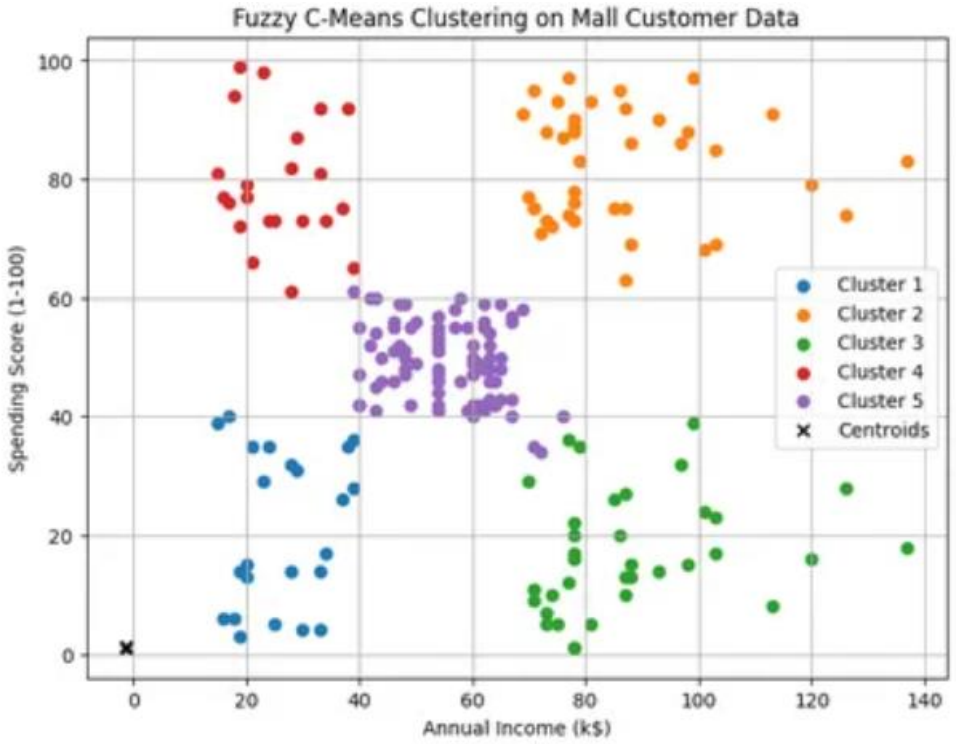
Clusterização Fuzzy

O que é Fuzzy C Means?

- As técnicas de clusterização fuzzy (ou clusterização difusa) são métodos de agrupamento que permitem que um dado pertença a mais de um cluster com diferentes graus de associação. A abordagem mais conhecida é o Fuzzy C-Means (FCM). Essa flexibilidade é útil em problemas onde não há fronteiras claras entre os grupos.



Como executar o algoritmo FCM

1. **Inicialização**: Escolha e inicialize aleatoriamente os centroides do cluster do conjunto de dados e especifique um parâmetro de imprecisão (m) para controlar o grau de imprecisão no cluster.
2. **Atualização de associação**: Calcule o grau de associação para cada ponto de dados para cada cluster com base em sua distância para os centroides do cluster usando uma métrica de distância (ex: distância euclidiana).
3. **Atualização de centroide**: Atualize o valor do centroide e recalcule os centroides do cluster com base nos valores de associação atualizados.
4. **Verificação de convergência**: Repita as etapas 2 e 3 até que um número especificado de iterações seja alcançado ou os valores de associação e centroides convirjam para valores estáveis.

Como executar o algoritmo FCM

- O objetivo é minimizar a função: $J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} ||\mathbf{x}_i - \mathbf{v}_j||^2$
- n = número de pontos de dados
- c = número de clusters
- x = ponto de dados ‘i’
- v = centroide do cluster ‘j’
- w = valor de associação do ponto de dados de i ao cluster j
- m = parâmetro de fuzziness (m>1)

Como executar o algoritmo FCM

- Atualize os valores de associacão usando a fórmula:

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||\mathbf{x}_i - \mathbf{v}_j||}{||\mathbf{x}_i - \mathbf{v}_k||} \right)^{\frac{2}{m-1}}}$$

- Atualizar valores do centroide do cluster usando uma média ponderada dos pontos de dados:

$$\mathbf{v}_j = \frac{\sum_{i=1}^{n} w_{ij}^m \cdot \mathbf{x}_i}{\sum_{i=1}^{n} w_{ij}^m}$$

- Continue atualizando os valores de associação e os centros de cluster até que os valores de associação e os centros de cluster parem de mudar significativamente ou quando um número predefinido de iterações for atingido.
- Atribua cada ponto de dados ao cluster ou a vários clusters para os quais ele tem o maior valor de associação.

Diferença entre FCM e K-Means

Fuzzy C Means	K-Means
Cada ponto de dados recebe um grau de associação a cada cluster, indicando a probabilidade ou verossimilhança do ponto pertencer a cada cluster.	Cada ponto de dados é atribuído exclusivamente a um e somente um cluster, com base no centroide mais próximo, normalmente determinado usando a distância euclidiana.
Não impõe nenhuma restrição à forma ou variância dos clusters. Ele pode lidar com clusters de diferentes formas e tamanhos, tornando-o mais flexível.	Assume que os clusters são esféricos e têm variância igual. Portanto, ele pode não ter um bom desempenho com clusters de formas não esféricas ou tamanhos variados.
É menos sensível a ruídos e valores discrepantes, pois permite atribuições de clusters suaves e probabilísticas.	É sensível a ruído e valores discrepantes nos dados

Aplicações

Segmentação de imagens: segmentação de imagens em regiões significativas com base em intensidades de pixels.

Reconhecimento de padrões: reconhecimento de padrões e estruturas em conjuntos de dados com relacionamentos complexos.

Imagem médica: análise de imagens médicas para identificar regiões de interesse ou anomalias.

Segmentação de clientes: segmentação de clientes com base em seu comportamento de compra.

Bioinformática: agrupamento de dados de expressão gênica para identificar genes co-expressos com funções semelhantes.

Vantagens e Desvantagens

Vantagens

- Robustez ao Ruído: FCM é menos sensível a outliers e ruído em comparação com algoritmos de agrupamento tradicionais.
- Atribuições Suaves: Fornece atribuições suaves e probabilísticas.
- Flexibilidade: Pode acomodar clusters sobrepostos e vários graus de associação de cluster.

Desvantagens

- Sensibilidade a Inicializações: O Desempenho é sensível ao posicionamento inicial dos centroides do cluster.
- Complexidade Computacional: A natureza iterativa do FCM pode aumentar a despesa computacional, especialmente para grandes conjuntos de dados.
- Seleção de Parâmetros: Escolher valores apropriados para parâmetros como o parâmetro de fuzziness (m) pode impactar a qualidade dos resultados do agrupamento.

Conclusão

Fuzzy C Means é um algoritmo de agrupamento muito diverso e bastante poderoso para descobrir significados ocultos (na forma de padrões) em dados, oferecendo flexibilidade no manuseio de conjuntos de dados complexos. Ele pode ser considerado um algoritmo melhor em comparação ao algoritmo k-means. Ao entender seus princípios, aplicações, vantagens e limitações, cientistas de dados e profissionais podem aproveitar esse algoritmo de clustering efetivamente para extrair insights valiosos de seus dados, tomando decisões bem informadas.

Referencias

ADITI V. **Understanding Fuzzy C Means Clustering**. Disponível em: <https://www.analyticsvidhya.com/blog/2024/05/understanding-fuzzy-c-means-clustering/#h-how-to-run-the-fcm-algorithm>. Acesso em: 22 jan. 2025.