K-MEANS

Roberta Pereira

APLICAÇÃO K-MEANS

O reconhecimento de padrões consiste em reconhecer similaridades entre as amostras do conjunto de dados em análise de forma a atribuir um rótulo a cada uma dessas, indicando a qual grupo ou classe ela pertence. Porém, nem sempre é possível saber previamente como os grupos estão separados ou em quantas classes o conjunto pode ser separado ou ainda qual o rótulo das amostras de treinamento. Para isso, existe o aprendizado não supervisionado que permite particionar inicialmente os dados.

Neste exercício foi abordado o algoritmo *K-means*, um algoritmo de partição, que permite agrupar os pontos do conjunto de dados de maneira que os pontos dentro de um mesmo cluster sejam o mais similar possível e pontos em clusters diferentes sejam o mais diferente possível. O parâmetro k é escolhido como a quantidade de grupos em que se quer separar os dados.

Para entender o comportamento do *K-means,* serão geradas quatro distribuições gaussianas concatenadas com diferentes desvios padrões e testadas com diferentes parâmetros k do algoritmo desenvolvido.

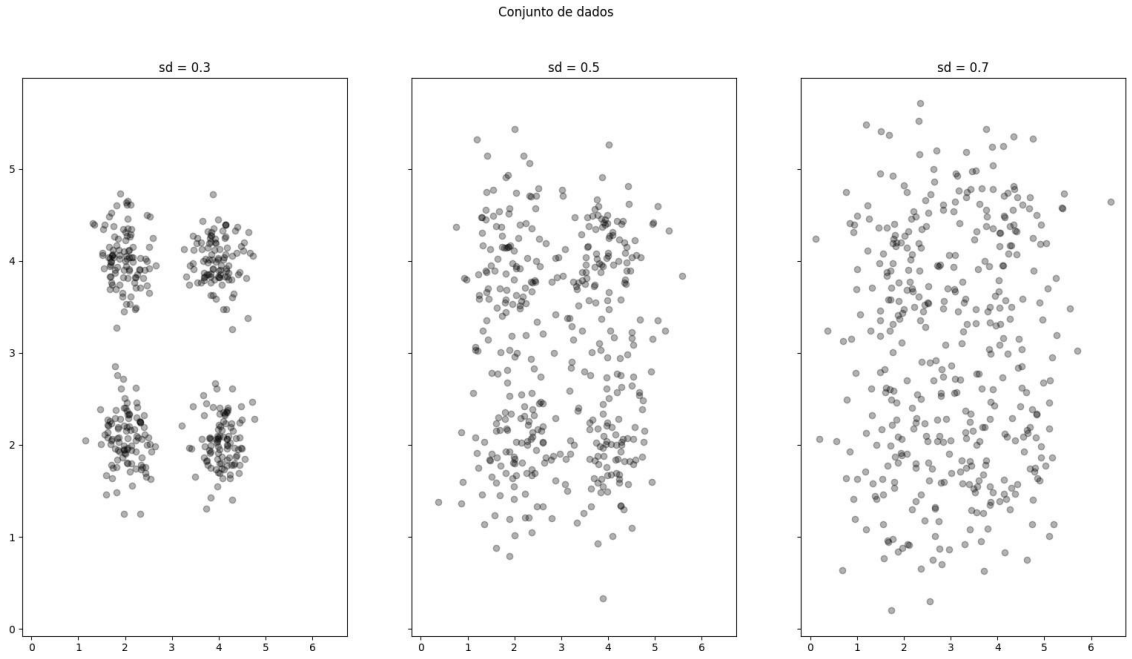
1. ALGORITMO *K-MÉDIAS*

O algoritmo desenvolvido recebe como parâmetros o valor k e os dados a serem analisados. Assim é escolhido aleatoriamente k pontos dentre os pontos do conjunto de dados para serem os centros dos k agrupamentos. Após esse procedimento, é calculada a distância de todos os demais pontos para todos os centros escolhidos e assim, foi atribuído para cada ponto o cluster mais próximo. Depois da atribuição dos pontos e com os agrupamentos completos, foi definido o novo centro de cada cluster como a média dos valores dos pontos pertencentes a ele. Foi repetido os passos anteriores, ou seja, a distância dos pontos aos centros foi novamente calculada e uma nova atribuição foi realizada. O loop termina quando não há mais diferenças entre os agrupamentos atuais e os definidos anteriormente.

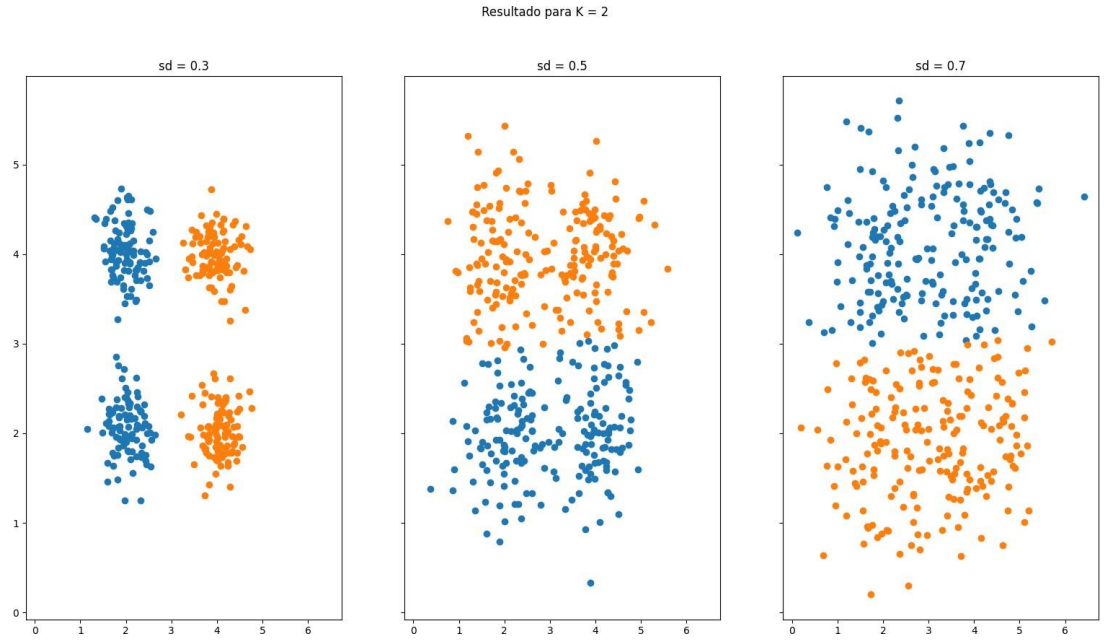
O algoritmo foi desenvolvido em Python e a distância utilizada foi a distância euclidiana adequada para o conjunto utilizado nesse exercício.

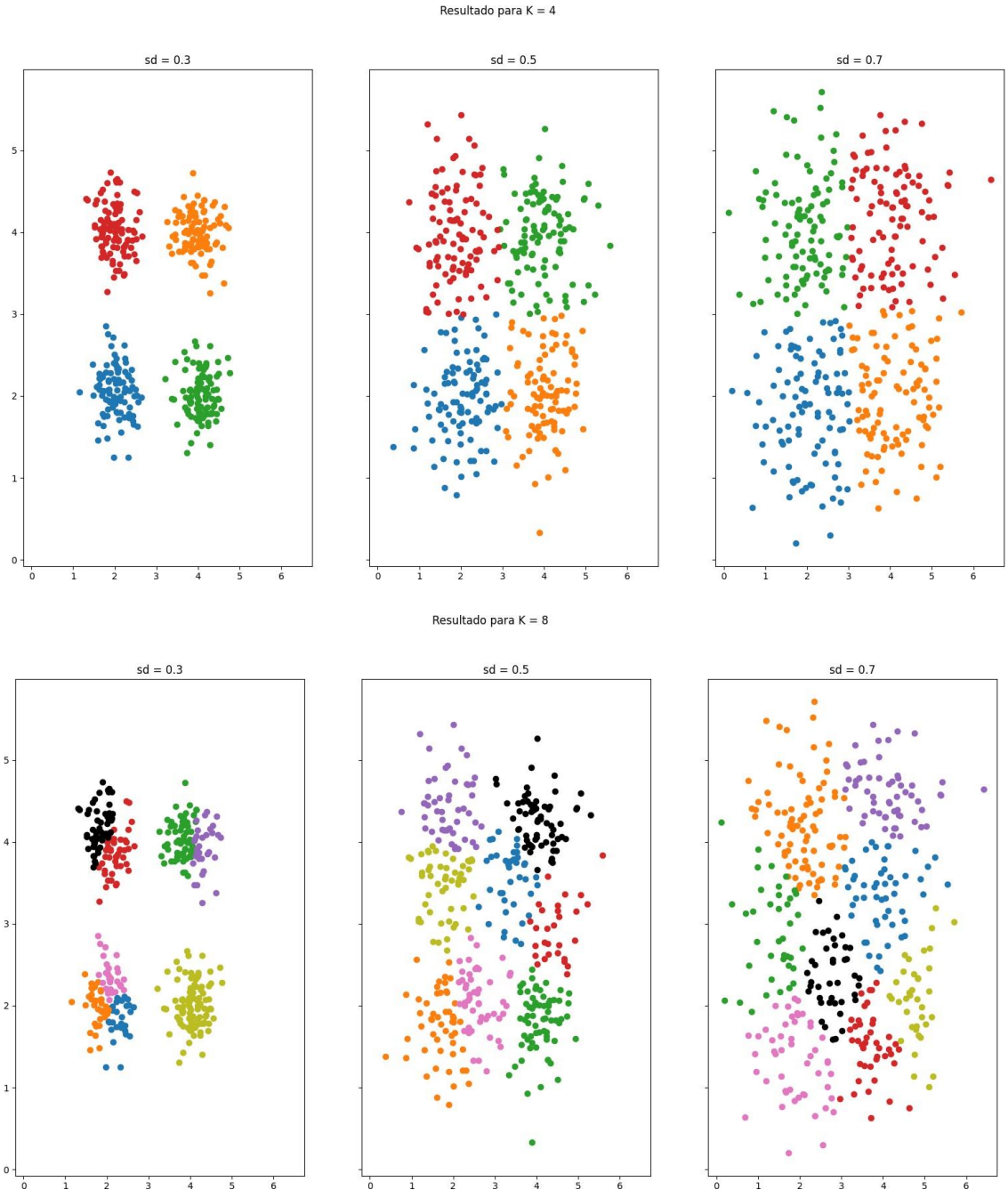
2. GAUSSIANAS BIDIMENSIONAIS

Para teste do algoritmo foram geradas quatro distribuições gaussianas bidimensionais com 100 pontos cada uma e que foram concatenadas para representar o conjunto de dados. O *K-means* foi testado para diferentes valores de desvio-padrão das distribuições gaussianas e diferentes valores de k.



3. RESULTADOS





4. DISCUSSÃO

Analisando os resultados de acordo com a dispersão distinta das amostras, percebe-se que para um desvio-padrão igual a 0,3, ou seja, com as amostras de cada gaussiana com mais concentradas sendo possível assim visualizar cada gaussiana, temos que o algoritmo foi capaz de atribuir os agrupamentos correspondentes a cada gaussiana para K até 4. Para K=8 é observado que o algoritmo atribuiu 3 agrupamentos para uma mesma gaussiana (conjunto laranja, azul e rosa), quando o esperado, considerando que todas as gaussianas têm a mesma

quantidade de pontos e desvio-padrão, é que cada gaussiana fosse separado em dois agrupamentos.

Para um desvio-padrão igual a 0,5, apresentando pontos mais dispersos no conjunto, o *K-means* conseguiu separar os agrupamentos mais adequadamente para todos os valores de K testados. Assim como para um desvio-padrão de 0,7.

Em uma análise geral aos resultados obtidos, é possível observar que o algoritmo conseguiu encontrar agrupamentos distintos para cada K escolhido, sendo que esses agrupamentos possuem melhor atribuição de acordo com as gaussianas para K até 4.

5. CONCLUSÃO

O *K-means* é um algoritmo de aprendizado não supervisionado capaz de particionar os dados, escolhendo-se pontos representando o centro de cada agrupamento e separar os dados de acordo com a distância ou similaridade entre os demais pontos e os centros dos clusters. Para aplicação do algoritmo foram gerados um conjunto de dados composto por quatro gaussianas bidimensionais variando a dispersão dos dados.

O algoritmo utilizado nesse exercício permite separar os dados na quantidade de agrupamentos determinado pelo parâmetro K. O *K-means* foi aplicado para separar o conjunto em 2, 4 e 8 agrupamentos considerando que as gaussianas tinham um desvio-padrão de 0,3, 0,5 e 0,7. Dessa maneira, foi plotado um gráfico para cada par do parâmetro K e desvio-padrão, demonstrando as partições realizadas pelo algoritmo.

Considerando que cada gaussiana fosse um conjunto de dados distinto com a mesma quantidade de pontos cada, temos que para um K maior do que 4 e menor desvio-padrão, o algoritmo, em geral, possui maior dificuldade para realizar os agrupamentos de acordo com os dados de cada gaussianas.

Os resultados encontrados são influenciados também pela métrica escolhida para o cálculo da distância entre os pontos e o centro dos clusters. Assim, se fosse escolhido uma métrica diferente da distância euclidiana, os resultados encontrados seriam diferentes e essa escolha depende do conjunto de dados utilizado.

Conclui-se que o algoritmo construído atendeu a expectativa ao conseguir separar o conjunto na quantidade de agrupamentos escolhidos, sendo que para um maior valor do parâmetro K foi mais difícil fazer com que os clusters coincidissem com as gaussianas, mas essa situação poderia ser diferente se a métrica escolhida para a distância fosse outra.