Prática knn - Modelos baseados em distâncias

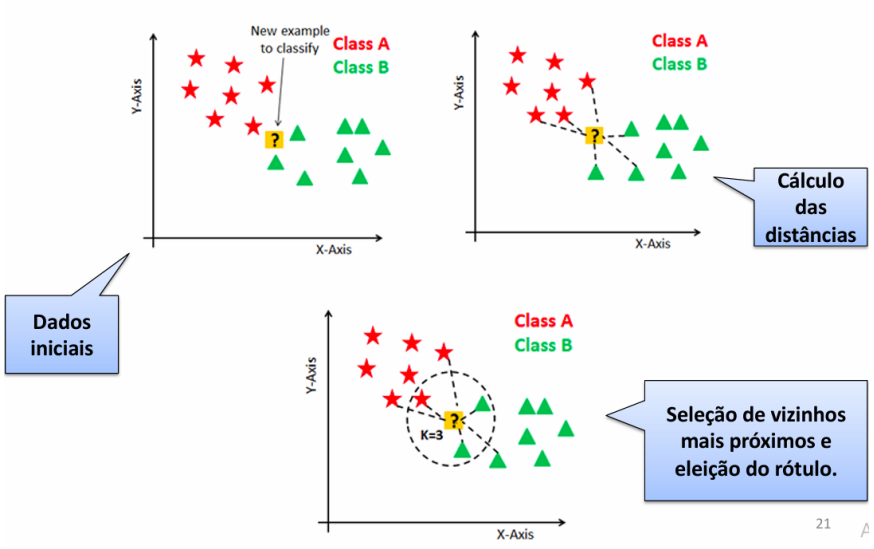
KNN → k-Nearest Neighbors

- Modelos que consideram a proximidade de objetos para a realização de predições.
- Supõe-se que objetos similares geralmente se concentram em uma mesma região de espaço.
- Não há aprendizado no sentido de indução de uma função hipótese. Há uma memorização do conjunto de treinamento.
    - Esses modelos não generalizam criando uma função explícita a partir dos dados, eles mantêm uma cópia completa do conjunto de treinamento e usam a comparação direta para fazer predições. Isso pode ser computacionalmente intenso para datasets muito grandes. Mas acaba sendo eficaz quando o padrão de similaridade entre os dados é um bom indicador das predições desejadas.
- Muito útil em situações onde a definição explícita de uma função de mapeamento é difícil de ser aplicada mas onde a similaridade local dos dados é um forte indicador de comportamento ou classificação.

O Algoritmo

Pode ser dividido em três etapas:

- Calcular as distâncias
- Encontrar os K vizinhos mais próximos
- Eleger o rótulo mais adequado



A distância pode ser calculada com 3 fórmulas: euclidiana, manhattan, minkowski.

- *Para p = 1, temos a distância de Manhattan.*
- *Para p = 2, a distância euclidiana.*
- *Para p = ∞, a dimensão dominante é destacada.*

A distância euclidiana é mais sensível a pequenas modificações do que a distância Manhattan.

- Não é adequada para dimensões com muito ruído.
- Um valor de p alto pode dar muita ênfase a dimensões com outliers.

> 💡 **Outliers**: São valores de dados que são significativamente diferentes dos outros pontos de dados em uma determinada dimensão. Em outras palavras, são valores extremos ou anômalos.
>
> **Dimensões com Outliers**: Em um espaço multidimensional, uma dimensão pode ter outliers, ou seja, valores que são muito maiores ou menores em comparação com a maioria dos dados.

Considerações

- A escolha do K não é trivial. Normalmente é um valor pequeno e ímpar.

    - A integração com algoritmos evolutivos é uma alternativa
- É considerado um algoritmo preguiçoso.
    - Maior parte da computação ocorre no momento da classificação
- Vantagens:
    - É intuitivo e simples de implementar.
    - É incremental. Novos exemplos são adicionados sem gerar esforço incremental.
    - Poucos parâmetros a serem ajustados.
- Desvantagens:
    - Necessário recalcular as distâncias para cada novo ponto a ser rotulado.
    - Susceptível a atributos redundantes e irrelevantes.
    - Com o aumento no número de dimensões, há um salto na magnitude das distâncias. A distância do vizinho mais próximo aproxima-se da do mais afastado.
    - Necessidade de normalização dos valores das dimensões.
- Aplicações:
    - Reconhecimento facial
    - Identificação de padrões de fraude na utilização de cartões de crédito.
    - Identificação de padrões de compra em lojas varejistas.
    - Sistemas de recomendação.
    - Benchmark para modelos mais sofisticados.