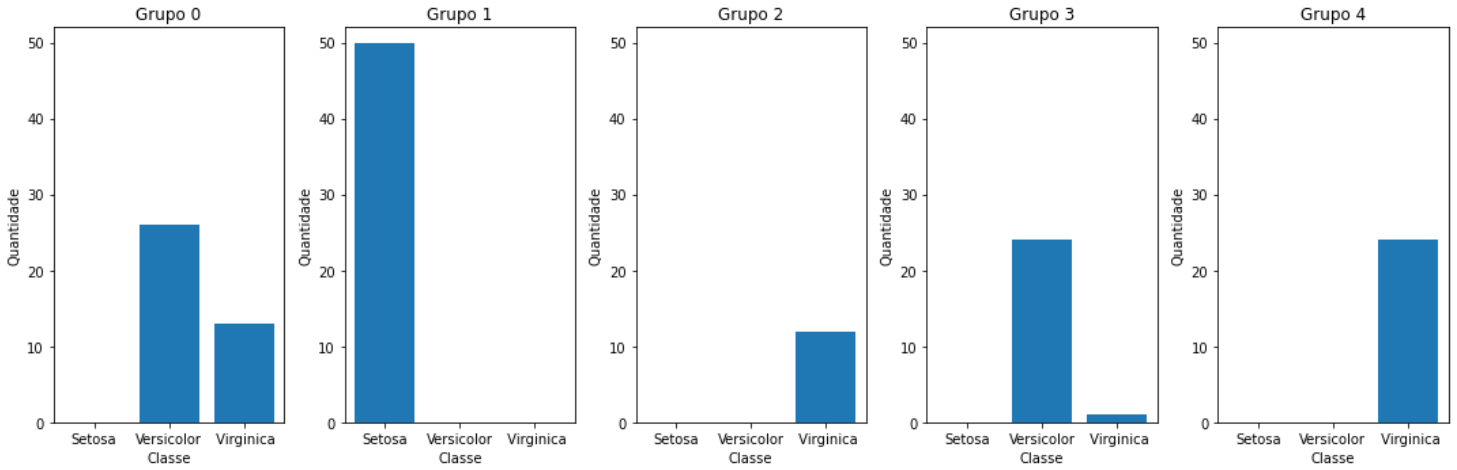
Reconhecimento de Padrões: Atividade 07 − Agrupamento

Carlos Emmanuel Pereira Alves
Curso de Bacharelado em Ciência da Computação
Universidade Federal do Agreste de Pernambuco (UFAPE)
Garanhuns, Brasil
carlos.emmanuel.236@gmail.com

1) Utilizado utilizando o algoritmo de agrupamento k-médias com todos os 150 exemplo da base Iris, https://archive.ics.uci.edu/ml/datasets/Iris:

a) Para o valor de k = 5, calcule um gráfico de barras para cada grupo (cinco gráficos). Cada gráfico deverá ter três barras: quantidade de elementos de cada classe (setosa, versicolor e virginica). Isto é, cada um dos cinco gráficos de ter: a quantidade de elementos da classe setosa no grupo, a quantidade de elementos da classe versicolor no grupo, a quantidade de elementos da classe virginica no grupo.



b) Mostre a média e o desvio padrão das distâncias de cada exemplo ao centróide mais próximo durante 1, 2, 3, . . . , 10 iterações. Monte uma tabela com as duas colunas (média e desvio padrão) e 10 linhas (número de iterações).

Iteração	Média	Desvio padrão
1	0.5098	0.2246
2	0.5105	0.2229
3	0.5105	0.2229
4	0.5105	0.2229
5	0.5105	0.2229
6	0.5105	0.2229
7	0.5105	0.2229
8	0.5105	0.2229
9	0.5105	0.2229
10	0.5105	0.2229

2) Utilizado a base Iris, https://archive.ics.uci.edu/ml/datasets/Iris, em um experimento do tipo Holdout 50/50 Estratificado realize seleção de protótipos da seguinte forma, para k = 9:

a) Execute o k-medóides no apenas no conjunto de treino.

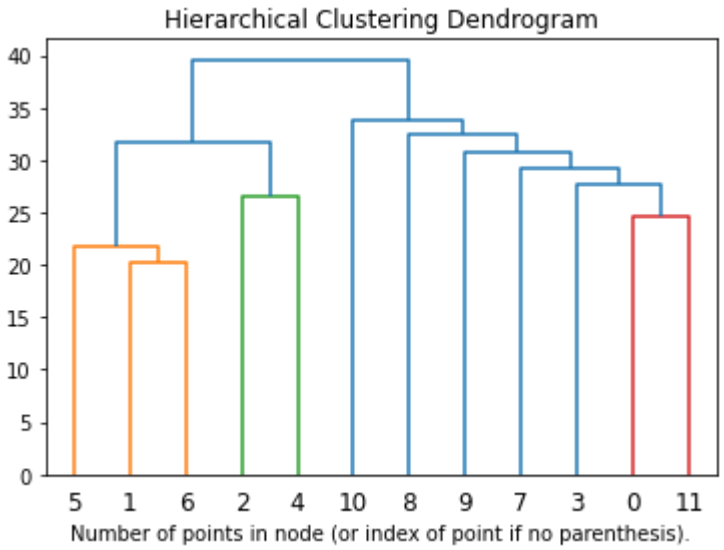
```
[7 8 2 4 4 3 0 4 4 4 1 1 2 0 1 5 8 4 4 4 4 4 0 2 4 6 8 4 4 4 3 8 8 1 3 8 7
 4 8 8 7 4 8 4 8 3 4 7 4 1 1 5 4 8 3 0 4 4 0 8 5 4 8 6 6 4 4 7 5 1 1 8 8 8
 1]
```

b) Remova do conjunto de treino todos os elementos que não são centroide de um grupo, isto é, mantém apenas os centróides no conjunto de treino. O conjunto de treino terá apenas k exemplos. A redução do conjunto de treino é chamada seleção de protótipos. O conjunto de teste permanecer inalterado, isto é, os 50% dos dados original (75 exemplos). Utilizando o conjunto de treino reduzido (composto apenas pelos centróides), calcule a taxa de acerto POR CLASSE para o conjunto de teste utilizando o classificador 1-NN com distância Euclidiana.

```
    setosa       1.00
versicolor       0.81
 virginica       1.00
```

3) O arquivo maisAssistidos.csv contém a avaliação dos usuários com notas de 1 a 5 da base Movie Lens 100k https://grouplens.org/datasets/movielens/ para os 12 filmes mais avaliados na base. São 943 usuários e cada filme foi avaliado por mais de 400 destes. Como cada atributo de um filme é a nota de um dos usuários existem vários valores de atributos omissos.

a) Realize agrupamento hierárquico aglomerativo.
b) Gere o dendrograma do agrupamento.



c) Pesquise qual o gênero de cada filme (comédia, policial, ficção científica etc.). Analise se o faz sentido, em relação ao gênero, cada vez que dois grupos são unidos no dendrograma.
Utilizando classificações de gênero, segundo o IMDB, eles ficaram assim:
0- Toy Story - Animação, Aventura, Comédia
1- Star Wars - Ação, Aventura, Fantasia
2- Fargo - Policial, Suspense
3- Independence Day: The ID4 Invasion - Documentário, Ficção Científica
4- The Godfather - Policial, Drama
5- Raiders of the Lost Ark - Ação, Aventura
6- Star Wars: Episode VI - Return of the Jedi - Ação, Aventura, Fantasia
7- Contact - Drama, Mistério, Ficção Científica
8- The English Patient - Drama, Romance, Guerra
9- Scream - Terror, Mistério
10- Liar Liar - Comédia, Fantasia
11- Air Force One - Ação, Drama, Suspense

No começo as uniões fazem até sentido com ele unindo os dois filmes de Star Wars e depois ligando a Raiders of the Lost Ark que tem o mesmo gênero deles. Depois ele liga Toy Story a Air Force One, que não faz sentido segundo o gênero. Após isso ele conecta os dois filmes de policial, Fargo e The Godfather, o que faz muito sentido. Mas depois disso as ligações começam a perder o sentido se formos olhar o gênero.

Ele conecta a ligação de Toy Story e Air Force One, ao documentário Independence Day: The ID4 Invasion, por exemplo, as outras conexões feitas depois perdem o sentido quando olhamos o gênero.

4) Utilizando uma base construída a partir de uma imagem digital de uma fotografia da natureza (escolha uma imagem, exemplo abaixo), cada pixel da imagem é um elemento do conjunto de dados.

a) Carregue cada pixel como um vetor de atributos. Cada pixel é representado pelo código RGB, ou seja, o primeiro atributo é o valor de intensidade R, o segundo é o valor de intensidade G e o último o valor de intensidade B.

b) Execute o algoritmo de agrupamento k-médias para k = 8, 64 e 512. Considere limitar o número máximo de interações se a convergência demorar muito. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

c) Ao final da convergência do k-médias arredonde os valor da posição de cada centroide para o inteiro mais próximo, exemplo, (15.3; 134.9; 9.4333) para (15; 135; 9). Reconstrua a imagem substituindo cada intensidade original de pixel pela intensidade do seu centroide, você obterá uma imagem na qual o número de cores distintas é k. Veja o exemplo das figuras abaixo.