

ANÁLISE INTELIGENTE DE DADOS (COB 754)

K-NEAREST NEIGHBOR (K-NN)

LETÍCIA MARTINS RAPOSO

**“Diga-me com quem andas e
te direi quem és!”**

ALGORITMO DOS VIZINHOS MAIS PRÓXIMOS

CARACTERÍSTICAS

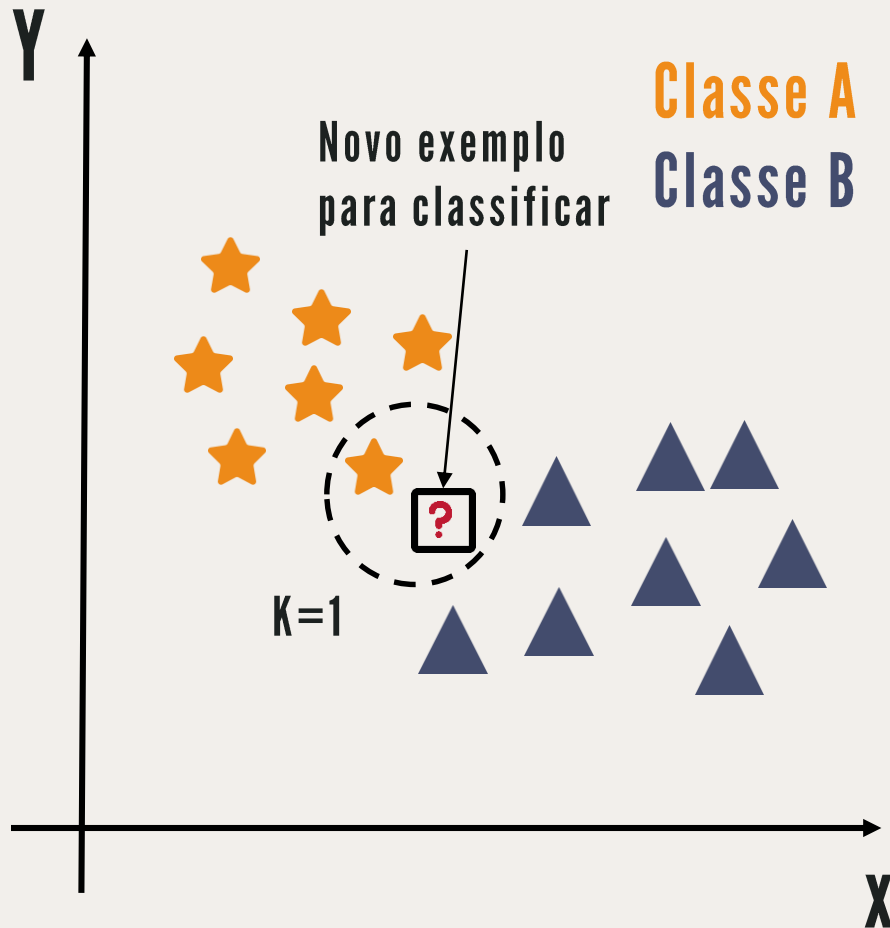
- ALGORITMO
SUPERVISIONADO
- NÃO PARAMÉTRICO
- PREGUIÇOSO
- BASEADO EM
INSTÂNCIAS
- USADO PARA
CLASSIFICAÇÃO E
REGRESSÃO

IDEIA

MEMORIZAR O CONJUNTO DE TREINAMENTO E DEPOIS PREDIZER O RÓTULO DE QUALQUER NOVA INSTÂNCIA COM BASE NOS RÓTULOS DE SEUS VIZINHOS MAIS PRÓXIMOS NO CONJUNTO DE TREINAMENTO



1-VIZINHO MAIS PRÓXIMO (1-NN)



CONSIDERA APENAS O VIZINHO MAIS PRÓXIMO

- CALCULA AS DISTÂNCIAS ENTRE CADA DOIS PONTOS.
- UM PONTO É O ROTULADO DO TREINAMENTO E O OUTRO É O QUE DESEJAMOS ROTULAR.
- O PONTO A SER ROTULADO RECEBE O RÓTULO DO EXEMPLO DE TREINAMENTO MAIS PRÓXIMO.

1-VIZINHO MAIS PRÓXIMO (1-NN)



O ALGORITMO PRESSUPÕE QUE OS ATRIBUTOS SÃO NUMÉRICOS

- QUALITATIVOS: CONVERTER!
- QUANTITATIVOS, MAS COM ESCALAS DIFERENTES: NORMALIZAR.
 - MEDIDAS DE DISTÂNCIAS SÃO AFETADAS PELA ESCALA DOS ATRIBUTOS.

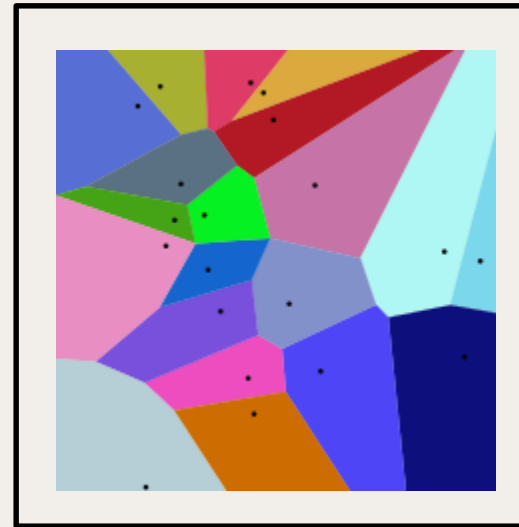
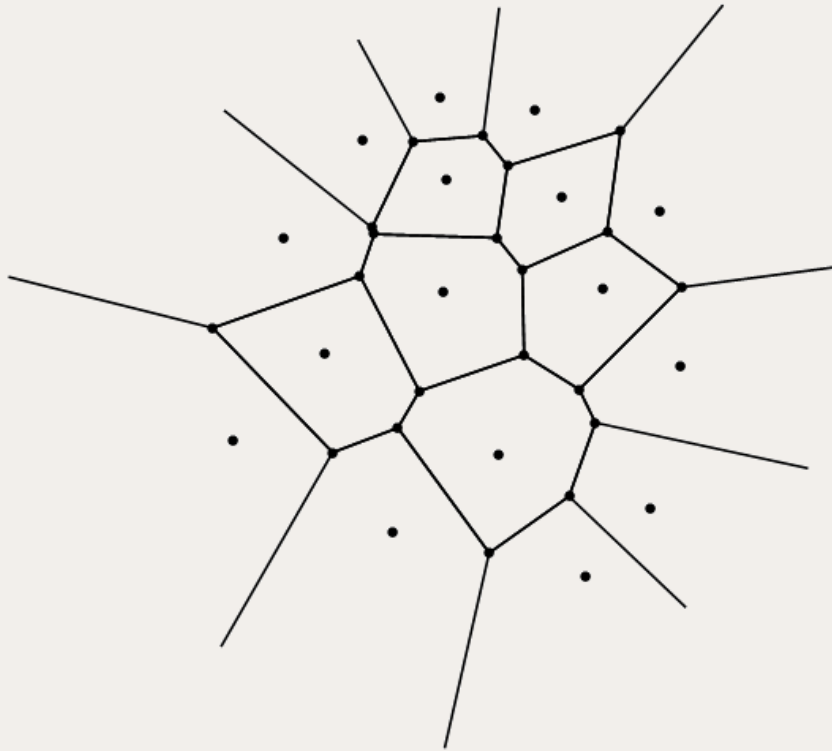


DIAGRAMA DE VORONOI

ILUSTRAÇÃO DOS LIMITES DE DECISÃO DA REGRA 1-NN. OS PONTOS DESCRITOS SÃO OS PONTOS DE AMOSTRA, E O RÓTULO PREVISTO DE QUALQUER NOVO PONTO SERÁ O RÓTULO DO PONTO DE AMOSTRA NO CENTRO DA CÉLULA À QUAL ELE PERTENCE.

K-NN

■ PODE SER USADO PARA PROBLEMAS PREDITIVOS DE CLASSIFICAÇÃO E REGRESSÃO

■ COMUMENTE USADO POR SUA FACILIDADE DE INTERPRETAÇÃO E BAIXO TEMPO DE CÁLCULO



CONJUNTO DE
EXEMPLOS DE
TREINAMENTO




DEFINIR UMA
MÉTRICA PARA
CALCULAR A
DISTÂNCIA ENTRE OS
EXEMPLOS DE
TREINAMENTO

K

DEFINIR O VALOR DE
K (O NÚMERO DE
VIZINHOS MAIS
PRÓXIMOS)

UTILIZAÇÃO DO K-NN

COMO O ALGORITMO K-NN FUNCIONA?

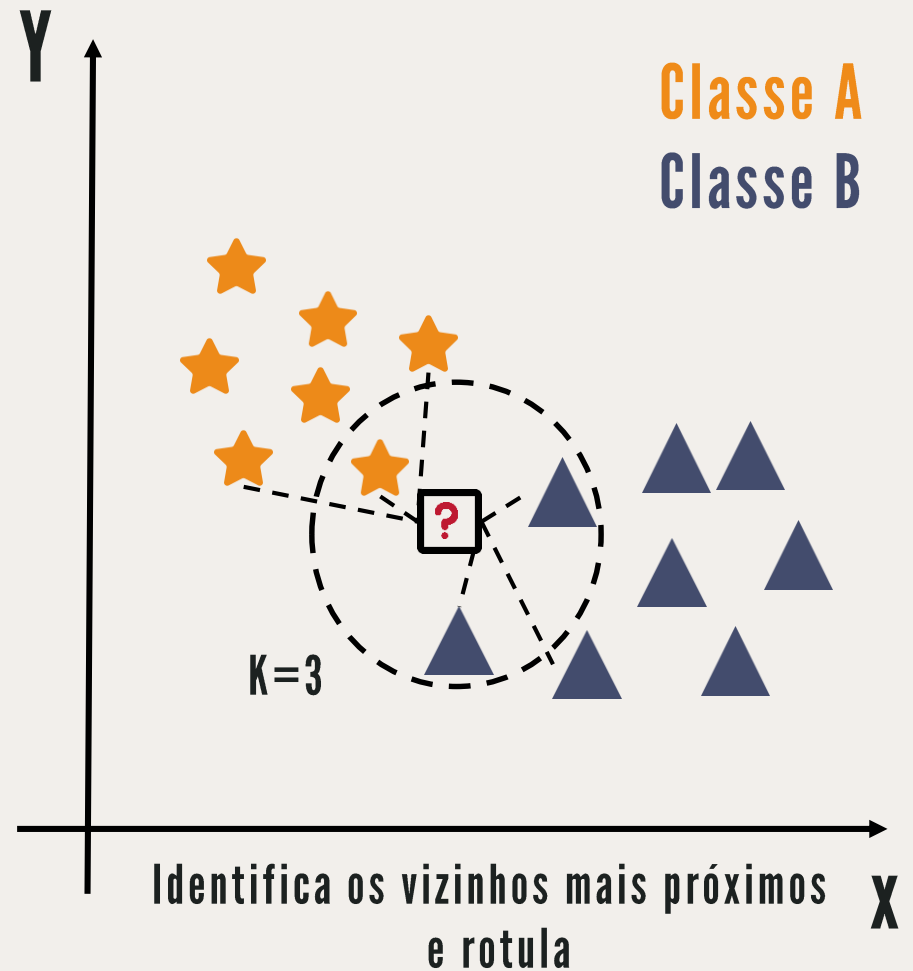
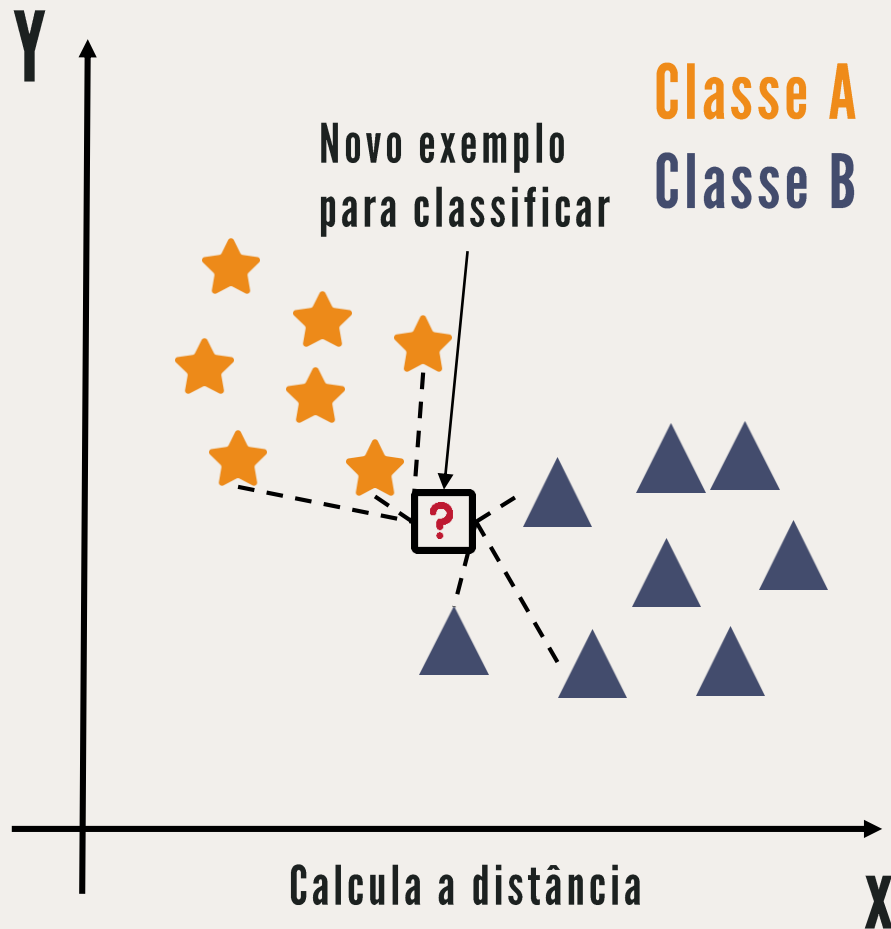


CALCULAR A
DISTÂNCIA ENTRE O
EXEMPLO
DESCONHECIDO E OS
OUTROS EXEMPLOS DO
CONJUNTO DE
TREINAMENTO

IDENTIFICAR OS K
VIZINHOS MAIS
PRÓXIMOS

UTILIZAR O RÓTULO
DOS VIZINHOS MAIS
PRÓXIMOS.
REGRESSÃO: MÉDIA
CLASSIFICAÇÃO: VOTO
MAJORITÁRIO

COMO O ALGORITMO K-NN FUNCIONA?



PREPARAÇÃO DOS DADOS PARA O K-NN



NORMALIZAÇÃO DOS DADOS

- Funciona muito melhor se todos os dados tiverem a mesma escala.
- Normalizar os dados para o intervalo $[0, 1]$ é uma boa ideia.

LIDAR COM DADOS AUSENTES

- Dados ausentes significam que a distância entre as amostras não pode ser calculada.
- Essas amostras podem ser excluídas ou os valores ausentes podem ser imputados.

REDUÇÃO DA DIMENSIONALIDADE

- Mais adequado para dados com dimensão reduzida.
- Pode se beneficiar da seleção de variáveis.

MEDIDAS DE DISTÂNCIA

- Determinar quais das K instâncias do conjunto de dados de treinamento são mais semelhantes a uma nova entrada.
- Para variáveis de entrada de valor real, a medida de distância mais popular é a **distância euclidiana**.

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^d (x_i^l - x_j^l)^2}$$

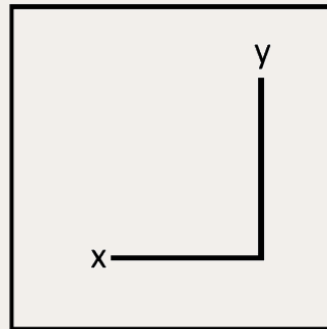
x_i, x_j : dois objetos representados por vetores do espaço \mathbb{R}^d ;
 x_i^l, x_j^l : são elementos desses vetores, que correspondem aos valores da coordenada l (atributos).

MEDIDAS DE DISTÂNCIA



DISTÂNCIA DE HAMMING = 3

1	1	0	1	1	1	0	0
1	1	1	1	0	1	1	0
XOR							
0	0	1	0	1	0	1	0



DISTÂNCIA DE HAMMING

Distância entre os vetores binários de igual comprimento: n° de posições em que os símbolos correspondentes são diferentes.

DISTÂNCIA DE MANHATTAN

$$d(x_i, x_j) = \sum_{i=1}^d |x_i^l - x_j^l|$$

DISTÂNCIA MINKOWSKI

$$d(x_i, x_j) = \sqrt[q]{\sum_{i=1}^d (x_i^l - x_j^l)^q}$$

MEDIDAS DE DISTÂNCIA

EXISTEM OUTRAS MEDIDAS DE DISTÂNCIA QUE PODEM SER USADAS, COMO A DISTÂNCIA DE JACCARD, MAHALANOBIS E COSSENO.

- A escolha da melhor medida de distância pode ser feita com base nas propriedades de seus dados.
- Se não tiver certeza, pode-se experimentar diferentes medidas e diferentes valores de K e ver qual mistura resulta nos modelos mais acurados.

COMO ESCOLHEMOS O K?

Nº PAR DE CLASSES - K IGUAL A UM NÚMERO ÍMPAR

- Empates: utilizar a classe da instância mais próxima.

K MUITO GRANDE

- Vizinhos podem ser muito diferentes;
- Predição tendenciosa para classe majoritária;
- Mais resistente a *outliers*.

K MUITO PEQUENO

- Apenas os objetos muito parecidos serão considerados;
- Predição pode ser instável;
- Sensível a *outliers*.

COMO ESCOLHEMOS O K?

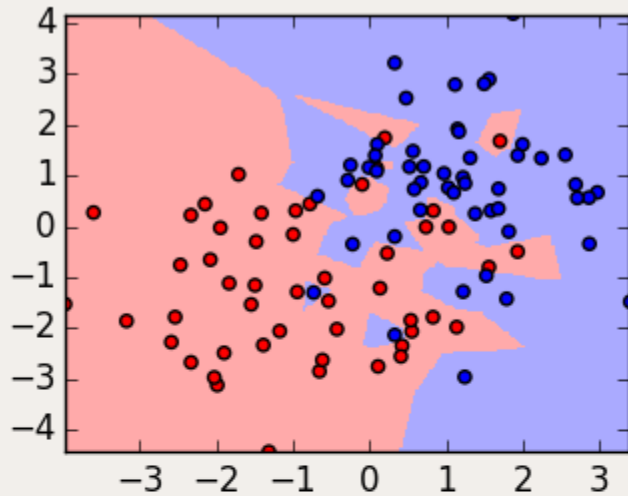
$$K = \sqrt{N}$$

TENTAR DIFERENTES VALORES
DE K

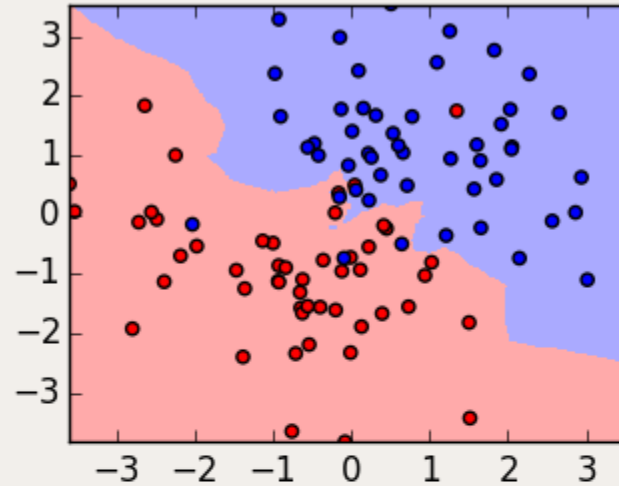
ESTIMAR K POR VALIDAÇÃO
CRUZADA

- Esse valor de K deve ser usado para todas as previsões.

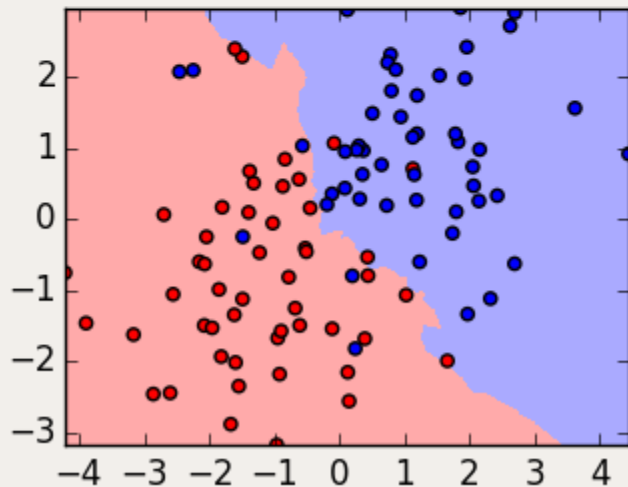
K=1



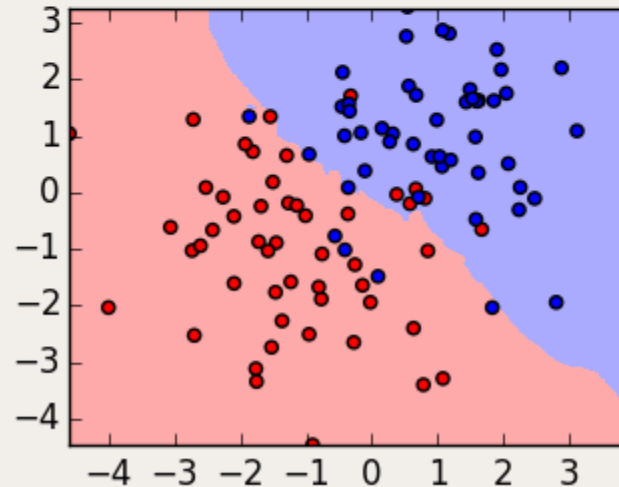
K=3



K=7



K=13



Com $k = 1$, a variância é muito grande. O modelo é sensível a todos os dados.

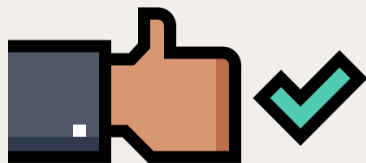
K-NN COM DISTÂNCIA PONDERADA

- Ponderar a contribuição de cada um dos K vizinhos de acordo com suas distâncias até o ponto x_j que queremos classificar, dando maior peso aos vizinhos mais próximos.
- Podemos ponderar o voto de cada vizinho de acordo com o quadrado do inverso de sua distância de x_j .

$$w_i = \frac{1}{d(x_i, x_j)^2}$$

- Porém, se $x_i = x_j$, o denominador torna-se zero. Neste caso, o ponto que queremos classificar receber a classe/valor do exemplo de treinamento.

VANTAGENS



NÃO TEM UM TREINAMENTO
PROPRIAMENTE DITO

APLICÁVEL EM PROBLEMAS
COMPLEXOS

FÁCIL DE ENTENDER

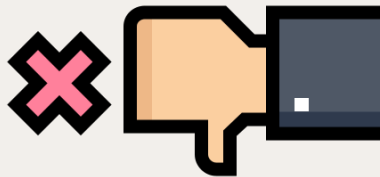
NENHUMA SUPOSIÇÃO
SOBRE DADOS

PODE SER APLICADO TANTO
À CLASSIFICAÇÃO QUANTO À
REGRESSÃO

FUNCIONA FACILMENTE EM
PROBLEMAS DE VÁRIAS
CLASSES

PODE SER PARALELIZADO

DESVANTAGENS



NÃO TEM UM MODELO
EXPLÍCITO

PODE SER CUSTOSO POR
COMPARAR O NOVO
EXEMPLO A CADA UMA DAS
INSTÂNCIAS DE
TREINAMENTO

AFETADO PELA PRESENÇA DE
VARIÁVEIS IRRELEVANTES OU
REDUNDANTES

SENSÍVEL À ESCALA DE
DADOS

ESCOLHER O MELHOR K
PODE SER DIFÍCIL

EXEMPLO REAL

Livros > Volta às Aulas > Universitários, Técnicos e Profissionais

Dê uma olhada



Introdução a ciência de dados (Português) Capa Comum – 31 dez 2015

por Fernando Amaral (Autor)

★★★★☆ 10 avaliações de clientes

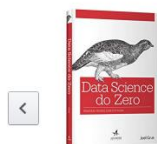
> Ver todos os 2 formatos e edições

eBook Kindle

Capa Comum

Clientes que visualizaram este item também visualizaram

Página 1 de 8



Data Science do zero
Joel Grus
★★★★☆ 56
Capa comum
R\$75,00



Data Science para negócios
Tom Fawcett
★★★★☆ 28
Capa comum
R\$68,56



Introdução à mineração de dados
Leandro Augusto Silva
★★★★☆ 29
Capa comum
R\$59,90



Big Data: O futuro dos dados e aplicações
Felipe Nery Rodrigues...
★★★★☆ 3
Capa comum
R\$66,22



Data Science Para Leigos
Lillian Pierson
Capa comum
R\$66,43



Aprenda mineração de dados
Amaral Fernando
★★★★☆ 8
Capa comum
R\$30,90



Data mining
Ronaldo Goldschmidt
★★★★☆ 8
Capa comum
R\$87,90



Data Smart. Usando Data Science Para Transformar Informação em Insight
John W. Foreman
★★★★☆ 6
Capa comum
R\$58,90



Ver todas as 2 imagens

Frequentemente comprados juntos



Preço total: R\$ 185,46

Adicionar os três ao carrinho

Estes itens são enviados e vendidos por vendedores diferentes. Ver detalhes

- ✓ Este item: Introdução a ciência de dados por Fernando Amaral Capa comum R\$ 41,90
- ✓ Data Science do zero por Joel Grus Capa comum R\$ 75,00
- ✓ Data Science para negócios por Tom Fawcett Capa comum R\$ 68,56

RESUMO



- ARMAZENA TODO O CONJUNTO DE DADOS DE TREINAMENTO.
- NÃO POSSUI UM MODELO EXPLÍCITO.
- FAZ PREDIÇÕES CALCULANDO A SIMILARIDADE ENTRE UMA OBSERVAÇÃO DE ENTRADA A CADA OBSERVAÇÃO DE TREINAMENTO.
- EXISTEM MUITAS MEDIDAS DE DISTÂNCIA, SENDO A EUCLIDIANA A MAIS FAMOSA.
- É UMA BOA IDEIA REDIMENSIONAR OS DADOS AO USAR O K-NN.