INTRODUÇÃO À CIÊNCIA DE DADOS

Algoritmo KNN (K-Nearest Neighbors) e Avaliação de Modelos Preditivos

KNN (K-Nearest Neighbors)

- KNN, também chamado de K-Vizinhos mais Próximos é um algoritmo bastante utilizado em vários projetos de DS.
- Veja que a Biblioteca Scikit-Learn disponibiliza uma grande variação de aplicações para o algoritmo.

KNN (K-Nearest Neighbors)

- Pode ser utilizado nos mais variados segmentos (áreas de negócios), veja que pode ser utilizado tanto para classificação quanto para regressão.
- Na classificação a máquina irá dizer a que grupo determinado registro faz parte, dentro obviamente de um contexto de negócio.
- Já a regressão irá nos fornecer um número/valor, por exemplo o valor de mercado de uma determinada casa que será colocada à venda.

KNN (K-Nearest Neighbors)

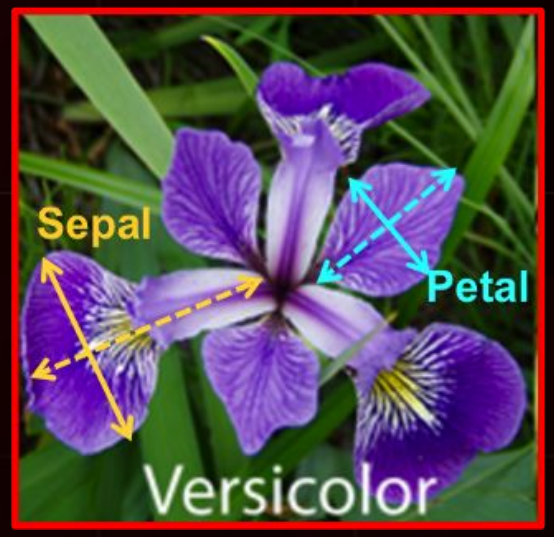
- O KNN é um dos modelos preditivos mais simples que existem. Ele não possui premissas matemáticas e não requer nenhum tipo de maquinário pesado. Ele apenas requer:
  - Noção de distância
  - Premissa de que pontos que estão perto um do outro são similares.

KNN - FUNCIONAMENTO

- O KNN determina pontos para cada registro e estabelece unidades de distância entre esses pontos.
- Algumas medidas de distância podem ser utilizadas:
  - Distância Euclidiana
  - Distância de Hamming
  - Distância Manhattan
  - Distância de Markowski
- Posteriormente, para cada novo ponto, ele calcula a distância entre os K vizinhos mais próximos. Procure sempre usar um K ímpar. E de acordo com os mais próximos ele é capaz de decidir sobre o novo ponto.

EXEMPLO KNN

- Um dos datasets mais conhecidos para aprender ML é o Iris Flower. Ele é formado por features (atributos) que indicam a largura e o comprimento da Pétala e da Sépala das flores indicando qual é a classificação delas.



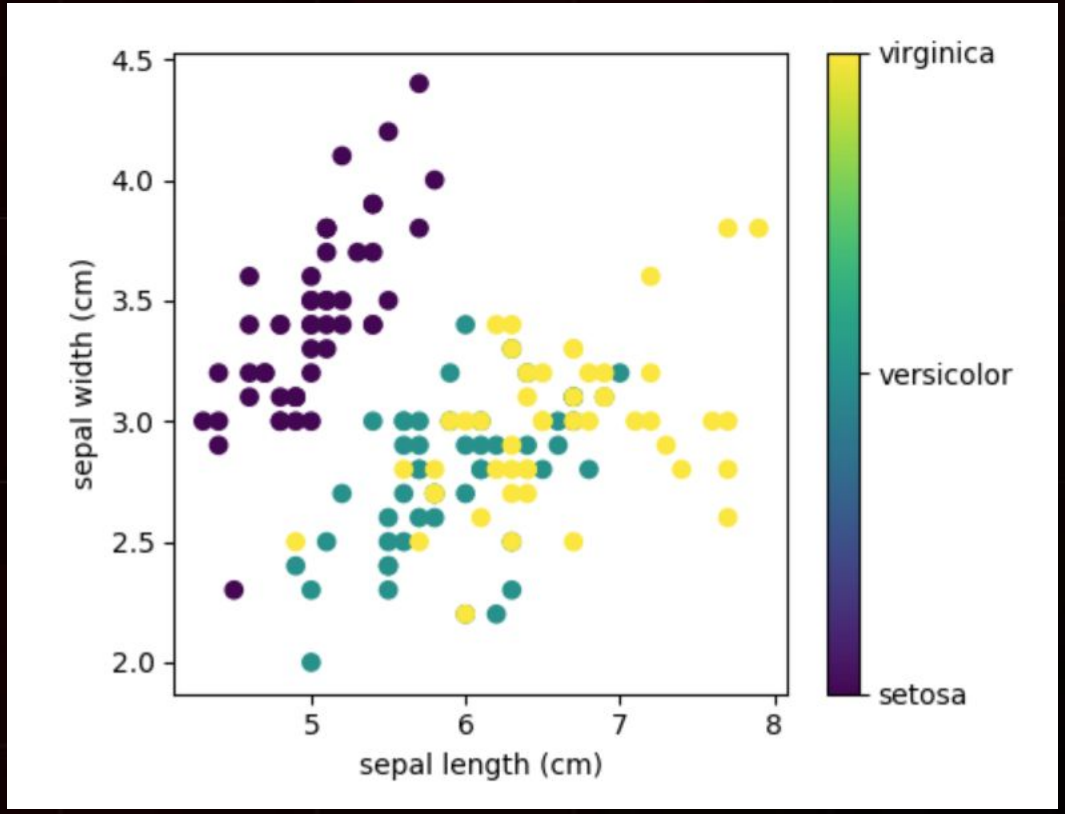
EXEMPLO KNN

- Estrutura básica do Dataset Iris Flower.

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target_name
5.8	4.0	1.2	0.2	setosa
7.3	2.9	6.3	1.8	virginica
6.4	2.7	5.3	1.9	virginica
4.9	2.4	3.3	1.0	versicolor
5.1	3.8	1.9	0.4	setosa
6.4	3.2	4.5	1.5	versicolor
5.4	3.7	1.5	0.2	setosa
5.9	3.0	4.2	1.5	versicolor
4.6	3.4	1.4	0.3	setosa
6.3	3.4	5.6	2.4	virginica
6.2	3.4	5.4	2.3	virginica
6.7	3.0	5.0	1.7	versicolor

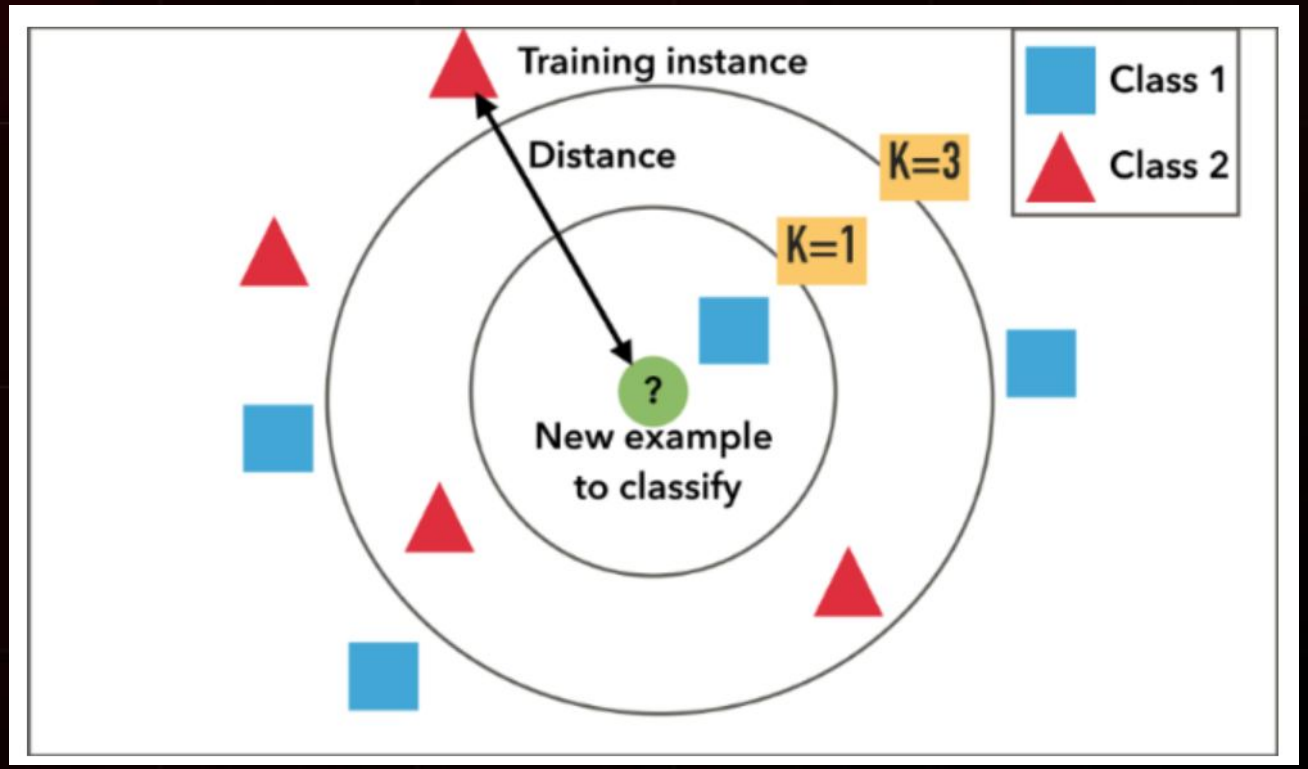
EXEMPLO KNN

- Pontos plotados (aqui apenas 2 variáveis).



EXEMPLO KNN

- Escolha, veja a quantidade de vizinhos para decidir (K).
- Se K=1, o novo elemento verde (círculo) será classificado como azul (quadrado), se o K=3, o novo será vermelho (triângulo)



KNN DEMONSTRAÇÃO

- Demonstração, apenas para apresentar como é o comportamento.

AVALIAÇÃO DE MODELOS

- De maneira geral, pode-se afirmar que não existe técnica universal, ou seja, não é possível estabelecer, a priori, que uma técnica de ML em particular se sairá melhor na resolução de qualquer tipo de problema.

- Por exemplo, em domínios em que os exemplos possuem alta dimensionalidade, as SVMs são boas candidatas, enquanto o algoritmo KNN usando a distância euclidiana pode, a princípio, não parecer uma escolha adequada.

AVALIAÇÃO DE MODELOS

- Mesmo com o uso dessas heurísticas, diversos algoritmos podem ser candidatos à solução de um problema.

- Ainda que um único algoritmo seja escolhido, pode-se realizar ajustes em seus parâmetros livres, o que leva à obtenção de múltiplos modelos para os mesmos dados.

AVALIAÇÃO DE MODELOS

- Fica clara a necessidade de experimentação.
- Dessa forma, é recomendável seguir procedimentos que garantam a correção, a validade e a reprodutibilidade dos experimentos realizados e, mais importante, das conclusões obtidas a partir de seus resultados.
- Essa avaliação experimental de um algoritmo de ML pode ser realizada segundo diferentes aspectos, tais como acurácia do modelo gerado, compreensibilidade do conhecimento extraído, tempo de aprendizado, requisitos de armazenamento do modelo, desempenho obtido nas predições realizadas, entre outros.

MÉTRICAS

- A avaliação de um algoritmo de ML supervisionado é normalmente realizada por meio da análise do desempenho do preditor gerado por ele na rotulação de novos objetos, não apresentados previamente em seu treinamento.

- Métricas diferentes podem ser aplicadas para classificação, para regressão e também várias outras alternativas como amostragem, bootstrap, validação cruzada, curva roc, entre muitos outros.

MÉTRICAS

- Para classificação pode-se utilizar a acurácia, que é uma métrica que informa a porcentagem de vezes que o modelo acertou. Há outras alternativas mais complexas e que podem ser utilizadas também, como matriz de confusão, que permite determinar revocação e precisão por exemplo.

- Para regressão é comum usar métricas que determinem a diferença de valor entre o valor previsto e o valor correto. Algumas medidas como distância média absoluta (Mean Absolute Distance - MAD) e erro quadrático médio (Mean Square Error - MSE), além das suas variações normalizadas, podem ser utilizadas.

AMOSTRAGEM

- Uma métrica bastante utilizada é dividir os dados disponíveis em duas partes. Utiliza-se muito a proporção 75/25, mas não é regra (vai depender dos seus dados).

- Neste caso, utiliza-se a parte maior, chamada TREINO, para treinar o modelo e depois realiza-se a predição com a outra parte dos dados, chamada TESTE. Esse é um passo interessante antes de se colocar dados novos para testar o modelo.

VALIDAÇÃO CRUZADA

- Essa métrica é conhecida como *r-fold cross-validation*.
- Métrica muito interessante, que usa o processo semelhante ao de amostragem, entretanto ele divide em vários subconjuntos de dados e repete o treinamento nos vários conjuntos.
- O desempenho final do preditor é dado pela média dos desempenhos observados entre cada conjunto de teste.

FINALIZANDO

- Abordamos o KNN, que é um algoritmo simples e que permite realizar vários tipos de operação (classificação, regressão…).
- Já conseguiu entender a diferença entre classificação e regressão? Pense sempre em um valor categórico (label) para classificação e um valor numérico para a regressão.
- As medidas para avaliação dos modelos são muito importantes.
- Na última semana nos dedicaremos a um projeto completo, com abordagem sobre todos esses elementos!

REFERÊNCIAS

- Parte do conteúdo desta aula é baseado nos livros:

CARVALHO, André Carlos Ponce de Leon Ferreira et al. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Disponível em: Minha Biblioteca, (2nd edição). Grupo GEN, 2021.

Grus, Joel. Data Science do Zero. Disponível em: Minha Biblioteca, (2nd edição). Editora Alta Books, 2021.

INTRODUÇÃO À CIÊNCIA DE DADOS

Algoritmo KNN (K-Nearest Neighbors) e Avaliação de Modelos Preditivos