Aprendizaje no supervisado

VC01: Medidas de evaluación de agrupamientos

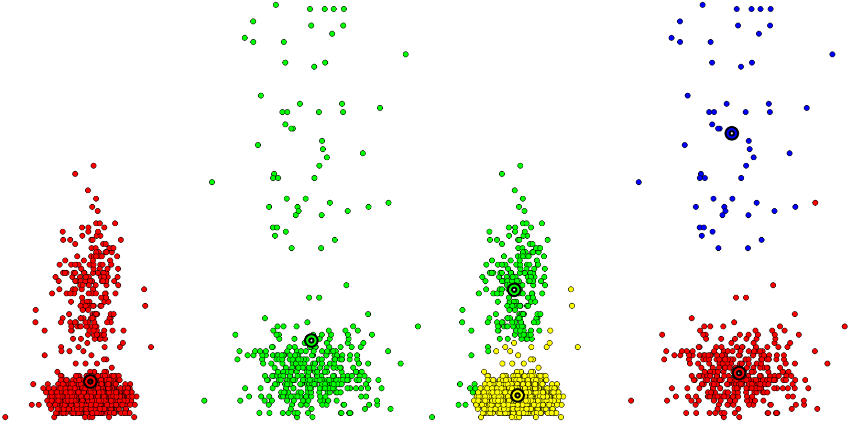
Félix José Fuentes Hurtado
felixjose.fuentes@campusviu.es

Universidad Internacional de Valencia

Universidad Internacional de Valencia

Evaluación

Extrínseca vs. Intrínseca



Evaluación extrínseca

- Se conoce el agrupamiento real (**verdad básica**)
- Se compara el resultado del algoritmo con la verdad básica

Evaluación extrínseca

- Se conoce el agrupamiento real (**verdad básica**)
- Se compara el resultado del algoritmo con la verdad básica
- No existe el concepto de etiqueta:
  búsqueda de la correspondencia entre clúster real y predicho

Evaluación intrínseca

- No se conoce el agrupamiento real (**verdad básica**), ni se sabe si existe
- Se mide la congruencia del agrupamiento
- Diferentes criterios posibles

Evaluación

$\{B_l\}_{l=1}^{K'}$: Verdad básica

$\{C_k\}_{k=1}^{K}$: Agrupamiento resultante de un algoritmo de *clustering*

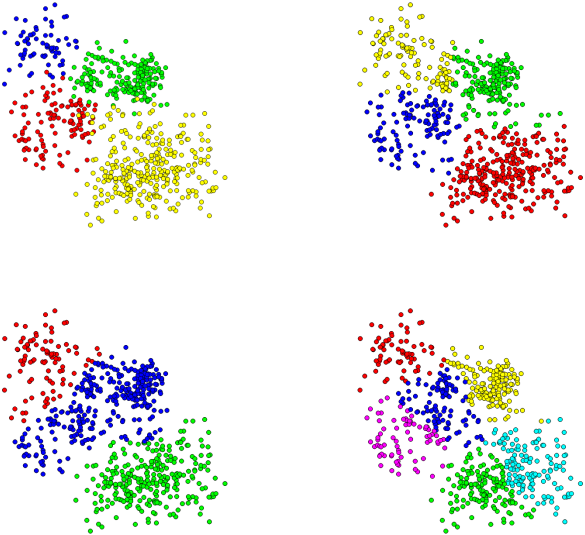
$\{\boldsymbol{c}_k\}_{k=1}^{K}$: Centro(ide)s de los clústeres resultantes

$n_l = |B_l|$: Tamaño de un clúster verdadero

$n_k = |C_k|$: Tamaño de un clúster resultante

$n_{kl} = |C_k \cap B_l|$: Número de ejemplos que comparten un clúster resultante y otro verdadero

Evaluación



Error:

$$E = 1 - \frac{1}{n} \max_{\sigma} \sum_{l=1}^{K'} n_{\sigma(l) l}$$

donde σ es una función de $\sigma : \{1, \ldots, K'\} \to \{1, \ldots, K\}$

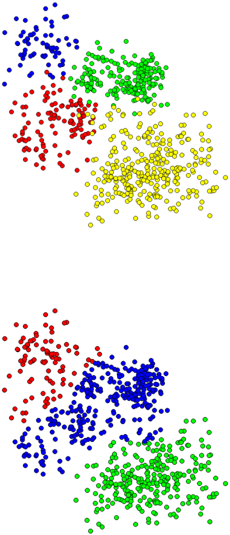
Error:

$$E = 1 - \frac{1}{n} \max_{\sigma} \sum_{l=1}^{K'} n_{\sigma(l)l}$$

donde σ es una función de $\sigma : \{1, \ldots, K'\} \rightarrow \{1, \ldots, K\}$

- Recorrido sobre los clústeres reales
- Máximo (optimista) para identificar la correspondencia C-B

Evaluación



Evaluación

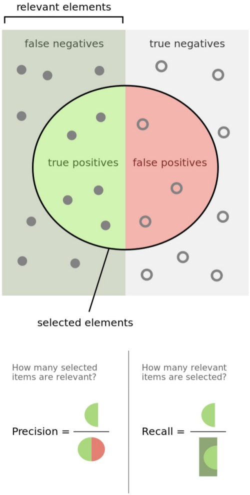
Extrínseca

Precisión:

$$P_{kl} = \frac{n_{kl}}{n_{k\cdot}}$$

Recall:

$$R_{lk} = \frac{n_{kl}}{n_{\cdot l}}$$



Precisión:

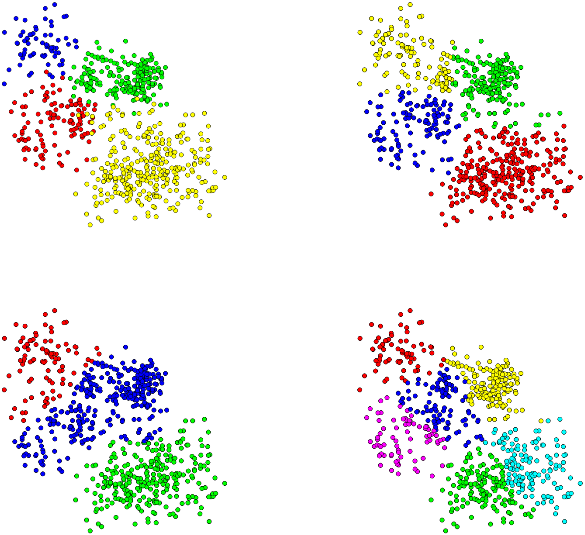
$$P_{kl} = \frac{n_{kl}}{n_{k\cdot}}$$

Recall:

$$R_{lk} = \frac{n_{kl}}{n_{\cdot l}}$$

- Medidas entre un clúster real y otro resultante
- Precisión: ¿Cuántos de los elementos del clúster resultante k lo son también del clúster real l?
- *Recall*: ¿Cuántos de los elementos del clúster real l lo son también del clúster resultante k?

Evaluación



Pureza:

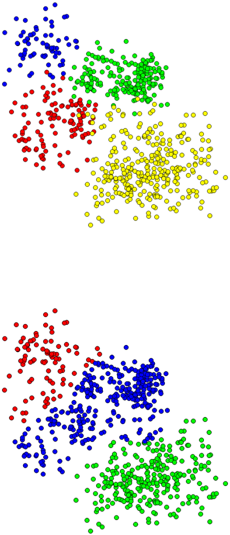
$$Pu = \sum_{k=1}^{K} \frac{n_{k\cdot}}{n} \max_{l \in \{1,\ldots,K'\}} P_{kl}$$

Pureza:

$$Pu = \sum_{k=1}^{K} \frac{n_{k\cdot}}{n} \max_{l \in \{1,\ldots,K'\}} P_{kl}$$

- Media ponderada de la precisión
- Recorrido sobre los clústeres resultantes
- Máximo (optimista) para identificar la correspondencia C-B

Evaluación



Medida F:

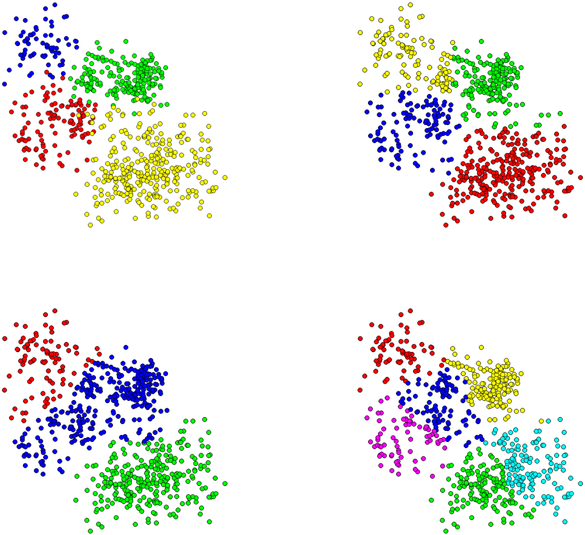
$$F1 = \sum_{l=1}^{K'} \frac{n_{\cdot l}}{n} \max_{k \in \{1,\ldots,K\}} \left(\frac{2 P_{kl} R_{lk}}{P_{kl} + R_{lk}} \right)$$

Medida F:

$$F1 = \sum_{l=1}^{K'} \frac{n_{.l}}{n} \max_{k \in \{1,\ldots,K\}} \left(\frac{2 P_{kl} R_{lk}}{P_{kl} + R_{lk}} \right)$$

- Media ponderada de la media harmónica de la precisión y el *recall*
- Recorrido sobre los clústeres reales
- Máximo (optimista) para identificar la correspondencia C-B

Evaluación



Entropía:

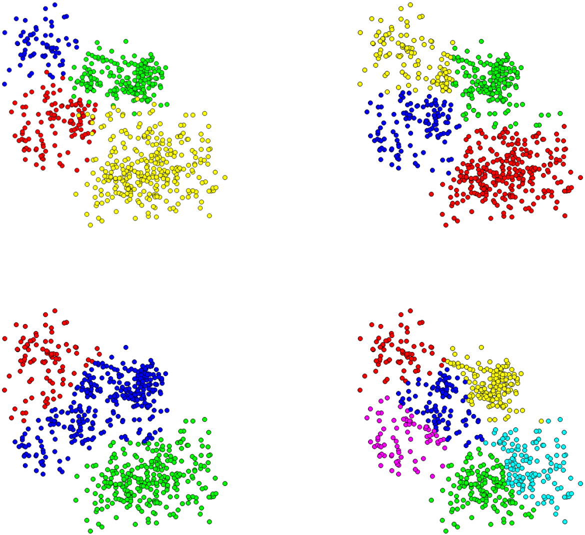
$$H = -\sum_{k=1}^{K} \frac{n_{k\cdot}}{n} \sum_{l=1}^{K'} \frac{n_{kl}}{n_{k\cdot}} \log \frac{n_{kl}}{n_{k\cdot}}$$

Entropía:

$$H = -\sum_{k=1}^{K} \frac{n_{k\cdot}}{n} \sum_{l=1}^{K'} \frac{n_{kl}}{n_{k\cdot}} \log \frac{n_{kl}}{n_{k\cdot}}$$

- Media ponderada de la entropía de cada clúster resultante
- Entropía: mide cómo se distribuyen los ejemplos de un clúster resultante entre los clústeres reales (crece a mayor desorden)
- Recorrido (principal) los clústeres resultantes

Evaluación



Información mutua:

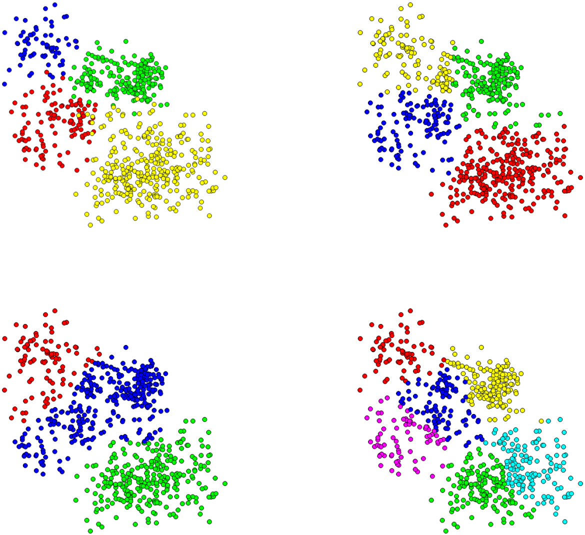
$$I = \sum_{k=1}^{K} \sum_{l=1}^{K'} \frac{n_{kl}}{n} \log \frac{n \cdot n_{kl}}{n_{k\cdot} \cdot n_{\cdot l}}$$

Información mutua:

$$I = \sum_{k=1}^{K} \sum_{l=1}^{K'} \frac{n_{kl}}{n} \log \frac{n \cdot n_{kl}}{n_{k.} \cdot n_{.l}}$$

- Información mutua entre dos agrupamientos (real y resultante)
- I: mide cómo se explican mutuamente ambos agrupamientos

Evaluación



¿Es razonable asumir la existencia de la verdad básica?

¿Es razonable asumir la existencia de la verdad básica?

¿Para qué queremos entonces un algoritmo de *clustering*?

La raíz del cuadrado de la media de la desviación típica:

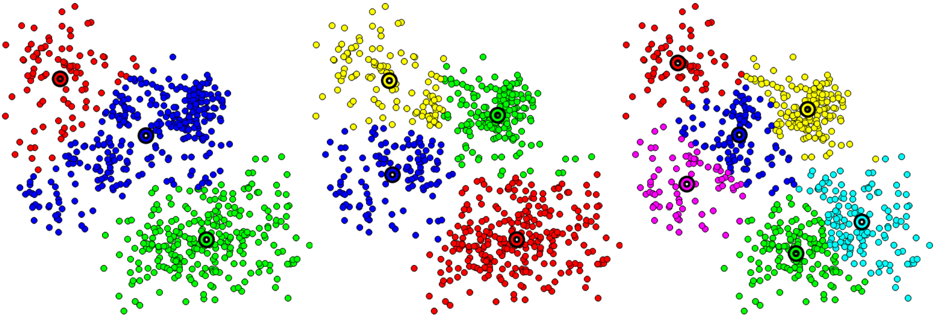
$$RMSSTD = \sqrt{\frac{\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in C_k} ||\boldsymbol{x}_i - \boldsymbol{c}_k||^2}{v \cdot \sum_{k=1}^{K} (|C_k| - 1)}}$$

La raíz del cuadrado de la media de la desviación típica:

$$RMSSTD = \sqrt{\frac{\sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in C_k} ||\boldsymbol{x}_i - \boldsymbol{c}_k||^2}{v \cdot \sum_{k=1}^{K} (|C_k| - 1)}}$$

- Mide la heterogeneidad de los clústeres
- Se reduce fácilmente aumentando el número de clústeres resultantes K

Evaluación



Evaluación
Intrínseca

Medida R-cuadrado

$$R^2 = \frac{\sum_{\boldsymbol{x}_i} ||\boldsymbol{x}_i - \boldsymbol{c}||^2 - \sum_{k=1}^{K} \sum_{\boldsymbol{x}_{i'} \in C_k} ||\boldsymbol{x}_{i'} - \boldsymbol{c}_k||^2}{\sum_{\boldsymbol{x}_i} ||\boldsymbol{x}_i - \boldsymbol{c}||^2}$$

donde $\boldsymbol{c}$ es el centro de todo el *dataset*.

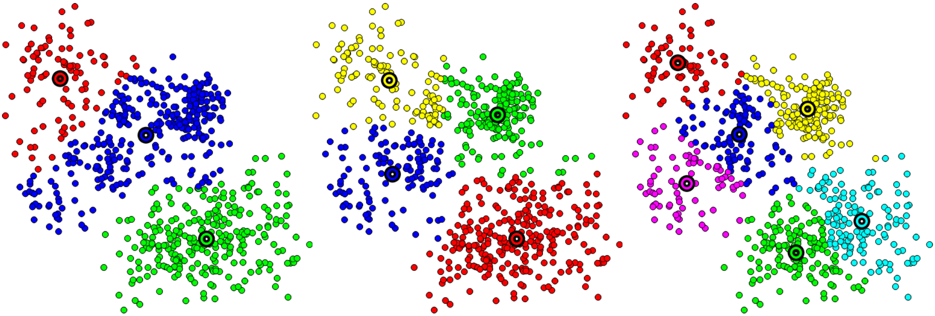
Medida R-cuadrado

$$R^2 = \frac{\sum_{\boldsymbol{x}_i} ||\boldsymbol{x}_i - \boldsymbol{c}||^2 - \sum_{k=1}^{K} \sum_{\boldsymbol{x}_{i'} \in C_k} ||\boldsymbol{x}_{i'} - \boldsymbol{c}_k||^2}{\sum_{\boldsymbol{x}_i} ||\boldsymbol{x}_i - \boldsymbol{c}||^2}$$

donde $\boldsymbol{c}$ es el centro de todo el *dataset*.

- Mide la homogeneidad de los clústeres
- Acotada entre 0 (sólo un clúster) y 1 ($K = n$)
- Se incrementa fácilmente aumentando el número de clústeres resultantes K

Evaluación



Silueta

$$S = \frac{1}{n} \sum_{\boldsymbol{x}_i} \frac{b_k(\boldsymbol{x}_i) - a_k(\boldsymbol{x}_i)}{\text{máx}\{b_k(\boldsymbol{x}_i), a_k(\boldsymbol{x}_i)\}}$$

donde

$$a_k(\boldsymbol{x}_i) = \frac{1}{n_k - 1} \sum_{\boldsymbol{x}_j \in C_k : \boldsymbol{x}_j \neq \boldsymbol{x}_i} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

y

$$b_k(\boldsymbol{x}_i) = \min_{h \neq k} \frac{1}{n_h} \sum_{\boldsymbol{x}_j \in C_h} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Evaluación

Intrínseca

Silueta

$$S = \frac{1}{n} \sum_{\boldsymbol{x}_i} \frac{b_k(\boldsymbol{x}_i) - a_k(\boldsymbol{x}_i)}{\max\{b_k(\boldsymbol{x}_i), a_k(\boldsymbol{x}_i)\}}$$

donde

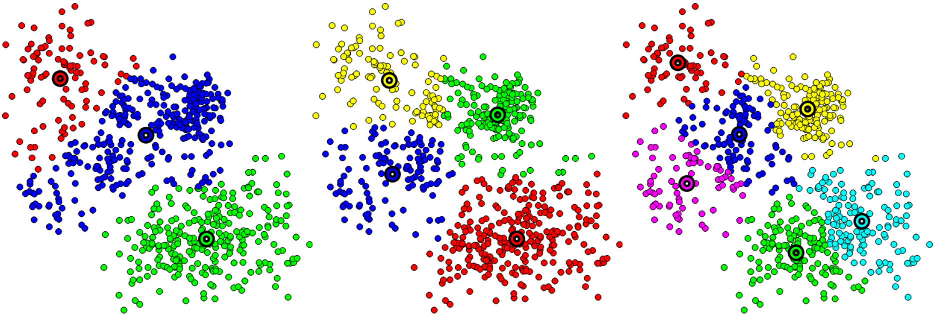
$$a_k(\boldsymbol{x}_i) = \frac{1}{n_k - 1} \sum_{\boldsymbol{x}_j \in C_k : \boldsymbol{x}_j \neq \boldsymbol{x}_i} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

y

$$b_k(\boldsymbol{x}_i) = \min_{h \neq k} \frac{1}{n_h} \sum_{\boldsymbol{x}_j \in C_h} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- Diferencia normalizada entre la distancia intraclúster y la interclúster
- Acotada entre -1 y 1

Evaluación



Índice Calinski-Harabasz:

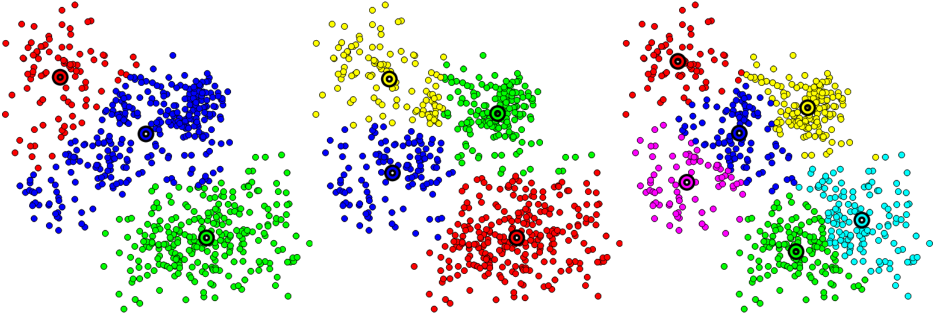
$$CH = \frac{(n-K)\sum_{k=1}^{K} n_k \cdot d(\boldsymbol{c}_k, \boldsymbol{c})^2}{(K-1)\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, \boldsymbol{c}_k)^2}$$

Índice Calinski-Harabasz:

$$CH = \frac{(n-K)\sum_{k=1}^{K} n_k \cdot d(\boldsymbol{c}_k, \boldsymbol{c})^2}{(K-1)\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, \boldsymbol{c}_k)^2}$$

- Suma promedio de las distancias inter e intraclúster al cuadrado
- A mayor valor, mejor agrupamiento

Evaluación



Índice I:

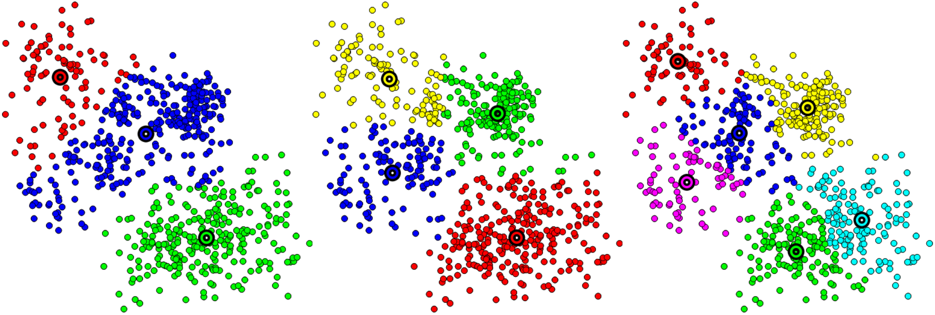
$$I = \left(\frac{\sum_{\boldsymbol{x}_i} d(\boldsymbol{x}_i, \boldsymbol{c})}{K \cdot \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, \boldsymbol{c}_k)} \cdot \max_{i,j \in \{1,\ldots,K\}} d(\boldsymbol{c}_i, \boldsymbol{c}_j) \right)^p$$

Índice I:

$$I = \left(\frac{\sum_{\boldsymbol{x}_i} d(\boldsymbol{x}_i, \boldsymbol{c})}{K \cdot \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in C_k} d(\boldsymbol{x}_i, \boldsymbol{c}_k)} \cdot \max_{i,j \in \{1,\ldots,K\}} d(\boldsymbol{c}_i, \boldsymbol{c}_j) \right)^p$$

- Mide la separación interclúster con respecto a la homogeneidad intraclúster
- A mayor valor, mejor agrupamiento

Evaluación



Aprendizaje no supervisado

VC01: Medidas de evaluación de agrupamientos

Félix José Fuentes Hurtado
felixjose.fuentes@campusviu.es

Universidad Internacional de Valencia