Aprendizagem de Máquina

Aprendizagem não-supervisionada

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA

Sumário

Introdução

- O algoritmo de aprendizagem recebe um conjunto com exemplos cujos rótulos não são conhecidos
- O objetivo é encontrar padrões previamente desconhecidos nos dados

Introdução

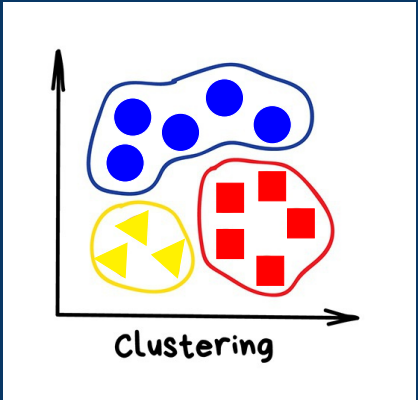
- Atividades não-supervisionadas incluem:
  - Agrupamento
  - Estimação de densidades
  - Descoberta de variáveis mais importantes

Agrupamento

Agrupamento

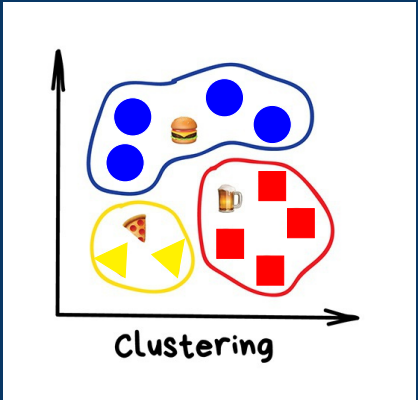
- Na tarefa de agrupamento, buscamos dividir um conjunto de dados em grupos cujos elementos são:

- Parecidos entre si
- Diferentes dos outros grupos



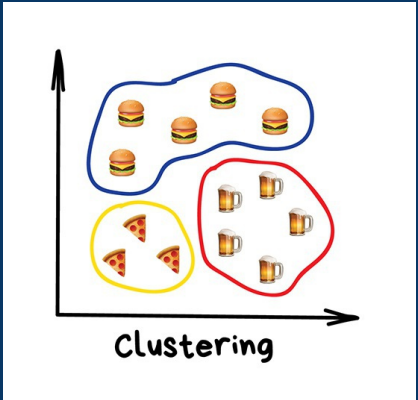
Agrupamento

- Após a realização do agrupamento, os grupos encontrados podem ser investigados para que os padrões sejam descobertos de fato



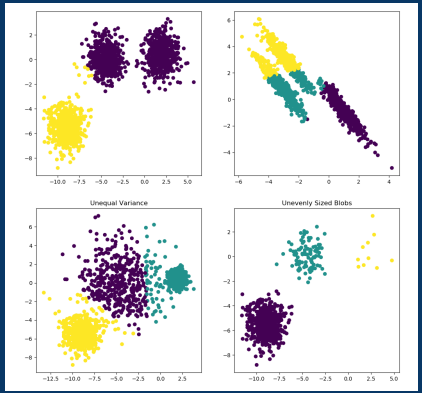
Agrupamento

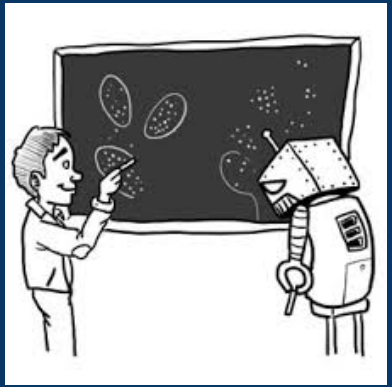
- Após a realização do agrupamento, os grupos encontrados podem ser investigados para que os padrões sejam descobertos de fato



Agrupamento

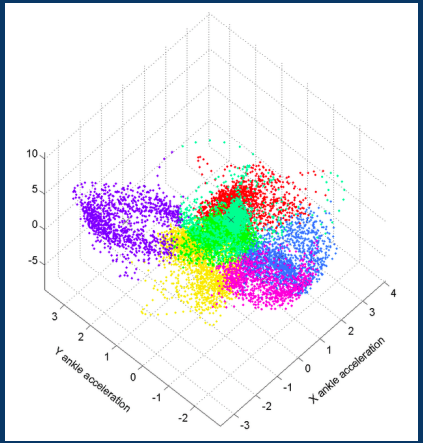
- Em conjuntos de dados 2D, essa tarefa pode ser visualmente resolvida por humanos até com mais sucesso do que a máquina:

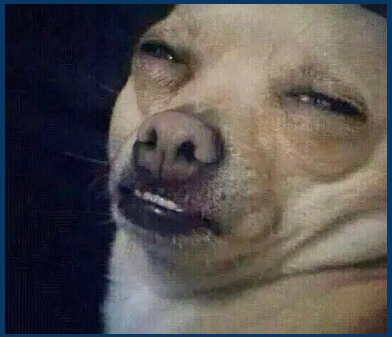




Agrupamento

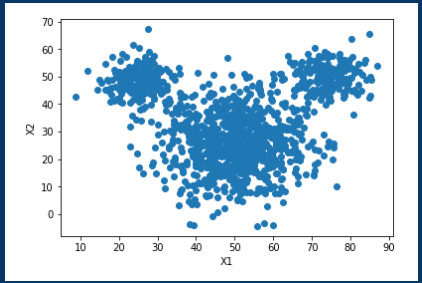
- As coisas começam a complicar quando vamos para 3 dimensões ou mais:





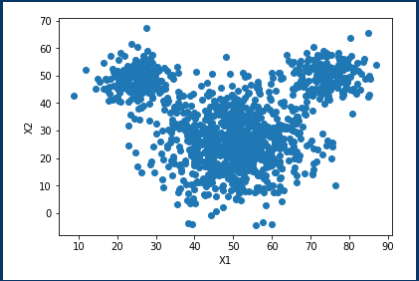
Agrupamento

- Vamos começar com um exemplo simples:
  - O conjunto ao lado foi gerado de forma simulada



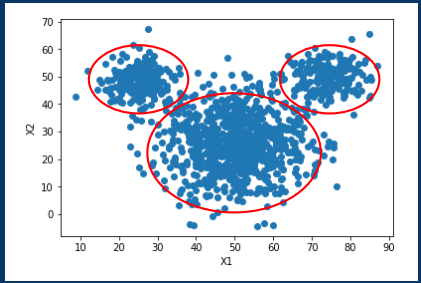
Agrupamento

- Vamos começar com um exemplo simples:
  - Podemos identificar 3 grupos ao observar os dados



Agrupamento

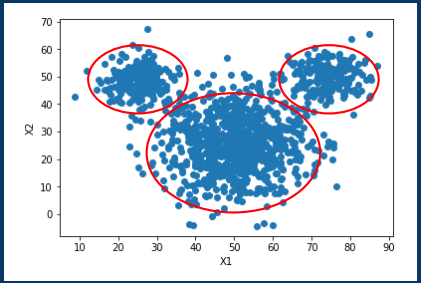
- Vamos começar com um exemplo simples:
  - Podemos identificar 3 grupos ao observar os dados



Agrupamento

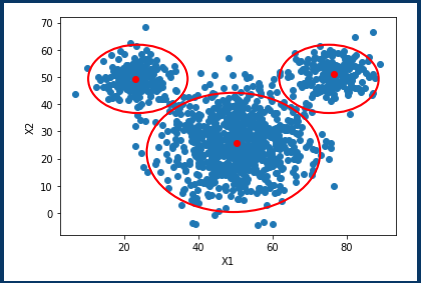
- Vamos começar com um exemplo simples:
  - Podemos identificar 3 grupos ao observar os dados

- Mas como automatizar esse processo?



Como automatizar esse processo?

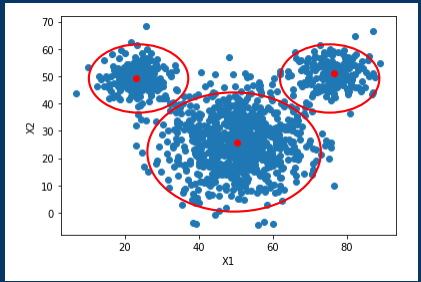
- Podemos definir cada grupo como o conjunto dos objetos mais próximos a um elemento central:



Como automatizar esse processo?

- Fazemos isso facilmente, mas o computador precisará buscar esses elementos

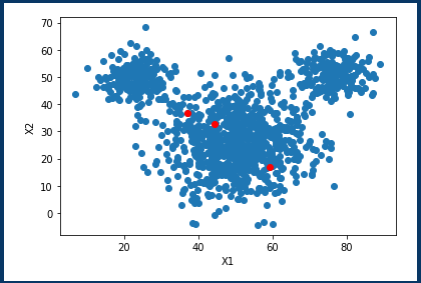
- E toda busca precisa de um ponto de partida



Como automatizar esse processo?

- Um bom ponto de partida é escolher aleatoriamente 3 elementos do nosso conjunto

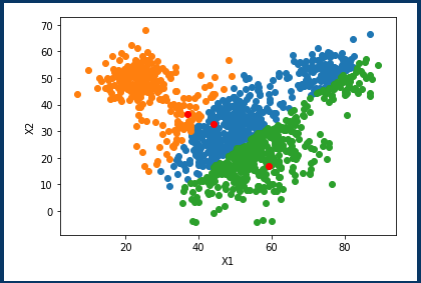
- Vamos chamar esses elementos de médias



Como automatizar esse processo?

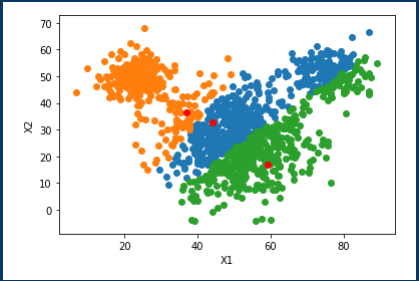
- Um bom ponto de partida é escolher aleatoriamente 3 elementos do nosso conjunto

- Temos como resultado o seguinte agrupamento inicial:



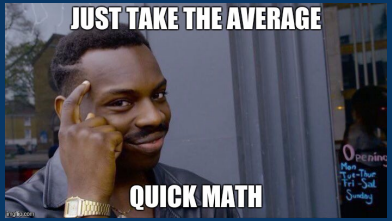
Como automatizar esse processo?

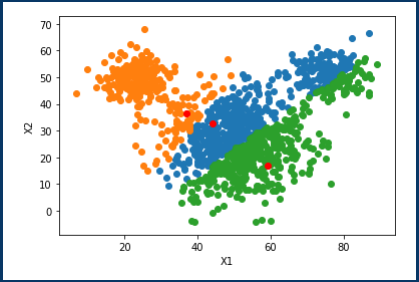
- Esse resultado parece bem ruim. O que faremos para melhorá-lo?



Como automatizar esse processo?

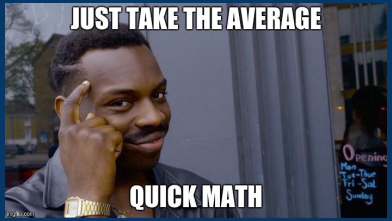
- Parece que se substituirmos cada média pelo ponto médio de cada área colorida, nos aproximaremos dos grupos desejados

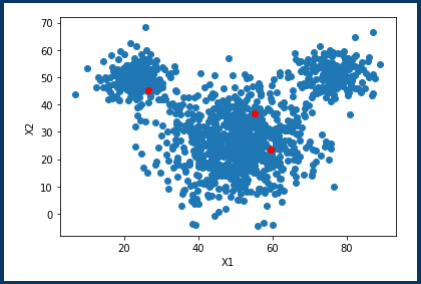




Como automatizar esse processo?

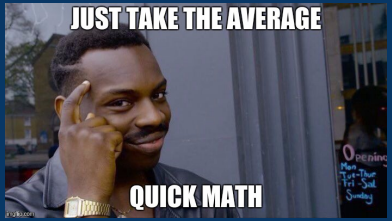
- Parece que se substituirmos cada média pelo ponto médio de cada área colorida, nos aproximaremos dos grupos desejados

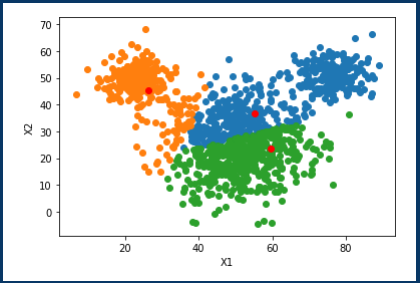




Como automatizar esse processo?

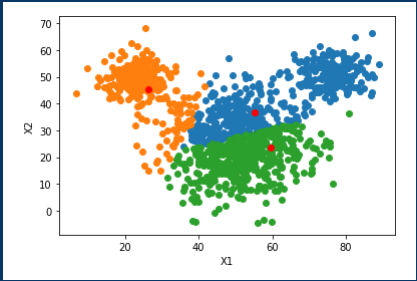
- Parece que se substituirmos cada média pelo ponto médio de cada área colorida, nos aproximaremos dos grupos desejados





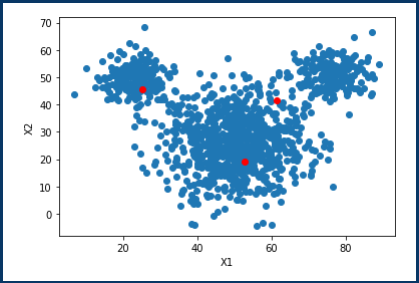
Como automatizar esse processo?

- Já melhorou, então vamos seguir com esses passos
- Novamente, vamos substituir nossas médias atuais pelos pontos médios das áreas coloridas correspondentes



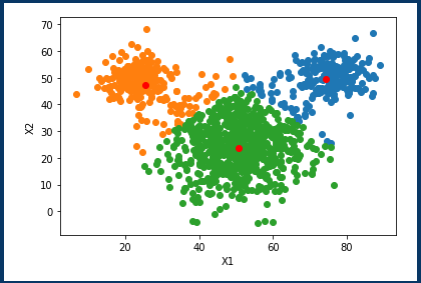
Como automatizar esse processo?

- Já melhorou, então vamos seguir com esses passos
- Novamente, vamos substituir nossas médias atuais pelos pontos médios das áreas coloridas correspondentes



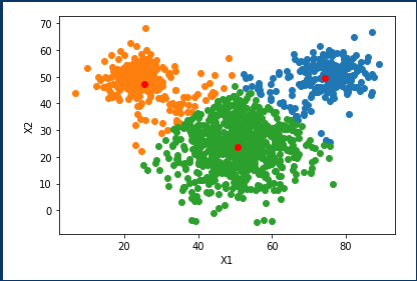
Como automatizar esse processo?

- Seguindo esses passos por mais algumas repetições, chegamos ao seguinte resultado:



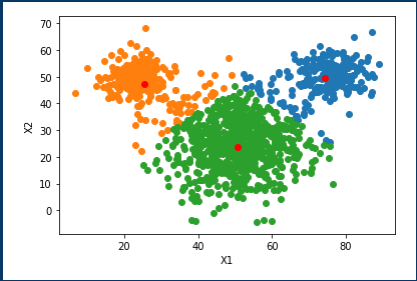
Como automatizar esse processo?

- Note que os grupos não ficaram separados como esperávamos visualmente
- Os elementos estão separados pelas suas distâncias em relação às médias



O algoritmo K-médias

- O algoritmo que acabamos de desenvolver é chamado de K-médias
- K é um parâmetro que indica o número de grupos que a máquina tentará encontrar nos dados



A matemática do K-médias

- Dado um conjunto de N observações $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, o algoritmo K-médias particiona as observações em K grupos $S = \{S_1, \ldots, S_K\}$ de forma a minimizar a soma da distâncias Euclidianas quadráticas:

$$J = \sum_{k=1}^{K} \sum_{\mathbf{x} \in S_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2$$

- Para encontrar o centroide $\boldsymbol{\mu}_k$ do grupo S_k que minimiza J, igualamos a derivada de J em relação a $\boldsymbol{\mu}_k$ a 0

$$2 \sum_{\mathbf{x} \in S_k} (\mathbf{x} - \boldsymbol{\mu}_k) = 0$$

A matemática do K-médias

- Para encontrar o centroide $\boldsymbol{\mu}_k$ do grupo S_k que minimiza J, igualamos a derivada de J em relação a $\boldsymbol{\mu}_k$ a 0

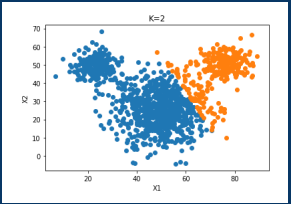
$$2 \sum_{\mathbf{x} \in S_k} (\mathbf{x} - \boldsymbol{\mu}_k) = 0$$

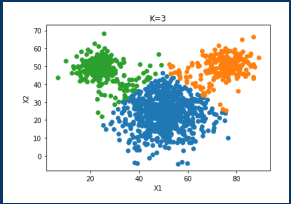
$$-N_k \boldsymbol{\mu}_k + \sum_{\mathbf{x} \in S_k} \mathbf{x} = 0$$

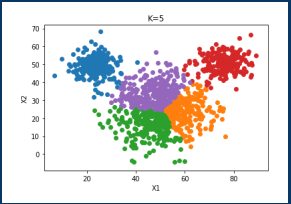
$$\boldsymbol{\mu}_k = \frac{\sum_{\mathbf{x} \in S_k} \mathbf{x}}{N_k}$$

A importância do parâmetro K

- O valor de K é extremamente importante e pode ser conhecido a priori por especialistas ou otimizado a partir dos dados

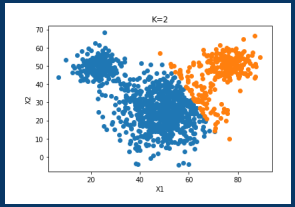


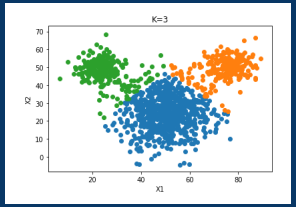


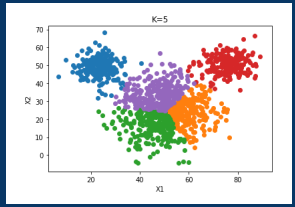


A importância do parâmetro K

- Além disso, não importa o real número de grupos nos dados, o algoritmo vai tentar encontrar o número de grupos especificado por K

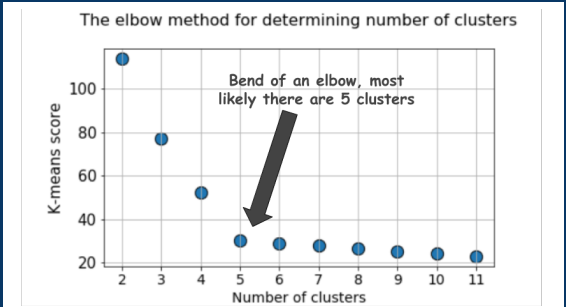






O método do “joelho”

- Para selecionar o valor ótimo de K, podemos usar o método do joelho ou do cotovelo (em inglês é *elbow method*)
- A ideia é que J vai continuar diminuindo se formos aumentando o número de grupos, porém em algum momento os ganhos não vão ser mais tão significativos
- Esse momento é o cotovelo da curva, que provavelmente representa o número correto de grupos. Daí pra frente é overfitting



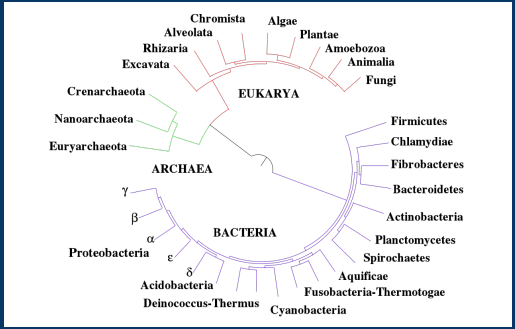
Agrupamento Hierárquico

Agrupamento Hierárquico

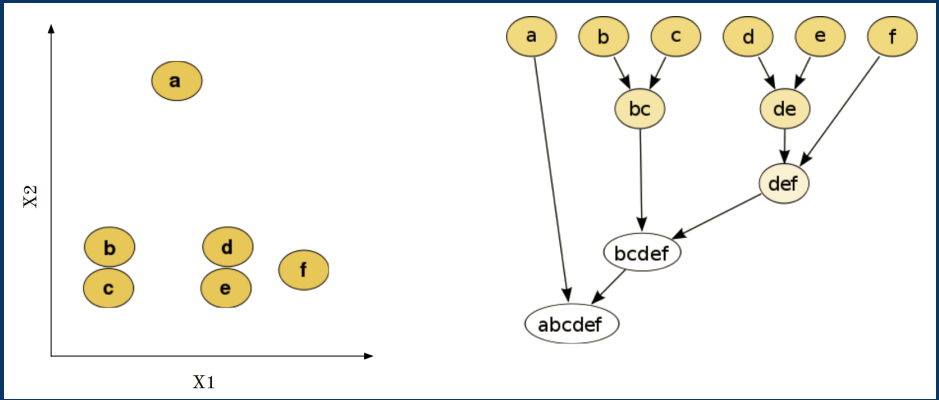
- Busca construir uma hierarquia com os grupos
- Dividido em dois tipos:
  - **Aglomerativo**: começa com cada elemento em um grupo separado, unindo-os gradativamente até que todos estejam juntos
  - **Divisivo**: começa com todos os elementos em um único grupo e os divide gradativamente até que cada elemento esteja em um grupo separado

Agrupamento Hierárquico

Os resultados do agrupamento hierárquico são geralmente apresentados na forma de um **dendrograma**



Como funciona o agrupamento hierárquico?



Prós e contras do agrupamento hierárquico

- **Prós:**
  - Não é necessário informar o número de grupos a priori
  - Visualização dos resultados no dendrograma
- **Contras:**
  - Treinamento muito custoso (lento)

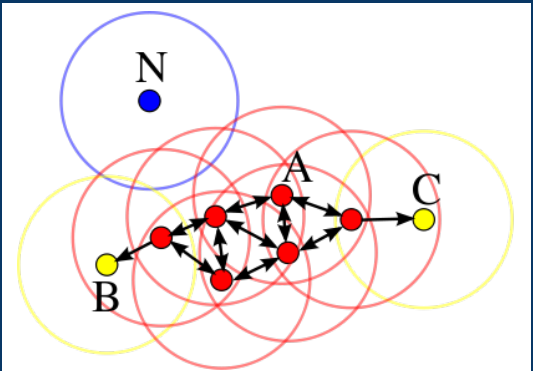
DBSCAN

DBSCAN

- O DBSCAN é um algoritmo baseado em densidade
- Para encontrar os grupos, ele atribui certas “funções” aos pontos do conjunto:
  - Um ponto p é um ponto principal se pelo menos **minPts** estiverem a um distância ϵ dele (incluindo ele mesmo)
  - Um ponto q é diretamente alcançável de p se q estiver dentro no máximo a uma distância ϵ do ponto principal p
  - Um ponto q é alcançável de p se houver um caminho $p_1, \ldots, p_n$, em que $p_1 = p$ e $p_n = q$ em que cada $p_i + 1$ é diretamente alcançável de p_i (ou seja todos os pontos no caminho são principais, menos q)
  - Por último se um ponto não for alcançável por ponto algum, ele é outlier ou ruído

DBSCAN

- Se *p* é um ponto principal, ele forma um cluster com todos os pontos (principais ou não) que são alcançáveis a partir dele. Todo cluster contém pelo menos um ponto principal. Pontos não-principais formam a “borda do cluster”, pois não podem ser usados para alcançar outros pontos



Prós e contras do DBSCAN

- **Prós:**
  - Não é necessário especificar o número de cluster, como no K-médias
  - DBSCAN consegue encontrar grupos com formas variadas, inclusive um grupo que envolve outro completamente (desde que não sejam conectadas)
  - DBSCAN é robusto contra outliers
  - Os parâmetros minPts e ϵ podem ser definidos por um expert no domínio do problema

Prós e contras do DBSCAN

- **Contras:**
  - DBSCAN não é totalmente determinístico: pontos não principais na borda de dois clusters podem trocar de grupo dependendo da ordem que os dados forem processados. No entanto, isso costuma afetar poucos os resultados e pontos principais e outliers são determinísticos
  - A qualidade do modelo resultante depende da medida de distância escolhida. Essa desvantagem não é exclusiva do DBSCAN
  - DBSCAN tem dificuldade de agrupar dados com densidades muito diferentes, porque fica difícil escolher valores para minPts e ϵ que funcionem bem para todos os clusters
  - Se os dados ou a escala dos dados não forem bem compreendidos, pode ser difícil escolher ϵ

Para Terminar

- Outros algoritmos de agrupamento que não precisam saber o número de grupos a priori incluem
  - Mean shift: encontra centroides que melhor representam os dados e depois elimina centroides desnecessários
  - OPTICS: modifica o DBSCAN para resolver o problema de agrupar dados com densidades muito diferentes
  - HDBSCAN: versão hierárquica do DBSCAN, forma clusters apenas com pontos principais e é mais rápido que o OPTICS

Para Terminar

- A avaliação de modelos de agrupamento não tem como levar em consideração uma informação de valores observados de uma variável alvo (*ground truth*)
- Portanto, costuma-se usar métricas que comparam os elementos dos grupos com os centroides (critério *J*) ou o quão similar os elementos de um grupo são entre si (coesão) em comparação aos outros clusters (separação), como é o caso da medida de silhueta (*silhouette*)
- Quando há um ground truth (geralmente quando queremos avaliar modelos de agrupamento usando conjuntos de classificação), podemos usar o índice de Rand, que avalia quantos elementos que compartilham grupos originalmente também estão no mesmo grupo no modelo de agrupamento

Sugestão de Atividade

- Implemente o K-médias, depois leia sobre o K-means++ (uma forma diferente de inicializar os centroides) e implemente-o também
- Use seus modelos implementados para agrupar conjuntos de classificação e veja se o método do cotovelo/joelho vai mostrar que o número ótimo de clusters é igual ao número de classes

FIM

Aprendizagem de Máquina

Aprendizagem não-supervisionada

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA