

Curso: Aprendizagem de Máquina em Inteligência Artificial

Disciplina: Aprendizado Não Supervisionado

Prof. Marcelo Novaes de Rezende



Processo de Clustering

- (1) Feature extraction and selection: extract and select the most representative features from the original data set;
- (2) Clustering algorithm design: design the clustering algorithm according to the characteristics of the problem;
- (3) Result evaluation: evaluate the clustering result and judge the validity of algorithm;
- (4) Result explanation: give a practical explanation for the clustering result;

Fonte: A Comprehensive Survey of Clustering Algorithms
Dongkuan Xu· Yingjie Tian 2015



**Faculdade de Computação e Informática
Mackenzie**

Clustering (Classic Methods)

Conglomerados (Clusters)

Métodos Hierárquicos

Aglomerativos

Divisivo

Métodos não hierárquicos : Ex: **K-Means**



Faculdade de Computação e Informática
Mackenzie

Algoritmos de Clustering (Classificação usual)

Algoritmos Hierárquicos

Criam uma hierarquia de conjuntos de classes por fusão de classes menores em classes maiores (ascendente) ou por divisão de classes maiores em classes menores (descendente).

O resultado de um algoritmo hierárquico é uma árvore ou dendrograma.

Cortando a árvore num determinado nível é obtida uma partição dos indivíduos em k classes.

Hierárquicos aglomerativos: Partem de n indivíduos agrupados em n classes, cada classe com 1 indivíduo. Agrupam as classes sucessivamente até se obter uma única classe.

Hierárquicos divisivos: Partem de uma única classe que inclui os n indivíduos. As classes são sucessivamente divididas em classes menores até se obterem n classes, cada uma com um indivíduo.



Faculdade de Computação e Informática
Mackenzie

K-Means

The core idea of **K-means** is to **update the center of cluster** which is represented by the center of data points, by iterative computation and the iterative process will be continued **until some criteria for convergence is met**. **K-medoids** is an improvement of K-means to deal with discrete data, which takes the data point, most near the center of data points, as the Representative of the corresponding cluster.

Fonte: A Comprehensive Survey of Clustering Algorithms
Dongkuan Xu· Yingjie Tian 2015



Faculdade de Computação e Informática
Mackenzie

K-Means Advantages/Disadvantages

Advantages : relatively low time complexity and high computing efficiency in general;

Disadvantages : not suitable for non-convex data, relatively sensitive to the outliers, easily drawn into local optimal, the number of clusters needed to be preset, and the clustering result sensitive to the number of clusters

Fonte: A Comprehensive Survey of Clustering Algorithms
Dongkuan Xu· Yingjie Tian 2015



Faculdade de Computação e Informática
Mackenzie

Kmeans- Pseudocode

Como é o K-Means em “pseudocódigo”

Defina o número de clusters (k)

Defina os centróides iniciais dos k clusters

Faça

Forme os k clusters associando cada objeto a seu centróide mais próximo

Recompute o centróide de cada cluster

Enquanto mudarem os objetos dos clusters



Faculdade de Computação e Informática
Mackenzie

Clustering

Custo de uma partição em “n” clusters

$$\sum_{i=1}^n d(x_i, cx_i)^2$$

n: número de pontos

d=distância

Cx_i : cluster associado ao ponto x_i

Faça o gráfico número de cluster x custo com sklearn (inertia_)

kmeans_cost_sklearn.ipynb



Faculdade de Computação e Informática
Mackenzie

Ainda sobre clusters, qualidade da partição

- Há várias métricas para avaliar a “qualidade” de uma partição do dataset em clusters: Dunn, Silhouette etc...



Faculdade de Computação e Informática
Mackenzie

Dunn Index

O índice Dunn é dado pelo quociente entre a menor *distância entre pontos de clusters diferentes e a maior distância entre pontos do mesmo cluster.

$$\text{Dunn} = \frac{\min \text{dist pontos entre diferentes clusters}}{\max \text{dist pontos do mesmo cluster}}$$

Quanto maior o índice, melhor a partição (mais compacto e separado é o cluster)



Faculdade de Computação e Informática
Mackenzie

Dunn Index

Partindo de `dunn_index.ipynb`, vamos calcular o índice dunn para $k=2$ e $k=3$ com `kmeans`



Faculdade de Computação e Informática
Mackenzie

Silhouette

Para um ponto “i” de um cluster (com k pontos), a média das distâncias dele a cada um dos (k-1) outros pontos do cluster é dada por **a(i)**.

Para um ponto “i” de um cluster, a (menor) média das distâncias dele a cada um dos pontos de um outro cluster é dada por **b(i)**.

Silhouette para um ponto “i” é $S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$

Podemos calcular Silhouette para o cluster todo pela média dos pontos e para a partição toda, pela média de todos os pontos.

Quanto mais próximo de 1 o índice, melhor a partição.



Faculdade de Computação e Informática
Mackenzie

Silhouette

Atividade :

1)Partindo de silhouette_res.ipynb com k=2.

Obter (na mão) o índice Silhouette para o ponto 0 e compará-lo com o gerado pelo Scikit

2)Rodar com k=3. Comparar índices com k=3..qual é a melhor partição? Discussão.

3)Navegar em: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
(coisa linda!!!!)



Faculdade de Computação e Informática
Mackenzie

Silhouette

Silhouette para ponto 0:

$$a=1.414$$

$$b = 5.65+6.4+4+5.09/4=5.27$$

$$s=(5.27-1.414)/5.27=0.73$$



Faculdade de Computação e Informática
Mackenzie

Cluster Hierárquico

Já trabalhamos com o K-Means (não hierárquico usual). Vamos agora trabalhar com cluster hierárquico aglomerativo, partindo de `hierárquico.ipynb`. Vamos analisar o código.



Faculdade de Computação e Informática
Mackenzie

Cluster Hierárquico

Explique como foi a ordem de formação dos clusters.



Faculdade de Computação e Informática
Mackenzie

Cluster Hierárquico

Explique como foi o critério de decisão para a formação dos clusters.

Utilize Single Linkage

ver critérios de linkage em:

<https://www.youtube.com/watch?v=vg1w5ZUF5IA>



Faculdade de Computação e Informática
Mackenzie

Cluster: Caso prático

Cidades do Estado de São Paulo



Faculdade de Computação e Informática
Mackenzie

Cluster: Caso prático

Partindo de big_cluster.ipynb (dataset municípios.xlsx)

- 1) Eliminar séries irrelevantes
- 2) Criar uma variável categórica (por exemplo IDH > média no estado (1 ou 0))
- 3) Standardizar
- 4) Gerar a curva de custos para 2 a 20 clusters
- 5) Definir o número escolhido de clusters (n)
- 6) Verificar silhouette dos clusters para $n-1$, n e $n+1$..mudanças?
- 7) Criar em df (DataFrame) a série cluster com os labels obtidos com k-means



Faculdade de Computação e Informática
Mackenzie

Cluster: Caso prático

- 8) Criar para $n=i$ até número de clusters “df filtrado para o cluster i ”
- 9) Sumarizar os clusters
- 10) Comentar sobre os clusters



Faculdade de Computação e Informática
Mackenzie

Sobrou Tempo?

Começaremos KNN



Faculdade de Computação e Informática
Mackenzie

Até a próxima aula

OBRIGADO!

Prof Marcelo Rezende
email rezendemn@gmail.com