licenciatura en inteligencia artificial

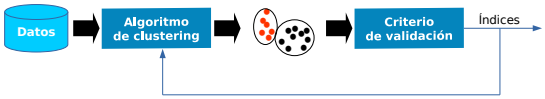
FACULTAD DE CIENCIAS EXACTAS

Introducción a IA

Aprendizaje No Supervisado

Clustering

Validación de clustering: se refiere a la evaluación de los resultados de clustering de una forma cuantitativa y objetiva



Clustering

Criterios de validación: un criterio de validación es una estrategia que emplea una serie de índices o medidas numéricas para medir distintos aspectos del agrupamiento

- **Índices internos**: se usan para medir la bondad de un agrupamiento sin información externa
- **Índices externos**: se usan para medir el grado en que las etiquetas en los clusters coinciden con etiquetas de clases dadas

Criterio Interno

Criterio interno: el **ground truth** (clases reales) rara vez está disponible, pero la validación debe realizarse de todos modos

- Minimiza o maximiza un **índice interno**:
  → Overall Similarity
  → Squared Error (SSE)
  → Silhouette Coefficient

Validación

Criterio Interno

Overall Similarity: cuanto mayor es el valor mejor el agrupamiento

$$OS = \sum_{j=1}^{k} \frac{n_j}{n} \sum_{x \in C_j}^{n} \sum_{y \in C_j}^{n} \frac{sim(x,y)}{|n_j|^2}$$

Validación

Criterio Interno

Cohesión: mide que tan cercanamente relacionados están los ejemplos en un cluster
Separación: mide que tan distintos o bien separados están los clusters entre si

- Un ejemplo es **Squared Error (SSE)** donde la cohesión se mide por la suma de errores cuadrados dentro del cluster:

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

con $|C_i|$ siendo el tamaño del cluster i, m es el centroide del conjunto de datos completo

- La separación se mide por la suma de errores cuadrados entre clusters:

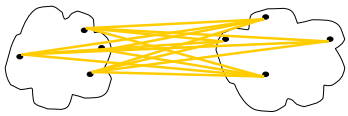
$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

Validación

Criterio Interno

Silhouette: combina las ideas de ambos cohesión y separación

cohesión



separación

Validación

Criterio Interno

$a(i)$: es la distancia promedio de i a todos los otros vectores en el mismo cluster
$b(i)$: es la distancia promedio de i a los vectores en otros clusters

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

silhouette $s(i)$:

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

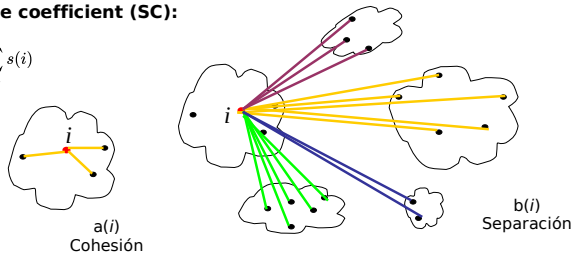
$s(x)$=[-1,+1] cuanto más cerca de 1 mejor, negativo indicaría que a>b, por lo cuál sería deseable un valor positivo

Validación

Criterio Interno

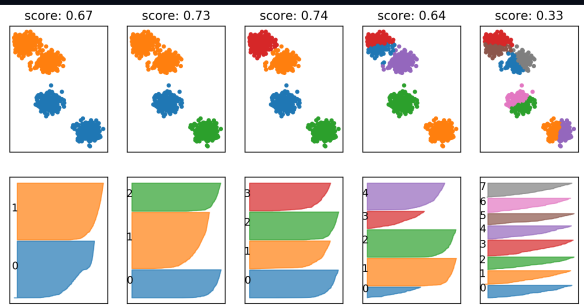
Silhouette coefficient (SC):

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(i)$$



Validación

Criterio Interno



Validación

Criterio Externo

Criterio externo: la calidad se mide por la capacidad de descubrir alguno o todos los patrones ocultos en los datos o clases latentes

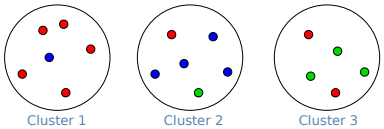
- Medir el clustering respecto de un **ground truth** requiere **datos etiquetados**
- Asume ejemplos pertenecientes a $C=\{c_1,\ldots,c_n\}$ clases, mientras que los algoritmos de clustering producen $K=\{k_1,\ldots,k_m\}$ clusters
- Minimiza o maximiza un **índice externo**:
  → Purity
  → Entropy
  → Rand Index

Validación

Criterio Externo

Purity: radio entre la clase dominante del cluster y su tamaño (sesgada porque más clusters maximizan la pureza)

$$Purity = \sum_{k=1}^{m} \frac{1}{n} \max_i(n_{ki})$$



Purity(cluster 1)= 1/6 * max{5,1,0}=5/6
Purity(cluster 2)= 1/6 * max{1,4,1}=4/6
Purity(cluster 3)= 1/5 * max{2,0,3}=3/5

Validación

Criterio Externo

Entropy:

- Distribución de las clases: p_{ij} es la probabilidad que un miembro del cluster *j* pertenezca a la clase *i*
- Entropía de un cluster *j*:

$$H_j = -\sum_{i=1}^{n} p_{ij} \log(p_{ij})$$

- Entropy total:

$$H(C,K) = \sum_{j=1}^{m} \frac{n_{ij}}{n} H_j$$

Validación

Criterio Externo

Homogeneity: cada cluster contiene solo miembros de una clase

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

Completeness: todos los miembros de una misma clase están en el mismo cluster

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

V-Measure: media armónica de homogeneidad y completitud

$$V - Measure = \frac{2 * homogeneity * completeness}{homogeneity + completeness}$$

Validación

Criterio Externo

Medidas basadas en pares: estadísticas para cada par de items

- SS= misma clase, mismo cluster
- SD=mismo cluster, diferente clase
- DS=diferente cluster, misma clase
- DD= diferente cluster, diferente clase

Número de ejemplos	Mismo cluster en el agrupamiento	Clusters diferentes en el agrupamiento
Misma clase en el ground truth	20	24
Diferentes clases en el ground truth	20	72

Validación

Criterio Externo

Rand Index:

$$Rand\ Index = \frac{SS + DD}{SS + SD + DS + DD}$$

Rand Index = 0.68

Número de puntos	Mismo cluster en el agrupamiento	Clusters diferentes en el agrupamiento
Misma clase en el ground truth	20	24
Diferentes clases en el ground truth	20	72

Validación

Clustering

Validación del clustering:

- Es una tarea compleja, involucra cierta subjetividad
- Varios índices internos y externos para considerar en forma conjunta
- En casos donde el clustering no es la tarea primaria, es posible evaluarlo a través de la aplicación