Implementação e Uso do Algoritmo *K-Means*

Prof. Ryan R. de Azevedo & Alunos de IA

Bacharelado em Ciência da Computação – Universidade Federal do Agreste de Pernambuco (UFAPE) – Garanhuns – PE– Brasil

{ryan.azevedo}@ufape.edu.br

1. Introdução e Conceitos Fundamentais

A atividade realizada consiste em implementar o algoritmo *K-means* e verificar os possíveis resultados da sua aplicação nos agrupamentos de instâncias em um determinado problema.

Problemas:

ATIVIDADE PARTE 1

O **primeiro** → classificar atletas de acordo com duas características dadas, peso e altura, em três modalidades possíveis (basquete, vôlei ou futebol). Para esse problema construímos nossa própria base de dados (*dataset)* com características reais de atletas profissionais (BaseRS).

1) Implementar em Python, observar os resultados e preparar uma apresentação ppt com os resultados obtidos e como chegaram nos resultados.

2) Implementar em Orange, observar os resultados e preparar uma apresentação ppt com os resultados obtidos e com o modelo visual utilizado para chegar nos resultados.

3) O que você faria para melhorar os resultados do agrupamento usando o dataset BaseRS? Faça! Melhore a base e apresente os novos resultados no mesmo relatorio ppt.

ATIVIDADE PARTE 2

O **segundo** → O segundo problema, um clássico da comunidade de estatística, Inteligência Artificial e Aprendizado de Máquina. classificar os tipos de Flores Íris. Nesse experimento usaremos um *dataset* (Iris) pronto e disponível para uso e mantido pela Universidade da California[1].

4) Implementar em Python, observar os resultados e preparar uma apresentação ppt com os resultados obtidos e como chegaram nos resultados.

5) Implementar em Orange, observar os resultados e preparar uma apresentação ppt com os resultados obtidos e com o modelo visual utilizado para chegar nos resultados.

[1] https://uci.edu/

2. Material e Métodos

Apresentamos na Tabela 1 as configurações que devem ser utilizadas no exercício (experimentos).

Experimento	***Tecnologia Utilizada na Configuração dos Ambientes Computacionais***
***Setup* 1**	*Python + Scitik-Learn + Seaborn*
***Setup* 2**	*Software Orange*

Tabela 1 – Configuração do ambiente para os experimentos.

O aluno, se desejar, pode usar outras bibliotecas *Python* para executar seus experimentos, mas deve usar as indicadas na Tabela 1 também. As bases não estão tratadas, como temos um algoritmo não supervisionado, a coluna ***target*** deve ser excluída durante o aprendizado do algoritmo.

2.1 Bases de Dados

Utilizamos dois *datasets* (**ds1** e **ds2**). O **ds1** possui 30 instâncias com características de atletas profissionais. As características usadas são: peso e altura. Os esportes praticados pelos indivíduos são: B - Basquete, V - vôlei e F - futebol. Na Figura 1 apresentamos parte (12 instâncias) do *dataset* criado.

Altura	Peso	target
1.9	100	B
1.9	80	V
1.75	67	F
1.60	70	F
1.71	60	F
2.1	109	B
2.0	85	V
1.88	80	V
2.15	115	B
1.95	113	V
1.8	73	F
1.82	76	F

Figura 1 – Parte do Dataset 1 (ds1) - Caracteristicas dos Atletas.

Criamos o **ds1** propositalmente com duas características (Altura e Peso) para que fossem expostas, nos resultados, algumas características negativas do *k-means*. Ressaltamos que não fizemos o mesmo para o experimento ***Setup* 2** utilizando o **ds2**. O **ds2** possui 150 instâncias com quatro características a respeito das flores Íris. Na Figura 2 apresentamos parte (12 instâncias) do **ds2**.

	1	2	3	4	5
1	sepal length(cm)	sepal width(cm)	petal length(cm)	petal width(cm)	target
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3.0	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5.0	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5.0	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa

Figura 2 – Parte do Dataset 2 (ds2) - Caracteristicas dos Atletas.

3. Fundamentos Teóricos

Os fundamentos teóricos necessários para um entendimento específico, coadunando com o tema da atividade realizada, são apresentados nesta seção de forma suscinta. Embora resumida, apresentamos, com referências importantes para um aprofundamento dos leitores.

Os conceitos básicos são:

- **Aprendizado de Máquina**: "*A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência*" [Mitchell, T. M., 1997]. Para isso usamos o *Scitik-Learn*.
- **Aprendizado de Máquina Não Supervisionado**: "*Aprendizado automático sem a presença, durante o treinamento, de um professor (exemplo de saída)*" [Géron, A., 2019]. Usamos o K-means, descrito a seguir.
- ***K-Means***: é um método de clusterização baseado em particionamento que utiliza um critério de realocação interativa. O objetivo deste tipo de agrupamento é encontrar os k pontos do espaço de busca que minimizem uma função erro ou função objetivo, a qual determina a similaridade global entre os objetos dos clusters. O *k-means* corresponde a um dos algoritmos de clusterização mais utilizados devido a sua simplicidade de implementação, rápida convergência para uma configuração estável, eficiência e sucesso empírico [Borges, F. A. S.,2017].

Referências

BORGES, F. A. S. **Método Híbrido Baseado no Algoritmo k-means e Regras de Decisão para Localização das Fontes de Variações de Tensões de Curta Duração no Contexto de Smart Grid**. Tese (Doutorado) - Programa de Pós-Graduação em Engenharia Elétrica e Área de Concentração em Sistemas Dinâmicos - Escola de Engenharia de São Carlos da Universidade de São Paulo, 2017.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts**, **Tools, and Techniques to Build Intelligent Systems**. (2nd ed.), 2019.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. In: In 5-th Berkeley Symposium on Mathematical Statistics and Probability. [S.l.: s.n.], 1967.

MITCHELL, T. M. Machine Learning. McGraw-Hill, New York, 1ª ed, 1997.