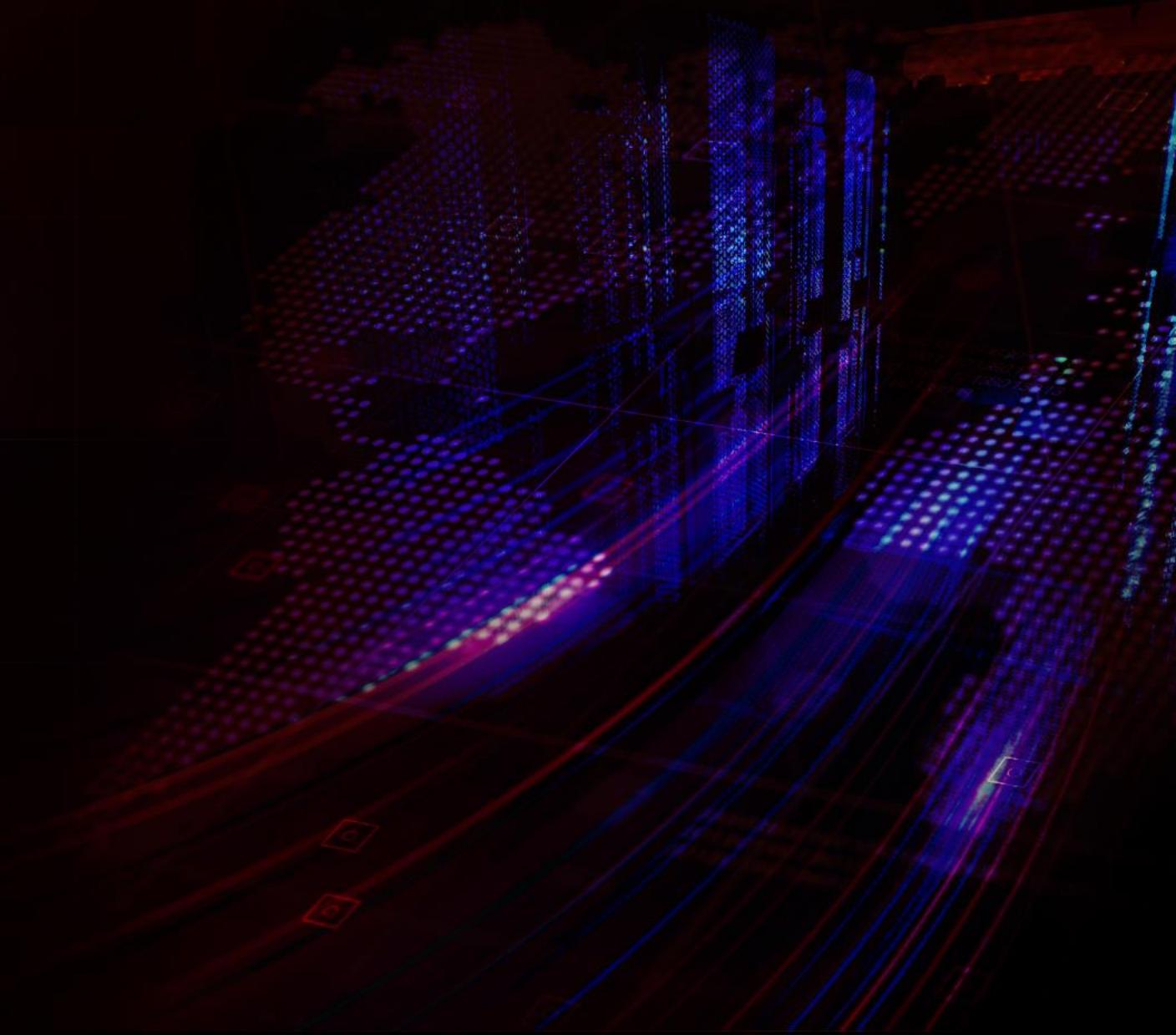


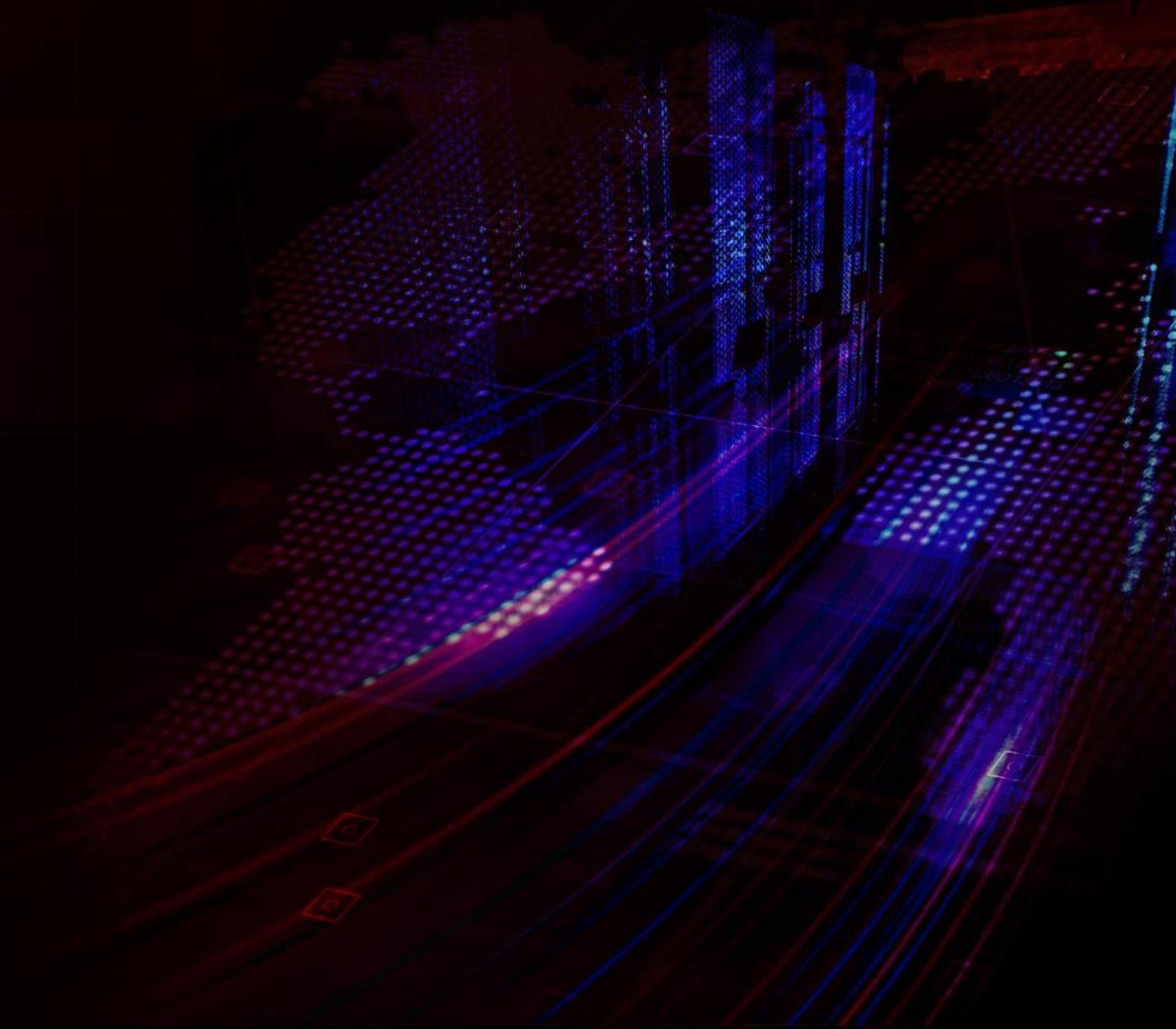
APRENDIZADO DE MÁQUINAS

Avaliação de Modelos Descritivos



TÓPICOS

- Objetivos de validação
- Critérios de validação
- Índices de validação



MODELOS DESCRIPTIVOS

Não existe resultado correto

Não existe um único algoritmo melhor em todos os casos

Características do conjunto de dados

Características do problema

Parâmetros dos algoritmos



OBJETIVOS DE VALIDAÇÃO

Avaliação e comparação de algoritmos

Validação das estruturas encontradas



CRITÉRIOS DE VALIDAÇÃO

Relativos

Internos

Externos



CRITÉRIOS DE VALIDAÇÃO

Relativos

Internos

Externos

Diversos índices para avaliar / comparar agrupamentos

CRITÉRIOS DE VALIDAÇÃO

Relativos

Internos

Externos

Comparam diversos agrupamentos com algum aspecto
Ex: estabilidade, número de clusters

Comparar algoritmos diferentes

Comparar parâmetros do mesmo algoritmo

CRITÉRIOS DE VALIDAÇÃO

Relativos

Internos

Externos

Dados originais

Ex: matriz de similaridade



CRITÉRIOS DE VALIDAÇÃO

Relativos

Internos

Externos

Estrutura previamente estabelecida

Ex: agrupamento sugerido por um especialista, número de clusters

Testes de hipóteses

NOTAÇÃO

- C : clusters
- π : partição
- x_i : objeto
- n : conjunto de objetos
- $d(x_1, x_2)$: distância entre x_1 e x_2
- \bar{x} : centróide de um cluster
- k : cluster (instância)

CRITÉRIOS RELATIVOS

Variância intracluster: $\text{var}(\pi)$

Meude a compactação dos clusters

$$var(\pi) = \sqrt{\frac{1}{n} \sum_{C_k \in \pi} \sum_{x_i \in C_k} d(x_i, \bar{x}^{(k)})}$$

0.0 

CRITÉRIOS RELATIVOS

Conectividade: $\text{con}(\pi)$

Grau com que os vizinhos pertencem ao mesmo cluster

$$\text{con}(\pi) = \sum_{x_i \in n} \sum_{j=1}^v f(x_i, nn_{ij})$$

$$f(x_i, nn_{ij}) = \begin{cases} 1/j & \text{se } x_i \in C_k, nn_{ij} \notin C_k \\ 0 & \text{caso contrário} \end{cases}$$

nn_{ij} : j-ésimo vizinho

v: vizinhos

melhor 0.0

∞ pior

CRITÉRIOS RELATIVOS

Família de Índices Dunn: $D(\pi)$

Mede a razão de separação intracluster e intercluster

$$D(\pi) = \min_{a=1,\dots,k} \left\{ \min_{b=a+1,\dots,k} \left\{ \frac{d(C_a, C_b)}{\max_{l=1,\dots,k} d(C_l)} \right\} \right\}$$

$$d(C_a, C_b) = \min_{x_i \in C_a, x_j \in C_b} d(x_i, x_j)$$

$$d(C_a) = \max_{x_i, x_j \in C_a} d(x_i, x_j)$$

pior 0.0

∞ melhor

CRITÉRIOS RELATIVOS

Silhueta: $sil(\pi)$

Mede a proximidade intracluster e a distância do cluster mais próximo

$$sil(x_i) = \begin{cases} 1 - a(x_i, C_i) / b(x_i), \\ 0, \\ b(x_i) / a(x_i, C_i) - 1, \end{cases}$$

$$a(x_i, C_k) = \frac{1}{|C_k|} \sum_{\substack{x_i, x_j \in C_k \\ x_i \neq x_j}} d(x_i, x_j)$$

$$b(x_i) = \min_{\substack{x_i \in C_i, \\ C_i \neq C_j}} a(x_i, C_j)$$

pior -1.0

1.0 melhor

CRITÉRIOS INTERNOS

Estatística Gap

Mede a dispersão intracluster em relação à sua esperança (E^*) sob uma distribuição de referência nula

$$Gap_n(k) = E_n^* \{ \log(W_k) \} - \log(W_k)$$

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} D_r$$

$$D_r = \sum_{x_i, x_j \in C_r} d(x_i, x_j)$$

D: dispersão intracluster, W: peso dos k clusters



CRITÉRIOS EXTERNOS

Notação:

π^c : partição resultante do algoritmo

π^r : partição conhecida *a priori*

(x_i, x_j) : par de objetos

a1: $(x_i, x_j) \in C_k$ de π^c e $(x_i, x_j) \in C_k$ de π^r

a2: $(x_i, x_j) \in C_k$ de π^c e $x_i \in C_k$, $x_j \in C_l$ de π^r

a3: $x_i \in C_k$, $x_j \in C_l$ de π^c e $(x_i, x_j) \in C_k$ de π^r

a4: $x_i \in C_k$, $x_j \in C_l$ de π^c e $x_i \in C_l$, $x_j \in C_k$ de π^r



CRITÉRIOS EXTERNOS

Índice Rand

Probabilidade de dois objetos pertencerem ao mesmo cluster ou a clusters diferentes em π^c e π^r

$$Rand(\pi^c, \pi^r) = \frac{(a_1 + a_4)}{a_1 + a_2 + a_3 + a_4}$$

0.0
pior

1.0
melhor

CRITÉRIOS EXTERNOS

Índice Jaccard

Probabilidade de dois objetos pertencentes ao mesmo cluster
em π^c e também em π^r

$$Jaccard(\pi^c, \pi^r) = \frac{a_1}{a_1 + a_2 + a_3}$$

0.0
pior

1.0
melhor

CRITÉRIOS EXTERNOS

Índice Fowlkes e Mallows (FM)

Semelhança entre as partições

$$FM(\pi^c, \pi^r) = \frac{a_1}{\sqrt{(a_1 + a_2)(a_1 + a_3)}}$$

0.0
pior

1.0
melhor

REFERÊNCIAS

Inteligência Artificial - Uma Abordagem de Aprendizado de
Máquina. Katti Facelli et al. SAGAH, 2021.

APRENDIZADO DE MÁQUINAS

Avaliação de Modelos Descritivos

