# EST-297

## Métricas de Evaluación Internas en Clustering

Juan Zamora O.

Junio, 2025.

INSTITUTO DE ESTADÍSTICA
IES PUCV

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

# ¿Qué es la validación interna?

- Evalúa la calidad del clustering sin usar etiquetas externas.
- Se basa en la compacidad (dentro del clúster) y separación (entre clústeres).
- Nos ayuda a decidir el número óptimo de clústeres.

# Índice de Silueta

- Evalúa cuán bien está asignado cada punto a su clúster.
- Fórmula:

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

- Donde:
  - $a(i)$: distancia media de $i$ a puntos de su propio clúster.
  - $b(i)$: menor distancia media a puntos de otro clúster.
- Valores cercanos a 1 indican buen agrupamiento.

# Índice de Calinski-Harabasz

- Mide la relación entre la dispersión entre clústeres y dentro de clústeres.
- Fórmula:

$$CH = \frac{B_k/(k-1)}{W_k/(n-k)}$$

- Donde:
  - $B_k$: varianza entre clústeres.
  - $W_k$: varianza dentro de los clústeres.
  - $k$: número de clústeres, $n$: número total de puntos.
- Cuanto mayor el índice, mejor la separación.

# Índice de Davies-Bouldin

- Evalúa la similitud entre clústeres considerando dispersión y distancia.
- Fórmula:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left( \frac{s_i + s_j}{d_{ij}} \right)$$

- Donde:
  - $s_i$: dispersión intra-clúster $i$.
  - $d_{ij}$: distancia entre centroides de $i$ y $j$.
- Valores más bajos indican mejor separación.

# WCSS y Método del Codo

- WCSS: Suma de distancias cuadradas entre los puntos y su centroide.
- Fórmula:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

- Se usa en el método del codo para elegir el mejor número de clústeres.
- Busca el punto donde la reducción de WCSS se vuelve marginal.

# Gap Statistic

- Compara el cambio en dispersión intra cluster con el esperado bajo una distribución uniforme.
- Fórmula:

$$Gap(k) = \mathbb{E}[\log(W_k^{\text{ref}})] - \log(W_k)$$

- Donde $W_k = \sum_{r=1}^{k} \frac{D_r}{2|C_r|}$ con $D_r = \sum_{i,i' \in C_r} d_{ii'}$.
- Se elige $k$ donde el Gap es máximo.

# Índice de Dunn

- Mide la relación entre la mínima distancia entre clústeres y la máxima dispersión interna.
- Fórmula:

$$D = \frac{\min_{i \neq j} d(C_i, C_j)}{\max_k \operatorname{diam}(C_k)}$$

- Donde:
  - $d(C_i, C_j)$: distancia entre clústeres.
  - $\operatorname{diam}(C_k)$: diámetro del clúster $k$.
- Valores altos indican clústeres bien separados y compactos.
- Diseñado para estimar el número de clusters

# Resumen

- Cada métrica tiene fortalezas y limitaciones.
- Usar varias medidas da una visión más completa.
- Idealmente, combinar con validación externa si hay etiquetas.