Implementação do k-Nearest Neighbor (k-NN) para o problema de reconhecimento de dígitos (0-9)

Rafael RampimSoratto[1]

[1] Universidade Tecnológica Federal do Paraná - UTFPR// (Campus Campo Mourão) - Brasil - Campo Mourão - PR

sorattol@alunos.utfpr.edu.br

Abstract.

Resumo.

1 Requisitos do trabalho

1. Seu algoritmo deve avaliar o desempenho para diferentes valores de k 1,3,5,7,9,11,13,15,17,19 ;
2. Gerar a matriz de confusão ;
3. Usar a distância Euclidiana e Manhattan ;
4. Normalizar os dados com Min-Max e Z-score ;
5. Separar o conjunto de treinamento (aleatoriamente) em 25%, 50% e 100% dos dados de treinamento.
6. Avaliar qual o impacto de usar mais e menos instâncias no conjunto de treinamento.

2 Introdução ao k-NN

O algoritmo k-NN(k-Nearest Neighbor ou k Vizinhos mais próximos) trata-se de um algoritmo de classificação de dados clássico e muito simples. Ele assume que todas as instâncias correspondem a pontos em um espaço n-dimensional. de Aprendizagem Supervisionada, onde se encontra a 'boa resposta' durante o treinamento.

As vantagens de se utilizar o algoritmo k-NN é que trata-se de uma técnica simples e de fácil implementação, que em alguns casos apresenta ótimos resultados. Pode ser aplicada a problemas complexos, como: Análise de Crédito, Diagnósticos Médicos, Detecção de Fraudes, entre outros.

A desvantagens são: tempo; e ruíıdos nos dados ou características irrelevantes podem "enganar" o algoritmo.

Na aprendizagem supervisionada:

- E possível ajustar os pesos em função das respostas corretas;
- O desafio é capacitar o sistema a atuar de acordo com o padrão observado nos exemplos de entrada e saída ;

3 Funcionamento do k-NN

Protocolo para funcionamento do algoritmo.

3.1 Entradas do algoritmo

1. Um elemento x no qual deseja-se classificar;
2. Um conjunto para treinamento ;
3. Uma métrica para calcular a distância entre x e as demais amostras;
4. Definir um valor para k, ou seja, quantos vizinhos iremos considerar (1,3,5,7,9,11,13,15,17,19).

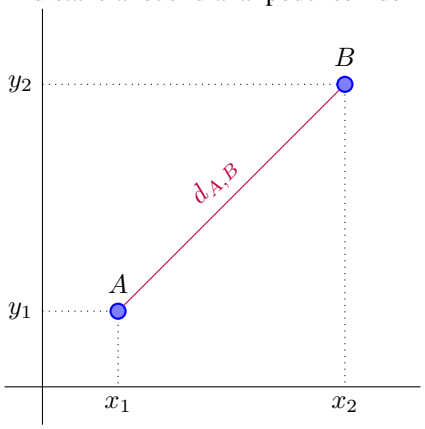
3.2 Funcionamento do algoritmo

1. Inicialmente, calcula-se a distância entre o exemplo desconhecido x e todos os exemplos do conjunto de treinamento ;
2. Identifica-se os k vizinhos mais próximos;
3. A classificação é feita associando o exemplo desconhecido x à classe que for mais frequente, entre os k exemplos mais próximos de x;

   Utiliza o voto majoritário para definir a classe mais frequente.

3.3 Distância euclidiana

A distância euclidiana pode ser definida pelo gráfico:



$$d\left(A,B\right)=\sqrt{\sum_{i=1}^{n}\left(A_i-B_i\right)^2}$$

3.4 Distancia de manhattan

The manhattan distance between two points is defined as:

$$d(A,B)\equiv|A_x-B_x|+|A_y-B_y|$$

4 Normalizacao dos dados

Os termos padronizar e normalizar são usados indistintamente no pré-processamento de dados, embora nas estatísticas, o último termo também tem outras conotações.

A normalização dos dados tenta proporcionar a todos os atributos um peso igual. Normalização é particularmente útil para algoritmos de classificação envolvendo redes neurais ou medições de distância, como classificação de vizinho mais próximo (k-NN) e "clustering".

Para métodos baseados em distância, a normalização ajuda a prevenir atributos com intervalos inicialmente grandes de superação de atributos com

intervalos inicialmente menores (por exemplo, atributos binários). Também é útil quando não é fornecido conhecimento dos dados.

Existem diversos métodos de normalização, neste trabalho serão utilizados os métodos Min-Max e Z-score. Para isto, utilizaremos um vetor com n elementos de $V_1 ... V_n$.

4.1 Normalização Min-Max

Normalização que executa uma transformação linear nos dados originais. Cada elemento do vetor é normalizado utilizando o valor máximo e mínimo do vetor. De acordo com a fórmula para definir cada elemento de um vetor A normalizado dentro de um intervalo $[0.0, 1.0]$:

$$v_i = \frac{v_i - min_a}{max_a - min_a} * 1$$

4.2 Normalização z-score

Considera a média e o desvio padrão durante a normalização de acordo com a formula.

$$v_i = \frac{v_i - \overline{A}}{\sigma_A}$$

sendo $\overline{A}$ a média e σ_A o desvio padrão.

Bibliographic references must be unambiguous and uniform. They must be numbered in order of appearance, e.g. [1], [2]. Self-citations can be anonymized using the model [3] or use constructions as: “previous work by Author et al.” instead of: “our previous work”.

References

[1] Donald E. Knuth. *The TEX Book*. Addison-Wesley, 15th edition, 1989.

[2] A. Smith and B. Jones. On the complexity of computing. In A. B. Smith-Jones, editor, *Advances in Computer Science*, pages 555–566. Publishing Press, 1999.

[3] Leslie Lamport. *LaTeX User's Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.