

INF 493/792
Tópicos especiais III
Introdução à mineração de dados

Aula IX: k-means

Avisos gerais

Dúvidas sobre a lista?

Não esqueçam dos horários agendados (projetos)

Objetivo de hoje

O problema de Agrupamento

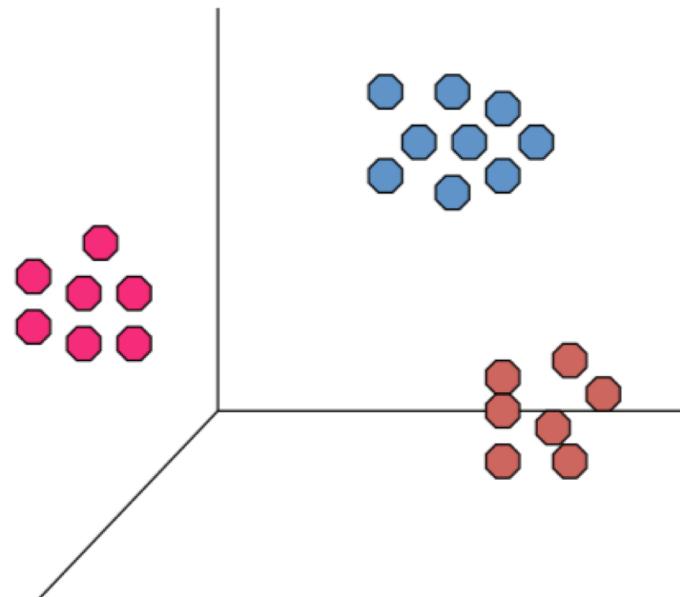
O problema k-means

Complexidade

Heurística

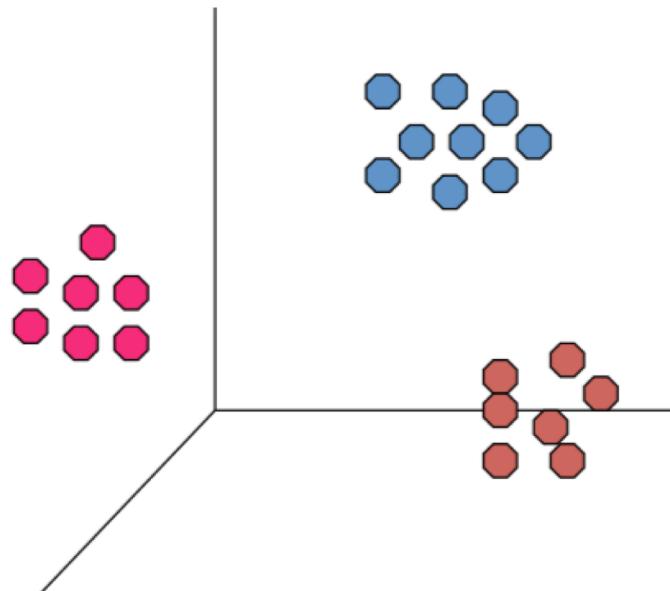
Como escolher o k?

O que é o problema de agrupamento?



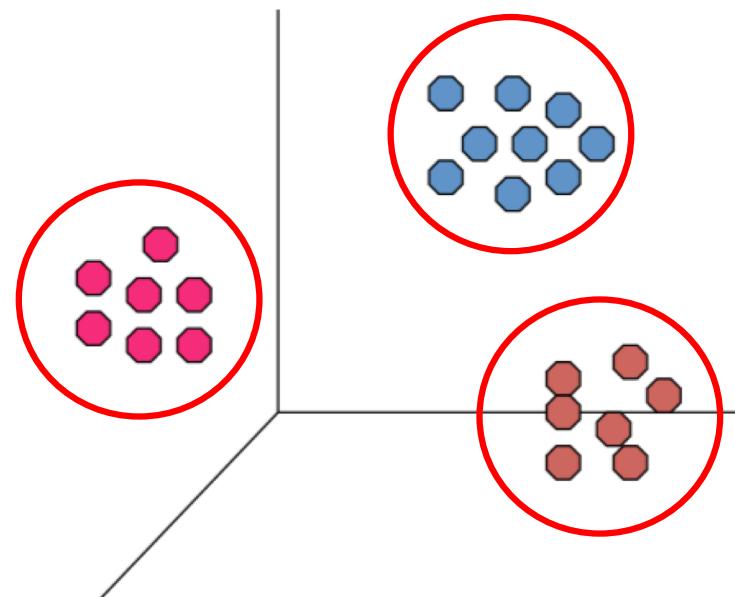
O que é o problema de agrupamento?

Encontrar grupos de objetos tais que objetos do **mesmo grupo** são **similares** (relacionados) e objetos de grupos **diferentes** são **não similares** (não relacionados)



O que é o problema de agrupamento?

Encontrar grupos de objetos tais que objetos do **mesmo grupo** são **similares** (relacionados) e objetos de grupos **diferentes** são **não similares** (não relacionados)



Por que fazemos agrupamento?

Nos permite ganhar informação sobre a distribuição dos dados

- Visualização pode revelar padrões importantes

- Ferramenta de pré-processamento para outros algoritmos

- Indexação ou compressão

Aplicações

Processamento de imagens

- Agrupar imagens com conteúdos similares

Web

- Encontrar grupos de usuários com comportamento similar
- Encontrar páginas com conteúdo similar

Bioinformática

- Encontrar proteínas similares (estrutura ou funcionalidade)

Muitas outras...

Questões básicas

O que similaridade significa?

O que é uma boa **partição**? I.e., como medir qualidade?

Como encontrar uma boa **partição** do conjunto de pontos?

Algoritmos de partição

Construir uma partição de **n** objetos em **k** grupos

Cada objeto pertence a **exatamente um** grupo

O número de grupos, **k**, é dado

O problema k-means

Dado um conjunto de pontos $X = \{x_1, \dots, x_n\}$, sendo cada x_i um ponto do espaço real d -dimensional, e um número inteiro k , o objetivo é encontrar pontos c_1, \dots, c_k que minimizem

$$\sum_{i=1}^n \min_{1 \leq j \leq k} \|x_i - c_j\|^2$$

Complexidade

É um problema NP-difícil

Para $d = 1$, o problema pode ser resolvido em tempo polinomial

Na prática, um algoritmo iterativo funciona bem

Propriedades do algoritmo

Encontra um ótimo local

Em geral, converge rápido

Implementado na *sklearn*

No entanto, cada execução pode ter uma resposta muito diferente das anteriores (**inicialização**)

Inicialização também tem impacto no tamanho e densidade dos grupos

Outliers podem ser um problema

Propriedades do algoritmo

Encontra um ótimo local

Em geral, converge rápido

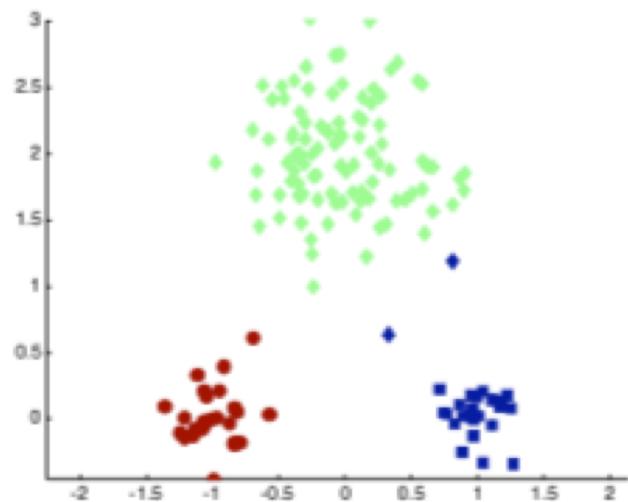
Implementado na *sklearn*

No entanto, cada execução pode ter uma resposta muito diferente das anteriores (**inicialização**)

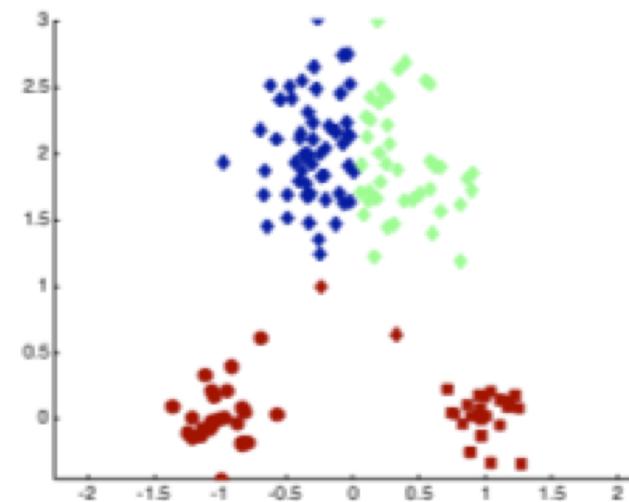
Inicialização também tem impacto no tamanho e densidade dos grupos (**k-means++**)

Outliers podem ser um problema (**k-means--**)

Devido a escolha dos pontos iniciais



Optimal Clustering

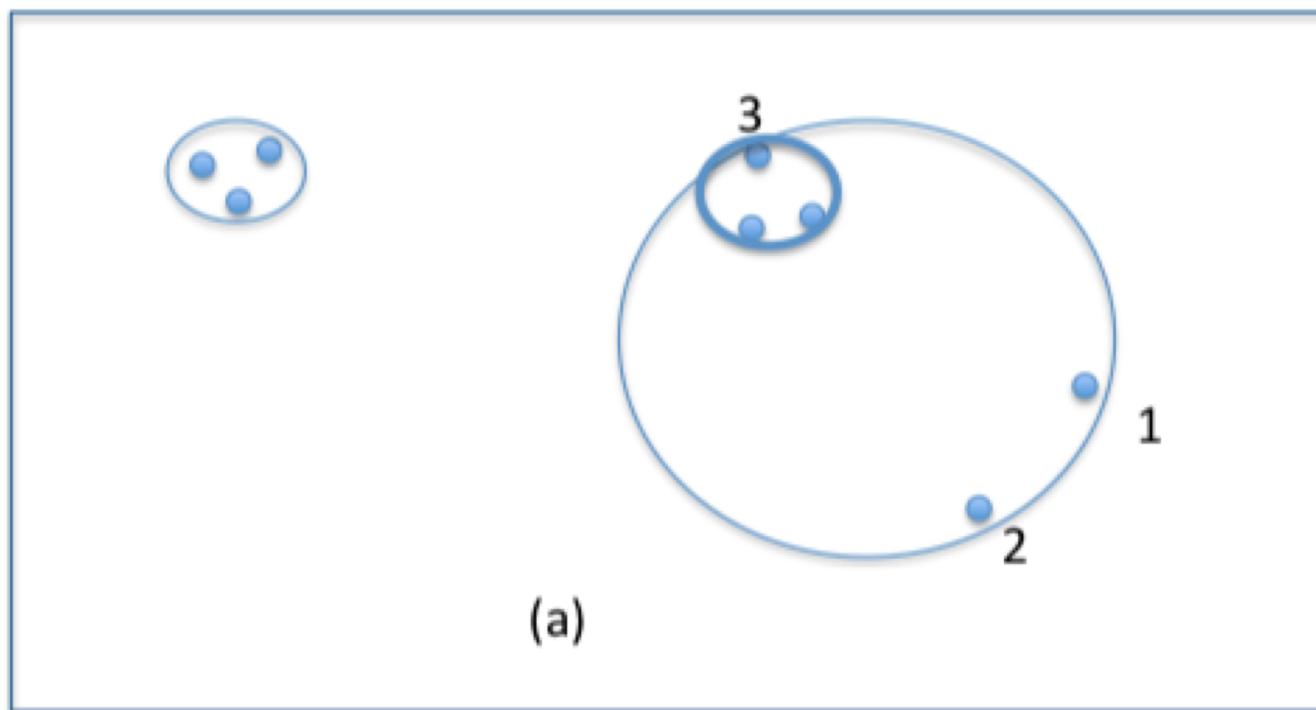


Sub-optimal Clustering

Sobre o k-means++

A ideia é escolher centroides iniciais que não estejam muito próximos dos outros

Sobre o k-means--



Como escolher k ?

O que acontece com o valor da função objetivo quando k varia de **1** para **n**?

Como escolher k?

O que acontece com o valor da função objetivo quando **k** varia de **1** para **n**?

Há várias heurísticas:

- Ponto em que a função objetivo tem mudança de comportamento

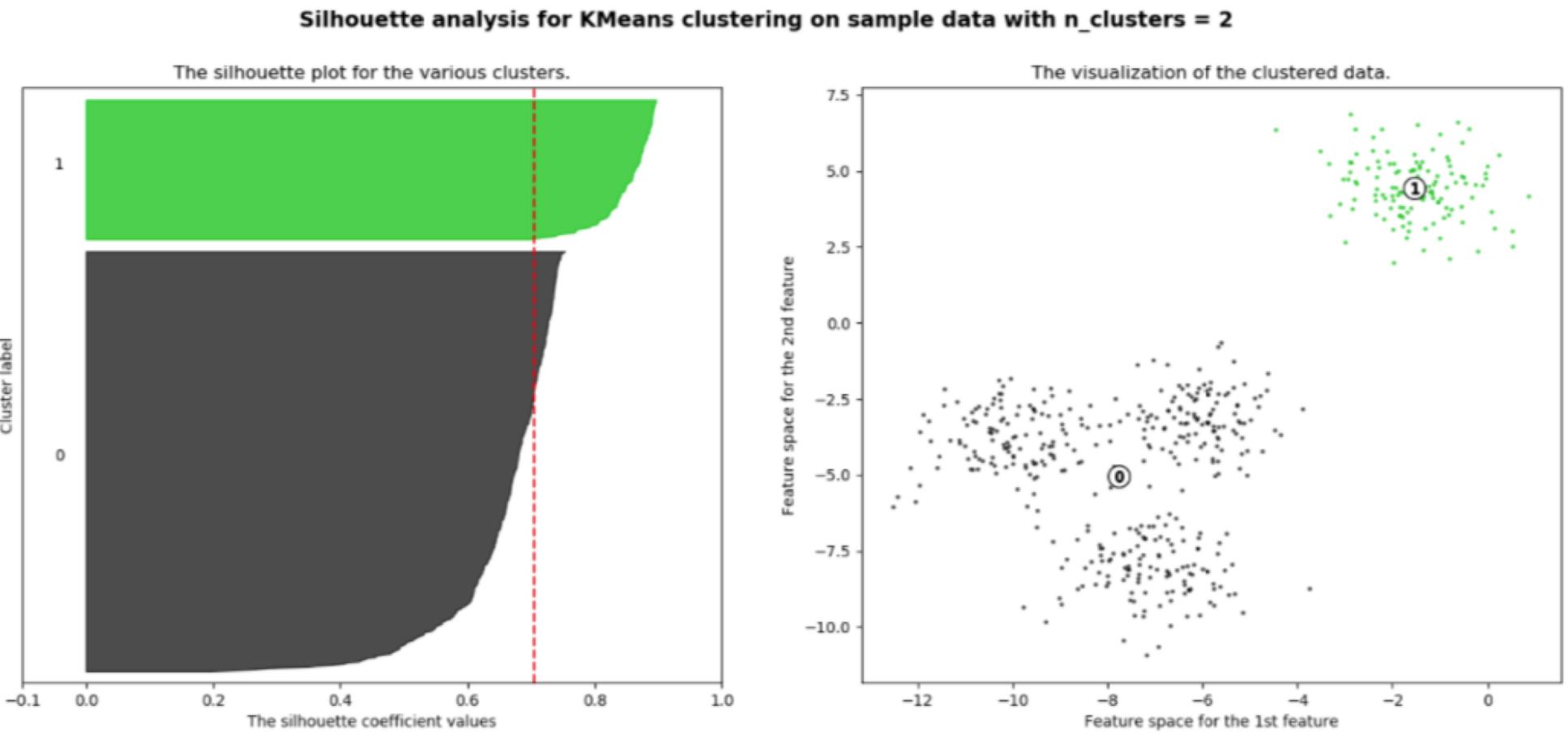
Como escolher k?

O que acontece com o valor da função objetivo quando **k** varia de **1** para **n**?

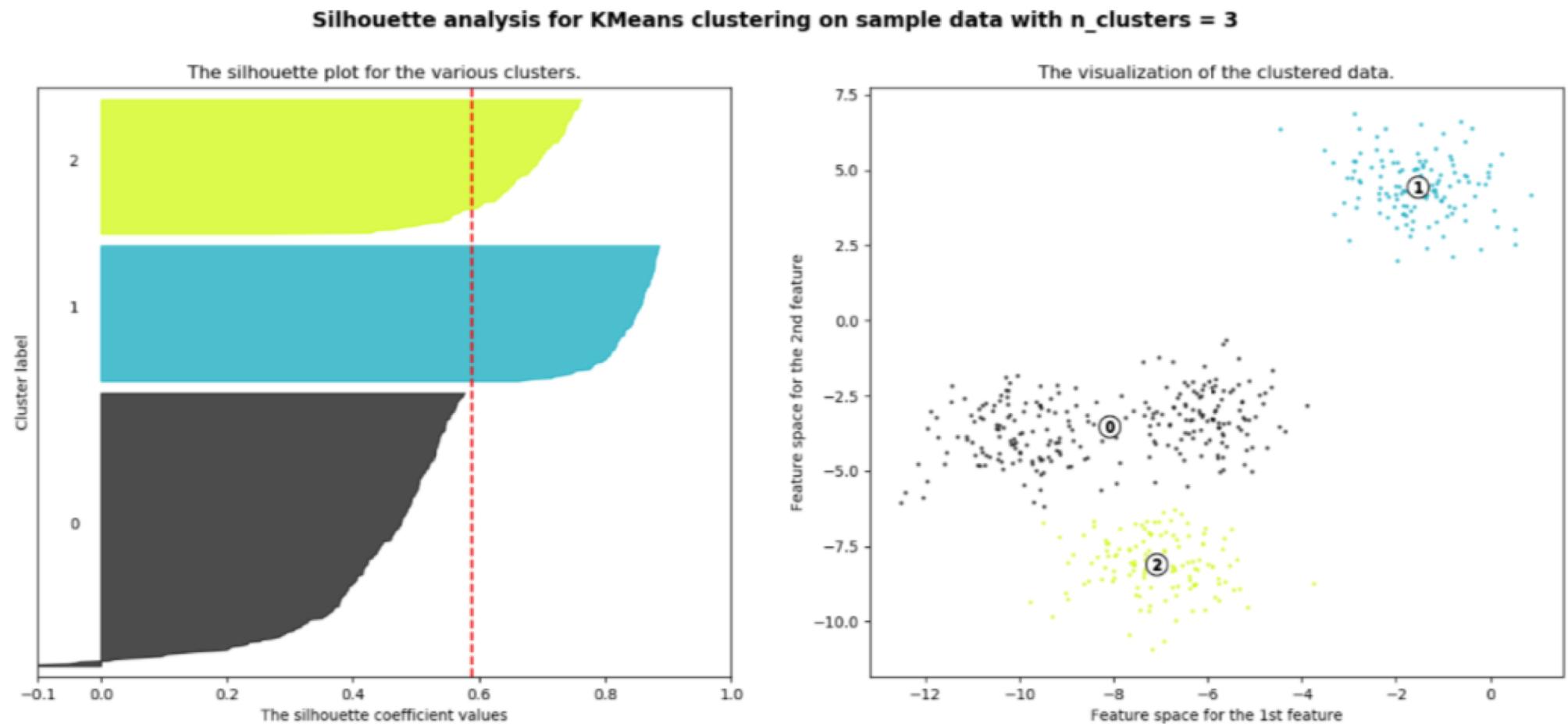
Há várias heurísticas:

- Ponto em que a função objetivo tem mudança de comportamento
- Coeficiente de silhueta

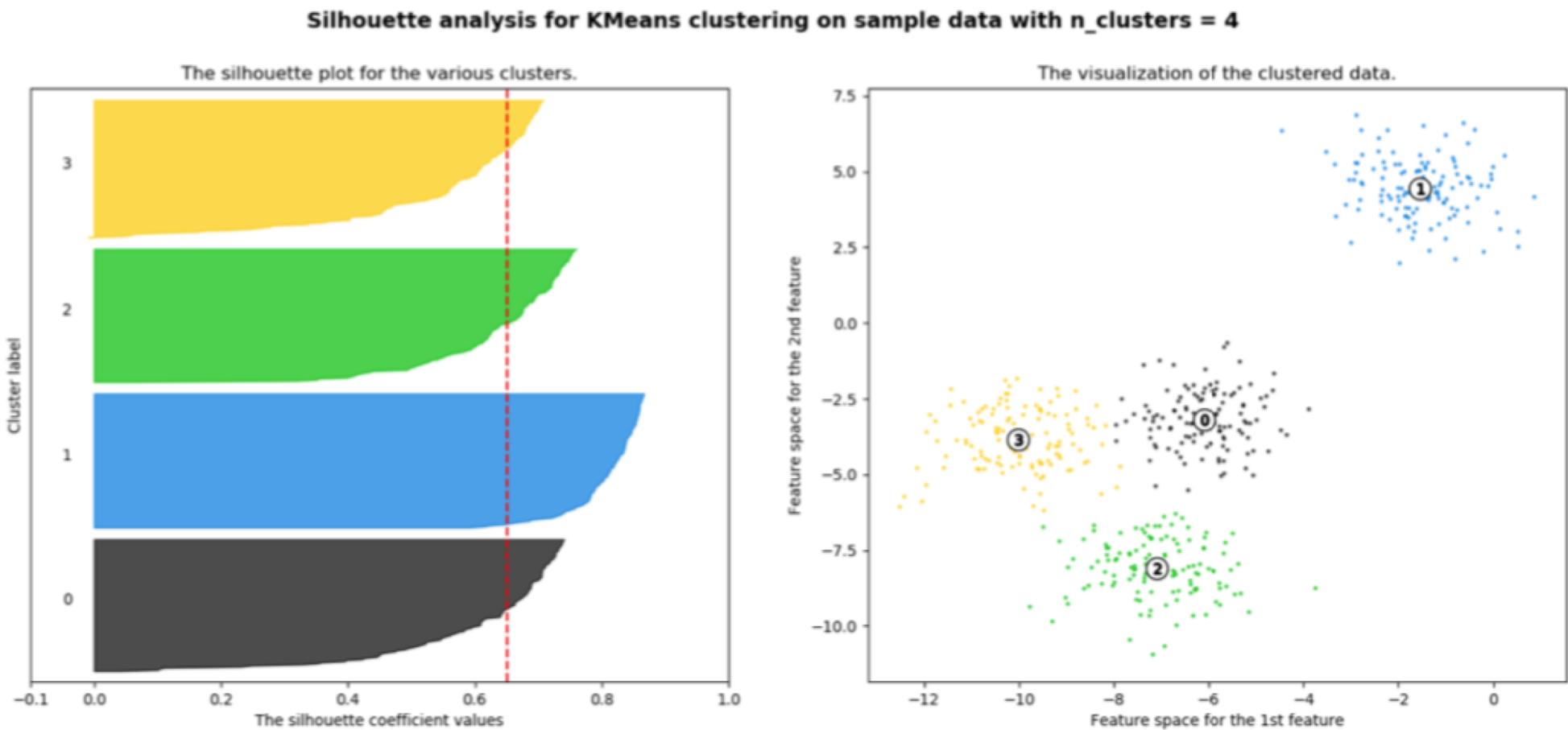
Exemplo com coeficiente de silhueta



Exemplo com coeficiente de silhueta



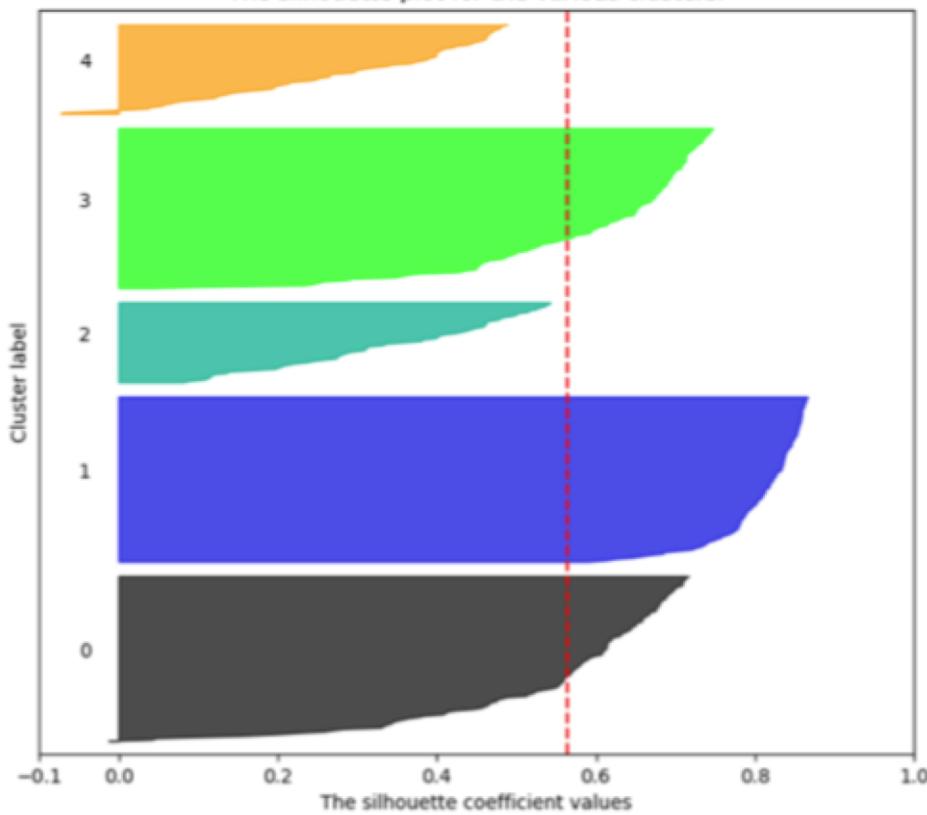
Exemplo com coeficiente de silhueta



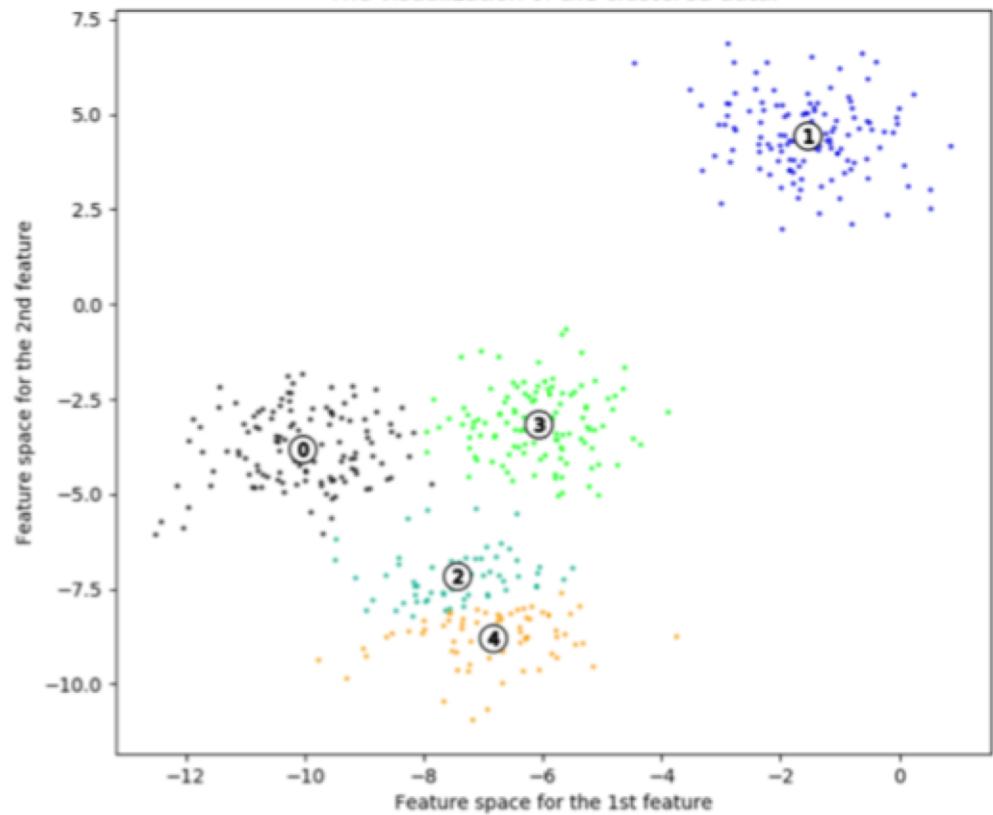
Exemplo com coeficiente de silhueta

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5

The silhouette plot for the various clusters.



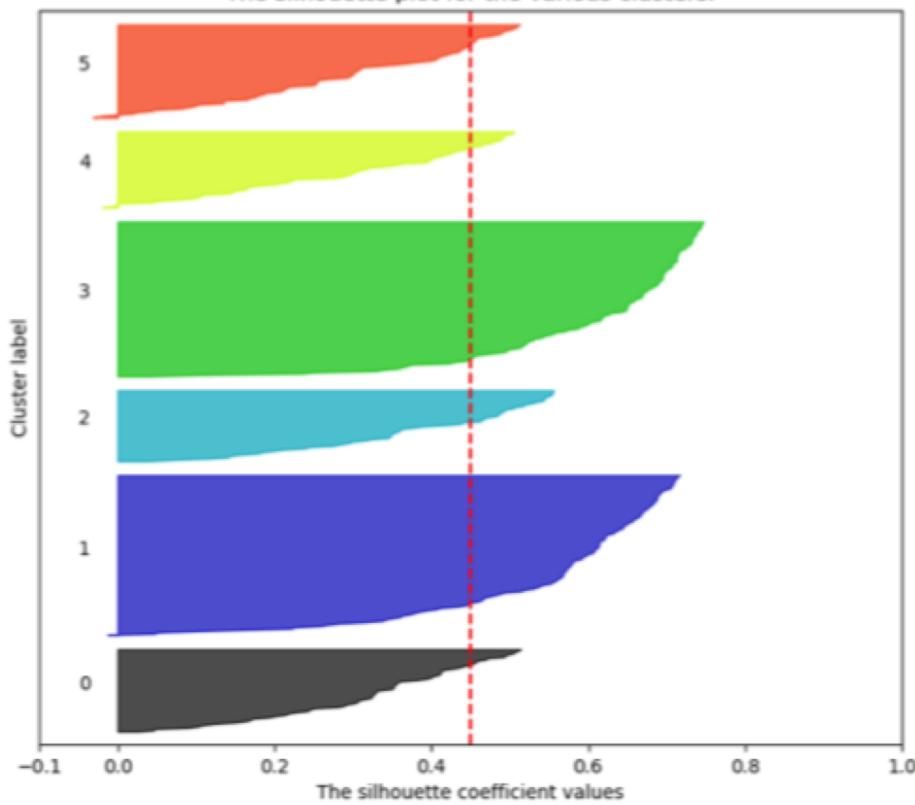
The visualization of the clustered data.



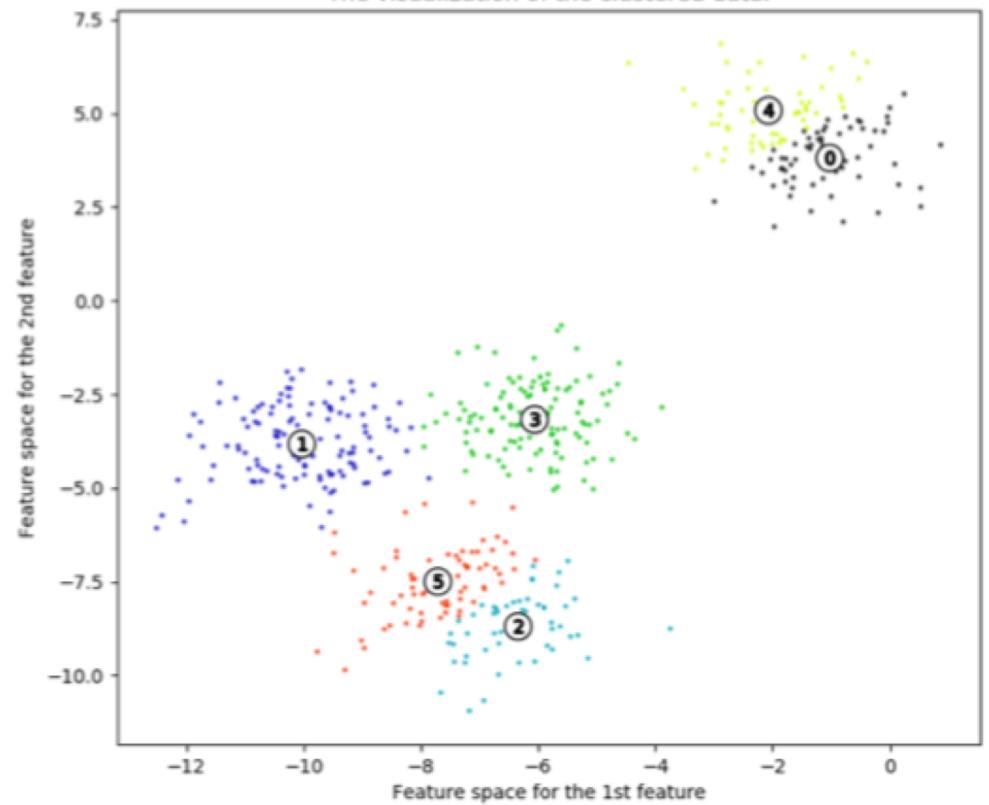
Exemplo com coeficiente de silhueta

Silhouette analysis for KMeans clustering on sample data with `n_clusters = 6`

The silhouette plot for the various clusters.



The visualization of the clustered data.



Características

Técnica de propósito geral

Número de grupos relativamente pequeno

Não funciona bem quando grupos naturais tem forma não regular

Problemas quando o número de dimensões é muito grande (maldição da dimensionalidade). Use PCA

Como resolver o problema
quando $d = 1$?

Vide lista =)

Variações comuns

k-medoids

k-center

Baseado em

<http://cs-people.bu.edu/evimaria/cs565-13/clustering.pdf>