

Aprendizado de Máquina (Machine Learning) e Ciência de Dados

João Pinheiro

@joaomh

30 de maio de 2024

Conteúdo do Curso I

1 Introdução e noções de Machine Learning

- Passos para Construção de um Modelo

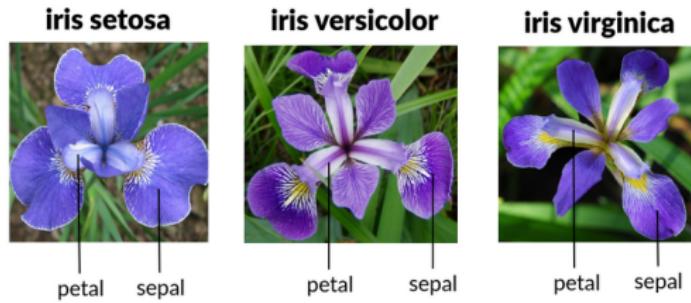
2 Aprendizado Supervisionado

- Viés-Variância Tradeoff
- Pré-processamento
- Treino, Validação, Teste
- Data Scaling
- Métricas de Regressão
- Métricas de Classificação
- Árvore de Decisão
- Random Forest
- LightGBM
- Tuning do Modelo, Otimização de Hiperparâmetro
- Interpretabilidade do modelo
- Shapley Values
- Aprendizado Não Supervisionado

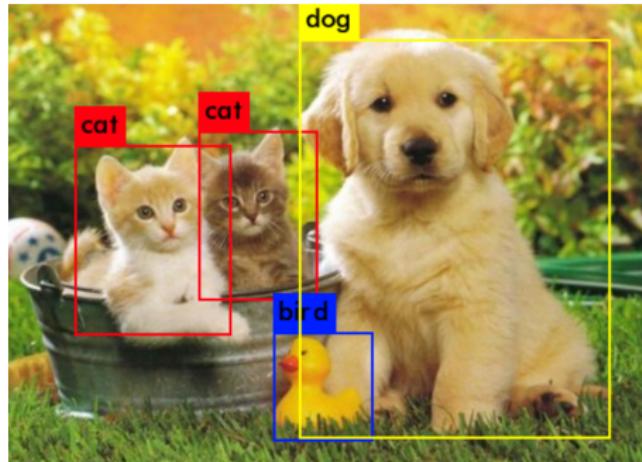
Predição de Preço de Imóveis



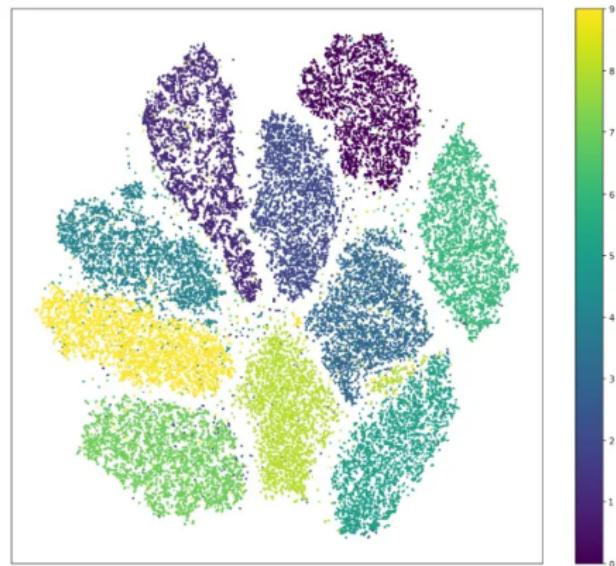
Classificação de Flores



Detecção e Classificação de Objetos



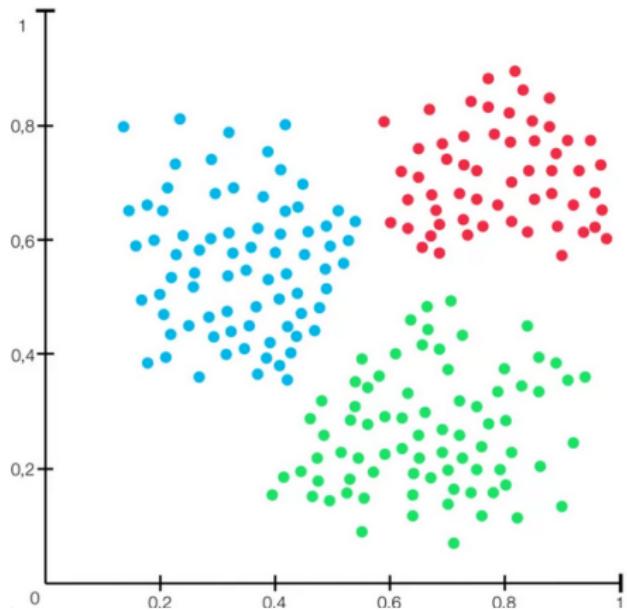
Reconhecimento de Dígitos



Clusterização de Músicas no Spotify



Clusterização e Segmentação de Clientes



Análise de Sentimentos



IA jogando jogos



State:247

R:	:	:	G
:			
:			
Y	:	:	B
+			+

State:99

R:	:	:	G
:			
:			
Y	:	:	B
+			+

State:475

R:	:	:	G
:			
:			
Y	:	:	B
+			+

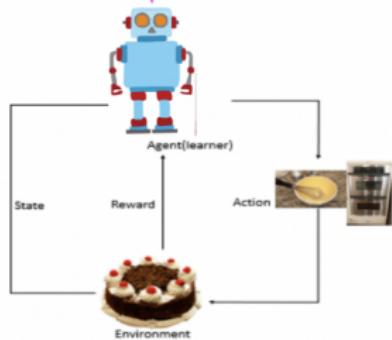
Machine Learning



$$V = \frac{4}{3}\pi r^3$$



Supervised Learning



Reinforcement Learning

Diferentes tipos de Aprendizagem

Aprendizado Supervisionado

Modelo aprende a partir de dados rotulados, ou seja as saídas são bem definidas com uma variável resposta é como se tivesse um 'professor'. Exemplos: prever o preço de imóvel, classificar imagens

Aprendizado Não Supervisionado

Envolve uma exploração intrínsecas nos dados, como agrupamentos ou distribuições, aqui não temos uma variável resposta.

Exemplos: agrupamento de tipos musicais, segmentação de clientes

Aprendizado por Reforço

Já o aprendizado por reforço é uma abordagem na qual um agente 'aprende' a realizar ações em um ambiente por meio de um estímulo (recompensa)

Exemplos: robótica, jogos

Métodos e ferramentas para resolver problema



Passos para Construção de um Modelo

- Entendimento do problema a ser resolvido (supervisionado, não supervisionado etc)
- Coleta dos dados
- Preparação dos dados, preenchimento de nulos, normalização (se necessário)
- Escolha do algorítimo
- Treinamento do modelo
- Validação da performance
- Tuning do modelo
- Modelo produtizado e realizando predições

Visão Geral dos Problemas Supervisionados

Objetivo: achar uma função que descrevam algum comportamento dado um conjunto de dados $(\mathbf{x}^{(i)}, y^{(i)})$, em que $y^{(i)} \in \mathbb{R}$ é a saída desejada

Dado um **conjunto de treino** com N pares de entrada-saída

$$\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

$\underbrace{\text{tamanho, cidade}}_{\text{input/features}} \rightarrow \underbrace{\text{valor}}_{\text{target/saída/variável resposta}}$

$$\text{input} \rightarrow \mathbf{x}^{(i)} \in \mathbb{R}$$

$$\text{output} \rightarrow y^{(i)} \in \mathbb{R}$$

$$f'(\mathbf{x}) \rightarrow y$$

Aprendizado Supervisionado: Regressão × Classificação

Regressão: a variável resposta é contínua e $y \in \mathbb{R}$

Exemplos: predição de preço de imóveis

Classificação: a variável resposta é discreta, sendo dividida em 'classes'.

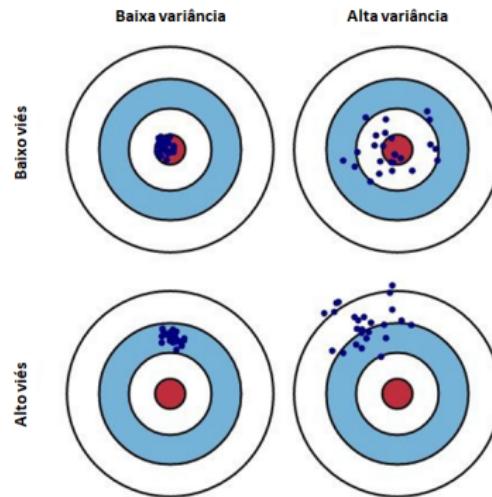
Geralmente, abordamos a classificação binária (0 ou 1)

Exemplos: classificação de flores, classificação de objetos em visão computacional, reconhecimento de dígitos

Viés-Variância Tradeoff

Viés: é o erro devido à diferença entre as previsões médias e os valores corretos que estamos tentando prever,

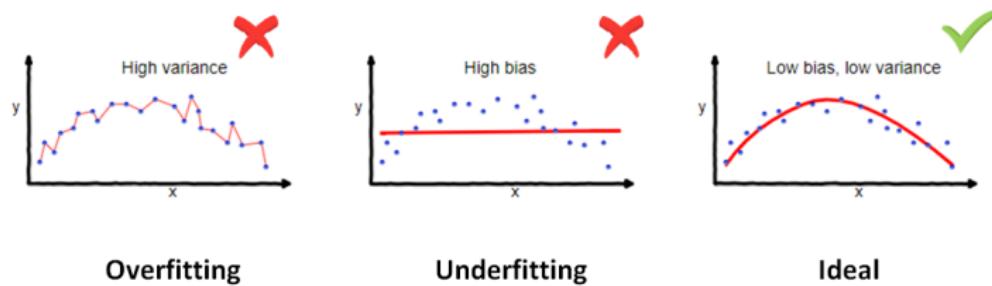
Variância: é o erro devido à variabilidade de uma previsão do modelo para um determinado ponto de dados



Viés-Variância Tradeoff

Overfitting: ocorre quando um modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados. Isso resulta em alta precisão nos dados de treinamento, mas baixa precisão nos dados de teste.

Underfitting: ocorre quando um modelo é muito simples para capturar os padrões subjacentes nos dados, resultando em baixa precisão tanto nos dados de treinamento quanto nos de teste.



Pré-processamento

Limpeza de Dados: Remoção de dados inconsistentes, incompletos ou irrelevantes. Isso pode incluir a correção de valores ausentes, eliminação de duplicatas e correção de erros.

Divisão dos Dados: Separação dos dados em conjuntos de treinamento e teste para validar a performance do modelo.

Transformação de Dados: Alteração da forma ou estrutura dos dados para adequar aos requisitos dos algoritmos. Isso pode incluir normalização, padronização e transformação logarítmica.

Codificação de Variáveis Categóricas: Conversão de dados categóricos em uma forma numérica que os algoritmos de aprendizado de máquina possam processar, como codificação one-hot ou label encoding.

Redução de Dimensionalidade: Eliminação de variáveis redundantes ou irrelevantes para simplificar o modelo e melhorar sua performance.

Tipo de Variáveis

Numéricas ou quantitativas: Contínuas e Discretas.

Categóricas ou qualitativas Ordinal (escolaridade, meses do ano, notas de provas) e Nominal (cor de olhos, profissão, região).

Treino, validação e teste



Conjunto de Treinamento: A amostra de dados usada para ajustar o modelo. O modelo vê e aprende com esses dados.

Conjunto de Validação: A amostra de dados usada para fornecer uma avaliação imparcial de um modelo ajustado no conjunto de treinamento enquanto ajusta os hiperparâmetros do modelo.

Conjunto de Teste: A amostra de dados usada para fornecer uma avaliação imparcial de um modelo final ajustado no conjunto de treinamento.

Data Scaling

É a normalização dos dados.

StandardScaler: Bom quando a distribuição segue a normal $x_{scaled} = \frac{x-\mu}{\sigma}$

MinMaxScaler: Uma transformação linear, geralmente quando tem média zero os dados $x_{scaled} = \frac{x-x_{min}}{x_{max}-x_{min}}$

RobustScaler: Escala conforme o intervalo interquartil, é bom quando se tem outliers $x_{scaled} = \frac{x-x_{med}}{x_{75}-x_{25}}$

Métricas de Regressão

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

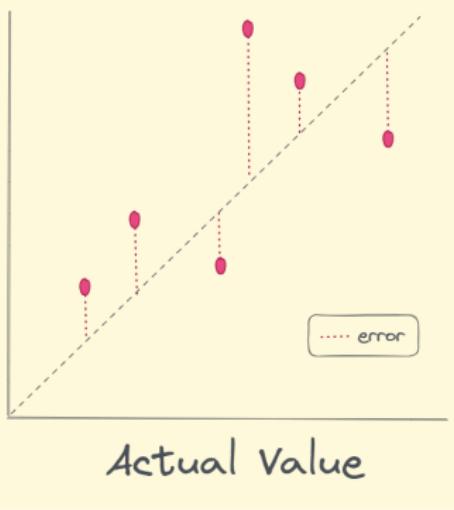
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

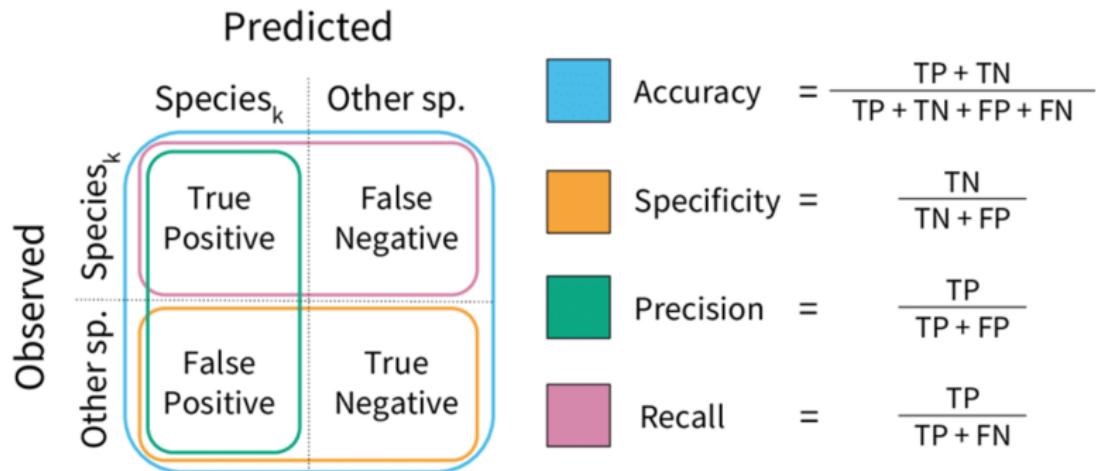
$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Predicted Value



Métricas de Classificação



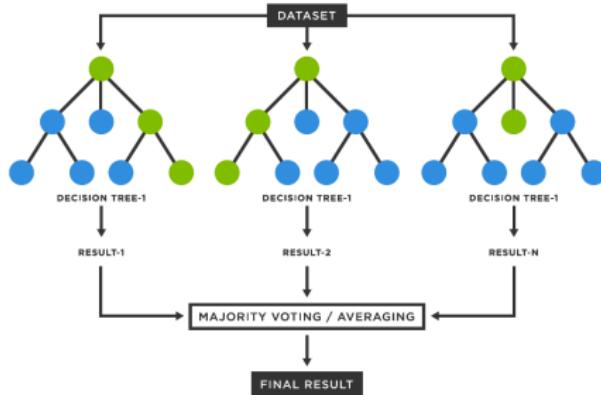
Arvore de Decisão



Random Forest

Bagging

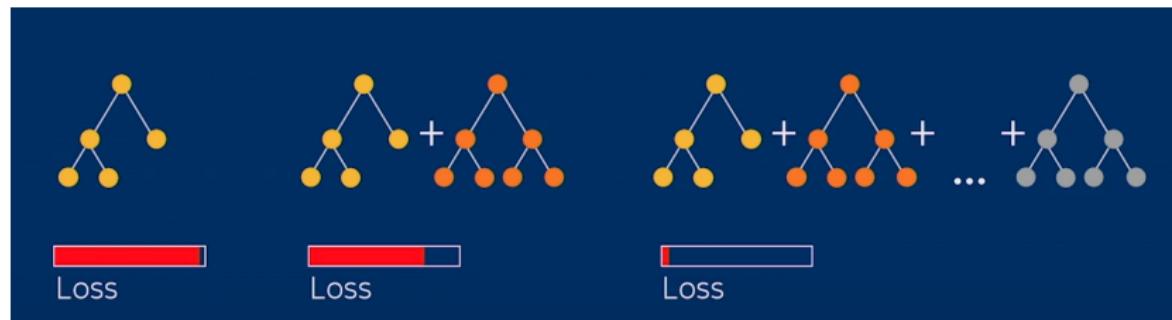
Bagging é um método de melhorar a precisão de modelos de aprendizado de máquina combinando múltiplos aprendizes fracos para criar um aprendiz forte. No bagging, várias amostras aleatórias do conjunto de dados original são usadas para treinar modelos individuais, e suas previsões são combinadas



Boosting

Boosting

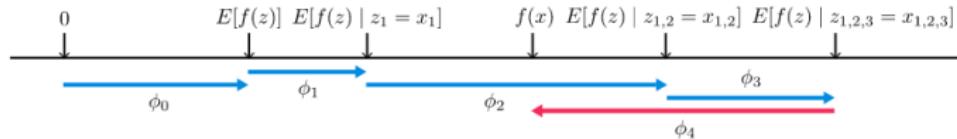
Boosting é um método de conversão de aprendizes fracos em aprendizes fortes. No boosting, cada nova árvore é ajustada em uma versão modificada do conjunto de dados original.



Shapley Values

SHAP (SHapley Additive exPlanations)

SHAP é uma abordagem baseada na teoria dos jogos para explicar a saída de qualquer modelo de aprendizado de máquina.



Aprendizado Não Supervisionado

Aprendizado Não Supervisionado

Envolve uma exploração intrínsecas nos dados, como agrupamentos ou distribuições, aqui não temos uma variável resposta.

© 2010梦工厂电影公司有限公司



Agrupamentos

Considerando um conjunto de N objetos a serem agrupados $X = \{x_1, x_2, \dots, x_N\}$, uma **partição** (rígida) é uma coleção de k grupos não sobrepostos $P = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k\}$ tal que:

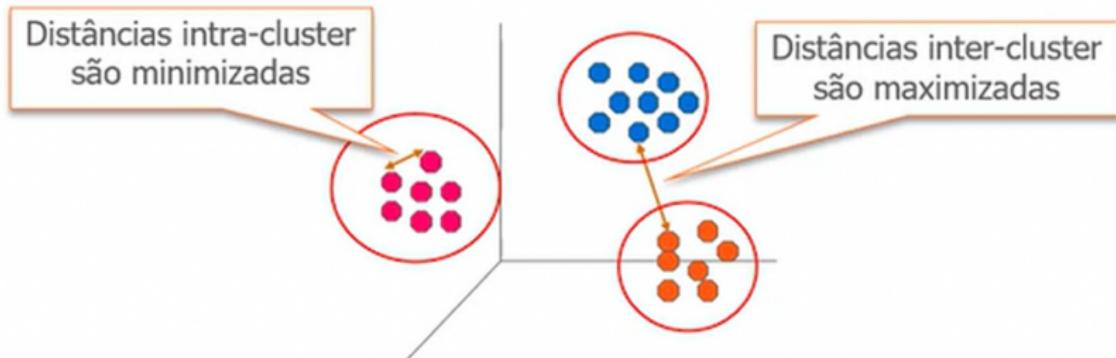
$$\begin{aligned}\mathcal{C}_1 \cup \mathcal{C}_2 \cup \dots \cup \mathcal{C}_k &= X \\ \mathcal{C}_i &\neq \emptyset \\ \mathcal{C}_i \cap \mathcal{C}_j &= \emptyset \text{ para } i \neq j\end{aligned}$$

Exemplo: $P = \{(x_1), (x_3, x_4, x_6), (x_2, x_5)\}$

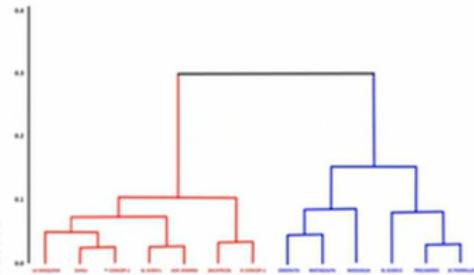
Agrupando MM's



Definição geométrica

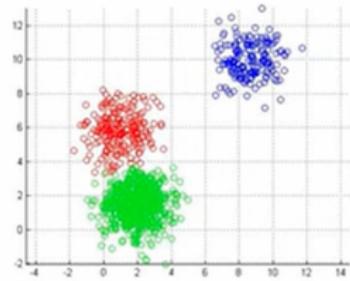


Tipos de Agrupamento



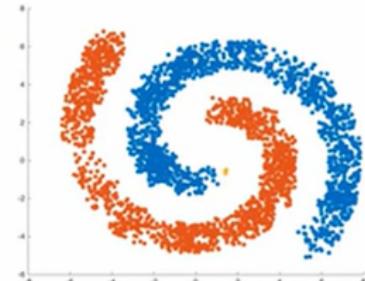
Hierárquico:

- Single Linkage
- Completed Linkage
- Average Linkage
- ...



Particionais:

- K-Means
- K-Median
- K-Medoid
- ...

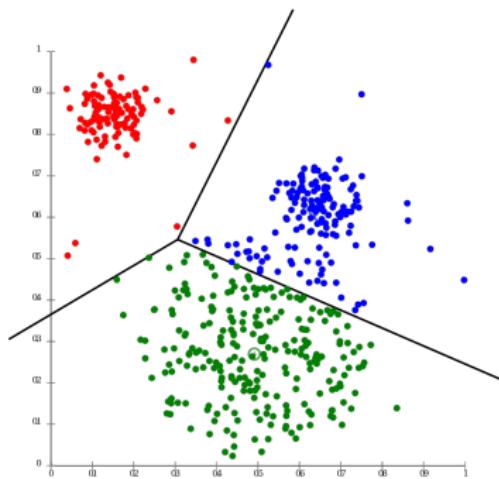


Densidade:

- DBScan
- ...

K-means

O algoritmo K-means é um método de clustering, ou agrupamento, que tem como objetivo partitionar um conjunto de dados em k grupos distintos. Cada ponto de dado pertence ao cluster cujo centróide (centro do cluster) é o mais próximo.

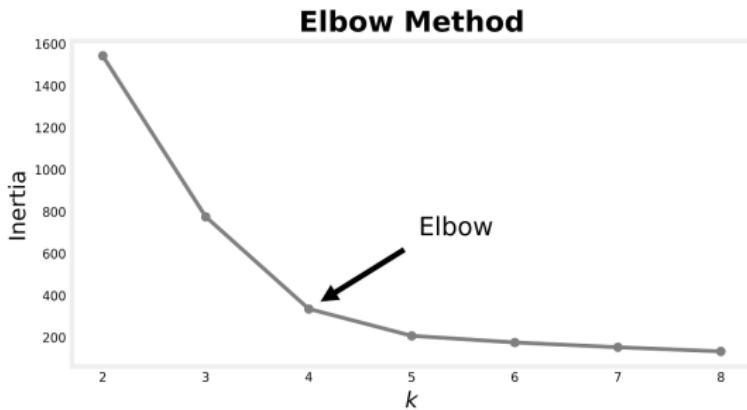


K-means

- 1 - Inicialização: Escolha o número de clusters k e Inicialize aleatoriamente k centróides.
- 2 - Atribuição de Clusters: Para cada ponto de dado, calcule a distância entre o ponto e cada um dos k centróides. Atribua o ponto ao cluster cujo centróide estiver mais próximo.
- 3 Atualização dos Centróides: Após todos os pontos terem sido atribuídos a um cluster, recalcular a posição dos k centróides. Isso é feito calculando a média de todos os pontos em cada cluster.
- 4 Repetição: Repita os passos de atribuição de clusters e atualização dos centróides até que os centróides não mudem significativamente entre as iterações ou até um número máximo de iterações ser atingido.

Método Elbow

A ideia central é executar o algoritmo K-means para diferentes valores de k e calcular a soma das distâncias quadradas dentro do cluster (WSS, Within-Cluster Sum of Squares). O WSS é a soma das distâncias quadradas de cada ponto até o centro do seu cluster.



Esse ponto sugere o número ideal de clusters. O "cotovelo" é onde a taxa de diminuição do WSS começa a se suavizar.

Silhouette Score

Para avaliar a qualidade do clustering, uma métrica comum é o Silhouette Score. Ele mede quão similar um ponto de dado é ao seu próprio cluster em comparação a outros clusters (separação). O Silhouette Score varia de -1 a 1:

Valores próximos de 1 indicam que os pontos estão bem agrupados.

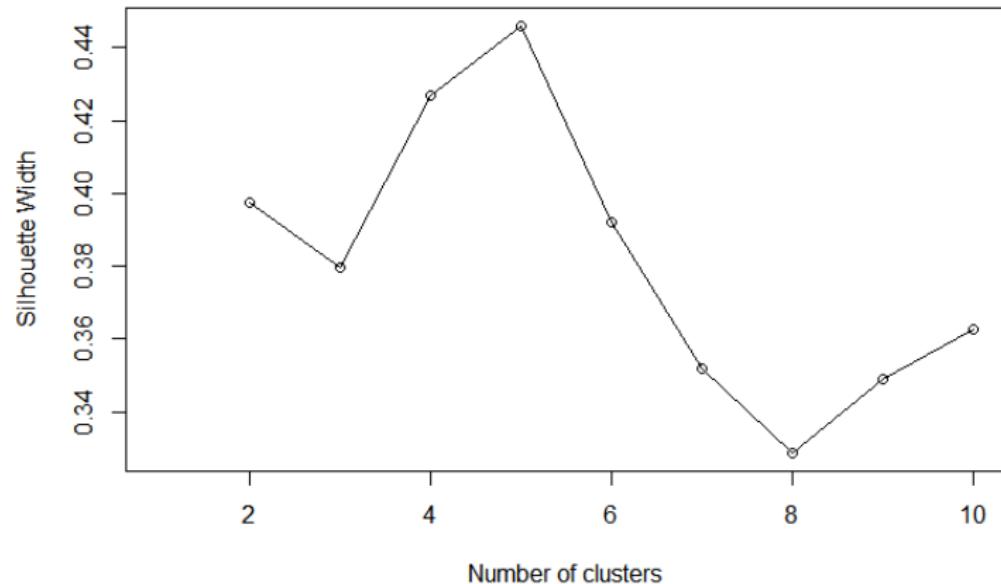
Valores próximos de 0 indicam que os pontos estão no limite entre dois clusters.

Valores negativos indicam que os pontos podem estar mal agrupados.

Calcule a distância média de $a(i)$ de i até todos os pontos do próprio cluster e calcule a distância média de $b(i)$ de i ao cluster mais próximo ao qual i não pertence

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Silhouette Score



Referencias



Mitchell T. M, (1997)

Machine Learning



Trevor Hastie and Robert Tibshirani and Jerome Friedman (2009)

The Elements of Statistical Learning



Stuart J. Russell and Peter Norvig (2021)

Artificial Intelligence: A Modern Approach, Global Edition



Ian Goodfellow, (2016)

Deep Learning



Sutton, Richard S. and Barto, Andrew G, (2018)

Reinforcement Learning: An Introduction