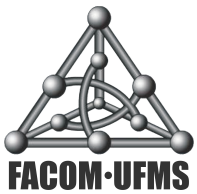

Inteligência Artificial - T03

Prof. Bruno M. Nogueira
Faculdade de Computação - UFMS

Trabalho III - Agrupamento de Dados

Neste trabalho, o objetivo é a implementação do algoritmo K-Means e sua aplicação em datasets bidimensionais. A implementação deve ser feita na linguagem Python e não deve-se utilizar implementações do K-Means disponíveis em bibliotecas. Nas seções a seguir, detalhes do que é esperado neste trabalho são apresentados.

1 Algoritmo K-Means

O algoritmo de agrupamento de dados *K-Means* [MacQueen, 1967] agrupa um conjunto de instâncias em k partições, sendo k um número pré-estabelecido. O arranjo dos elementos é feito de maneira que um elemento pertença a um *cluster* a cujo centro o elemento é mais próximo. O Algoritmo 1 mostra o procedimento do *K-Means*.

Input: $X = \{x_1, x_2, ..., x_n\}$: conjunto de n dados de entrada; k: número de clusters a ser encontrado
Output: Conjunto K de k partições disjuntas de elementos
Inicialize os centroides μ_j, $j = 1, ..., k$ com um elemento aleatoriamente escolhido;
while *Atingir critério de convergência* **do**
  for *Todo elemento* $x_i \in X$ **do**
    Associe x_i ao cluster h_j, tal que $h_j = argmin_{h_j} \left\|x - \mu_{h_j}\right\|^2$;
  end
  for *Todo cluster* $h_j \in K$ **do**
    Calcule o novo centroide do cluster h_j, tal que $\mu_j = \frac{1}{|h_j|} \sum_{x \in h_j}$;
  end
end

Algorithm 1: Algoritmo *K-Means*

Dessa maneira, o algoritmo *K-Means* consegue encontrar k partições disjuntas, buscando sempre minimizar a variância *intra-cluster* e maximizar a variância *inter-cluster*. Como critérios de convergência usuais, pode-se citar o número de interações que o algoritmo executa e o número de realocações de *clusters*.

2 Conjuntos de dados

Para este trabalho, 9 datasets serão utilizados. Todos estes datasets são bi-dimensionais (isto é, têm 2 atributos), com o número de clusters variando de 2 a 10 (o primeiro dataset tem 2 clusters, o

segundo tem 3 clusters, etc.). Todos estes datasets apresentam uma partição de referência (grupo de cada um dos pontos), sendo perfeitamente balanceados (30 exemplos por grupo).

Cada um dos datasets está disposto em um arquivo no formato CSV ("*comma separated values*"). Em cada linha há um exemplo (instância) da base, no formato:

Valor_Atributo_1,Valor_Atributo_2,Partição_de_Referência

Lembrando que, para fins de agrupamento e cálculo das distâncias, deve-se desprezar o valor da partição de referência, devendo-se usar somente os atributos 1 e 2.

3 Procedimento experimental

Dado que a inicialização do algoritmo *K-Means* não é exata (as sementes para a inicialização dos centroides são geradas aleatoriamente), o algoritmo deve ser aplicado 10 vezes sobre cada base de dados, a fim de obter uma média de desempenho com diferentes configurações de inicialização. Em todas as execuções, o número de grupos formados deve ser igual ao número de partições de referência ($k = 2$ para o primeiro dataset, $k = 3$ para o segundo dataset e assim por diante). Para os algoritmos, deve ser utilizado como critério de parada um limite de 50 iterações e como função de distância a função euclideana.

Em cada uma das execuções do algoritmo, deve-se armazenar o valor de Soma do Erro Quadrático (SSE) obtido, de acordo com a fórmula a seguir:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x) \tag{1}$$

considerando x um ponto pertencente ao cluster C_i e m_i o centróide do cluster C_i. Ao final, deve-se escolher o melhor agrupamento como aquele que apresenta o menor valor de Soma do Erro Quadrático.

4 Entrega

Deverão ser entregues um relatório e todos os scripts construídos para a experimentação. Neste relatório, devem ser apresentados:

- Uma introdução, apresentando o problema e o objetivo do trabalho;
- Uma breve descrição do algoritmo de agrupamento utilizado, com referências;
- Uma explicação do procedimento experimental, incluindo uma explicação acerca da Soma de Erro Quadrático (SSE);
- Descrição dos datasets utilizados, contendo: nome do dataset, número de exemplos e número de grupos presentes;
- Resultados obtidos;
- Discussão dos resultados (não somente transcrição!);
- Considerações finais.

Os resultados obtidos deverão ser apresentados por meio de tabelas e gráficos. Nas tabelas, deverão ser mostrados, para todos os datasets, o valor da média e desvio padrão de valor de SSE obtido para aquele dataset. Nos gráficos, deve-se plotar os resultados obtidos, mostrando os

pontos e os grupos aos quais foram atribuídos. Usem símbolos e cores diferentes para cada um dos grupos obtidos.

O relatório deve ser formatado em formato de artigo, com no máximo 15 páginas no modelo disponibilizado no Moodle. Outros modelos podem ser utilizados, a critério do grupo, desde que o professor seja consultado.

O prazo para entrega do trabalho será as **23:55** do dia **03/07**. Todos os scripts com os códigos das soluções devem ser devidamente comentados e identificados com os nomes de todos os integrantes do grupo. Uma única entrega por grupo deve ser feita, via Moodle (EAD). Entregas fora do prazo, ou feitas por outros meios, serão desconsideradas. Não será tolerado plágio de quaisquer fontes, mesmo que parcial. Quando detectado plágio, o trabalho terá nota zero.

Referências

[MacQueen, 1967] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.