MBA USP ESALQ

UNSUPERVISED MACHINE LEARNING: CLUSTERING

Prof. Dr. Wilson Tarantin Junior

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Contextualização

- Quando aplicar a análise de cluster?
  - **Quando o objetivo for agrupar as observações em grupos homogêneos internamente e heterogêneos entre si**
    - Dentro do grupo: observações semelhantes com base nas variáveis utilizadas na análise
    - Entre grupos distintos: observações diferentes com base nas variáveis utilizadas na análise

Contextualização

- Técnica exploratória (não supervisionada)
  - A análise de agrupamentos caracteriza-se por ser uma técnica exploratória, de modo que não tem caráter preditivo para observações de fora da amostra
  - Se novas observações forem adicionadas à amostra, novos agrupamentos devem ser realizados, pois a inclusão de novas observações pode alterar a composição dos grupos
  - Se forem alteradas variáveis da análise, novos agrupamentos devem ser realizados, pois a inclusão/retirada de uma variável pode alterar os grupos

Métodos

- Analisaremos dois métodos para a obtenção de agrupamentos
  - **Método Hierárquico Aglomerativo**
    - A quantidade de clusters é definida ao longo da análise (passo a passo)
  - **Método Não Hierárquico K-means**
    - Define-se a priori quantos cluster serão formados

Método Hierárquico Aglomerativo

Tratamento inicial dos dados

- Análise das variáveis que serão estudadas
  - Antes de iniciar os procedimentos, é importante realizar uma análise das unidades de medidas das variáveis
  - Se estiverem em unidades de medidas distintas, é importante realizar a padronização das variáveis antes de iniciar a análise de cluster
  - Aplica-se o ZScore (torna variáveis com média = 0 e desvio padrão = 1)

$$ZX_{ji} = \frac{X_{ji} - \overline{X}_j}{s_j}$$

Escolhas inerentes ao método

- A análise de cluster hierárquica depende de escolhas
  - Escolha da medida de dissimilaridade (distância)
    - Refere-se à distância entre as observações, com base nas variáveis escolhidas
    - Portanto, indica o quanto as observações são diferentes entre si
  - Escolha do método de encadeamento das observações
    - Refere-se à especificação da medida de distância quando houver cluster formados

Esquemas de aglomeração

- **Hierárquico aglomerativo: observações separadas → um único cluster**
  - Considerando *n* observações, inicia-se com *n* clusters (estágio 0)
  - Na sequência, une-se as duas observações com **menor distância** (n-1 clusters)
  - Em seguida, um novo grupo é formado pela união de duas novas observações ou pela inclusão de uma observação ao cluster formado na etapa anterior (**sempre pela menor distância**). **O método de encadeamento indica qual é a distância a ser considerada**
  - Repete-se a etapa anterior n-1 vezes, ou seja, até restar somente 1 cluster
  - O **dendrograma** é um gráfico que permite visualizar a formação dos clusters

Medidas de dissimilaridade

- Identifica a distância entre observações
  - Distância de Minkowski: $\boldsymbol{d_{pq} = \left[\sum_{j=1}^{k}(|ZX_{jp} - ZX_{jq}|)^{m}\right]^{\frac{1}{m}}}$
  - Distância euclidiana: $\boldsymbol{d_{pq} = \sqrt{\sum_{j=1}^{k}(ZX_{jp} - ZX_{jq})^{2}}}$
  - Distância euclidiana quadrática: $\boldsymbol{d_{pq} = \sum_{j=1}^{k}(ZX_{jp} - ZX_{jq})^{2}}$

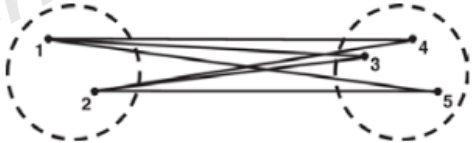
Medidas de dissimilaridade

- Identifica a distância entre observações
  - Distância de Manhattan (City Block): $\boldsymbol{d_{pq} = \sum_{j=1}^{k} |ZX_{jp} - ZX_{jq}|}$
  - Distância de Chebychev: $\boldsymbol{d_{pq} = max\ |ZX_{jp} - ZX_{jq}|}$
  - Distância de Canberra: $\boldsymbol{d_{pq} = \sum_{j=1}^{k} \frac{|ZX_{jp} - ZX_{jq}|}{(ZX_{jp} + ZX_{jq})}}$ → variáveis de valores positivos
  - A correlação de Pearson entre as observações também pode ser utilizada (mas é uma medida de semelhança, portanto ajusta-se sua interpretação)

Métodos de encadeamento

- Esquemas hierárquicos aglomerativos
  - Método de encadeamento: indica qual distância utilizar quando já existem clusters formados durante os estágios aglomerativos
    - *Nearest neighbor* (*single linkage*): privilegia menores distâncias, recomendável em casos de observações distintas
    - *Furthest neigbor* (*complete linkage*): privilegia maiores distâncias, recomendável em casos de observações parecidas
    - *Between groups* (*average linkage*): junção de grupos pela distância média entre todos os pares de observações do grupo em análise

Métodos de encadeamento

Método de Encadeamento	Ilustração	Distância (Dissimilaridade)
Único *(Nearest Neighbor ou Single Linkage)*		d_{23}
Completo *(Furthest Neighbor ou Complete Linkage)*		d_{15}
Médio *(Between Groups ou Average Linkage)*		$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$

Fonte: Fávero & Belfiore (2024, Capítulo 9)

Métodos de encadeamento

- Esquemas hierárquicos aglomerativos
  - *Nearest neighbor* (vizinho mais próximo): *single linkage*
    - **d(MN)W = mín{dMW ; dNW}**
  - *Furthest neighbor* (vizinho mais distante): *complete linkage*
    - **d(MN)W = máx{dMW ; dNW}**
  - *Between groups* (média das distâncias): *average linkage*
    - **d(MN)W = média entre dMW e dNW (distância média entre todos os pares)**

Quantos clusters escolher?

- Esquemas hierárquicos aglomerativos
  - Como critério para a escolha do número final de clusters em uma análise, pode-se adotar o tamanho dos saltos de distância para a incorporação seguinte
  - Saltos muito elevados podem indicar o agrupamento de observações com características mais distintas, isto é, há a união de observações mais distintas
  - Comparar dendrogramas obtidos por diferentes métodos de encadeamento

Análise dos agrupamentos

- Quais variáveis contribuem?
  - Após a clusterização, é importante comparar se a variabilidade dentro do grupo é menor do que a variabilidade entre grupos com base nas variáveis da análise
  - Aplica-se um teste F para análise de variância: $\boldsymbol{F} = \frac{\boldsymbol{Variabilidade\ entre\ grupos}}{\boldsymbol{Variabilidade\ dentro\ dos\ grupos}}$
    - Graus de liberdade no numerador: K – 1
    - Graus de liberdade no denominador: n – K

    K = nº de clusters
    n = tamanho da amostra
  - É possível analisar quais variáveis mais contribuíram para a formação de pelo menos um dos clusters: maiores valores da estatística F (em conjunto com a significância)

Método Não Hierárquico K-means

Tratamento inicial dos dados

- Análise das variáveis que serão estudadas
  - Também é importante realizar a análise das unidades de medidas das variáveis para a aplicação do K-means
  - Se estiverem em unidades de medidas distintas, é fundamental padronizar as variáveis antes de iniciar a análise
  - Padronização pelo ZScore (variáveis com média = 0 e desvio padrão = 1)

$$ZX_{ji} = \frac{X_{ji} - \overline{X}_j}{s_j}$$

Esquemas de aglomeração

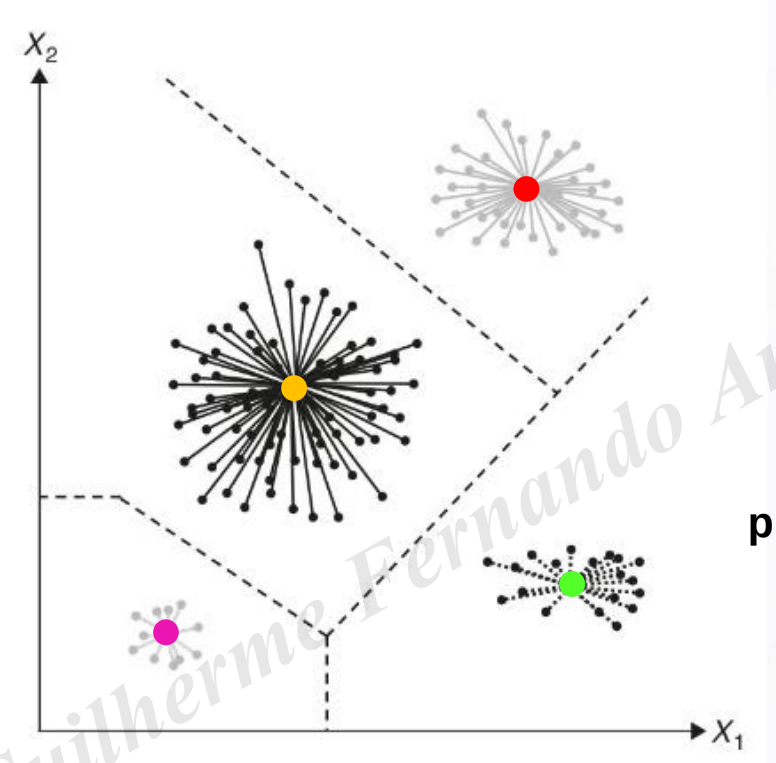
- Esquema não hierárquico K-means
  - A quantidade K de clusters é escolhida a priori e é usada como base para a identificação dos centros de aglomeração, de modo que as observações são arbitrariamente alocadas aos K clusters para o cálculo dos centroides iniciais
  - Nas etapas seguintes, as observações vão sendo comparadas pela proximidade aos centroides dos outros clusters. Se houver realocação a outro cluster por estar mais próxima, os centroides são recalculados (em ambos os clusters)
    - Trata-se de um processo iterativo

Esquemas de aglomeração

- Esquema não hierárquico K-means
  - O procedimento K-means encerra-se quando não for possível realocar qualquer observação por estar mais próxima do centroide de outro cluster: indica que a soma dos quadrados de cada observação até o centro do cluster alocada foi minimizada
  - A soma total dos quadrados dentro dos clusters pode ser representada por:

$$SS = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Esquemas de aglomeração



Fonte: Fávero & Belfiore (2024, Capítulo 9)

Representa a solução do K-*means*

Não ocorrerão outras realocações, pois não há observações que estejam mais próximas dos centroides de outros clusters

Identificação da quantidade de clusters

- Técnicas para a identificação da quantidade de clusters no K-means

  - **Método de Elbow**: calcula-se a soma total dos quadrados dentro dos clusters (WCSS) para várias opções de K (quantidade de clusters). No gráfico, busca-se a dobra (“cotovelo”), ou seja, o ponto a partir do qual a diminuição na WCSS não é mais tão expressiva, mesmo aumentando a quantidade de clusters

  - **Método da Silhueta**: para cada observação, calcula-se: **(b)** sua distância média para o cluster mais próximo onde não esteja alocada; **(a)** sua distância média dentro do cluster onde está alocada

$$\textbf{silhueta} = \frac{(\mathbf{b}-\mathbf{a})}{\mathbf{max}(\mathbf{a},\mathbf{b})}$$

Quanto mais próximo de 1, melhor a clusterização. Quanto mais próximo de -1, pior!

  - Em seguida, calcula-se o **coeficiente de silhueta médio** para todas as observações. O procedimento é realizado para várias opções de K

Considerações

- Alguns aspectos relevantes
  - A análise de cluster é bastante sensível à presença de outliers
  - Quando há variáveis categóricas, pode aplicar a Análise de Correspondência
  - O output do método hierárquico pode ser utilizado como input no método não hierárquico para a identificação inicial da quantidade de clusters
  - O método não hierárquico k-means pode ser aplicado em amostras maiores

Referência

Fávero, Luiz Paulo; Belfiore, Patrícia. (2024). Manual de análise de dados: estatística e machine learning com Excel®, SPSS®, Stata®, R® e Python®. 2 ed. Rio de Janeiro: LTC.

OBRIGADO!

linkedin.com/in/wilson-tarantin-junior-359476190