

Clustering - Fundamentos

Uma introdução abrangente à análise de agrupamento e suas aplicações em aprendizado de máquina e ciência de dados

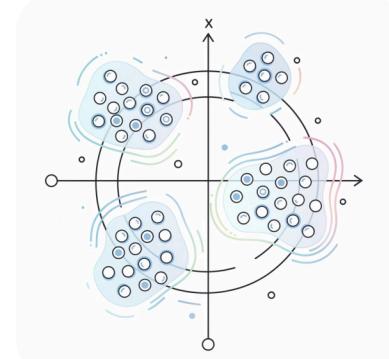
Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br
<https://eic.cefet-rj.br/~eogasawara>

O Que É Análise de Cluster?

Conceitos Fundamentais

Um **cluster** é uma coleção de objetos de dados que compartilham características semelhantes dentro do mesmo grupo, mas são dissimilares aos objetos em outros grupos. A análise de cluster é o processo de identificar similaridades entre dados com base nas características encontradas e agrupá-los em clusters significativos.



Aprendizado Não Supervisionado

Diferente de métodos supervisionados, não existem classes predefinidas. O algoritmo aprende por observação, descobrindo padrões naturais nos dados sem exemplos rotulados.

Aplicações Típicas

- Ferramenta independente para compreender a distribuição de dados
- Etapa de pré-processamento para outros algoritmos de análise
- Descoberta de estruturas ocultas em grandes conjuntos de dados

Aplicações da Análise de Cluster

A análise de cluster oferece soluções versáteis para diversos desafios em ciência de dados, desde redução dimensional até detecção de anomalias. Suas aplicações práticas transformam dados brutos em insights acionáveis.



Redução de Dados

Sumarização e compressão de conjuntos de dados para pré-processamento em regressão, PCA, classificação e análise de associação, tornando análises subsequentes mais eficientes.



Predição Baseada em Grupos

Identificação de características e padrões distintos para cada grupo, permitindo previsões mais precisas e segmentação estratégica de dados.



K-Vizinhos Mais Próximos

Localização eficiente de busca limitando-a a um ou poucos clusters, otimizando significativamente o tempo de processamento em grandes datasets.



Detecção de Outliers

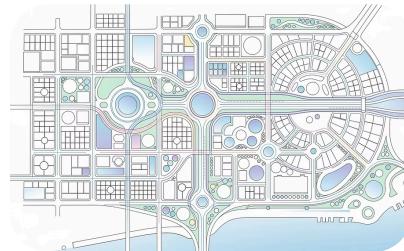
Identificação de valores atípicos como pontos distantes de qualquer cluster, essencial para garantir qualidade e integridade dos dados.

Exemplos Práticos de Clustering



Marketing

Descoberta de grupos distintos nas bases de clientes, permitindo que profissionais de marketing desenvolvam programas direcionados e personalizados para cada segmento identificado.



Planejamento Urbano

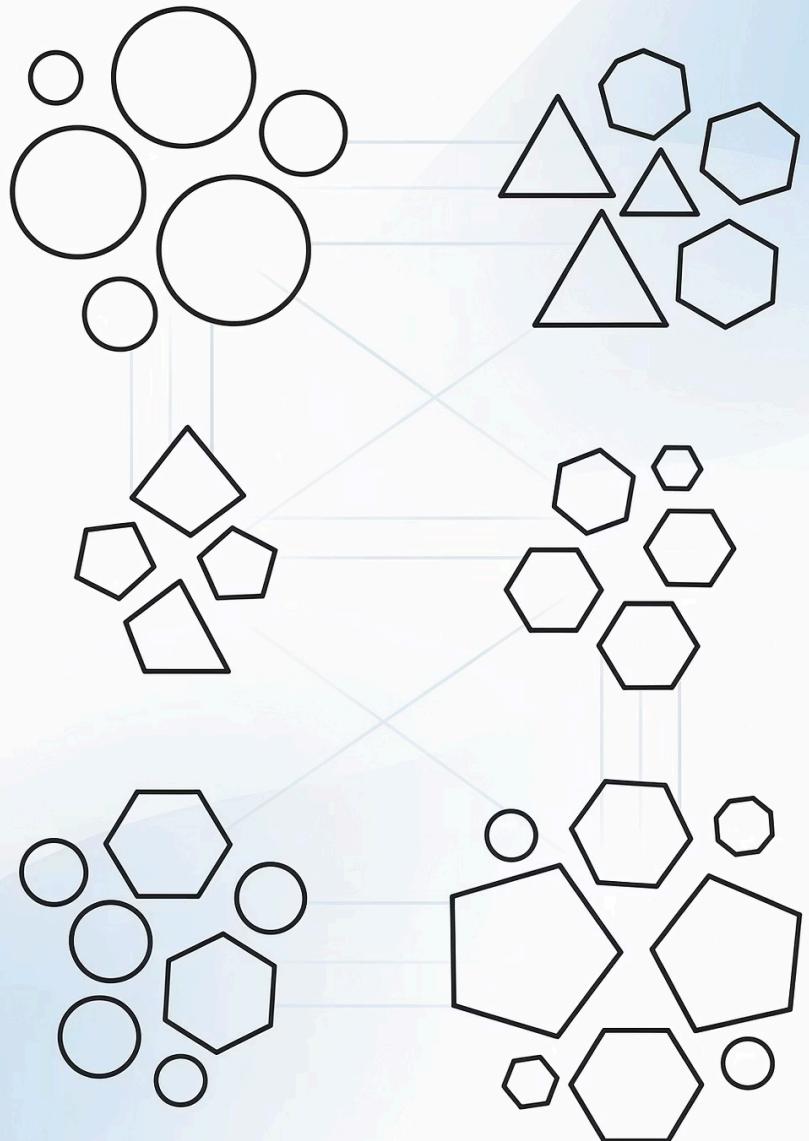
Identificação de grupos de residências de acordo com tipo, valor e localização geográfica, auxiliando no desenvolvimento de políticas públicas e infraestrutura adequada.



Climatologia

Compreensão do clima terrestre através da identificação de padrões atmosféricos e oceânicos, fundamentais para previsões meteorológicas e estudos de mudanças climáticas.

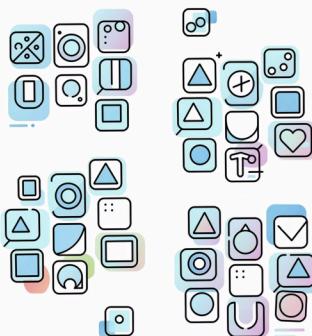
- ❑ **Insight:** Estas aplicações demonstram a versatilidade do clustering em transformar dados complexos em conhecimento açãoável em diversos domínios.



**O Que É Um
"Bom"
Agrupamento?**

Características de Um Clustering de Qualidade

"Um método de clustering eficaz produz clusters de alta qualidade com forte coesão interna e clara distinção entre grupos."



Princípios Fundamentais

- **Alta similaridade intra-classe:** Objetos dentro do mesmo cluster devem ser altamente coesos e semelhantes entre si
- **Baixa similaridade inter-classe:** Clusters diferentes devem ser distintivos e bem separados
- **Padrões ocultos:** Capacidade de descobrir estruturas não óbvias nos dados



Medida de Similaridade

A escolha da métrica apropriada impacta diretamente a qualidade dos resultados

Implementação

A eficiência e robustez do algoritmo utilizado

Descoberta de Padrões

Capacidade de revelar estruturas significativas e açãoáveis

Medindo a Qualidade do Clustering

Métricas de Dissimilaridade/Similaridade

A similaridade é expressa através de funções de distância, tipicamente métricas como $d(i, j)$. As definições variam significativamente conforme o tipo de variável:

- Variáveis de intervalo
- Variáveis booleanas e categóricas
- Variáveis ordinais e de razão
- Variáveis vetoriais

Ponderação de Variáveis

Pesos devem ser associados a diferentes variáveis baseando-se nas aplicações específicas e na semântica dos dados, refletindo a importância relativa de cada característica no contexto do problema.

Função de Qualidade

Geralmente existe uma função de qualidade separada que mede a "bondade" de um cluster. Esta medida quantifica quanto bem os dados foram agrupados.

Desafio da Subjetividade

É difícil definir "suficientemente similar" ou "bom o suficiente". A resposta é tipicamente altamente subjetiva e dependente do contexto da aplicação.

Considerações Para Análise de Cluster

A escolha adequada de parâmetros e abordagens é crucial para o sucesso da análise de cluster. Diversos fatores devem ser cuidadosamente avaliados antes da implementação.

1

Critérios de Particionamento

Nível único vs. hierárquico: Particionamento hierárquico multi-nível é frequentemente desejável, oferecendo diferentes níveis de granularidade na análise.

2

Separação de Clusters

Exclusivo vs. não-exclusivo: Clusters exclusivos (um cliente pertence a apenas uma região) ou não-exclusivos (um documento pode pertencer a múltiplas classes).

3

Medida de Similaridade

Baseada em distância vs. conectividade: Distância euclidiana, rede viária, vetorial versus densidade ou contiguidade espacial.

4

Espaço de Clustering

Espaço completo vs. subespaços: Espaço completo em baixa dimensionalidade versus subespaços em clustering de alta dimensão.

Requisitos e Desafios



Escalabilidade

Capacidade de realizar clustering em todos os dados, não apenas em amostras, mantendo eficiência computacional mesmo com datasets massivos.



Tipos de Atributos

Habilidade para lidar com atributos numéricos, binários, categóricos, ordinais, vinculados e misturas destes diferentes tipos de dados.



Clustering Baseado em Restrições

Usuários podem fornecer inputs sobre restrições, utilizando conhecimento de domínio para determinar parâmetros de entrada.



Interpretabilidade e Usabilidade

Os resultados devem ser compreensíveis e aplicáveis, traduzindo descobertas técnicas em insights açãoáveis.



Formas Arbitrárias

Descoberta de clusters com formas complexas e não-convexas



Dados Ruidosos

Robustez para lidar com ruído e outliers



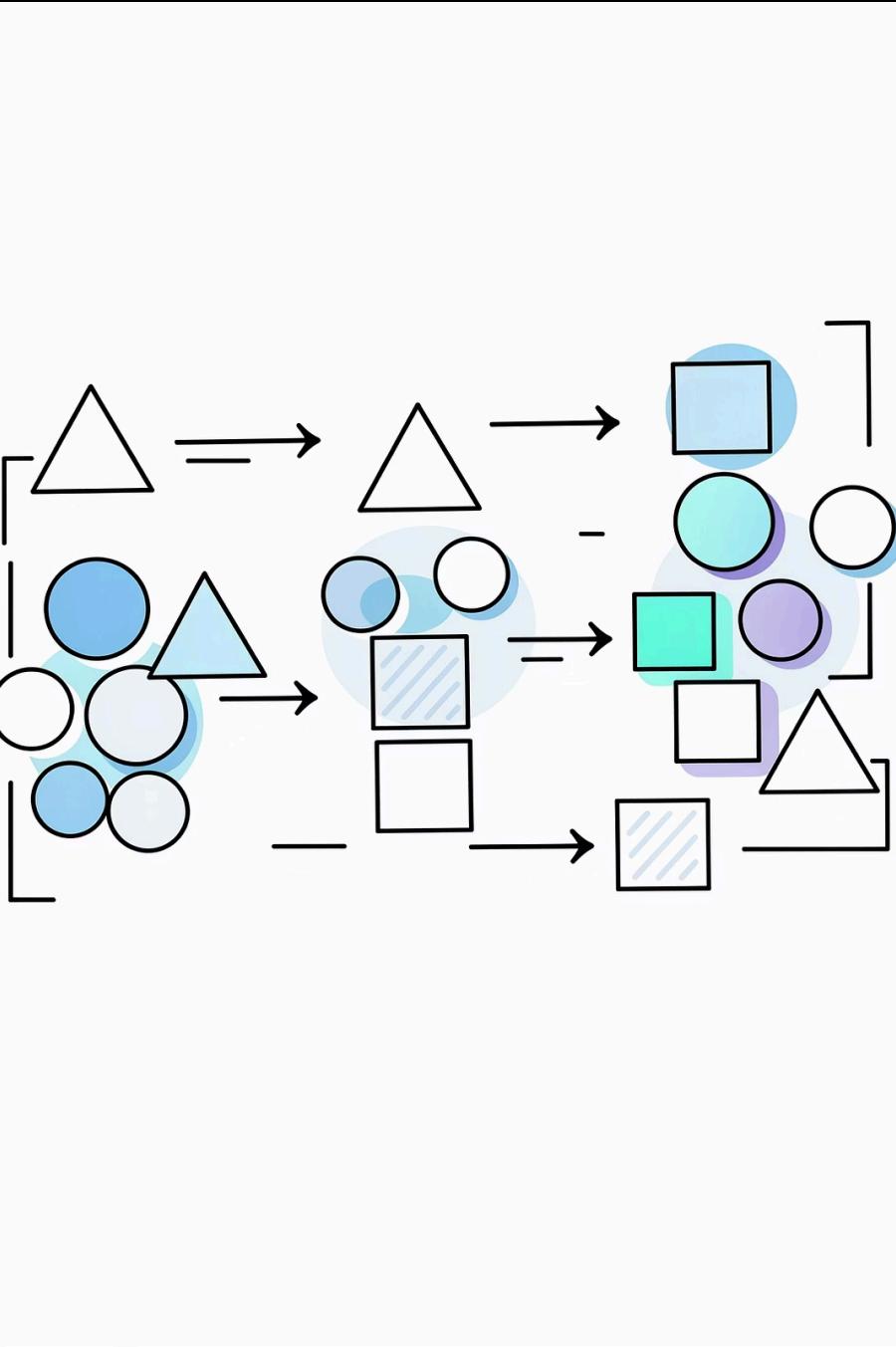
Clustering Incremental

Insensibilidade à ordem de entrada dos dados



Alta Dimensionalidade

Eficácia em espaços de alta dimensão



Formalização do Problema de Clustering

Etapas Básicas Para Desenvolver Uma Tarefa de Clustering

O desenvolvimento sistemático de uma solução de clustering requer seguir etapas bem definidas, desde a seleção inicial de características até a validação e interpretação final dos resultados.



Seleção de Características

Selecionar informações relevantes à tarefa de interesse, minimizando redundância e maximizando o poder discriminativo dos dados.



Medida de Proximidade

Definir como calcular a similaridade entre dois vetores de características, escolhendo métricas apropriadas ao tipo de dados.



Critério de Clustering

Expressar objetivos através de uma função de custo ou conjunto de regras que guiem o processo de agrupamento.



Algoritmos de Clustering

Escolher algoritmos apropriados considerando escalabilidade, tipos de dados e requisitos específicos da aplicação.



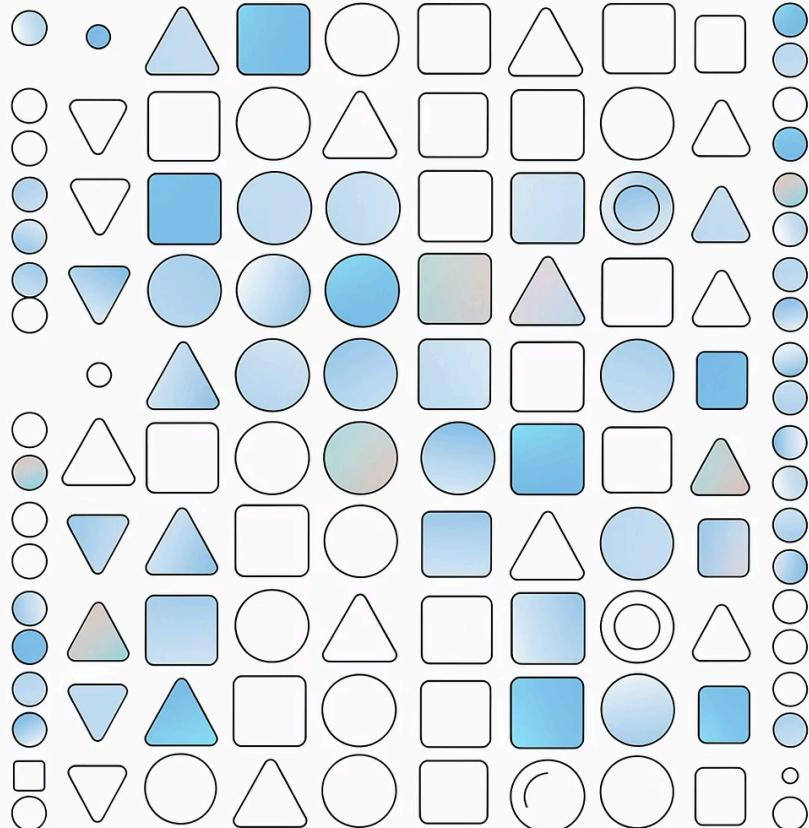
Validação dos Resultados

Aplicar testes de validação e tendência de clustering para avaliar objetivamente a qualidade dos agrupamentos obtidos.



Interpretação dos Resultados

Integrar descobertas com aplicações práticas, traduzindo resultados técnicos em insights de negócios acionáveis.



Fundamentos de Similaridade

Compreender como medir a semelhança entre objetos de dados é fundamental para técnicas avançadas de análise e mineração de dados. Estas medidas formam a base para algoritmos de clustering, classificação e recuperação de informação.

Similaridade e Dissimilaridade

Similaridade

Medida numérica de quanto semelhantes são dois objetos de dados. O valor é mais alto quando os objetos são mais parecidos, frequentemente variando no intervalo [0,1]. Esta métrica quantifica a proximidade conceitual entre elementos.

Dissimilaridade (Distância)

Medida numérica de quanto diferentes são dois objetos de dados. Valores mais baixos indicam maior semelhança. A dissimilaridade mínima é frequentemente 0, enquanto o limite superior varia conforme a métrica utilizada.

- ❑ **Proximidade:** Termo genérico que se refere tanto à similaridade quanto à dissimilaridade entre objetos de dados.

Distância de Minkowski para Dados Numéricos

A distância de Minkowski é uma família generalizada de métricas de distância amplamente utilizada para dados numéricos. Esta fórmula fornece uma estrutura flexível que engloba várias medidas de distância comuns através da variação do parâmetro h .

Fórmula Geral

A distância de Minkowski entre dois objetos i e j é definida como:

$$d(i, j) = \sqrt[h]{\sum_{f=1}^p |x_{if} - x_{jf}|^h}$$

onde p é o número de dimensões e h é o parâmetro de ordem.

Propriedades

- Métrica positiva e simétrica
- Satisfaz a desigualdade triangular
- Parametrizável através de h
- Aplicável a espaços multidimensionais

Variações da Distância de Minkowski

Diferentes valores do parâmetro h na fórmula de Minkowski produzem métricas de distância com propriedades e aplicações distintas. Cada variação possui características únicas que a tornam adequada para diferentes cenários analíticos.



Distância de Manhattan ($h=1$)

$$d(i, j) = \sum_{f=1}^p |x_{if} - x_{jf}|$$

Também conhecida como distância L₁ ou City Block. Mede a distância percorrida ao longo de eixos em ângulos retos, como em uma grade urbana.



Distância Euclidiana ($h=2$)

$$d(i, j) = \sqrt{\sum_{f=1}^p (x_{if} - x_{jf})^2}$$

A métrica de distância mais comum, representando a linha reta entre dois pontos no espaço euclidiano.



Distância de Chebyshev ($h \rightarrow \infty$)

$$d(i, j) = \max_{f=1}^p |x_{if} - x_{jf}|$$

Também chamada de distância L_∞ ou distância do máximo. Define distância como a maior diferença em qualquer dimensão.

Medidas de Proximidade para Atributos Binários

Atributos binários requerem abordagens especializadas para cálculo de similaridade. A tabela de contingência organiza as concordâncias e discordâncias entre dois objetos, permitindo o cálculo de diversos coeficientes de similaridade.

Tabela de Contingência

Objeto i = 1	q	r	q+r
Objeto i = 0	s	t	s+t
Soma	q+s	r+t	p

Coeficientes

Coeficiente de Correspondência Simples (SMC):

$$SMC = \frac{q + t}{q + r + s + t}$$

Coeficiente de Jaccard:

$$J = \frac{q}{q + r + s}$$

- ❑ **Nota:** O coeficiente de Jaccard é equivalente à medida de "coerência" e ignora concordâncias negativas (t).

Dissimilaridade entre Variáveis Binárias

EXEMPLO APLICADO

Para ilustrar o cálculo de dissimilaridade, consideremos atributos com diferentes características de simetria. O gênero representa um atributo simétrico, enquanto os demais atributos são binários assimétricos. Atribuímos valor 1 para Y e P, e valor 0 para N.

01

Identificar Tipo de Atributo

Classifique cada atributo como simétrico ou assimétrico. Atributos simétricos tratam ambos os estados igualmente, enquanto assimétricos dão maior peso a uma condição.

02

Construir Tabela de Contingência

Organize os dados em uma matriz 2×2 , contabilizando concordâncias positivas (q), discordâncias (r, s) e concordâncias negativas (t).

03

Aplicar Fórmula Apropriada

Para atributos assimétricos, use Jaccard. Para simétricos, use SMC. Calcule a dissimilaridade como 1 menos a similaridade obtida.

Medidas de Proximidade para Atributos Nominais

Atributos nominais representam categorias sem ordem intrínseca, como cores, marcas ou tipos. Para estes dados, utilizamos abordagens baseadas em correspondência de categorias, contando o número de atributos que coincidem entre objetos.

Método de Correspondência Simples

A dissimilaridade entre dois objetos i e j pode ser calculada como:

$$d(i, j) = \frac{p - m}{p}$$

onde p é o número total de atributos e m é o número de correspondências (atributos com valores iguais).

Alternativamente, a similaridade pode ser expressa como:

$$sim(i, j) = \frac{m}{p}$$

Ponderação de Atributos

Podemos atribuir pesos diferentes a cada atributo conforme sua importância, resultando em uma medida ponderada:

$$d(i, j) = \frac{\sum_{f=1}^p w_f \delta_{if,jf}}{\sum_{f=1}^p w_f}$$

Tratamento de Variáveis Ordinais

Variáveis ordinais possuem uma ordem natural entre suas categorias, como classificações (baixo, médio, alto) ou rankings. Para calcular distâncias, transformamos estas categorias em valores numéricos preservando a ordenação, depois normalizamos para o intervalo [0,1].



Mapeamento de Ranks

Atribua ranks sequenciais às categorias ordinais: 1, 2, 3, ..., M_f , onde M_f é o número de categorias do atributo f .

Normalização

Transforme cada rank para o intervalo [0,1] usando a fórmula:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

onde r_{if} é o rank do objeto i no atributo f .

Cálculo de Distância

Aplique qualquer métrica de distância (como Euclidiana ou Manhattan) aos valores normalizados z_{if} , tratando-os como dados numéricos contínuos.

Similaridade do Cosseno

A similaridade do cosseno mede o cosseno do ângulo entre dois vetores em um espaço multidimensional. Esta métrica é especialmente útil em mineração de texto e recuperação de informação, pois é invariante à magnitude dos vetores, focando apenas em sua direção.

Interpretação Geométrica

Dois vetores com orientação similar terão um cosseno próximo de 1, independentemente de suas magnitudes. Vetores perpendiculares resultam em cosseno 0, enquanto vetores opostos produzem -1.

Fórmula da Similaridade

$$\cos(A, B) = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i^2)} \times \sqrt{\sum_{i=1}^n (B_i^2)}}$$

O numerador calcula o produto interno dos vetores, enquanto o denominador normaliza pelo produto de suas magnitudes, garantindo que o resultado permaneça em [-1, 1].



Aplicação em Documentos

Frequentemente usada para comparar documentos representados como vetores de termos, onde cada dimensão corresponde a uma palavra e o valor à sua frequência.



Sistemas de Recomendação

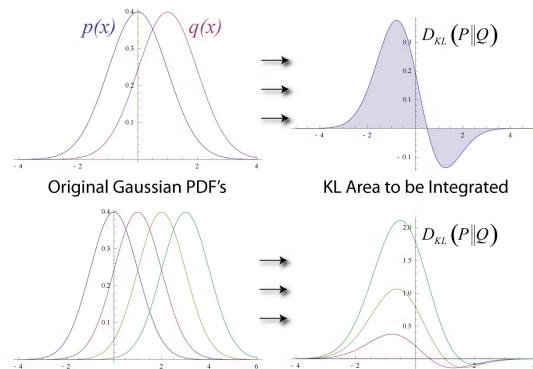
Útil para encontrar itens similares baseando-se em perfis de usuários ou características de produtos em espaços de alta dimensionalidade.

Divergência KL: Comparando Distribuições de Probabilidade

A divergência de Kullback-Leibler (KL) é uma medida fundamental da teoria da informação que quantifica a diferença entre duas distribuições de probabilidade sobre a mesma variável. Também conhecida como entropia relativa ou divergência de informação, ela mede a informação perdida quando uma distribuição é usada para aproximar outra.

Conceito e Definição

$D_{KL}(p(x) \parallel q(x))$ representa a divergência de $q(x)$ em relação a $p(x)$, quantificando quanta informação é perdida quando $q(x)$ é usada para aproximar $p(x)$. Esta medida é assimétrica: $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$.



Formas Matemáticas

Forma Discreta:

$$D_{KL}(p(x) \parallel q(x)) = \sum_{x \in X} p(x) \ln \left(\frac{p(x)}{q(x)} \right)$$

Utilizada para distribuições de probabilidade discretas, somando sobre todos os valores possíveis.

Forma Contínua:

$$D_{KL}(p(x) \parallel q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

Aplicada a distribuições contínuas, integrando sobre todo o domínio.

Propriedades Importantes

- Sempre não-negativa: $D_{KL} \geq 0$
- Zero apenas quando $p = q$
- Assimétrica
- Não satisfaz desigualdade triangular

Aplicações Práticas

- Seleção de modelos em machine learning
- Compressão de dados
- Análise de modelos probabilísticos
- Otimização de redes neurais

Combinando Tipos Mistos de Dados

Conjuntos de dados reais frequentemente contêm atributos de tipos diversos: numéricos, nominais, ordinais e binários. Para calcular distâncias entre objetos com atributos heterogêneos, precisamos de uma abordagem unificada que respeite as características de cada tipo de dado.

Normalização por Tipo

Cada tipo de atributo requer tratamento específico: atributos numéricos são normalizados para $[0,1]$, ordinais são convertidos em ranks normalizados, e nominais usam correspondência binária (0 ou 1).

Ponderação Adaptativa

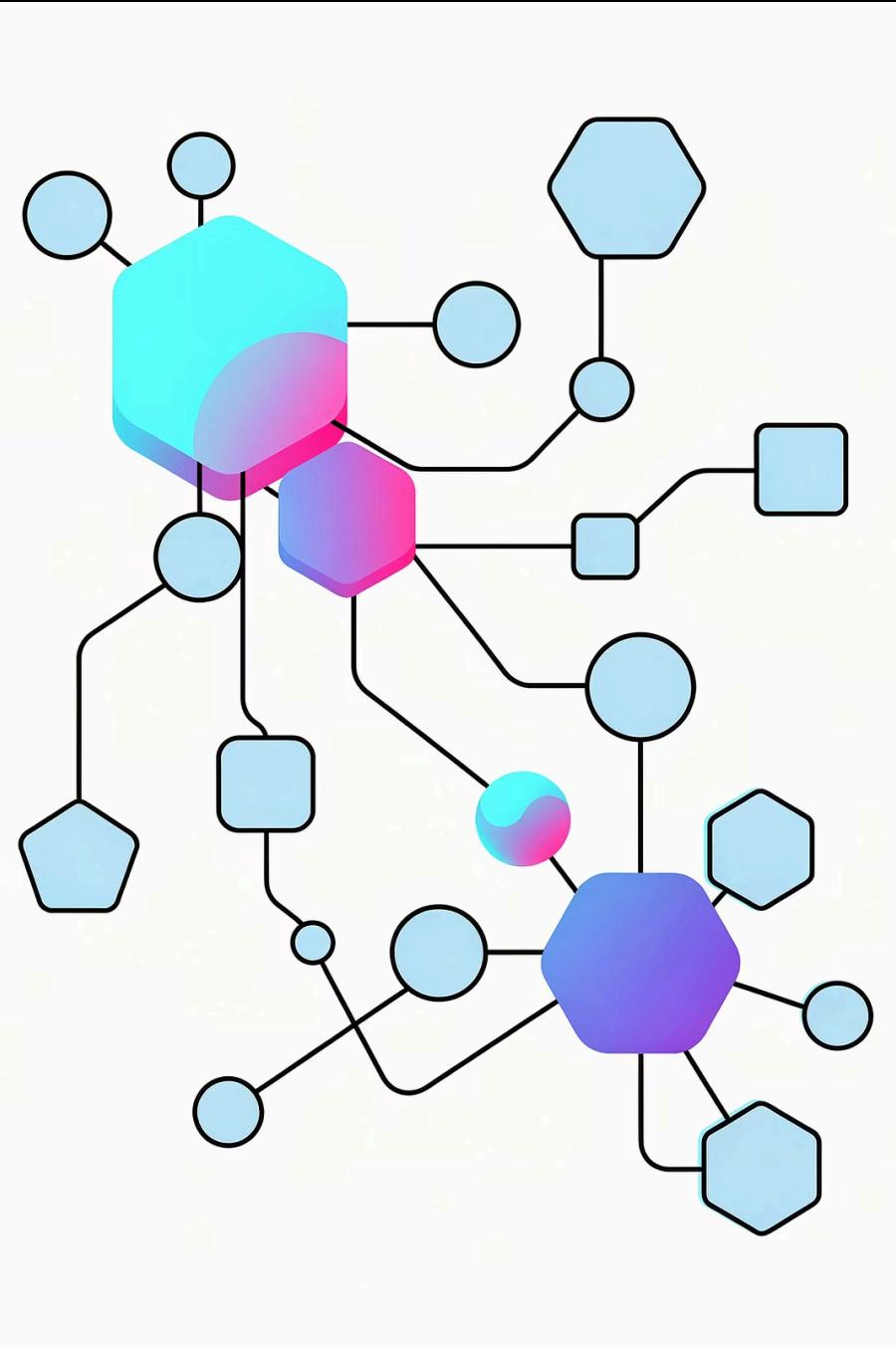
Atributos podem receber pesos diferentes baseados em sua relevância ou confiabilidade. Atributos com valores ausentes podem ter peso zero para aquele par de objetos específico.

Fórmula Generalizada

A dissimilaridade combinada usa uma soma ponderada das contribuições de cada atributo:

$$d(i, j) = \frac{\sum_{f=1}^p w_f \cdot \delta_f(i, j)}{\sum_{f=1}^p w_f}$$

- ❑ **Importante:** A função $\delta_f(i, j)$ é escolhida apropriadamente para cada tipo de atributo, e o peso w_f pode ser ajustado para refletir a importância ou disponibilidade do atributo.



Panorama dos Métodos de Clustering

Após compreender as medidas de similaridade, podemos explorar como agrupar objetos similares. O clustering é uma técnica fundamental de aprendizado não supervisionado com diversas abordagens metodológicas, cada uma adequada para diferentes tipos de dados e objetivos analíticos.

Principais Abordagens de Clustering (I)

As técnicas de clustering dividem-se em várias categorias baseadas em seus princípios fundamentais e estratégias algorítmicas. Cada abordagem oferece vantagens distintas para diferentes cenários de análise de dados.



Abordagem por Particionamento

Constrói várias partições do conjunto de dados e as avalia segundo critérios como minimização da soma dos erros quadrados. Métodos eficientes para grandes conjuntos de dados.

- **K-means:** Algoritmo clássico baseado em centroides
- **K-medoids:** Mais robusto a outliers que k-means
- **CLARANS:** Eficiente para dados espaciais



Abordagem Hierárquica

Cria uma decomposição hierárquica do conjunto de dados usando algum critério, resultando em uma estrutura de árvore (dendrograma) que pode ser cortada em diferentes níveis.

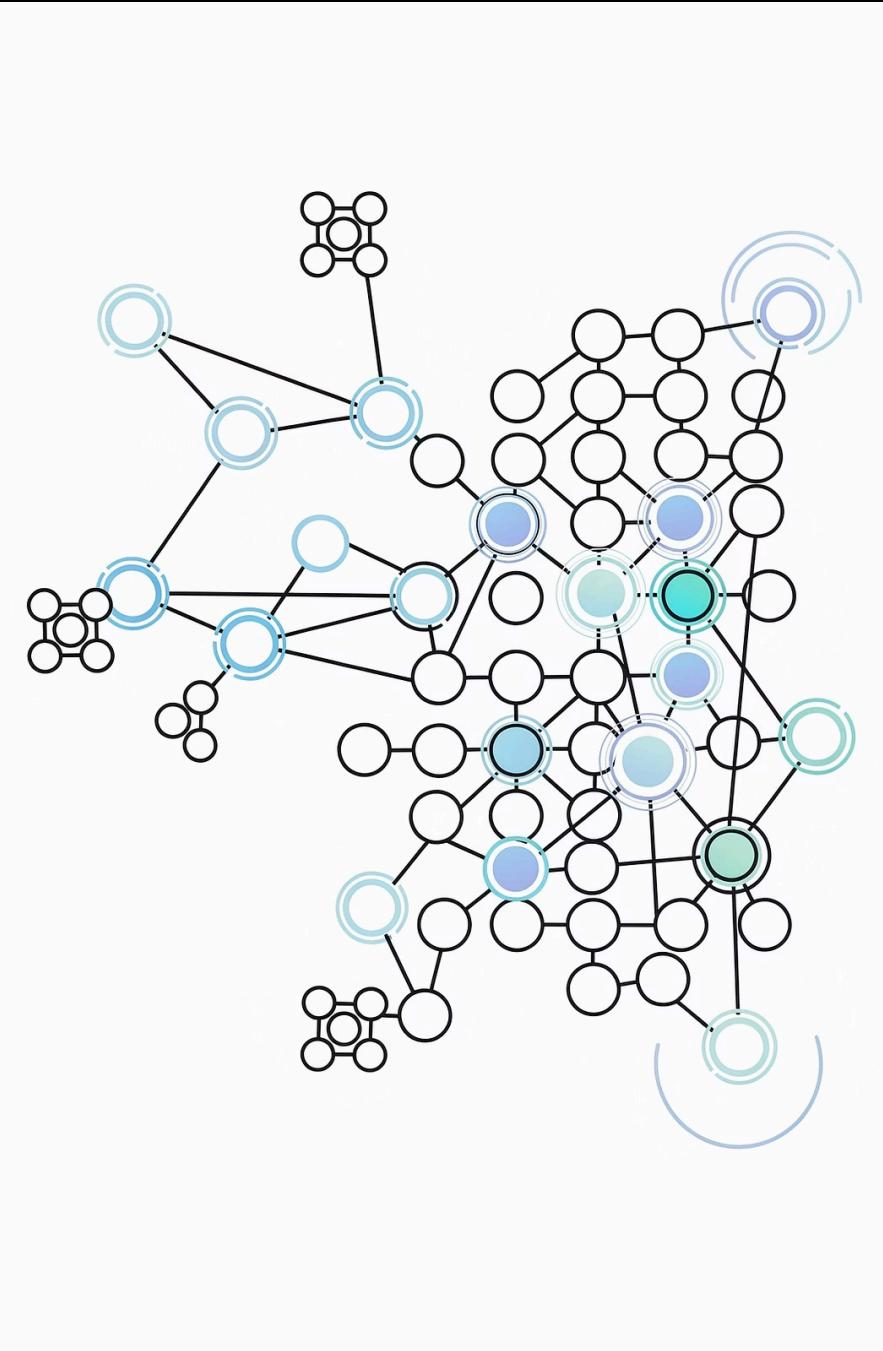
- **DIANA:** Abordagem divisiva (top-down)
- **AGNES:** Abordagem aglomerativa (bottom-up)
- **BIRCH, CAMELEON:** Métodos escaláveis



Abordagem Baseada em Densidade

Baseia-se em funções de conectividade e densidade para identificar clusters de formas arbitrárias. Especialmente útil para detectar outliers e clusters não-esféricos.

- **DBSCAN:** Density-Based Spatial Clustering
- **OPTICS:** Ordering Points To Identify Clustering
- **DENCLUE:** Baseado em estimativa de densidade



Clustering por Particionamento

Os algoritmos de particionamento representam uma das abordagens mais fundamentais em clustering. Eles dividem um conjunto de dados em k grupos distintos, onde cada objeto pertence ao cluster com o centroide mais próximo.

A ideia central é criar partções que minimizem a variação intra-cluster enquanto maximizam a separação entre clusters. Este processo iterativo continua até que um critério de convergência seja satisfeito.

O Método K-Means de Clustering

O K-Means é um dos algoritmos de clustering mais populares e amplamente utilizados devido à sua simplicidade e eficiência. Dado um valor k (número de clusters), o algoritmo funciona através de um processo iterativo bem definido.

- 1 Particionamento Inicial**
Dividir os objetos em k subconjuntos não vazios, geralmente de forma aleatória ou usando heurísticas específicas
- 2 Cálculo dos Centroides**
Computar os pontos sementes como os centroides dos clusters da partição atual (o centroide é o ponto médio do cluster)
- 3 Reatribuição de Objetos**
Atribuir cada objeto ao cluster com o ponto semente mais próximo, utilizando uma medida de distância apropriada
- 4 Iteração e Convergência**
Retornar ao Passo 2 e parar quando a atribuição não mudar mais, indicando convergência do algoritmo

Um Exemplo de Clustering K-Means

Vamos visualizar o processo do K-Means com k=2, demonstrando como o algoritmo evolui desde a partição inicial até a convergência final.



Conjunto de Dados Inicial

Pontos de dados distribuídos no espaço aguardando agrupamento

Particionamento Arbitrário

Objetos divididos arbitrariamente em k grupos não vazios

Atualização de Centroides

Cálculo do centroide (ponto médio) para cada partição

Reatribuição

Objetos reatribuídos ao cluster do centroide mais próximo

Convergência

Processo continua até que não haja mudanças nas atribuições

Comentários sobre o Método K-Means

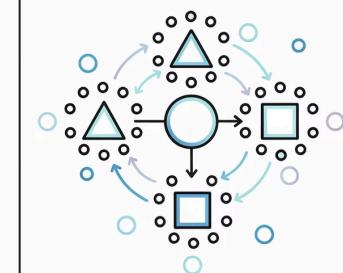
Vantagens do K-Means

- Algoritmo relativamente eficiente com complexidade $O(tkn)$, onde n é o número de objetos, k é o número de clusters e t é o número de iterações
- Implementação simples e intuitiva, facilitando sua aplicação em diversos contextos
- Geralmente converge rapidamente para um mínimo local
- Funciona bem quando os clusters têm forma aproximadamente esférica

Limitações e Desafios

- Necessita especificar k antecipadamente, o que nem sempre é conhecido
- Sensível à escolha dos centroides iniciais, podendo convergir para diferentes soluções
- Aplicável apenas quando a média é definida, limitando-se a dados numéricos
- Sensível a outliers e ruído nos dados
- Dificuldade em descobrir clusters com formas não convexas ou tamanhos muito diferentes

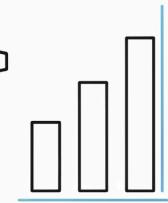
K-means Clustering



Advantages

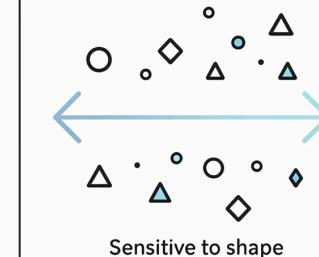


Efficient



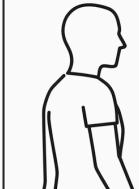
Scalable

Limitations



Sensitive to shape

Use with Care



Requires insight

Qual é o Problema do Método K-Means?

Média sem Significado

A média de um cluster pode não ter significado real ou interpretável no contexto dos dados. Por exemplo, em dados categóricos ou quando os atributos têm escalas muito diferentes, o centroide calculado pode não representar um objeto real ou significativo.

Considere um cluster com objetos muito dispersos: a média pode cair em uma região vazia do espaço de características, não representando adequadamente nenhum dos objetos do cluster.

Mínimo Local e Configuração Inicial

O algoritmo é altamente dependente da configuração inicial dos centroides. Diferentes inicializações podem levar a soluções completamente diferentes, todas representando mínimos locais da função objetivo.

Como mostra o exemplo visual, o mesmo conjunto de dados com diferentes pontos de partida pode resultar em agrupamentos distintos, cada um sendo um mínimo local válido, mas potencialmente não sendo a solução global ótima.

O Método K-Medoids de Clustering

O K-Medoids surge como uma alternativa robusta ao K-Means, especialmente adequada para situações onde a média não é apropriada ou quando há necessidade de maior resistência a outliers.

Conceito de Medoide

Em vez de usar o valor médio dos objetos em um cluster como ponto de referência, utiliza-se medoides - o objeto mais centralmente localizado no cluster. Este objeto é real, não uma construção matemática.

Busca por Representantes

O algoritmo busca objetos representativos (medoides) nos clusters através de um processo iterativo que avalia trocas potenciais entre medoides e não-medoides.

Processo Iterativo

Inicia com um conjunto de medoides e iterativamente substitui um dos medoides por um não-medoide se isso melhorar a distância total do clustering resultante.

O algoritmo PAM (Partitioning Around Medoids) funciona efetivamente para conjuntos de dados pequenos, mas não escala bem para grandes volumes devido à sua complexidade computacional $O(k(n-k)^2)$.

MÉTRICAS DE CLUSTER

Centroide, Raio e Diâmetro de um Cluster

Para conjuntos de dados numéricos, três medidas fundamentais caracterizam a geometria e dispersão dos clusters:



Centroide

O centroide C_m é o ponto médio de um cluster, calculado como a média de todos os pontos no cluster. É representado matematicamente como a média vetorial de todos os objetos pertencentes ao cluster m .



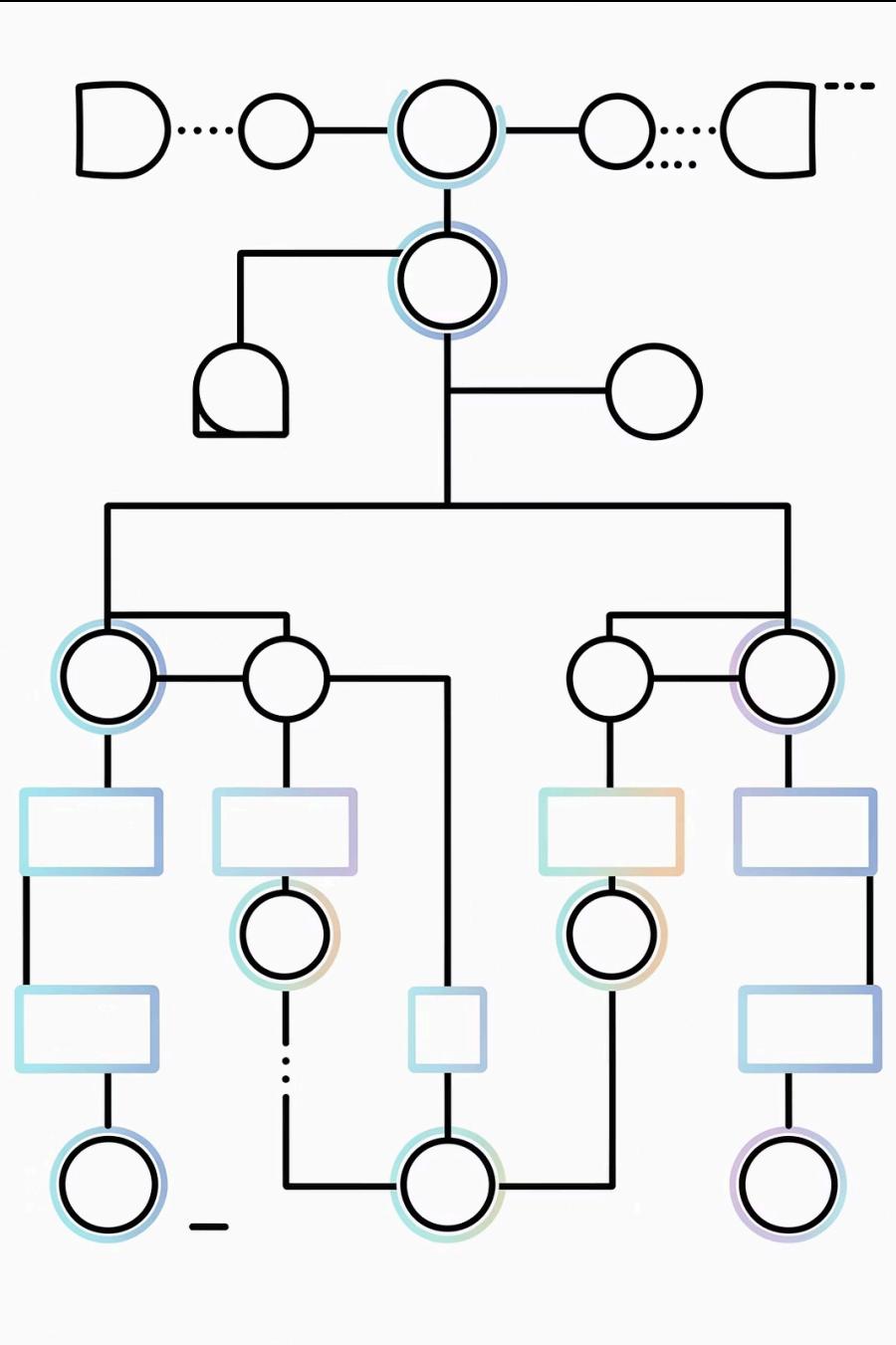
Raio

O raio R_m representa a dispersão média do cluster, calculado como a raiz quadrada da distância média ao quadrado de todos os objetos no cluster ao seu centroide. Menor raio indica cluster mais compacto.



Diâmetro

O diâmetro D_m é a raiz quadrada da distância média ao quadrado entre todos os pares de objetos no cluster. Mede a máxima extensão do cluster e sua coesão interna.



Clustering Hierárquico

Explorando métodos que criam estruturas hierárquicas de clusters, revelando relações em múltiplos níveis de granularidade

Distância entre Clusters

A escolha da medida de distância entre clusters é fundamental para o sucesso do clustering hierárquico. Diferentes métricas capturam diferentes aspectos da similaridade entre grupos.

Single Linkage (Vizinho Mais Próximo)

Distância mínima entre quaisquer dois pontos nos diferentes clusters. Tende a criar clusters alongados (efeito "corrente").

$$d_{min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p, q)$$

Complete Linkage (Vizinho Mais Distante)

Distância máxima entre quaisquer dois pontos nos diferentes clusters. Produz clusters mais compactos e balanceados.

$$d_{max}(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$$

Average Linkage (Ligaçāo Média)

Distância média entre todos os pares de pontos nos diferentes clusters. Oferece compromisso entre single e complete.

$$d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{p \in C_i} \sum_{q \in C_j} d(p, q)$$

Centroid Linkage (Ligaçāo por Centroide)

Distância entre os centroides dos clusters. Intuitiva mas pode sofrer com inversões no dendrograma.

$$d_{centroid}(C_i, C_j) = d(c_i, c_j)$$

Clustering Hierárquico

O clustering hierárquico utiliza uma matriz de distâncias como critério de agrupamento, construindo uma hierarquia de clusters através de fusões ou divisões sucessivas.

Características Principais

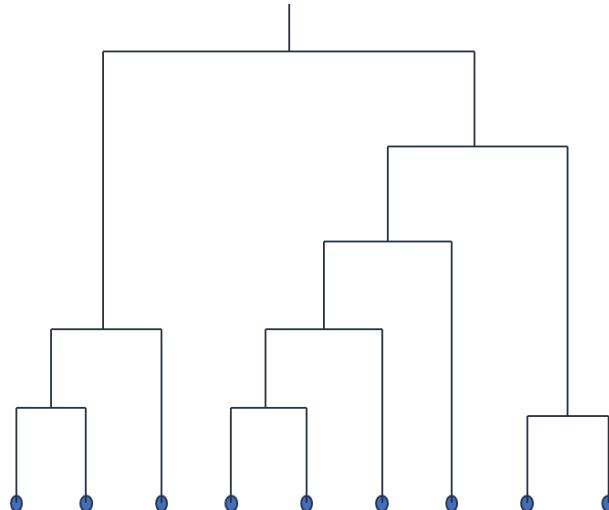
- **Não requer k predefinido:** Diferentemente do K-Means, não necessita especificar o número de clusters antecipadamente
- **Estrutura hierárquica:** Produz uma árvore de clusters (dendrograma) que captura relações em múltiplos níveis
- **Critério de terminação:** Necessita de uma condição de parada, como altura do corte no dendrograma ou número desejado de clusters
- **Flexibilidade:** Permite explorar diferentes níveis de granularidade no agrupamento

Abordagens

Aglomerativa (bottom-up): Inicia com cada objeto como um cluster e sucessivamente funde os clusters mais próximos.

Divisiva (top-down): Inicia com todos os objetos em um único cluster e sucessivamente divide em clusters menores.

Dendrograma: Visualizando a Fusão de Clusters



O dendrograma é uma representação visual em forma de árvore que mostra como os clusters são fundidos (ou divididos) em cada etapa do algoritmo hierárquico.



Níveis Aninhados

Decompõe os objetos de dados em vários níveis de particionamento aninhado, criando uma árvore de clusters com múltiplas camadas



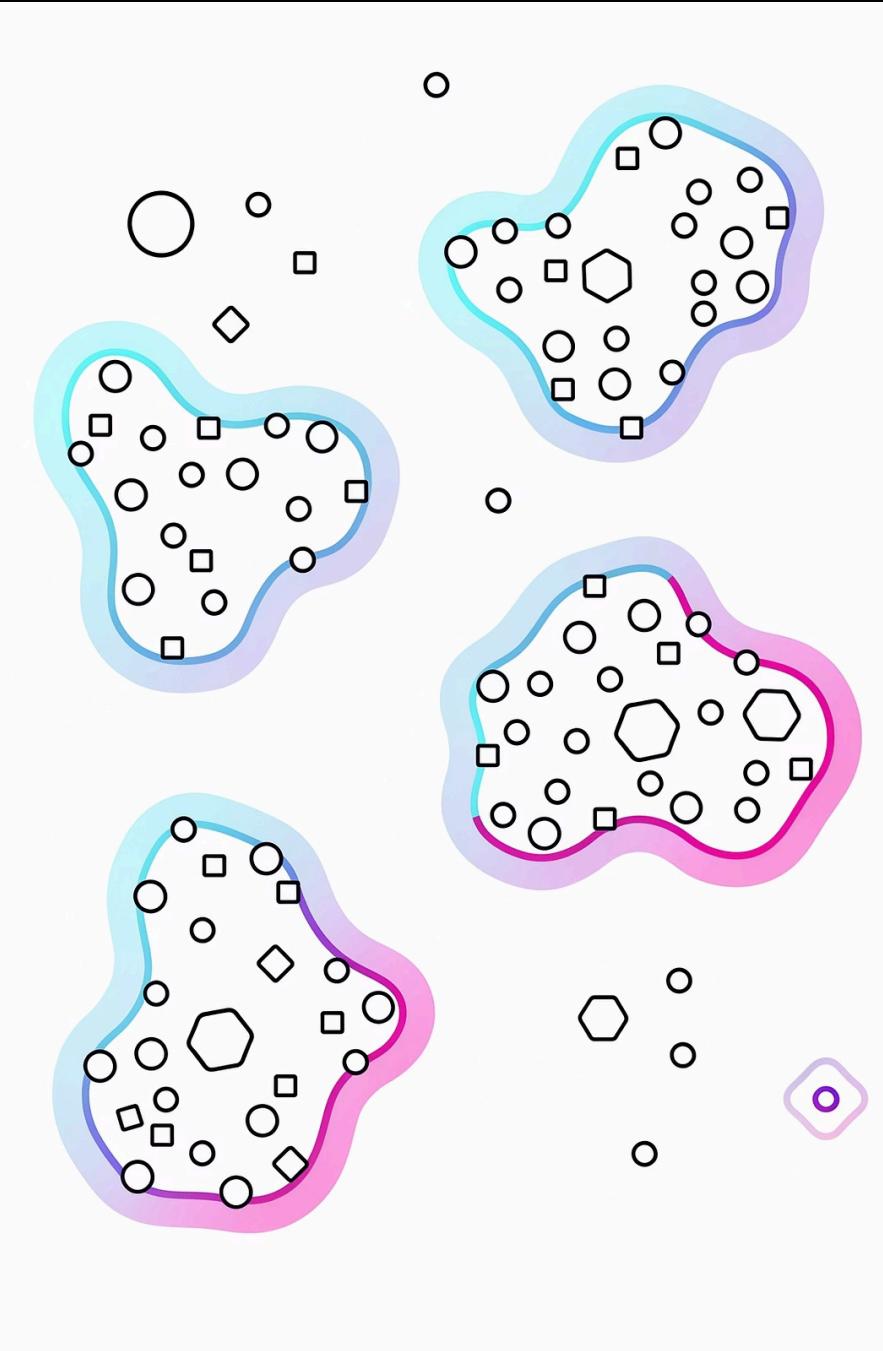
Corte no Dendrograma

Um clustering dos objetos é obtido cortando o dendrograma no nível desejado. A altura do corte determina o número e granularidade dos clusters



Componentes Conectados

Cada componente conectado após o corte forma um cluster independente, permitindo visualizar as relações de similaridade



Clustering Baseado em Densidade

Descobrindo clusters de formas arbitrárias através da análise de regiões densas no espaço de características

Métodos de Clustering Baseados em Densidade

Os métodos baseados em densidade identificam clusters como regiões de alta densidade separadas por regiões de baixa densidade. Essa abordagem oferece vantagens significativas sobre métodos tradicionais.

Clusters de Forma Arbitrária

Capacidade de descobrir clusters de qualquer forma geométrica, não se limitando a formas esféricas ou convexas como K-Means

Tratamento de Ruído

Identifica e isola pontos de ruído (outliers) que não pertencem a nenhum cluster denso, melhorando a qualidade dos resultados

Eficiência Computacional

Requer apenas uma varredura dos dados, tornando-o eficiente para grandes conjuntos de dados

Parâmetros de Densidade

Necessita de parâmetros de densidade como condição de terminação, definindo o que constitui uma região "densa"

Algoritmos Importantes

DBSCAN

Density-Based Spatial Clustering
of Applications with Noise

OPTICS

Ordering Points To Identify
Clustering Structure

DENCLUE

DENSity-based CLUStErIng

CLIQUE

CLustering In QUEst

Clustering Baseado em Densidade: Conceitos Básicos

Para entender os algoritmos baseados em densidade, precisamos compreender alguns conceitos fundamentais que definem quando pontos pertencem a regiões densas.



Parâmetros Principais

- **Eps (ϵ):** Raio máximo da vizinhança ao redor de um ponto.
No exemplo: Eps = 1 cm
- **MinPts:** Número mínimo de pontos necessários para formar uma região densa. No exemplo: MinPts = 5



Ponto Core (Núcleo)

Um ponto p é um **ponto core** se pelo menos MinPts pontos estão dentro da distância Eps de p (incluindo p). Esses pontos formam o centro de regiões densas.



Ponto Border (Fronteira)

Um ponto q é um **ponto border** se está dentro da distância Eps de um ponto core p, mas não tem MinPts em sua própria vizinhança Eps. Fica na periferia dos clusters.



Ponto Outlier (Ruído)

Um ponto que não é nem core nem border é considerado **ruído** ou **outlier**. Esses pontos não pertencem a nenhum cluster denso.

Density-Reachable e Density-Connected

Dois conceitos adicionais são essenciais para definir formalmente como pontos se relacionam em regiões densas e como clusters são formados.

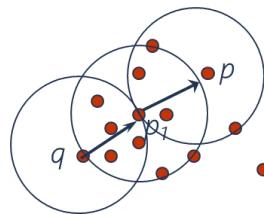
Density-Reachable (Alcançável por Densidade)

Um ponto p é **diretamente alcançável por densidade** a partir de q se:

- q é um ponto core
- p está dentro da Eps-vizinhança de q

Um ponto p é **alcançável por densidade** a partir de q se existe uma cadeia de pontos p_1, p_2, \dots, p_n onde $p_1 = q$ e $p_n = p$, e cada p_{i+1} é diretamente alcançável por densidade a partir de p_i .

Esta relação é assimétrica: p pode ser alcançável de q , mas q pode não ser alcançável de p .

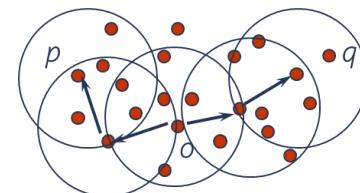


Density-Connected (Conectado por Densidade)

Dois pontos p e q são **conectados por densidade** se existe um ponto o tal que tanto p quanto q são alcançáveis por densidade a partir de o .

Esta relação é simétrica: se p é conectado por densidade a q , então q é conectado por densidade a p .

Pontos conectados por densidade formam a base para a definição de clusters em algoritmos como DBSCAN.



DBSCAN: Density-Based Spatial Clustering of Applications with Noise

O DBSCAN é um dos algoritmos de clustering baseado em densidade mais populares e influentes, introduzindo uma abordagem revolucionária para descoberta de clusters.

Noção de Cluster Baseada em Densidade

O DBSCAN define um cluster como um **conjunto maximal de pontos conectados por densidade**. Esta definição permite que o algoritmo descubra clusters de formas arbitrárias que não seriam detectados por métodos tradicionais baseados em distância.

Um cluster contém pelo menos um ponto core e todos os pontos alcançáveis por densidade a partir desse core. Pontos que não pertencem a nenhum cluster são classificados como ruído.

Descoberta de Formas Arbitrárias

Uma das principais vantagens do DBSCAN é sua capacidade de descobrir clusters de **formas arbitrárias** em bancos de dados espaciais, mesmo na presença de ruído significativo.

O algoritmo não assume nenhuma forma predefinida para os clusters (como esferas no K-Means), permitindo identificar estruturas complexas como clusters em forma de anel, serpentina ou qualquer outra geometria irregular.

"DBSCAN revolucionou o clustering ao demonstrar que a densidade local é um critério mais robusto do que a distância a centroides para muitas aplicações do mundo real."

IMPLEMENTAÇÃO

DBSCAN: O Algoritmo

O algoritmo DBSCAN opera através de um processo sistemático que examina cada ponto e expande clusters a partir de pontos core identificados.

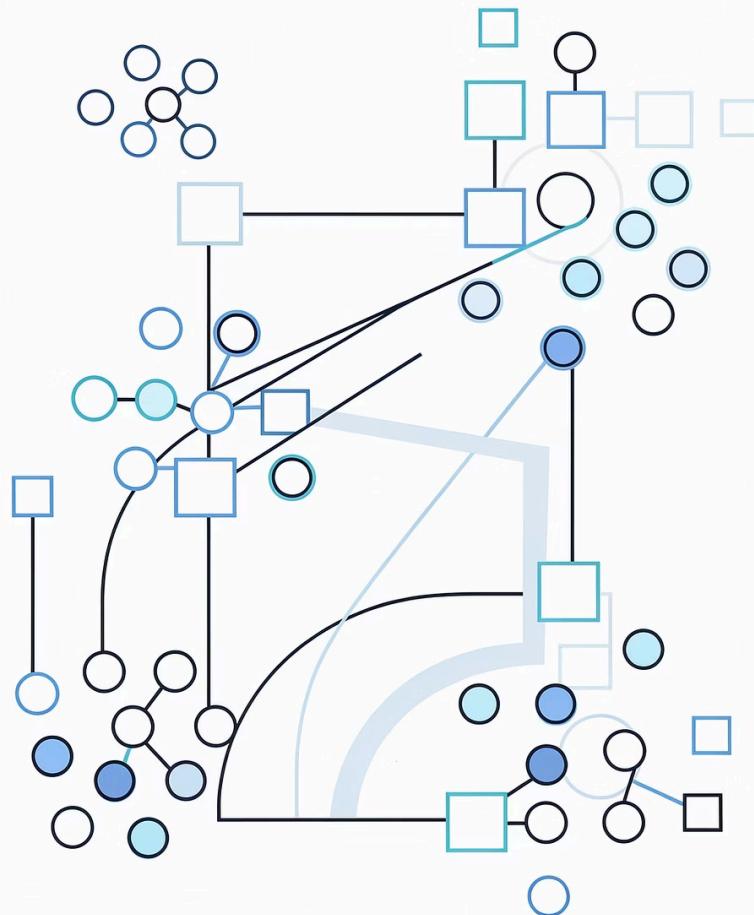
- 1 Inicialização**
Marcar todos os pontos como não visitados. Definir os parâmetros Eps e MinPts baseados no conhecimento do domínio ou análise exploratória.
- 2 Seleção de Ponto**
Selecionar um ponto arbitrário não visitado p do conjunto de dados e marcá-lo como visitado.
- 3 Verificação de Vizinhança**
Recuperar todos os pontos dentro da Eps-vizinhança de p . Se houver pelo menos MinPts pontos, p é um ponto core.
- 4 Expansão do Cluster**
Se p é core, criar um novo cluster e expandir recursivamente adicionando todos os pontos alcançáveis por densidade.
- 5 Classificação**
Se p não é core e não está na vizinhança de nenhum ponto core já processado, marcá-lo como ruído (pode ser reclassificado posteriormente).
- 6 Iteração**
Repetir os passos 2-5 até que todos os pontos tenham sido visitados e processados.

Complexidade e Eficiência

Complexidade temporal: $O(n \log n)$ com índices espaciais apropriados (como R*-tree), ou $O(n^2)$ no pior caso sem otimizações.

Complexidade espacial: $O(n)$ para armazenar os rótulos dos clusters e estruturas de dados auxiliares.

Vantagens práticas: Requer apenas uma varredura do banco de dados (com acesso eficiente à vizinhança), tornando-o escalável para grandes conjuntos de dados.



Avaliação e Escolha de Modelos de Clustering

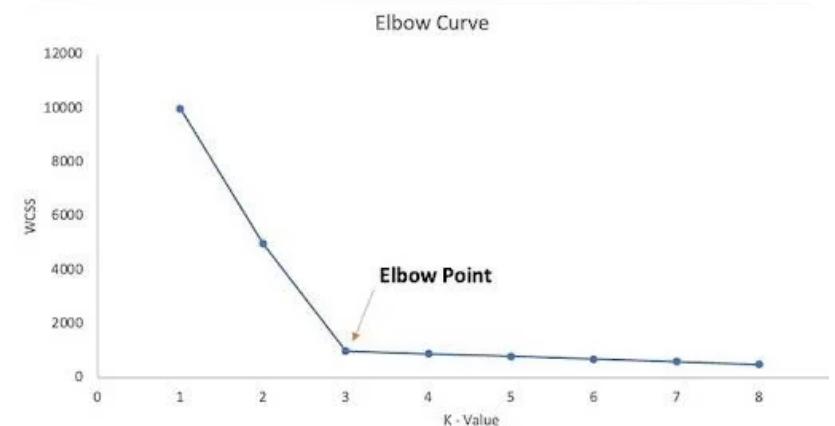
Explorando métricas, técnicas probabilísticas e estratégias avançadas para otimizar a qualidade e interpretação de agrupamentos em ciência de dados.

Determinando o Número Ideal de Clusters

Uma das decisões mais críticas em clustering é determinar quantos grupos são apropriados para seus dados. Esta escolha fundamental impacta diretamente a qualidade e interpretabilidade dos resultados.

Métodos populares incluem a análise do cotovelo (elbow method), índice de silhueta, e critérios estatísticos como BIC e AIC. Cada abordagem oferece perspectivas diferentes sobre a estrutura ótima de agrupamento.

A escolha do número de clusters deve equilibrar complexidade do modelo com capacidade de generalização, considerando também o contexto e objetivos do domínio de aplicação.



Medindo a Qualidade do Clustering

A avaliação de qualidade em clustering utiliza três paradigmas distintos, cada um com propósitos e aplicações específicas. Compreender estas abordagens é essencial para validar resultados de agrupamento.

Medidas Externas

Abordagem supervisionada que compara clustering contra conhecimento prévio ou ground truth.

- Entropia
- Normalized Mutual Information (NMI)
- Adjusted Rand Index

Medidas Internas

Abordagem não-supervisionada usando critérios derivados dos próprios dados.

- Coeficiente de Silhueta
- Índice Davies-Bouldin
- Compactação e separação

Medidas Relativas

Comparação direta entre diferentes clustering do mesmo algoritmo.

- Variação de parâmetros
- Análise comparativa
- Otimização iterativa

As medidas internas avaliam quão bem os clusters estão separados e quão compactos são internamente, enquanto medidas externas requerem labels verdadeiros para validação. Medidas relativas são úteis para fine-tuning de hiperparâmetros.

MÉTRICA EXTERNA

Entropia Condisional: Medida Baseada em Informação

A entropia condicional é uma métrica fundamental que quantifica a incerteza na classificação verdadeira dos dados dado o clustering obtido.

Conceito chave: Quanto menor a entropia condicional, melhor o clustering captura a estrutura real dos dados. Valores próximos a zero indicam alta concordância com o ground truth.

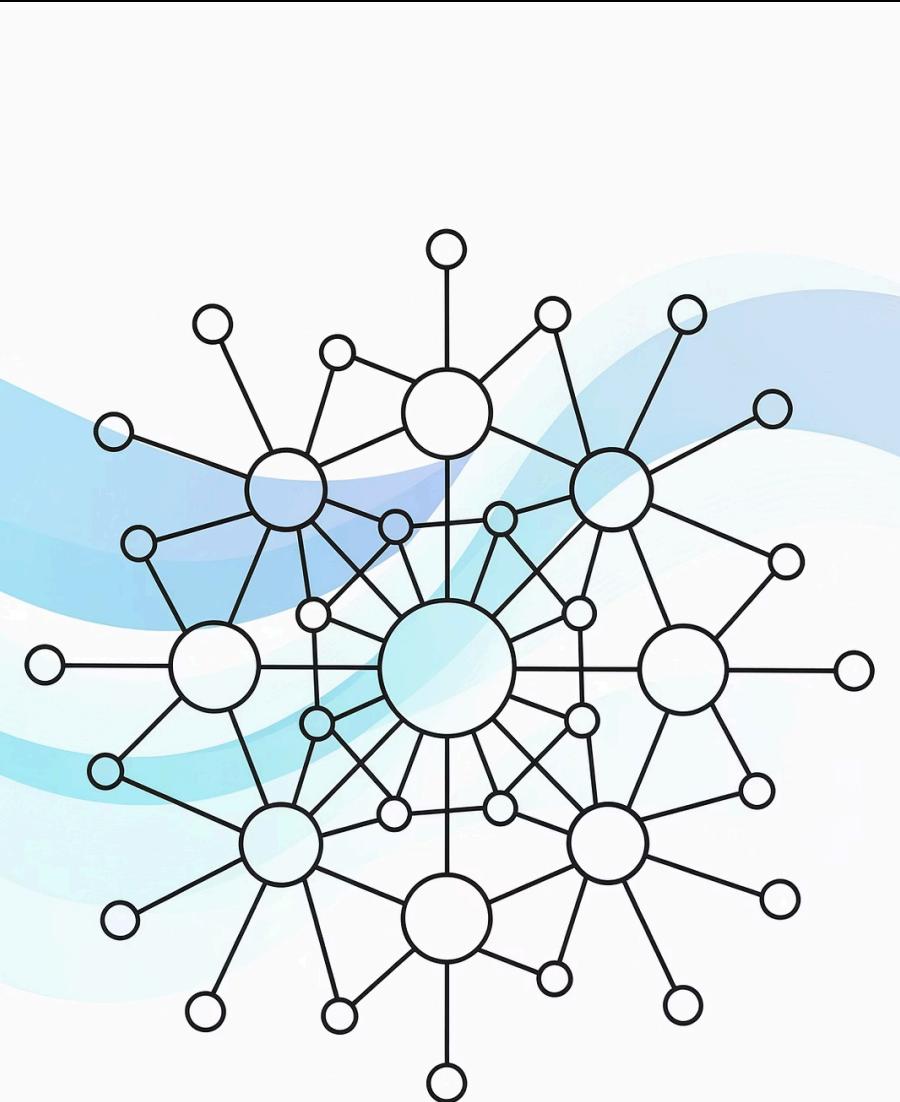
Esta medida é particularmente útil quando há labels conhecidos e desejamos avaliar o quanto bem nosso algoritmo recuperou a estrutura verdadeira latente nos dados.

Interpretação Matemática

$H(C|K)$ mede a entropia média das classes dentro de cada cluster, ponderada pela distribuição de objetos.

Aplicação Prática

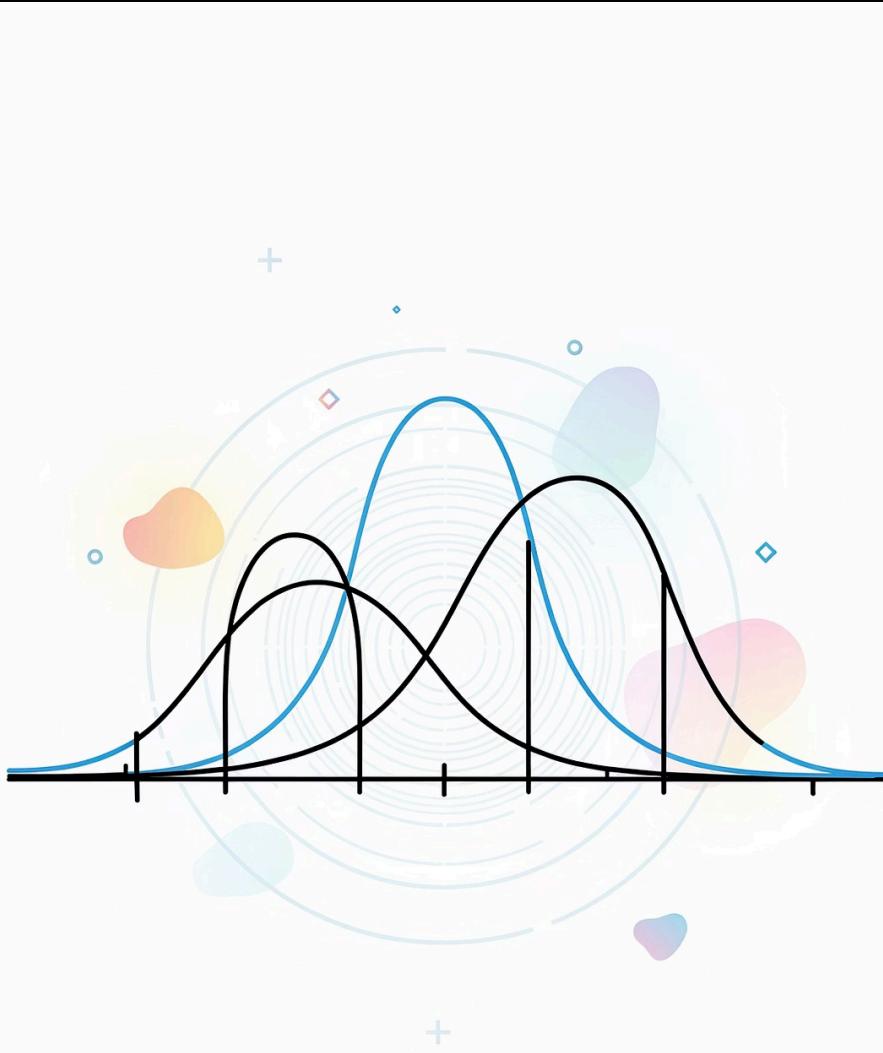
Permite comparar objetivamente diferentes algoritmos de clustering quando ground truth está disponível.



Tópicos Avançados

Paradigmas de Clustering Além de Métodos Baseados em Distância

Exploramos abordagens sofisticadas que superam as limitações de algoritmos tradicionais baseados apenas em distâncias euclidianas, introduzindo conceitos probabilísticos e modelagem estatística.



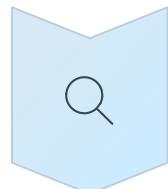
Clustering Probabilístico

Abordagem Baseada em Modelos

Uma mudança de paradigma que representa clusters como distribuições de probabilidade sobre o espaço de dados, proporcionando fundamentação matemática rigorosa e interpretabilidade estatística.

Clustering Probabilístico Baseado em Modelos

O clustering probabilístico adota uma perspectiva fundamentalmente diferente: clusters são categorias ocultas representadas por funções de densidade de probabilidade. Este framework matemático oferece maior flexibilidade e interpretabilidade.



Descobrir Categorias Ocultas

Identificar distribuições latentes que explicam a estrutura observada nos dados.



Representação Matemática

Cada cluster é uma função de densidade probabilística sobre o espaço de características.



Exemplo Aplicado

Segmentação de câmeras digitais: linha consumidor vs. linha profissional com densidades f_1 e f_2 .

Clusters probabilísticos permitem modelar a incerteza inerente à atribuição de objetos, reconhecendo que alguns pontos podem pertencer parcialmente a múltiplos grupos.

Conjuntos Fuzzy e Clustering Fuzzy

A teoria de conjuntos fuzzy estende a lógica binária tradicional, permitindo que elementos pertençam parcialmente a múltiplos conjuntos com graus de pertinência entre 0 e 1.

Diferença fundamental: Em clustering tradicional (hard), cada objeto pertence exclusivamente a um cluster. Em clustering fuzzy, cada objeto tem graus de pertinência para todos os clusters.



Hard Clustering

Pertinência = {0, 1}

Decisão binária e exclusiva



Fuzzy Clustering

Pertinência $\in [0, 1]$

Graus de pertinência contínuos

Esta abordagem é particularmente útil quando os limites entre clusters são naturalmente difusos ou quando objetos apresentam características de múltiplas categorias simultaneamente.

Clustering Fuzzy (Soft Clustering)

No clustering fuzzy, cada ponto de dados recebe um grau de pertinência para cada cluster, representando a probabilidade ou intensidade com que pertence àquele grupo.

Vantagens principais:

- Captura incerteza natural nos dados
- Permite transições suaves entre clusters
- Mais robusto a outliers e ruído
- Fornece informação rica sobre estrutura dos dados

O algoritmo Fuzzy C-Means (FCM) é a implementação mais popular, generalizando o K-means ao permitir pertinências parciais. O parâmetro de fuzziness (m) controla o grau de sobreposição entre clusters.

- ❑ A matriz de pertinência U fornece interpretação rica: cada linha representa um objeto e cada coluna um cluster, com valores indicando graus de associação.

Clustering Probabilístico: Modelos de Mistura

Um modelo de mistura (mixture model) assume que os dados observados são gerados por uma combinação de múltiplas distribuições probabilísticas subjacentes. Esta é uma abordagem poderosa e matematicamente elegante para clustering.

01

Premissa Fundamental

Cada objeto observado é gerado independentemente a partir de uma das k distribuições probabilísticas (clusters).

02

Processo Gerativo

Para cada objeto: (1) seleciona-se um cluster segundo probabilidades π_i , (2) gera-se o objeto segundo a distribuição daquele cluster.

03

Tarefa de Inferência

Dado o conjunto D de objetos observados, inferir os k clusters probabilísticos que maximizam a verossimilhança dos dados.

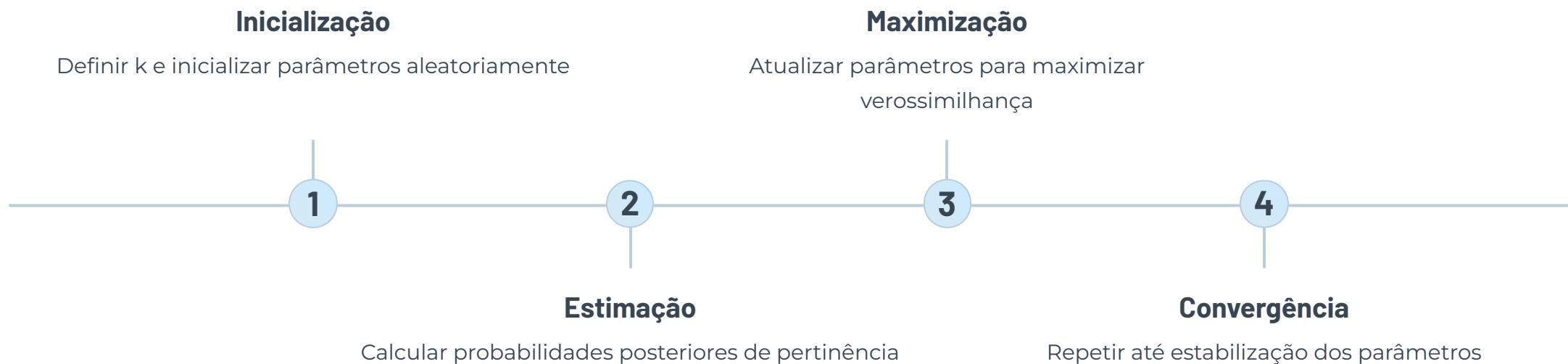
A beleza dos modelos de mistura está em sua fundamentação estatística: ao invés de apenas agrupar dados, inferimos um modelo gerativo que explica como os dados foram produzidos.

Clustering Baseado em Modelos: Fundamentos

O clustering baseado em modelos utiliza inferência estatística para descobrir a estrutura latente nos dados. A abordagem se fundamenta em maximizar a verossimilhança dos dados observados sob o modelo de mistura.

Componentes do Modelo

- **Parâmetros de mistura (π):** Probabilidades a priori de cada cluster
- **Parâmetros de distribuição (θ):** Caracterizam cada cluster (ex: μ e σ para Gaussianas)
- **Função de verossimilhança:** Probabilidade dos dados dado o modelo



Clustering Baseado em Modelos: Algoritmo EM

O algoritmo EM (Expectation-Maximization) é o método padrão para estimar parâmetros em modelos de mistura quando há variáveis latentes (as atribuições de cluster).

1

Passo E (Expectation)

Calcula as probabilidades posteriores de cada objeto pertencer a cada cluster, dados os parâmetros atuais do modelo. Estas são as "responsabilidades" de cada cluster por cada ponto.

Saída: Matriz de pertinências fuzzy para todos os objetos.

2

Passo M (Maximization)

Atualiza os parâmetros do modelo (médias, variâncias, pesos de mistura) para maximizar a verossimilhança esperada dados as pertinências calculadas no passo E.

Saída: Novos parâmetros ótimos dado o clustering soft atual.

- ❑ O algoritmo EM garante que a verossimilhança nunca diminui entre iterações, convergindo para um máximo local. Executar múltiplas vezes com inicializações diferentes ajuda a encontrar melhores soluções.

Modelo de Mistura Gaussiana Univariado

O Gaussian Mixture Model (GMM) é o modelo de mistura mais utilizado, assumindo que cada cluster segue uma distribuição normal (Gaussiana). Começamos com o caso univariado para ilustrar os conceitos fundamentais.

Componentes do Modelo

Cada cluster k é caracterizado por:

- μ_k : Média (centro do cluster)
- σ_k : Desvio padrão (dispersão)
- π_k : Peso da mistura (proporção)

A densidade de probabilidade para um ponto x é a soma ponderada das densidades Gaussianas individuais.

Interpretação intuitiva: Os dados são gerados ao selecionar aleatoriamente um dos k Gaussianos (com probabilidades π) e então amostrar um ponto daquela distribuição.

GMM Univariado: Estimação de Parâmetros

A estimação dos parâmetros de um GMM univariado via EM envolve fórmulas de atualização específicas que balanceiam as contribuições de cada ponto para cada cluster.

Passo E: Calcular Responsabilidades

- Para cada objeto x_i e cluster k , calcular $\gamma_{ik} = P(\text{cluster } k | x_i)$, a probabilidade posterior de x_i pertencer ao cluster k usando regra de Bayes.

Passo M: Atualizar Médias

- Nova média $\mu_k = \sum_i (\gamma_{ik} \cdot x_i) / \sum_i (\gamma_{ik})$. É a média ponderada dos pontos, onde os pesos são as responsabilidades.

Passo M: Atualizar Variâncias

- Nova variância $\sigma_k^2 = \sum_i (\gamma_{ik} \cdot (x_i - \mu_k)^2) / \sum_i (\gamma_{ik})$. Variância ponderada em torno da nova média.

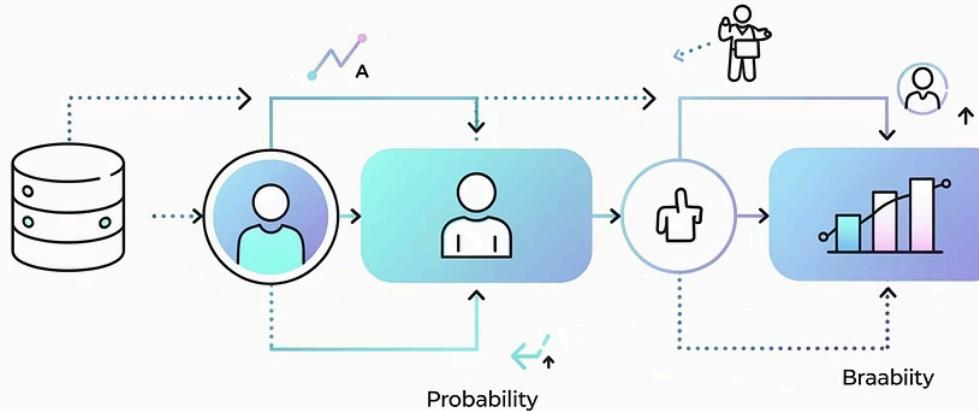
Passo M: Atualizar Pesos

- Novo peso $\pi_k = \sum_i (\gamma_{ik}) / n$, a proporção efetiva de pontos atribuídos ao cluster k .

O Algoritmo EM (Expectation-Maximization)

O EM é um framework geral para estimação de máxima verossimilhança quando há variáveis latentes. O K-means pode ser visto como um caso especial de EM com hard assignments.

Exreexoioity Step Prorjebarility

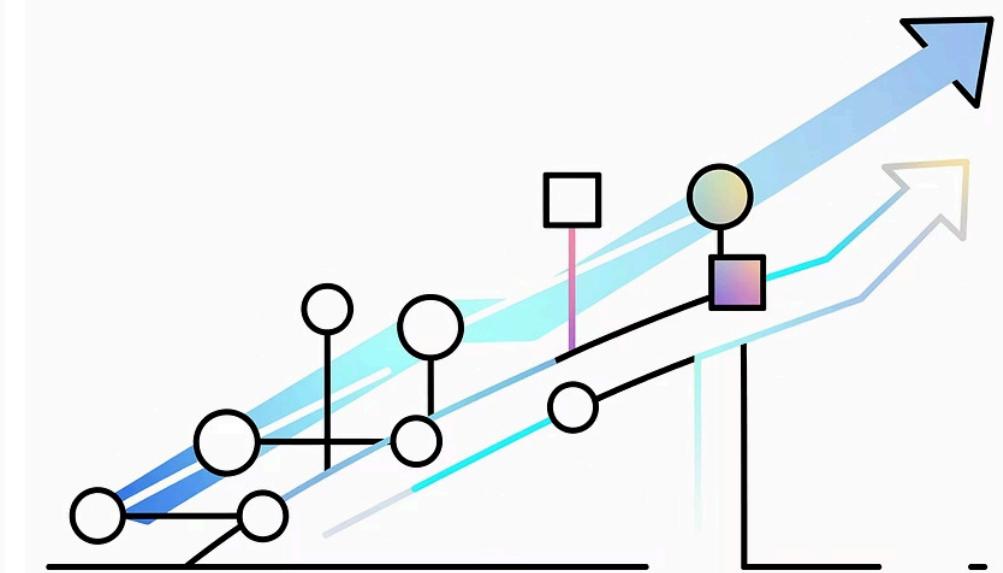


Passo E (Expectation)

Dado o clustering atual ou parâmetros, calcula-se a expectativa: cada objeto é atribuído a clusters com probabilidades que refletem quão provável é pertencer a cada um.

Paralelo com K-means

- **E-step do K-means:** Atribuição hard ao cluster mais próximo
- **M-step do K-means:** Recalcular centros como médias dos pontos atribuídos



Passo M (Maximization)

Dado o assignment probabilístico, encontra-se novos parâmetros que maximizam a verossimilhança esperada ou minimizam o erro quadrático esperado.

Generalização do EM

- **E-step do EM:** Atribuição soft com probabilidades
- **M-step do EM:** Maximizar verossimilhança esperada

□ O EM é amplamente aplicado além de clustering: modelos ocultos de Markov, análise factorial, e muitos outros contextos de inferência com variáveis latentes.

Modelos de Mistura: Vantagens e Limitações

Como toda técnica, modelos de mistura apresentam trade-offs importantes que devem ser considerados ao escolher esta abordagem para um problema específico de clustering.

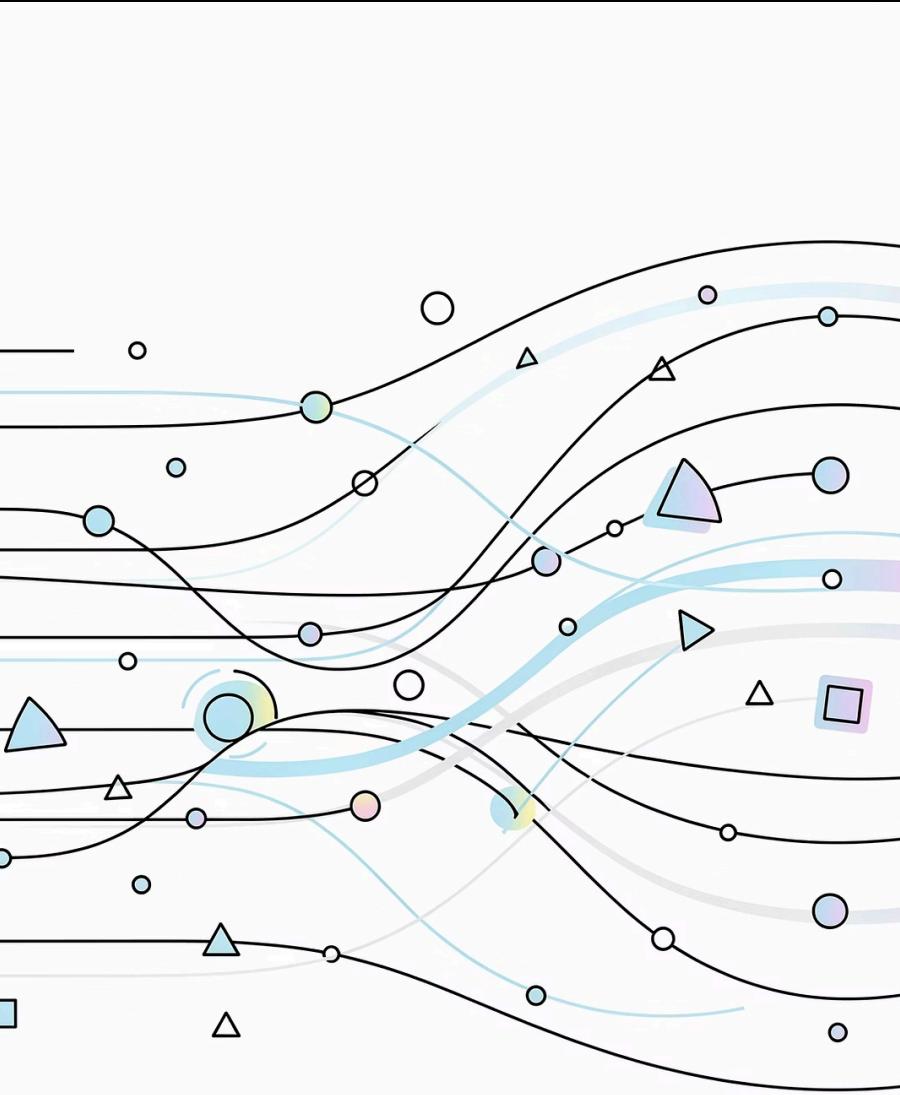
Pontos Fortes

- **Generalidade:** Mais flexíveis que particionamento e clustering fuzzy tradicional
- **Parcimônia:** Clusters caracterizados por poucos parâmetros interpretáveis
- **Fundação estatística:** Satisfazem pressupostos de modelos gerativos
- **Incerteza quantificada:** Probabilidades de pertinência têm interpretação clara

Desafios

- **Ótimos locais:** Convergência para máximos locais (mitigado com múltiplas inicializações)
- **Custo computacional:** Alto quando k é grande ou dados são escassos
- **Requisito de dados:** Necessita datasets grandes para estimativa confiável
- **Seleção de k:** Determinar número de clusters continua sendo desafiador

Recomendação prática: Modelos de mistura são ideais quando há justificativa teórica para distribuições Gaussianas e dados suficientes. Para problemas exploratórios com poucos dados, métodos mais simples podem ser preferíveis.



Clustering em Altas Dimensões e Subespaços

A análise de dados em espaços de alta dimensionalidade apresenta desafios únicos que exigem abordagens especializadas de clustering. Esta apresentação explora técnicas avançadas para descobrir estruturas ocultas em dados complexos e redes.

 DESAFIO FUNDAMENTAL

A Maldição da Dimensionalidade

Em espaços de alta dimensionalidade, os métodos tradicionais de clustering enfrentam obstáculos significativos que comprometem sua eficácia. O fenômeno conhecido como "maldição da dimensionalidade" manifesta-se através de múltiplos fatores que degradam o desempenho dos algoritmos.

Atributos irrelevantes ou altamente correlacionados introduzem ruído no processo de análise, mascarando padrões verdadeiros. A esparsidade dos dados aumenta exponencialmente com o número de dimensões, fazendo com que os pontos se tornem cada vez mais distantes uns dos outros.

O efeito de concentração de distâncias torna as medidas de similaridade praticamente inúteis, pois todos os pontos parecem igualmente distantes. Além disso, problemas de otimização tornam-se computacionalmente intratáveis à medida que a dimensionalidade cresce.

Esparsidade de Dados

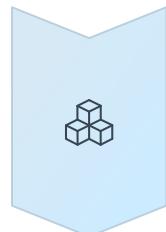
Distribuição dispersa no espaço

Concentração de Distâncias

Medidas de similaridade perdem significado

Modelos de Clustering em Alta Dimensionalidade

A descoberta de clusters em dados de alta dimensionalidade baseia-se em um princípio fundamental: clusters significativos frequentemente se manifestam em subespaços específicos, envolvendo apenas um subconjunto dramaticamente menor de atributos em relação ao espaço completo.



Clusters em Subespaços



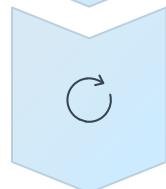
Estruturas ocultas reveladas através da análise de subconjuntos específicos de dimensões, onde padrões verdadeiros emergem claramente.



Subespaços Axis-Parallel



Identificação de clusters alinhados com os eixos originais dos atributos, simplificando interpretação e descoberta.



Subespaços Orientados Arbitrariamente



Detecção de estruturas em subespaços rotacionados, capturando padrões mais complexos e sofisticados nos dados.

- **Princípio Essencial:** O sucesso no clustering de alta dimensionalidade depende da capacidade de encontrar e explorar os subespaços corretos onde os clusters naturalmente existem.

Categorização de Métodos de Clustering em Alta Dimensionalidade

As técnicas para clustering em espaços de alta dimensionalidade podem ser organizadas em duas categorias principais, cada uma com suas próprias estratégias e abordagens especializadas.

1

Abordagens de Clustering

Identificam clusters em subespaços específicos através de três metodologias distintas:

- **Métodos de Subspace Clustering:** Encontram todos os clusters em todos os subespaços possíveis do espaço completo de dados
- **Métodos de Projected Clustering:** Particionam o conjunto de dados em subconjuntos não-sobrepostos, projetando-os em dimensões relevantes
- **Métodos de Bi-clustering:** Agrupam simultaneamente objetos e atributos, descobrindo estruturas bidimensionais

2

Métodos de Redução de Dimensionalidade

Transformam os dados de alta dimensionalidade em representações de menor dimensão, preservando as propriedades essenciais para clustering.

Estas técnicas incluem PCA (Principal Component Analysis), MDS (Multidimensional Scaling), e métodos de manifold learning que capturam a estrutura intrínseca dos dados.

Subspace Clustering: Métodos de Busca em Subespaços

Os métodos de subspace clustering exploram sistematicamente diferentes subespaços para descobrir clusters. Duas estratégias principais orientam essa busca, cada uma com suas vantagens computacionais e características específicas.

Abordagens Bottom-Up

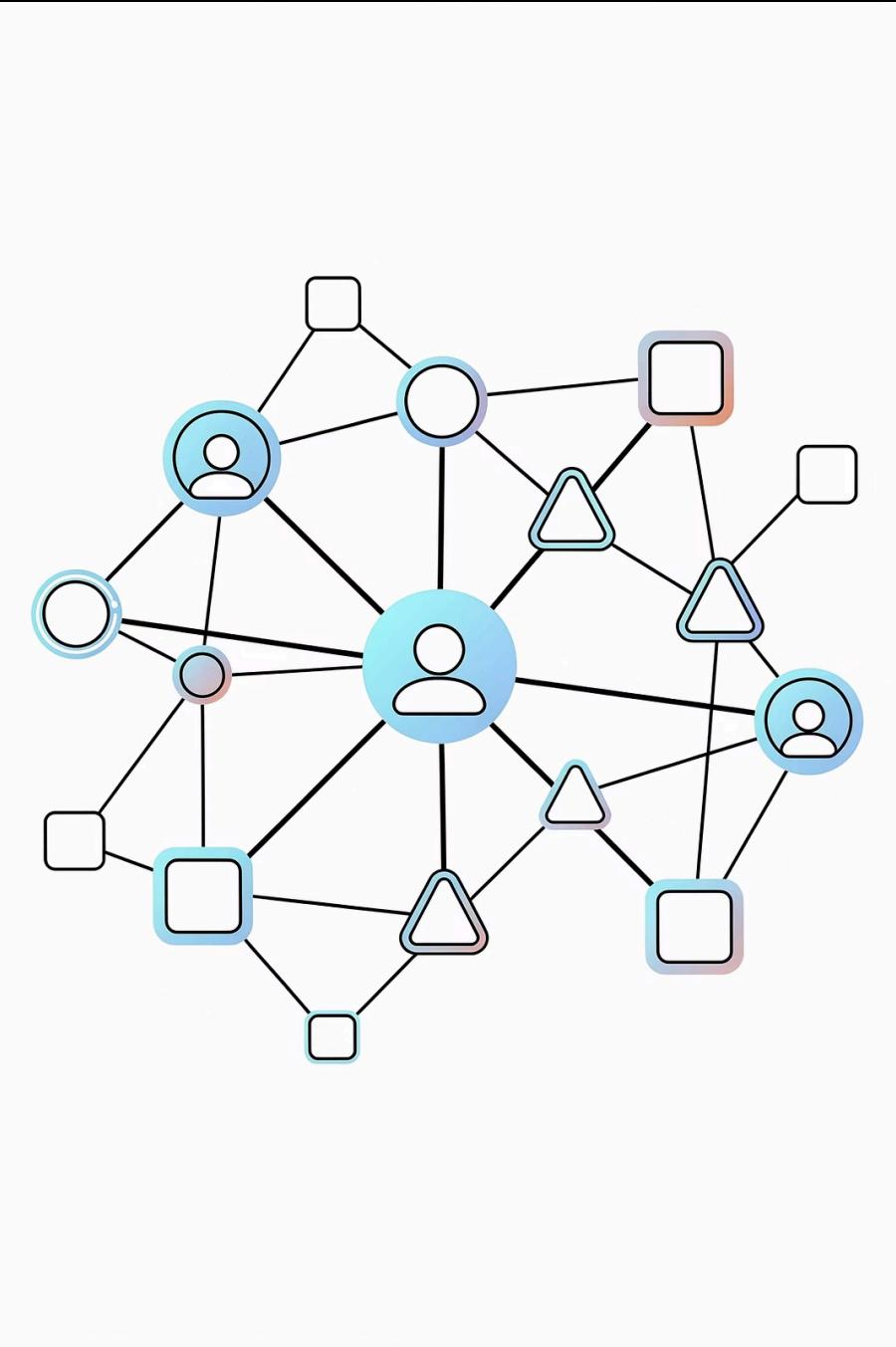
Iniciam a busca em subespaços de baixa dimensionalidade e progressivamente exploram subespaços de maior dimensão apenas quando há evidências de que clusters podem existir nessas dimensões superiores.

- Utilizam técnicas sofisticadas de poda para reduzir drasticamente o número de subespaços de alta dimensão a serem examinados
- Baseiam-se no princípio de anti-monotonicidade: se não há cluster denso em um subespaço, não haverá em seus superspaços
- **Exemplo clássico: CLIQUE** (Agrawal et al., 1998) - Pioneiro em clustering baseado em densidade para dados de alta dimensão

Abordagens Top-Down

Começam do espaço completo de atributos e recursivamente buscam subespaços menores, refinando progressivamente a seleção de dimensões relevantes.

- Efetivas principalmente quando a *suposição de localidade* é válida: o subespaço de um cluster pode ser determinado examinando a vizinhança local dos pontos
- Permitem descobrir diretamente os subespaços mais relevantes através de refinamento iterativo
- **Exemplo representativo: PROCLUS** (Aggarwal et al., 1999) - Método inspirado em k-medoid que projeta clusters em subespaços



Clustering de Grafos e Redes

A análise de estruturas em grafos e redes representa um domínio especializado de clustering com aplicações transformadoras em ciência de dados moderna. Estas técnicas revelam comunidades, padrões de conexão e estruturas latentes em dados relacionais complexos.

Clustering de Grafos e Dados de Rede

Aplicações



Grafos Bipartidos

Relacionamentos entre clientes e produtos, autores e conferências



Motores de Busca

Grafos de click-through e estrutura da Web



Redes Sociais

Grafos de amizade e colaboração científica

Métodos de Clustering em Grafos



Cortes Mínimos

FastModularity (Clauset, Newman & Moore, 2004): Otimiza modularidade para detectar comunidades através de particionamento eficiente do grafo

Medidas de Similaridade

Distâncias Geodésicas

Comprimento do caminho mais curto entre nós no grafo

Random Walk

SimRank: similaridade baseada em probabilidade de caminhadas aleatórias

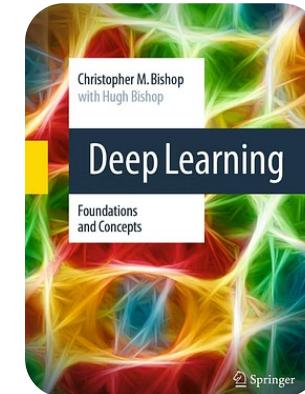
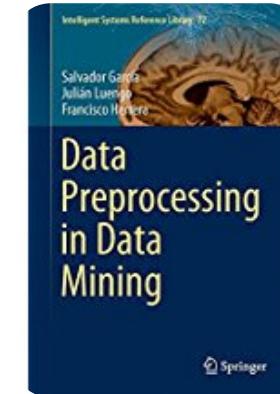
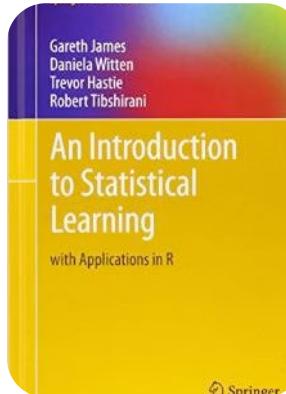
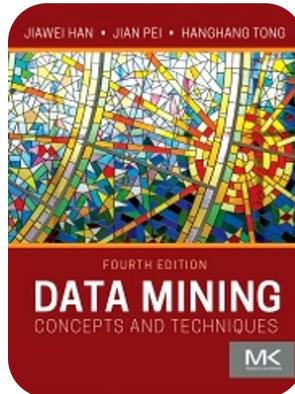


Clustering Baseado em Densidade

SCAN (Xu et al., KDD 2007): Structural Clustering Algorithm for Networks, identifica clusters, hubs e outliers em redes complexas

Referências Principais

Esta seleção de referências representa os pilares fundamentais para o estudo aprofundado de mineração de dados, cobrindo desde conceitos básicos até técnicas avançadas e aplicações contemporâneas.



1. **J. Han, J. Pei, and H. Tong**, *Data Mining: Concepts and Techniques*, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.
2. **G. M. James, D. Witten, T. Hastie, and R. Tibshirani**, *An Introduction to Statistical Learning: With Applications in R*. Springer Nature, 2021.
3. **S. Garcia, J. Luengo, and F. Herrera**, *Data Preprocessing in Data Mining*. Springer, 2014.
4. **C. M. Bishop and H. Bishop**, *Deep Learning: Foundations and Concepts*. Springer Nature, 2023.