Aprendizado de Máquina

Aula 7.2 - Avaliação de agrupamentos

Adriano Rivolli

rivolli@utfpr.edu.br

Especialização em Inteligência Artificial
Universidade Tecnológica Federal do Paraná (UTFPR)
Câmpus Cornélio Procópio
Departamento de Computação

Conteúdo

Visão geral

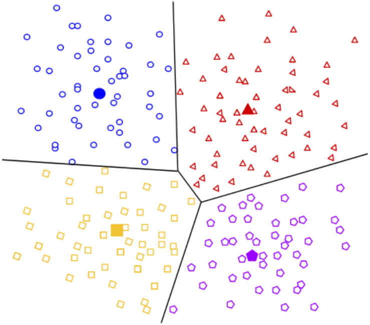
Introdução

- A avaliação não é trivial como ocorre na classificação
- Como definir que um grupo é bom ou não?
- As medidas avaliam perspectivas como separação, elementos no mesmo grupo, similaridade

Centroide

- Um ponto que representa o grupo
- É dado pela média dos pontos de um *cluster*

Centroide (visualização)



Fonte: https://developers.google.com/machine-learning/clustering/clustering-algorithms

Critério de validação

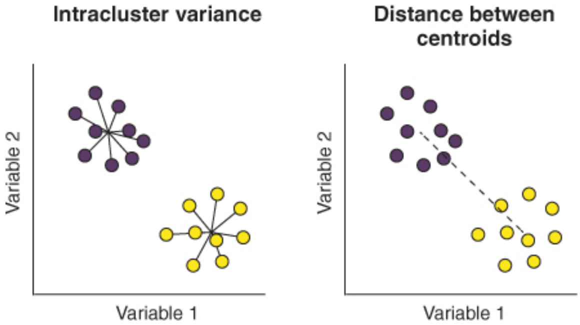
- **Avaliação Interna**
  - Avalia a qualidade dos grupos baseados apenas nos dados e na formação dos grupos pelo algoritmo
  - Não utiliza informações externas ao problema
  - Mede algum critério de qualidade do agrupamento
- **Avaliação Externa**
  - Avalia a qualidade dos grupos baseados em informações adicionais que define como os grupos deveriam ser formados
  - Estas informações não devem ser usadas durante o processo de agrupamento, apenas na avaliação

Validação Interna

Davies-Bouldin Index

- Mede a compactação e a separação de *clusters* em um conjunto de dados
- Valores próximos de 0 indicam melhores partições (mais compactas e separadas)
- Usado para definir o número de grupos

Davies-Bouldin Index (visualização)



Fonte: https://livebook.manning.com/concept/r/davies-bouldin-index

Davies-Bouldin Index (fórmula)

- s_i, the average distance between each point of cluster i and the centroid of that cluster – also know as cluster diameter.
- d_{ij}, the distance between cluster centroids i and j.

A simple choice to construct R_{ij} so that it is nonnegative and symmetric is:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Then the Davies-Bouldin index is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

Calinski-Harabasz Index

- A razão entre a dispersão *between-clusters* e da dispersão *within-cluster*
  - ***between-clusters*** ($\uparrow$) mede a separação entre *clusters*, a partir da distância de cada centroide para o centro dos dados
  - ***within-clusters*** ($\downarrow$) mede a compactação entre *clusters*, a partir da distância de cada ponto com o centroide
- Também conhecido por *Variance Ratio Criterion*
- Valores altos indicam grupos densos e bem separados

UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Calinski-Harabasz Index (fórmula)

$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

where $\mathrm{tr}(B_k)$ is trace of the between group dispersion matrix and $\mathrm{tr}(W_k)$ is the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

with C_q the set of points in cluster q, c_q the center of cluster q, c_E the center of E, and n_q the number of points in cluster q.

Silhouette Coefficient

- Calculado para cada ponto do dataset
- Mede o quão similar um objeto é para com seu grupo comparado com os outros grupos
- O valor da medida fica entre -1 e 1
- Valores próximos a 1 indica quais os objetos estão bem colocados em seu grupo e longe dos demais

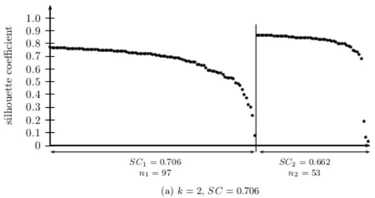
Silhouette (fórmula)

- **a**: The mean distance between a sample and all other points in the same class.
- **b**: The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

Silhouette (visualização)



Fonte: Zaki, M. J., Meira, Jr, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge: Cambridge University Press.

Validação Externa

Rand index

- Mede a similaridade de dois conjuntos de grupos
  - Proporção de pares de rótulos que estão no mesmo grupo ou que estão em grupos diferentes
- Um agrupamento perfeito gera o valor 1
- Existe uma versão ajustada que faz com que grupos aleatórios sejam próximos de zero
- Pode ser utilizado como uma medida de consenso entre diferentes algoritmos
  - A medida é simétrica

Rand index (fórmula)

- a, the number of pairs of elements that are in the same set in C and in the same set in K
- b, the number of pairs of elements that are in different sets in C and in different sets in K

The unadjusted Rand index is then given by:

$$\text{RI} = \frac{a + b}{C_2^{n_{samples}}}$$

where $C_2^{n_{samples}}$ is the total number of possible pairs in the dataset. It does not matter if the calculation is performed on ordered pairs or unordered pairs as long as the calculation is performed consistently.

Mutual Information

- Mede a concordância entre dois conjuntos de grupos usando o conceito de entropia
- O valor 1 indica que o agrupamento é perfeito em relação à referência
- Existe uma versão ajustada e uma versão normalizada da medida
- Não adequada quando o número de grupos (esperado e real) é diferente
- Pode ser utilizado como uma medida de consenso entre diferentes algoritmos
  - A medida é simétrica

Mutual Information (fórmula)

$$H(U) = -\sum_{i=1}^{|U|} P(i)\log(P(i))$$

where $P(i) = |U_i|/N$ is the probability that an object picked at random from U falls into class U_i. Likewise for V:

$$H(V) = -\sum_{j=1}^{|V|} P'(j)\log(P'(j))$$

With $P'(j) = |V_j|/N$. The mutual information (MI) between U and V is calculated by:

$$\mathrm{MI}(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)\log\left(\frac{P(i,j)}{P(i)P'(j)}\right)$$

where $P(i,j) = |U_i \cap V_j|/N$ is the probability that an object picked at random falls into both classes U_i and V_j.

Homogeneity, completeness e V-measure

- **homogeneity** cada grupo contem somente membros de uma única classe
- **completenesss** todos os membros de uma classe estão em apenas um grupo
- **V-measure** corresponde a média harmônica entre as 2 medidas anteriores
- O valor destas medidas estão entre 0 e 1, no qual 1 é o resultado perfeito

Homogeneity e completeness (fórmula)

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

where $H(C|K)$ is the **conditional entropy of the classes given the cluster assignments** and is given by:

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \cdot \log\left(\frac{n_{c,k}}{n_k}\right)$$

and $H(C)$ is the **entropy of the classes** and is given by:

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

V-measure (fórmula)

$$v = \frac{(1+\beta) \times \text{homogeneity} \times \text{completeness}}{(\beta \times \text{homogeneity} + \text{completeness})}$$

$$v = 2 \times \frac{h \times c}{h + c}$$

- $\beta = 1$ (segunda fórmula)
  - *homogeneity* e *completeness* têm o mesmo peso
- $\beta < 1$ atribui mais peso para a *homogeneity*
- $\beta > 1$ atribui mais peso para *completeness*

Matriz de contingência

- Calcula a intersecção para cada combinação dos grupos entre os esperados e preditos
- Permite examinar como as instâncias de uma classe foram distribuídas entre os grupos e vice-versa
- Não se trata de uma métrica, mas é utilizada para calcular a maioria das métricas de validação externa

Matriz de contingência (exemplo)

	G1	G2	G3	Total
Classe A	10	20	30	60
Classe B	5	15	25	45
Total	15	35	55	105

Tabela: Matriz de Contingência para duas Classes e três Grupos

Matriz de confusão de pares

- Combina os pares que estão presentes no mesmo grupo e na mesma classe
- O total de pares possíveis em um dataset é $N = \frac{n \times (n-1)}{2}$

$$C = \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix}$$

C_{00} Número de pares que **NÃO** estão na mesma classe e mesmo grupo (TN)

C_{10} Número de pares que estão na mesma classe, mas **NÃO** estão no mesmo grupo (FN)

C_{01} Número de pares que **NÃO** estão na mesma classe, mas estão no mesmo grupo (FP)

C_{11} Número de pares que estão na mesma classe e grupo (TP)

Matriz de confusão de pares (exemplo)

Suponha que temos o seguinte agrupamento real e previsto:

- Agrupamento real: {1, 2, 3}, {4, 5}
- Agrupamento previsto: {1, 2, 4}, {3, 5}

$$C = \begin{bmatrix} (1,2) & (1,4)(2,4)(3,5) \\ (1,3)(2,3)(4,5) & (1,5)(2,5)(3,4) \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 3 \\ 3 & 3 \end{bmatrix}$$

Fowlkes-Mallows scores

- Definido como a medida F1 utilizando a matriz de confusão de pares
- Média geométrica entre precisão e revocação dos pares:

$$FMI = \frac{C_{11}}{\sqrt{(C_{11} + C_{01}) + (C_{11} + C_{10})}}$$

Considerações finais

- A medida de avaliação usada para agrupamento depende do que se deseja aferir
- Medidas de validação interna são normalmente utilizadas para definir o número de grupos e encontrar valores para os hiperparâmetros do algoritmo
- Medidas de validação externas dependem do conhecimento dos grupos ou da validação de um especialista