UFES

CEUNES

K-Nearest Neighboor

Elyabe Alves, Luis H. G. Valim

Universidade Federal do Espírito Santo
CComp - TEBD120181

26 de Junho de 2018

K Vizinhos mais próximos

"Reflitão":

"Diga-me quem são seus amigos e eu lhe direi quem você é!"[2]

Definição:

KNN é um algoritmo simples que classifica novos casos baseando-se na similaridade entre casos já armazenados. [4]

Funcionamento Básico

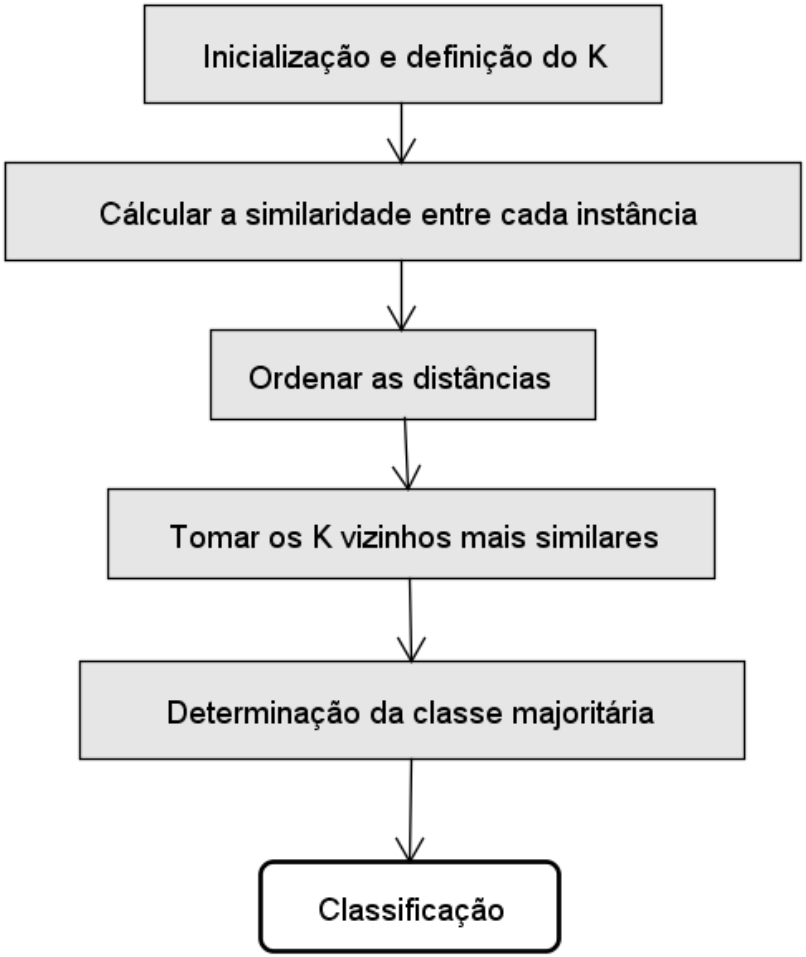

Figura: Ideia básica do KNN

Características relevantes

- **Simplicidade**: Fácil de entender e programar;
- **Lazzy**: Baseado na memória, sem treinamento ou modelo explícito;
- Opção de rejeição **explícita**;
- Fácil **pré-processamento** no tratamento de valores ausentes;
- Taxa de erro de classificação **assintótica**. [3]

Características relevantes

- Afetado pela estrutura local
  - **Sensível** a ruídos e características irrelevantes [3];
  - Computacionalmente **caro**;
  - Requisição de **memória**;
  - Classes mais **frequentes**, normalmente, predominam o resultado;
- O vizinho mais próximo pode não estar tão próximo;
- Se dim ≥ 4 perde-se o **sentido geométrico**.

Vamos brincar, lekes?

Simulador online

https://lettier.com/projects/knearestneighbors/

Enade 2012: Um caso de aplicação

Exame Nacional de Desempenho de Estudantes [1]

- Avaliação obrigatória;
- Métrica de rendimento:
  - Conteúdos programáticos;
  - Competências e habilidades necessárias para aprofundamento na formação;
  - Nível de atualização
- Indicadores de qualidade de educação superior:
  - Conceito Enade;
  - Conceito Preliminar de Curso; (CPC); e
  - Índice Geral de Cursos Avaliados da Instituição (IGC);

Configurações gerais

- $n > 180$ atributos:
  - Informações de cidadania;
  - Situação socieconômica;
  - Gabarito, respostas e notas;

Dados selecionados

Atributos selecionados	
Categoria	**Atributo(s)**
Inst. Ensino	cd_catad cd_orgac
Curso	co_regiao_curso co_uf_curso
Inscrito	tp_sexo sg_uf ano_fim_2g ano_in_gra tp_semest in_matut in_vesper in_noturno

Tabela: Seleção dos atributos.

Dados selecionados - Continuação

Atributos selecionados	
Categoria	**Atributo(s)**
Inscrito	in_grad
	tp_def_fis
	tp_def_vis
	tp_def_aud
Desempenho	nt_obj_fg
	nt_fg_d1
	nt_fg_d2
	nt_dis_fg
	nt_fg
	nt_obj_ce
	nt_ce_d1

Tabela: Seleção dos atributos. pt. 2

Dados selecionados - Continuação

Atributos selecionados	
Categoria	**Atributo(s)**
Desempenho	nt_ce_d2
	nt_ce_d3
	nt_dis_ce
	nt_ce
	nt_ger

Tabela: Seleção dos atributos. pt. 3

CLASSE REGIÃO (co_regiao_curso)

Pré-processamento

1. Transformação:
   1. **tp_sexo**
      $f : Caracter \mapsto i \in [0, 1]$

$$i = \begin{cases} 0, & \text{se tp\_sexo} = \text{'F'} \\ 1, & \text{se tp\_sexo} = \text{'M'} \end{cases}$$

   2. **sg_uf**
      $\{"AC", "AL", "AM", \ldots, "SP", "TO"\} \mapsto [0, 1, \ldots 25, 26]$

Normalização

$$x_s = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Determinação de parâmetros

Escolha do k

1. pequeno $\Longrightarrow$ maior variância (menos estabilidade); 5
2. grande $\Longrightarrow$ maior desvio (menos preciso);

E agora? Como escolher?

R: (Resposta preferia da prof. Maria)

Outras formas: Métodos adaptativos, heurísticas e validação cruzada

Determinação de parâmetros

Escolha da métrica de similaridade:

- Distância euclidiana

$$d(R_0, R_1) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \iff d^2(R_0, R_1) = \sum_{i=1}^{m}(x_i - y_i)^2$$

A Implementação

Linguagens

1. Matrix Laboratory (MATLAB);
    1. Abstração;
    2. Proprietário;
    3. "Levemente pesado";
2. Python;
    - Bibliotecas *pandas* e *numpy*;

Referências Bibliográficas

[1] INEP. Instituto nacional de estudos e pesquisas educacionais anísio teixeira.

[2] D. H. Izabela Moise, Evangelos Pournaras. K-nearest neighbour classifier. Technical report, ETHzürich.

[3] L. Kozma. k nearest neighbors algorithm. Technical report, Helsink University of Technology, 2008.

[4] S. Sayad. K nearest neighbors. Technical report, University of Toronto, 2010.