Nome: Fernanda Fernandes Borges

Período: 8

Inteligência Artificial

O que é KNN?

R: KNN: K-nearest-neighbor (K-vizinho-próximo)

Utilizado para classificar objetos com base em exemplos de treinamento que estão mais próximos no espaço de características;

O parâmetro k é o número de vizinhos mais próximos.

Para utilizar o KNN é necessário: – Um conjunto de exemplos para treinamento; – Definir uma métrica para calcular a distância entre os exemplos de treinamento; – Definir o valor de K.

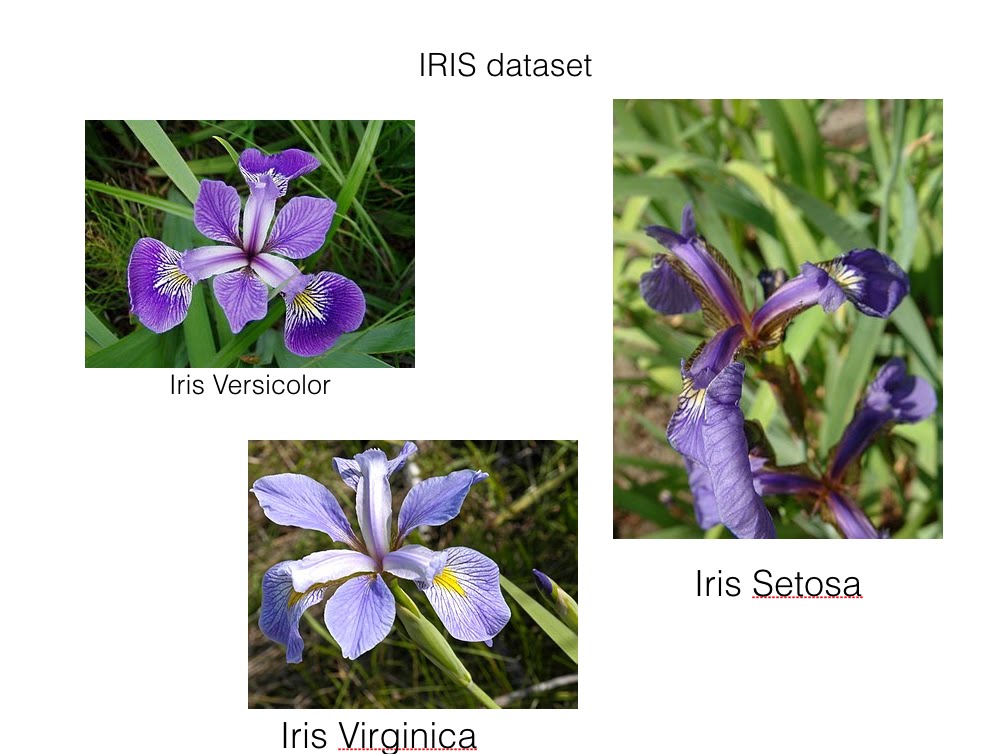
Classificar um exemplo desconhecido com o algoritmo KNN consiste em: – Calcular a distância entre o objeto desconhecido e os outros exemplos do conjunto de treinamento; – Identificar os K vizinhos mais próximos; – Utilizar o “rótulo” (nome) da classe dos vizinhos mais próximos para determinar o rótulo de classe do exemplo desconhecido (votação majoritária).

Calculo da distância entre dois pontos: – A mais simples é a distancia euclidiana: Outras formas de medir distâncias: – Distância de Mahalanobis, Minkowsky, Hamming, entre outros.

Como escolher o valor de K? – Se K for muito pequeno, a classificação fica sensível a pontos de ruído; – Se k for elevado, a vizinhança pode incluir elementos de outras classes. \*normalmente, escolhe-se um valor ímpar para K (evita empates desnecessários no cálculo).

O conjunto de dados de flores de Iris ou o conjunto de dados de Fisher Iris é um conjunto de dados multivariados introduzido pelo estatístico britânico e biologista Ronald Fisher em seu artigo de 1936. O uso de medidas múltiplas em problemas taxonômicos como exemplo de análise discriminante linear . Às vezes é chamado de conjunto de dados Iris de Anderson porque Edgar Anderson coletou os dados para quantificar a variação morfológica de flores de íris de três espécies relacionadas. Duas das três espécies foram coletadas na Península de Gaspé "todas da mesma pastagem, e colhidas no mesmo dia e medidas ao mesmo tempo pela mesma pessoa com o mesmo aparelho".

O conjunto de dados consiste em 50 amostras de cada uma das três espécies de íris



Iris setosa , Iris virginica e Iris versicolor ). Foram medidas quatro características de cada amostra: o comprimento e a largura das sépalas e pétalas , em centimetros. Com base na combinação destas quatro características, Fisher desenvolveu um modelo discriminante linear para distinguir as espécies umas das outras.

Este trabalho foi desenvolvido um programa, que através de um algoritmo estruturado têm o objetivo de identificar novas flores .O sistema classificará as novas flores, para isso foi utilizado um Iris Dataset que é um banco de dados com as informações sobre 3 tipos de flores.

-Setosa

-Virginica

-Setosa

Para isso foi utilizado arquivos como o Iris Dataset que foi dividido em duas partes. Sendo que uma seria usado como base de treinamento e uma base de teste. O programa irá receber os dois arquivos e salvar em duas listas para que possa ser utilizado durante o cálculo.

Para fazer o cálculo da distância foi utilizado a fórmula Euclidiana para o cálculo. Onde você terá dois loops sendo que o primeiro será da base de teste e o segundo loop interno será o de treinamento. Dentro do loop interno será feito o cálculo, onde será comparado cada posição da lista de testes com cada posição da lista de treinamento. No final esse resultado será salvo na tabela de treinamento.

Antes de interar o loop da lista de teste, ele irá ordenar a lista e irá classificar conforme o k digitado no inicio da execução conforme o resultado de K será possível classificar a classe do objeto testado.