Aprendizagem Computacional

202106775 Guilherme Vaz
202106968 Pedro Campião
202107547 Ricardo Costa

Trabalho Prático

Objetivos

- Compreender o funcionamento de um determinado algoritmo de Machine Learning.
- Como podemos alterar um algoritmo de ML de modo a melhorar o seu desempenho.
- Como avaliar o desempenho de um algoritmo de ML num determinado contexto.

Abordagem

Alteramos o método de medir distâncias do algoritmo K-nearest neighbors e avaliamos o seu impacto em diferentes tipos de conjuntos de dados.

Sumário de Resultados

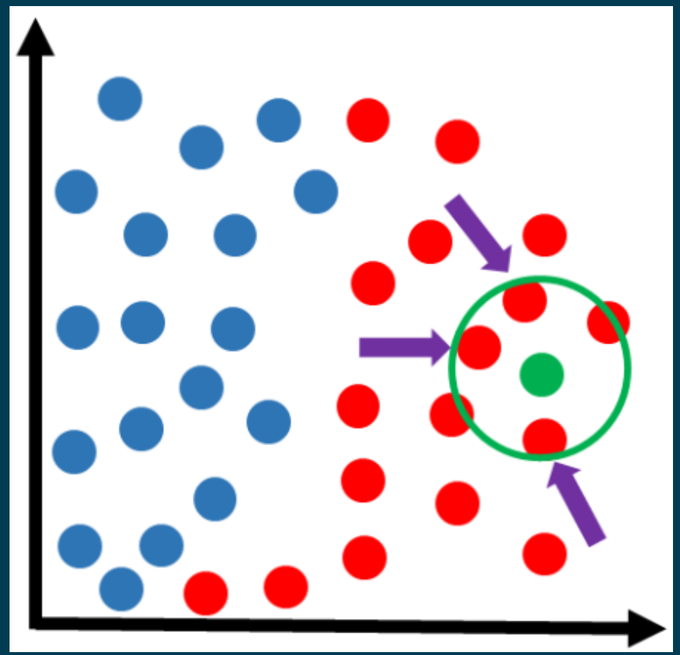
Concluímos que certos métodos de calcular distâncias conduzem a um melhor desempenho para certos tipos de dados.

Algoritmo KNN

O algoritmo KNN é uma técnica de Machine Learning utilizado e aplicado em problemas de classificação.

Cada amostra de um conjunto de dados é classificada, a partir das amostras vizinhas mais próximas
Essa aproximação é avaliada, através do calculo das distâncias entre amostras.

1) Recebe um dado por classificar e é medido a sua distância em relação a outros dados já classificados.
2) Seleciona-se as k menores distâncias.
3) Verifica-se as classes das amostras que tiveram as k menores distâncias e contabiliza-se a quantidade de vezes que cada classe apareceu.
4) Classifica-se o novo dado com a classe mais contabilizada.



Comportamento do KNN consoante as características dos dados

Este algoritmo tem um bom desempenho quando as classes do conjunto de dados têm regiões distintas e separadas no espaço de características.

O algoritmo KNN produz bons resultados quando o conjunto de dados não contém outliers.

Se certas classes do conjunto de dados estiverem desequilibradas, ou seja, uma classe tiver mais exemplos do que outras, o KNN pode ser influenciado negativamente.

O desempenho do KNN tende a piorar em conjuntos de dados de alta dimensionalidade, uma vez que a noção de vizinhança torna-se menos significativa.

PROPOSTA

O algoritmo padrão KNN utiliza o método das distâncias euclidianas para determinar as amostras vizinhas mais próximas.

Apesar deste método ter um bom desempenho para todo o tipo de dados, existem outras medidas de distância que aumentam a precisão do algoritmo KNN.

A nossa proposta baseia-se em explorar e implementar diferentes métodos de cálculo de distância no algoritmo KNN e analisar de que maneira afetará os resultados.

Distâncias

Euclidiana

Calculada como a raiz quadrada da soma dos quadrados das diferenças entre os componentes dos vetores.

É frequentemente usada para dados numéricos contínuos.

Hamming

Corresponde ao menor número de substituições necessárias para transformar um vetor binário no outro.

Adequado para conjuntos de dados que utiliza dados categóricos binários.

Manhattan

Determinada pela soma das diferenças absolutas entre os componentes dos vetores.

É uma alternativa à distância euclidiana para dados numéricos contínuos.

Jaccard

É uma medida de similaridade entre conjuntos e calculada como a diferença entre a interseção e a união dos conjuntos.

É adequada quando os dados são categóricos binários, multivalor ou uma combinação de ambos.

Chebyshev

Calculada a partir da maior diferença absoluta entre os componentes dos vetores.

Frequentemente utilizada para dados numéricos contínuos e quando os atributos têm escalas distintas.

Motivação da proposta

EXPLORAÇÃO DE TÉCNICAS	• Compreender a influência no desempenho dos diferentes métodos utilizados para calcular a distância.
COMPREENSÃO DO ALGORITMO	• Perceber a flexibilidade e versatilidade do algoritmo K-Nearest Neighbors.
EXTRAIR O MELHOR RESULTADO	• Modificar o algoritmo para produzir o melhor resultado possível para qualquer conjunto de dados.

Conjunto de dados e as suas características

Letras

Conjunto de dados que identifica a letra conforme as suas medidas

Conjunto só com valores numéricos.

6100 entradas e 17 colunas

Tic Tac Toe

Conjunto de dados que verifica se o jogo tem vencedor, a partir do estado atual do tabuleiro.

Conjunto só com valores categóricos binários.

959 entradas e 10 colunas

Riscos de Crédito

Conjunto de dados que avalia o risco de crédito.

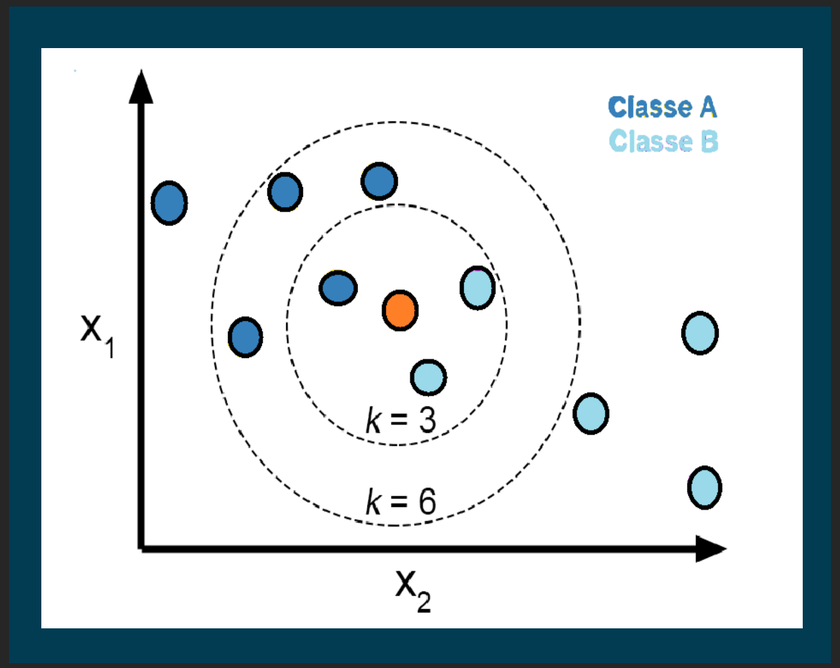
Conjunto dividido entre valores categóricos e numéricos

1000 entradas e 21 colunas

Hiperparâmetros

Para o desenvolvimento deste algoritmo, utilizamos um hiperparâmetro **K**, que determina o número de amostras vizinhas mais próximas que serão consideradas para a classificação da amostra teste.

A escolha do valor de **K** depende da natureza dos dados, do número de amostras de treino disponíveis, da dimensionalidade do espaço de recursos e do problema específico a resolver



Método de Estimativa de Desempenho

Para medir e analisar a performance dos diferentes métodos de calcular distâncias no algoritmo KNN, utilizamos a função “accuracy score” da biblioteca “sklearn”.

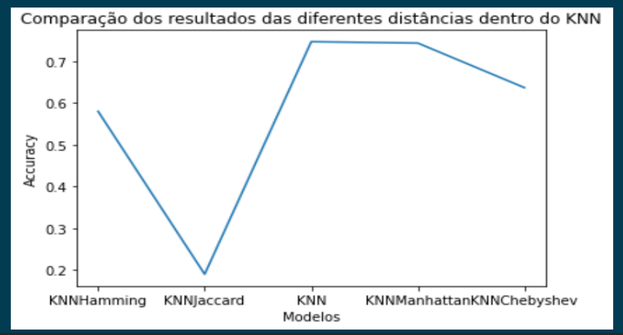
Esta função recebe os rótulos verdadeiros das amostras e os rótulos previstos pelo algoritmo e de seguida conta o número de previsões corretas. Posteriormente, calcula o valor de precisão ao dividir o número de previsões corretas peço número total de amostras.

Desempenho de cada medida de distância

Letras

Conjunto só com valores numéricos.

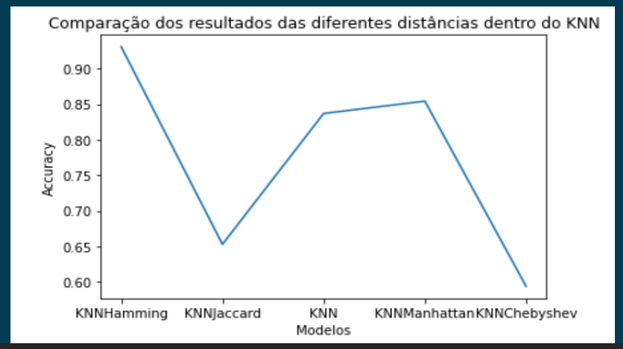
- KNN – **Padrão**: 0.747
- KNN – **Manhattan**: 0.743
- KNN –**Chebyshev**: 0.637
- KNN - **Hamming**: 0.58
- KNN – **Jaccard**: 0.19



Tic Tac Toe

Conjunto só com valores categóricos binários.

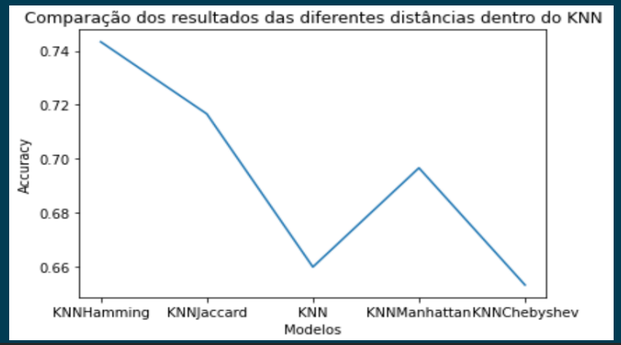
- KNN – **Padrão**: 0.8368
- KNN – **Manhattan**: 0.8542
- KNN –**Chebyshev**: 0.5938
- KNN - **Hamming**: 0.931
- KNN – **Jaccard**: 0.653



Riscos de Crédito

Conjunto de dados que avalia o risco de crédito.

- KNN – **Padrão**: 0.66
- KNN – **Manhattan**: 0.697
- KNN –**Chebyshev**: 0.653
- KNN - **Hamming**: 0.743
- KNN – **Jaccard**: 0.717



Algoritmo original vs algoritmo proposto

O algoritmo original superiorizou ligeiramente o algoritmo proposto com uma pontuação de precisão de 0.747, enquanto que o algoritmo com várias medidas implementadas teve 0.743.

O algoritmo KNN padrão teve uma precisão inferior comparando com o algoritmo proposto em relação a um conjunto de dados categóricos. O algoritmo original teve uma exatidão de 84%, enquanto que o algoritmo proposto teve 93%.

Em dados repartidos entre categóricos e numéricos também verificamos uma superioridade do algoritmo proposto.
O algoritmo alterado acertou 74% das suas previsões e por sua vez, o algoritmo padrão apenas acertou 66%.

Conclusão

Após implementar o algoritmo com as alterações propostas em três conjuntos de dados bastante distintos, verificamos que o método euclidiano de calcular distâncias (método padrão) teve um bom desempenho não só em dados numéricos, como também em dados categóricos.

Porém, em dois dos três casos, verificamos que **o maior** valor de precisão **não** surgiu do método padrão, mas sim de um dos métodos implementados.

Assim, provamos que é uma vantagem utilizar o algoritmo KNN com vários métodos de distância implementados, uma vez que nos permite obter um melhor desempenho.

QUESTÕES