Aula 13 – Exercício de Programação
Clustering

1) Qual tipo de problema você deve resolver?

Esse exercício de programação foi dado pelo professor Ryan Baker no MOOC Big Data in Education, disponibilizado pela plataforma http://edx.org, como **Assignment Week 7**.

Nesse exercício, ele disponibiliza uma base de dados simulada, que contém 6 *features* e 605 registros. O arquivo que contém a base de dados foi disponibilizado no e-aula e se chama "aula21.csv".

Nesse exercício, você deve usar a base de dados e algoritmo de *clustering* para responder às perguntas abaixo. Todas as explicações solicitadas devem ser providas como texto no notebook.

2) Usando o K-means

2.1) Realize a clusterização utilizando k-Means++ com K=2. Observe os centros dos clusters. Quais 2 atributos têm a maior diferença entre o cluster 0 e o cluster 1? Esses 2 atributos serão usados nas próximas questões.

2.2) É hora de plotar os dados. Defina o eixo X como a primeira resposta da questão 3.1 e defina o eixo Y como a segunda resposta da questão 3.1. Em seguida, defina a coluna de cores para refletir os números dos clusters (usar a variável colors a seguir que será útil para as próximas questões: `colors = ["red", "blue", "green", "purple", "orange", "yellow", "brown"])`. Existem sete grupos principais ("caroços") neste conjunto de dados. Quantos deles são vermelhos e quantos são azuis? Observe que essa é uma outra maneira da gente plotar os dados, selecionando apenas as duas features mais importantes.

2.3) Agora, execute o k-Means++ novamente com k=7. O k-Means encontrou os 7 "caroços" nos dados que você viu anteriormente? Comente sua resposta.

2.4) Plote o gráfico bidimensional para cada par possível de variável:

```
feature_pairs = [("a", "b"), ("a", "c"), ("a", "d"), ("a",
"e"), ("a", "f"),
                 ("b", "c"), ("b", "d"), ("b", "e"), ("b",
"f"),
                 ("c", "d"), ("c", "e"), ("c", "f"),
                 ("d", "e"), ("d", "f"),
                 ("e", "f")]
```

- Observe os gráficos gerados. Algum deles parece gerar clusters com algum significado?

2.5) Filtrar todas as variáveis exceto aquelas da questão 3.1, e reexecutar o k-Means++ usando apenas essas duas variáveis, com k=7. Todos os sete aglomerados de dados agora estão mais ou menos incorporados em sete clusters razoáveis? Explique sua resposta.

2.6) Execute o método de cotovelo para achar um ou mais valores de k. Existe um único valor de k que poderia ser considerado o cotovelo ou mais de um? Execute o K-means++ para diferentes valores de k que você acha que poderia ser o cotovelo.

3) Usando o DBSCAN

3.1) Execute o DBSCAN considerando todas as features. Plote o gráfico considerando apenas as features achadas na questão 3.1. Teste diferentes valores de eps e min_samples. Qual o melhor valor de eps e min_samples?

3.2) Execute o DBSCAN considerando apenas as features achadas na questão 3.1. Plote o gráfico considerando apenas as features achadas na questão 3.1. Teste diferentes valores de eps e min_samples. Qual o melhor valor de eps e min_samples?

3.3) Compare os dois gráficos. Você achou os mesmos clusters nos dois casos? O DBSCAN se mostrou melhor em dos casos? Qual? Explique.

3.4) Para esse exemplo, o DBSCAN se mostrou melhor, pior ou igual que o K-means++? Explique sua resposta.