade posterior para cada classe;
- ◎ A classe com maior probabilidade posterior é o resultado da previsão.

## Prós

- © É fácil e rápido para prever o conjunto de dados da classe de teste. Também tem um bom desempenho na previsão de classes múltiplas;
- © Quando a suposição de independência prevalece, um classificador Naive Bayes tem melhor desempenho em comparação com outros modelos como regressão logística, e você precisa de menos dados de treinamento;
- © O desempenho é bom em caso de variáveis categóricas de entrada comparada com a variáveis numéricas. Para variáveis numéricas, assume-se a distribuição normal (curva de sino, que é uma suposição forte).

## Contras

- © Se a variável categórica tem uma categoria (no conjunto de dados de teste) que não foi observada no conjunto de dados de treinamento, então o modelo irá atribuir uma probabilidade de 0 (zero) e não será capaz de fazer uma previsão. Isso é muitas vezes conhecido como “Zero Frequency”. Para resolver isso, podemos usar a técnica de alisamento. Uma das técnicas mais simples de alisamento é a chamada estimativa de Laplace;
- © Por outro lado naive Bayes é também conhecido como um mau estimador, por isso, as probabilidades calculadas não devem ser levadas muito a sério;
- © Outra limitação do Naive Bayes é a suposição de preditores independentes. Na vida real, é quase impossível que ter um conjunto de indicadores que sejam completamente independentes.

## Aplicações

- © **Previsões em tempo real:** Naive Bayes é um classificador de aprendizagem voraz e com certeza rápido. Assim, pode ser usado para fazer previsões em tempo real;
- © **Previsões multi-classes:** Este algoritmo também é conhecido pela funcionalidade de previsão multi-classes. Aqui podemos prever a probabilidade de múltiplas classes das variáveis-alvo;
- © **Sistema de Recomendação:** o classificador e a filtragem colaborativa Naive Bayes em conjunto constroem um sistema de recomendação que utiliza técnicas de machine learning e mineração de dados para filtrar a informação invisível e prever se um usuário gostaria de um determinado recurso ou não;
- © **Classificação de textos/Filtragem de spam/Análise de sentimento:** Naive Bayes também é utilizado em classificação de textos, têm maior taxa de sucesso em comparação com outros algoritmos. Como resultado, é muito utilizado na filtragem de spam (identificar spam) e Análise de Sentimento (em análise de mídia social, para identificar sentimentos positivos e negativos dos clientes).



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue and others in grey.

# K-means

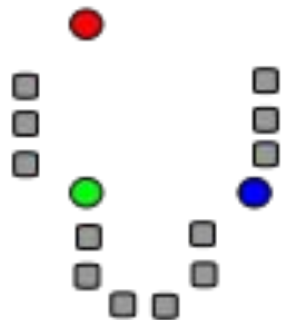
## Introdução

Um algoritmo de aprendizado não supervisionado que avalia e agrupa os dados de acordo com suas características.

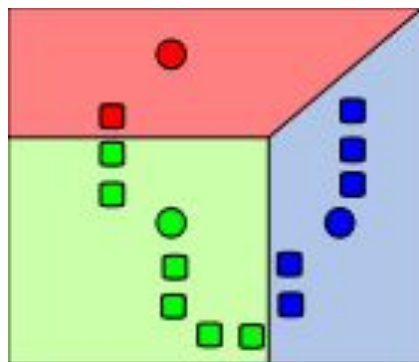
## Funcionamento

- ⦿ Definir um  $K$  (número de clusters);
- ⦿ Escolher aleatoriamente uma centróide para cada cluster (pontos centrais do grupo);
- ⦿ Calcular, para cada ponto, o centróide de menor distância. Cada ponto pertencerá ao centróide mais próximo;
- ⦿ Reposicionar o centróide. A nova posição do centróide deve ser a média da posição de todos os pontos do cluster;
- ⦿ Os dois últimos passos são repetidos, iterativamente, até obtermos a posição ideal dos centróides.

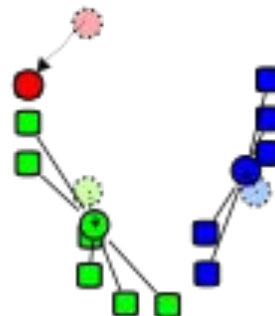
1



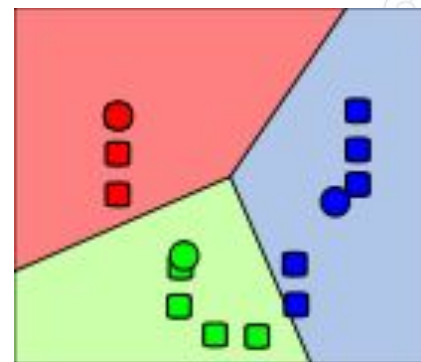
2



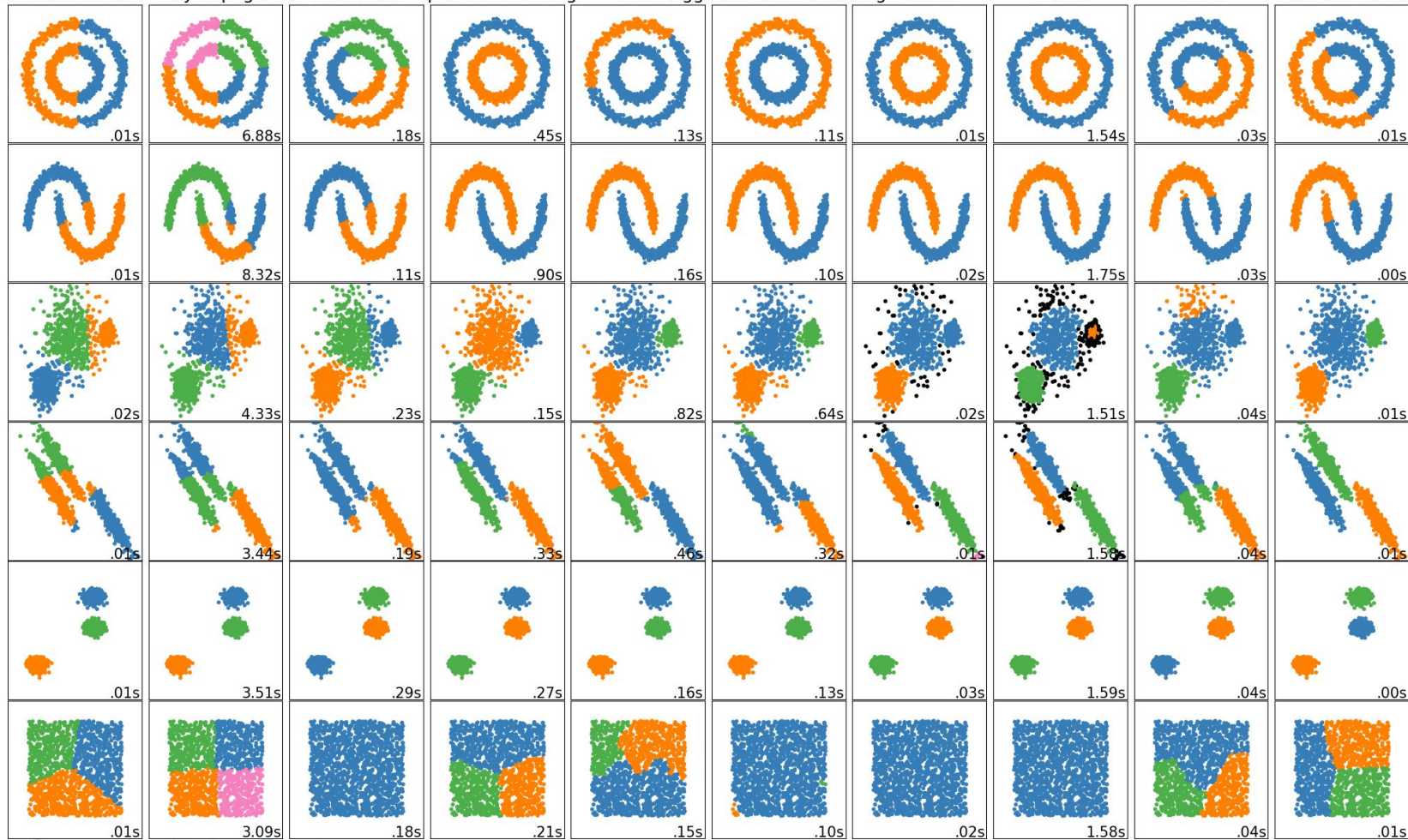
3



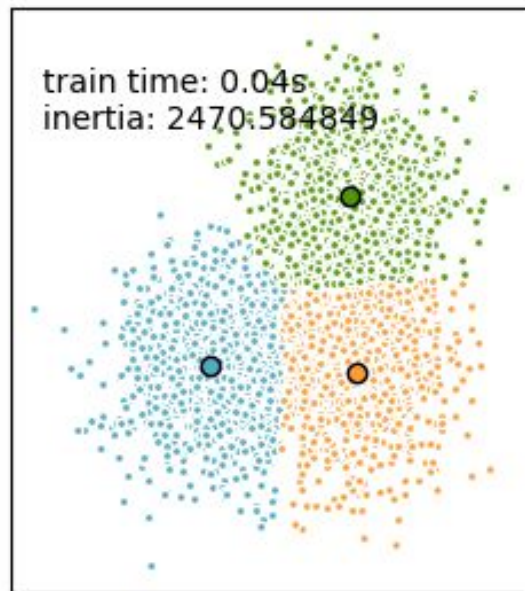
4



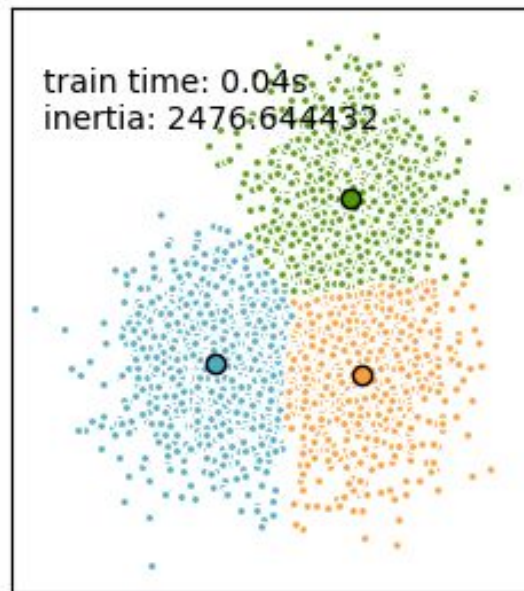
MiniBatchKMeans AffinityPropagation MeanShift SpectralClustering Ward AgglomerativeClustering DBSCAN OPTICS Birch GaussianMixture



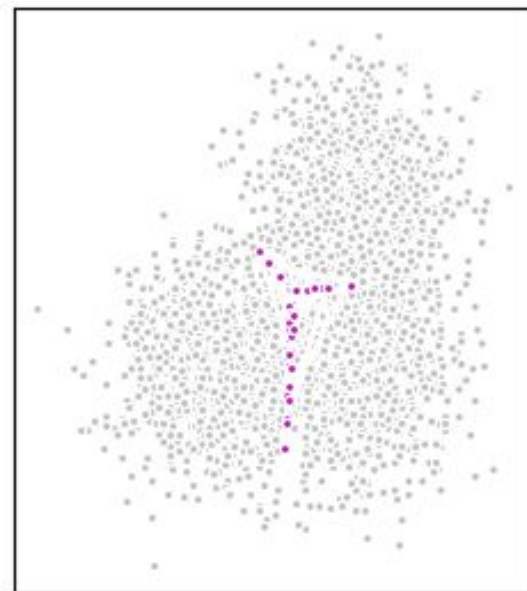
KMeans



MiniBatchKMeans



Difference



## Prós

- © Boa escalabilidade em grandes conjuntos de dados;
- © Generalização, funcionando para clusters de diferentes formas e tamanhos, como clusters elípticos.

## Contras

- ◎ K deve-ser escolhido manualmente;
- ◎ Dependente dos valores escolhidos inicialmente;
- ◎ Sensível a outliers;
- ◎ Aumenta muito o tempo de execução conforme número de dimensões vai aumentando.



## Aplicações

- © **Encontrar hotspot :** Utilizado para encontrar pontos onde um determinados eventos ocorrem com mais frequência, encontrando lugares com maior histórico de ocorrência de crimes e de queimadas por exemplo;
- © **Marketing:** Pesquisa e segmentação de mercado para determinar potenciais grupos homogêneos de consumidores para melhor definir a disposição de produtos que uma estratégia corporativa;
- © **Análise de redes sociais:** Dentro de um grande grupo de pessoas, reconhecer comunidades que compartilhem de alguma preferência ou opinião;
- © **Reconhecimento de imagens:** Clusterização pode ser aplicado para se reconhecer uma pessoa ou um objeto numa foto; os clusters seriam as regiões da imagem que contenham rostos, paisagem, vestimentas, etc.;
- © **Detecção de anomalias:** A análise de clusters pode ser adaptada para a detecção de outliers que destoam da maioria dos outros elementos baseada em alguma métrica de similaridade.





# Rregressão

## Introdução

Regressão é uma técnica que permite quantificar e inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes (variáveis explicativas). Essa relação é representado por um modelo matemático, sendo ele uma relação linear simples, ou uma relação linear múltipla caso exista várias variáveis independentes.

STATISTICS REGRESSION



COMPUTER  
SCIENCE



ARTIFICIAL  
INTELLIGENCE!

MACHINE  
LEARNING



## Funcionamento

O modelo de regressão linear simples consiste de 2 parâmetros, que correspondem aos coeficientes de uma equação da reta qualquer:


$$E(Y) = \alpha + \beta X$$

$E(Y)$  representa a variável de resposta.

Os parâmetros  $\alpha$  e  $\beta$  são estimados através do Método dos Mínimos Quadrados Ordinários. O objetivo deste método é obter uma reta que minimiza as distâncias entre os valores estimados e os valores observados.

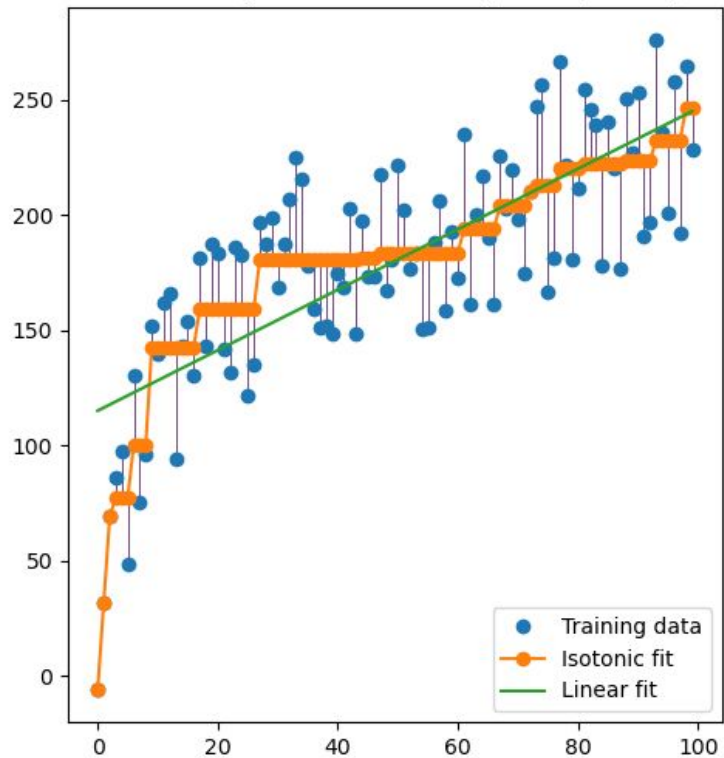
O parâmetro  $\alpha$  representa o intercepto da reta, onde ela cruza o eixo Y, ou seja, o valor de  $E(Y)$  para o qual  $X = 0$ .

Já o parâmetro  $\beta$  representa, neste caso, o efeito da variável explicativa sobre a variável resposta.

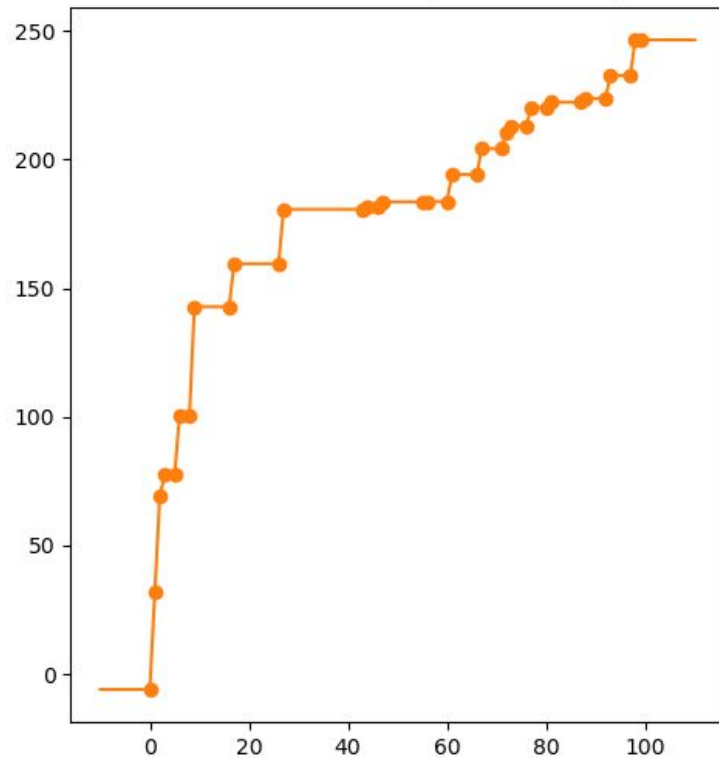
A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines, with some nodes highlighted in blue.

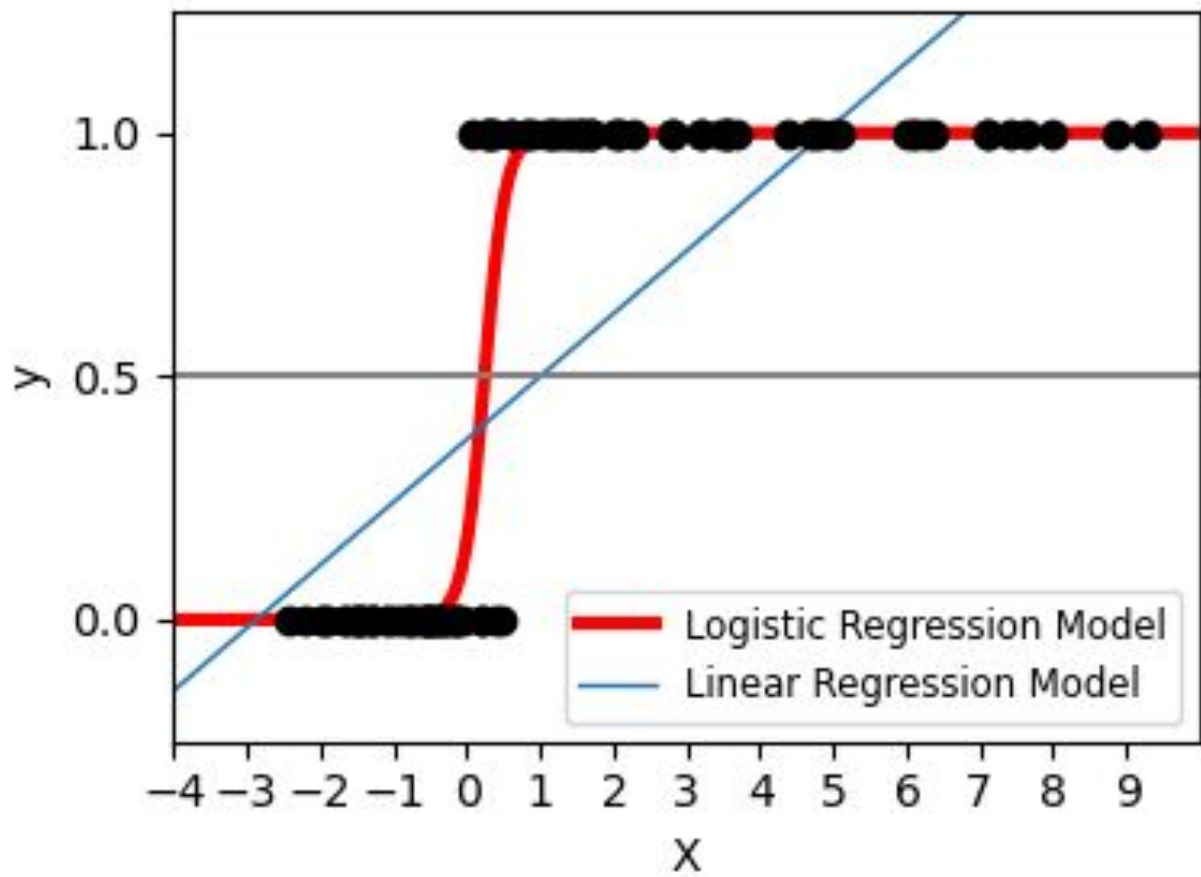
# **Alguns exemplos de tipos Regressões:**

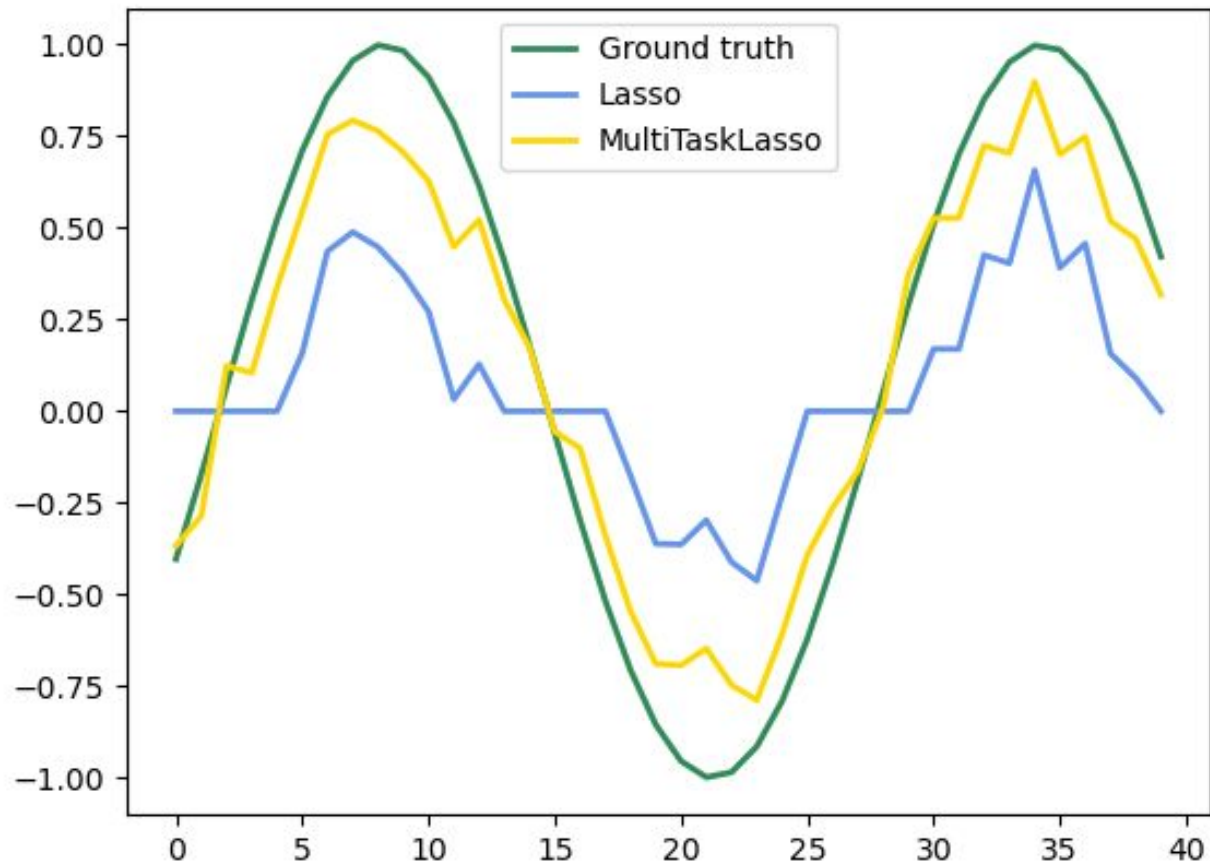
Isotonic regression fit on noisy data (n=100)

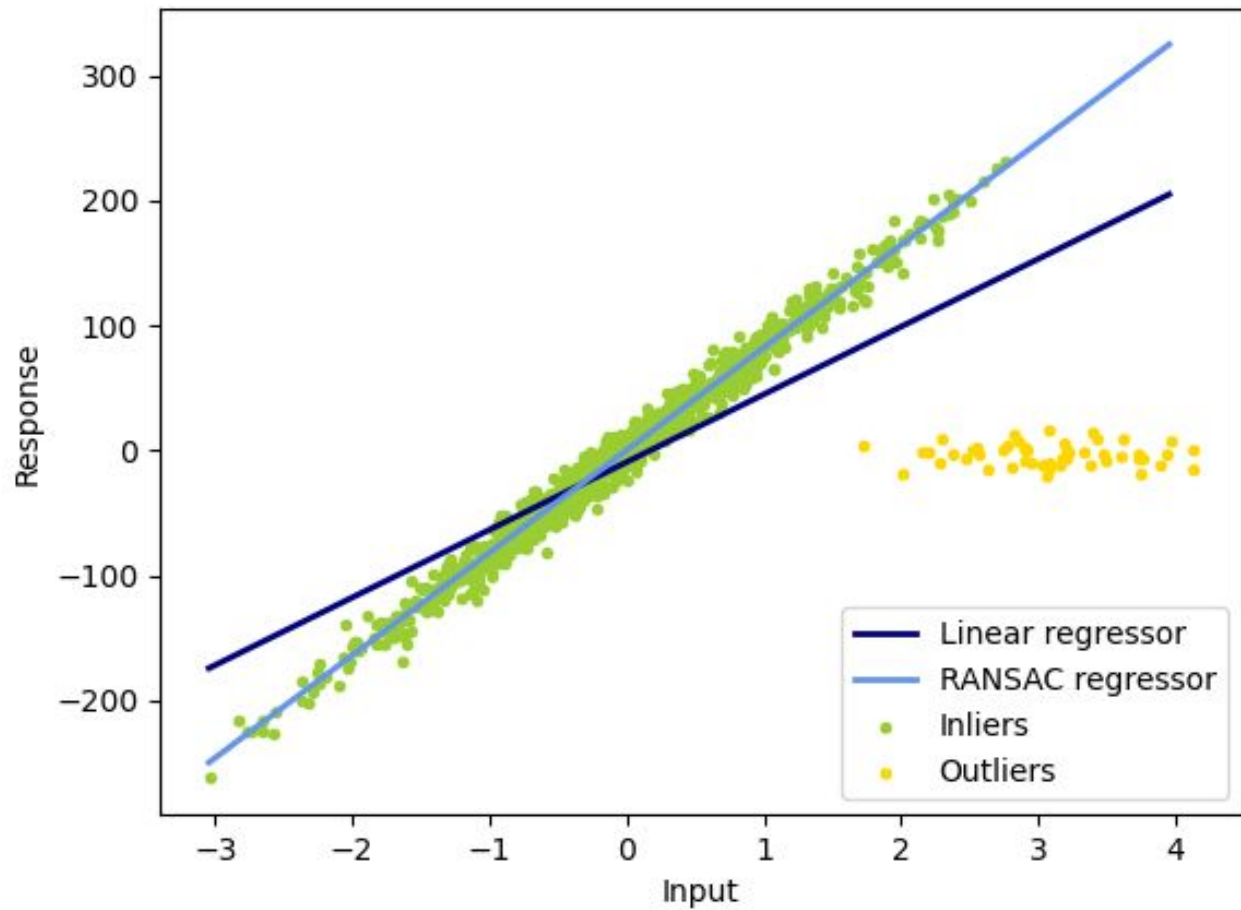


Prediction function (36 thresholds)



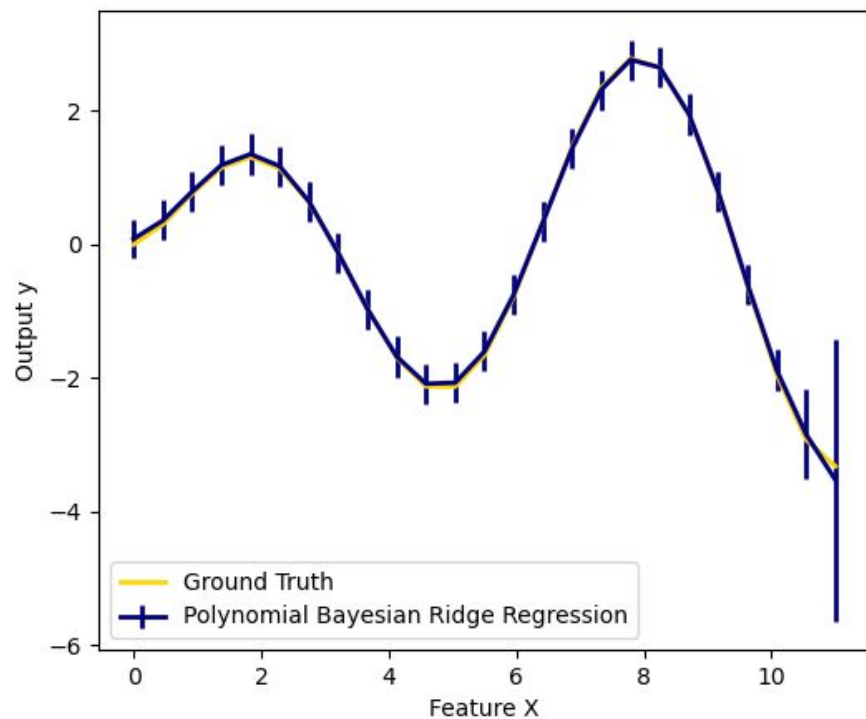
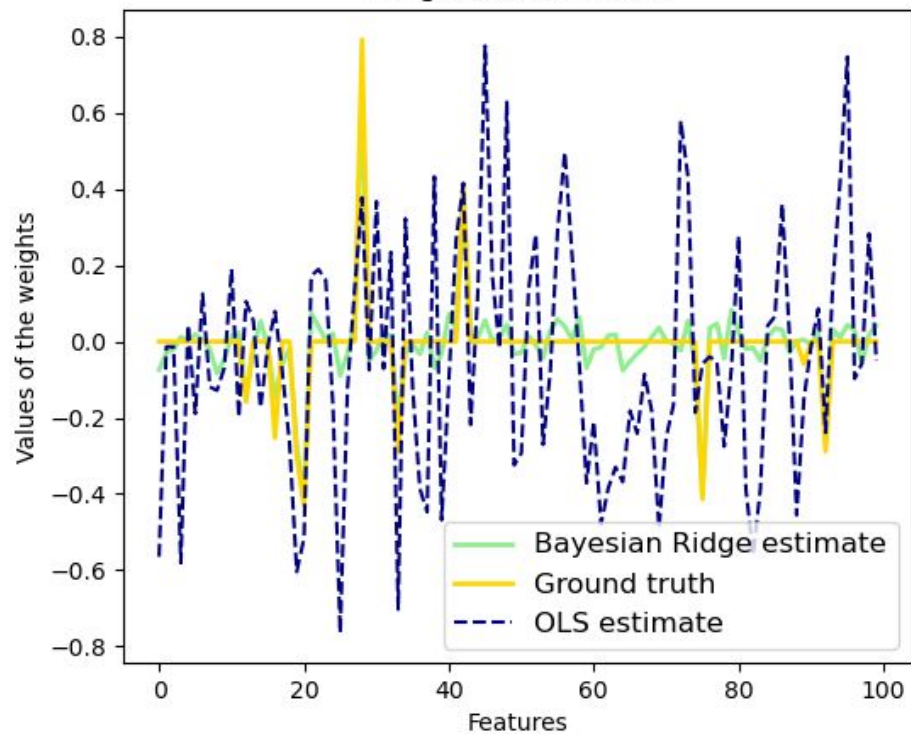








Weights of the model



## Prós

- ◎ **Implementação simples:** a regressão linear é um algoritmo muito simples que pode ser implementado muito facilmente para fornecer resultados satisfatórios. Além disso, esses modelos podem ser treinados de forma fácil e eficiente, mesmo em sistemas com poder computacional relativamente baixo quando comparado a outros algoritmos complexos;
- ◎ **Desempenho em conjuntos de dados separáveis linearmente:** Ajusta conjuntos de dados separáveis linearmente quase perfeitamente e é freqüentemente usada para encontrar a natureza da relação entre as variáveis;
- ◎ **Overfitting pode ser reduzido por regularização:** Overfitting é uma situação que surge quando um modelo de aprendizado de máquina se ajusta a um conjunto de dados e, portanto, captura os dados ruidosos também. Isso reduz sua precisão no conjunto de teste. A regularização é uma técnica de fácil implementação e capaz de reduzir efetivamente a complexidade de uma função de forma a diminuir o risco de sobreajuste.

## Contras

- ◎ Propenso a underfitting;
- ◎ Sensível a outliers;
- ◎ A regressão linear assume que os dados são independentes.

## Aplicações

- © Produtividade estimada de sacas de café dada a altitude de plantio;
- © Número de mortes por doenças em relação a quantidade de pessoas vacinadas de uma população;
- © Estimar a venda de pipoca em relação pessoas que vão ao cinema