FACULTAD de
CIENCIAS ECONÓMICAS

Métricas para Evaluar Algoritmos de Agrupamiento

Alan Reyes-Figueroa

Elements of Machine Learning (Aula 17) 23.marzo.2023

Métricas de evaluación

Queremos desarrollar diversas técnicas para evaluar algoritmos de agrupamiento de datos. Para ello, consideramos una matriz de datos $\mathbb{X} \in \mathbb{R}^{n \times d}$.

Mencionamos a continuación cuatro técnicas (internas) para evaluar algoritmos de *clustering*:

- Método del codo (*elbow method*)
- Índice de Dunn
- Índice de David-Bouldin
- Método de las siluetas (*silhouette*)
- Índice de Calinski-Harabasz

Métricas de evaluación

Existen además, otras técnicas (externas), las cuales hacen uso de un conjunto de test, para el cual se conocen las etiquetas o grupos ya definidos.

Entre ellas:

- Índice de Rand (*Rand index*)
- Información mutua
- Validacion cruzada (*cross validation*)
- Homogeneidad, completitud, V-*score*
- Puntuación de Fowlkes-Mallows
- Matriz de confusión y métricas derivadas

Métodos Externos

Estos métodos requieren que se tengan etiquetas *ground-truth* para ser evaluadas. Básicamente, estas métricas contrastan las etiquetas reales (*ground-truth*) contra las etiquetas o grupos estimados por el algoritmo de agrupamiento.

Imaginen que tenemos un conjunto de datos $\{\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$, donde las $\mathbf{y}_i$ son las etiquetas reales, y $\widehat{\mathbf{y}}_i$ son las etiquetas estimadas por el algoritmo. Estas metricas comparan coincidencias por pares de puntos $\mathbf{x}_i$ y $\mathbf{x}_j$:

- si $\mathbf{y}_i = \mathbf{y}_j$ (el dato *i* y el dato *j* están en el mismo clúster), y las predicciones $\widehat{\mathbf{y}}_i = \widehat{\mathbf{y}}_j$, tenemos una coincidencia.
- si $\mathbf{y}_i = \mathbf{y}_j$ (el dato *i* y el dato *j* están en el mismo clúster), y las $\widehat{\mathbf{y}}_i \neq \widehat{\mathbf{y}}_j$, no tenemos coincidencia.
- si $\mathbf{y}_i \neq \mathbf{y}_j$ (el dato *i* y el dato *j* no están en el mismo clúster), pero $\widehat{\mathbf{y}}_i = \widehat{\mathbf{y}}_j$, no tenemos coincidencia.
- si $\mathbf{y}_i \neq \mathbf{y}_j$ (el dato *i* y el dato *j* no están en el mismo clúster), y además $\widehat{\mathbf{y}}_i \neq \widehat{\mathbf{y}}_j$, tenemos una coincidencia.

Métodos Externos

Rand Index:

$$RI = \frac{\text{pares que coinciden}}{\text{número total de pares}} = \frac{TP+TN}{TP+FP+TN+FN}.$$

Adjusted Rand Index:

$X \backslash Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}$$

Métodos Externos

Mutual Information:

$$MI = -\sum_{i=1}^{r}\sum_{j=1}^{s} \mathbb{P}_{XY}(i,j)\, \log \frac{\mathbb{P}_{XY}(i,j)}{\mathbb{P}_X(i)\,\mathbb{P}_Y(j)} = \text{Entropía cruzada.}$$

Normalized Mutual Information:

$$NMI = \frac{2\,MI}{H(X) + H(Y)}.$$

Adjusted Mutual Information:

$$AMI(U,V) = \frac{MI(U,V) - E\{MI(U,V)\}}{\max\{H(U), H(V)\} - E\{MI(U,V)\}}.$$

Métodos Externos

Homogeneity:

$$H = 1 - \frac{H(X \mid Y)}{H(X)}, \qquad \text{donde } H(X \mid Y) = -\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{N}\log\frac{n_{ij}}{n_j}.$$

Completeness:

$$C = 1 - \frac{H(Y \mid X)}{H(Y)},$$

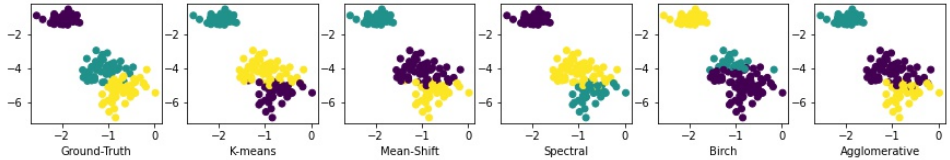
V-score ó **V-measure**: Depende de un parámetro o $< \beta < 1$

$$V_\beta = \frac{(1+\beta)HC}{\beta H + C},$$

Fowlkes-Mallows Index:

$$FM = \sqrt{PPV \cdot TPR} = \sqrt{\frac{TP}{TP+FP}\cdot\frac{TP}{TP+FN}}$$

Ejemplo



	K-means	Mean-Shift	Spectral	Birch	Agglomerative
RI	0.879732	**0.885906**	**0.885906**	0.819597	0.879732
ARI	0.730238	0.743683	**0.745504**	0.609625	0.731199
MI	0.825591	0.835346	**0.864524**	0.74124	0.835825
NMI	0.758176	0.766036	**0.797989**	0.705099	0.770084
AMI	0.755119	0.763083	**0.795421**	0.701217	0.767167
H	0.751485	0.760365	**0.786923**	0.674706	0.760801
C	0.764986	0.771792	**0.809369**	0.73836	0.779596
V	0.758176	0.766036	**0.797989**	0.705099	0.770084
FMI	0.820808	0.829449	**0.83205**	0.751487	0.82217
CHI	**561.627757**	560.13945	556.117692	458.472511	558.058041

Elbow method

Se cree que inició con los trabajos de Robert L. Thorndike (1953).

Es una heurística que se utiliza para determinar el número de clústers k en un conjunto de datos. El método consiste en graficar la variación explicada en función de k y elegir el codo de la curva como el número de clústers a utilizar. El mismo método se puede utilizar para elegir el número de parámetros en otros modelos basados en datos, como el número de componentes principales en PCA.

Como ya hemos mencionado antes, esta técnica sirve tanto para determinar k, el número óptimo de clusters, como para evaluar diferentes métodos de agrupamiento.

Elbow method

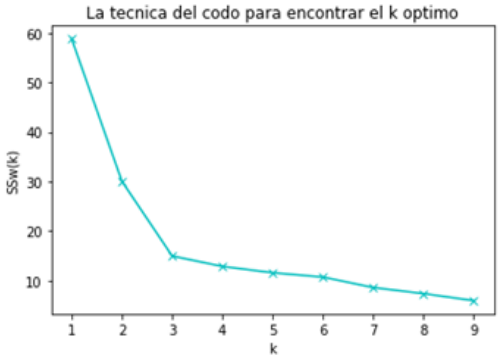
La idea principal consiste en medir el error intra-grupos SSE_w (*sum of squared errors within*),

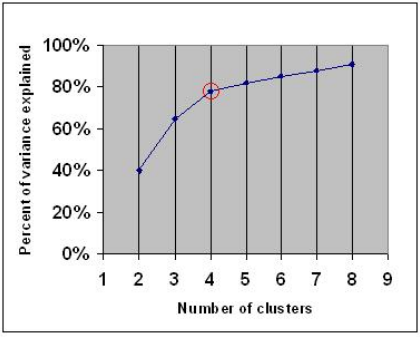
$$SSE_w = SSE_{within} = \sum_{j=1}^{k} \sum_{i:g(i)=j} (\mathbf{x}_i - \mathbf{c}_j)^2.$$

Otra alternativa consiste en medir el porcentaje de la varianza explicada por el método de agrupamiento.
Existen diversos criterior para el porcentaje de varianza explicada. En la versión más popular, se entiende como varianza explicada la relación entre la varianza intra-grupos contra la varianza total

$$\text{Varianza explicada} = 1 - \frac{SSE_{intra}}{SSE_{total}} = 1 - \frac{\sum_{j=1}^{k} \sum_{i:g(i)=j} (\mathbf{x}_i - \mathbf{c}_j)^2}{\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})^2}.$$

Elbow method





Método del codo: (a) comparando el error intra-grupos SSE_w, (b) comparando el porcentaje de varianza explicada.

Índice de Dunn

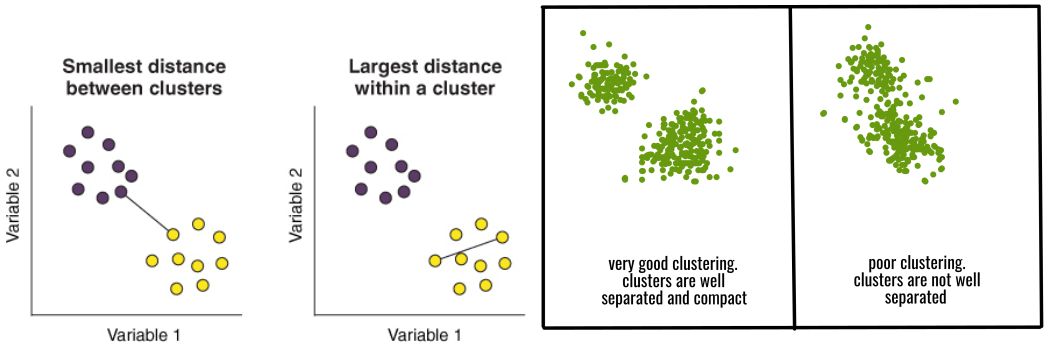
Desarrollado por Joseph Dunn (1973).

Es una medida de la calidad de una partición de un conjunto de datos. Consiste en medir la relación entre la distancia máxima que separa dos elementos clasificados juntos y la distancia mínima que separa dos elementos clasificados por separado.

Este es un índice que no se basa en una distancia en particular d, y que por lo tanto, se puede utilizar en una amplia variedad de situaciones y métricas.

Sea $\mathbb{X} \in \mathbb{R}^{n \times d}$ a la matriz de datos, y consideramos una métrica o disimilaridad $d(\mathbf{x}_i, \mathbf{x}_j)$. Denotamos por $\mathbf{c}_g$ el centroide de cada grupo $g \in \{1, 2, \ldots, k\}$.

Índice de Dunn



Índice de Dunn: (a) compara la mínima distancia entre centroides, respecto al máximo diámetro; (b) ejemplos de buen y mal agrupamiento.

Índice de Dunn

El índice de Dunn se calcula de la siguiente forma: Para cada grupo, construimos su diámetro (esto es, la máxima distancia entre dos de sus elementos)

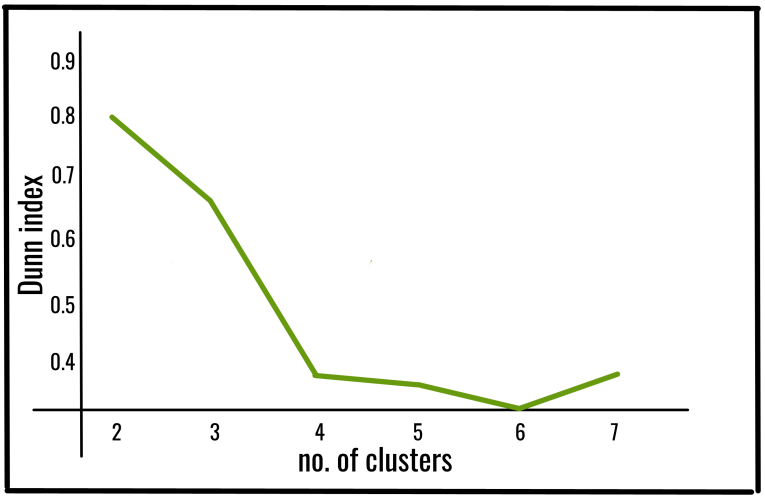
$$\Delta_j = \max_{r,s:g(r)=g(s)=j} d(\mathbf{x}_r, \mathbf{x}_s).$$

Luego, el índice de Dunn se define como

$$S_D = \frac{\min_{1\leq i<j\leq k} d(\mathbf{c}_i, \mathbf{c}_j)}{\max_{1\leq j\leq k} \Delta_j}.$$

Obs! Este índice puede variar un poco según las implementaciones (definición del diámetro de un grupo, distancia entre centros reemplazada por otra distancia entre grupos).

Índice de Dunn



Índice de Davies-Bouldin

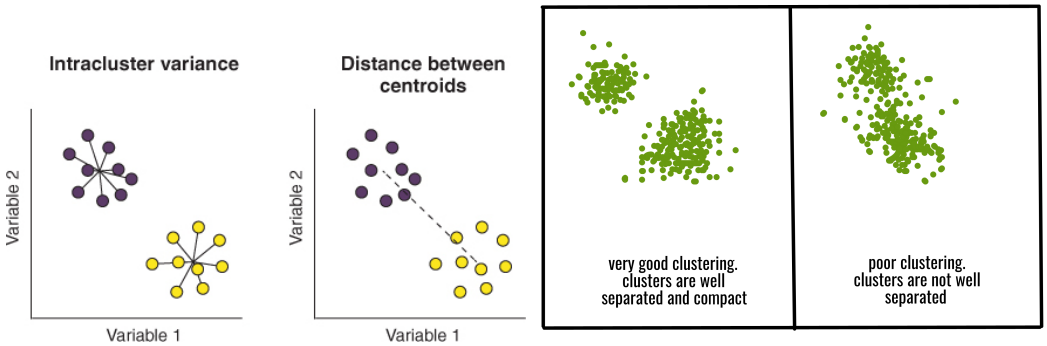
Una alternativa al índice de Dunn es el índice de Davies-Bouldin.

D. L. Davies y D. W. Bouldin, “A Cluster Separation Measure”, en *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no 2, (1979).

Similar a Dunn, este índice de Davies-Bouldin, compara la distancia promedio de cada grupo (respecto de su centroide), y la contrasta con la distancia entre grupos.

Al final toma la suma de los máximos de estas relaciones entre distancias.

Índice de Davies-Bouldin



Índice de Davies-Bouldin: (a) compara las distancias promedio de cada grupo, respecto a la distancia entre grupos; (b) ejemplos de buen y mal agrupamiento.

Índice de Davies-Bouldin

Para cada grupo, definimos la distancia media

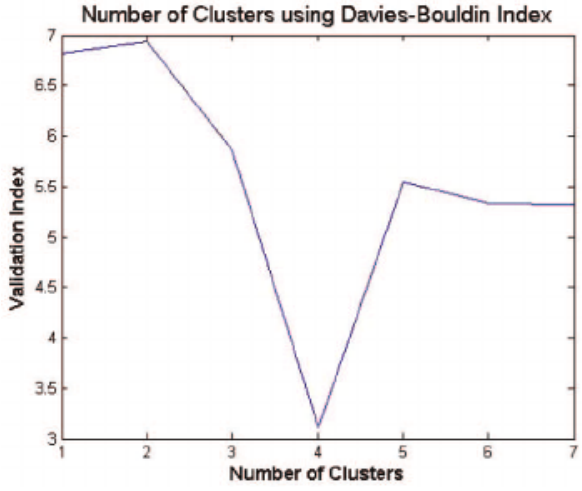
$$\delta_j = \frac{1}{|\{i : g(i) = j\}|} \sum_{i:g(i)=j} d(\mathbf{x}_i, \mathbf{c}_j).$$

Luego contrastamos estas distancias con las distancias extra-grupos. Definimos el índice de Davies-Bouldin por

$$S_{DB} = \frac{1}{k} \sum_{j=1}^{k} \max_{j' \neq j} \left(\frac{\delta_j + \delta_{j'}}{d(\mathbf{c}_j, \mathbf{c}_{j'})} \right).$$

Obs! Puede variar un poco según las implementaciones (distancia impuesta, representante, distancia entre grupos).

Índice de Davies-Bouldin



Método de las Siluetas

Propuesto por Peter J. Rousseeuw (1987). “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis” en *Computational and Applied Mathematics*. 20: 53-65.

Es un método de interpretación y validación de la coherencia dentro de clústers de datos. La técnica proporciona una representación gráfica de qué tan bien se ha clasificado cada objeto en un método de agrupamiento.

El valor de silueta es una medida de cuán similar es un objeto a su propio grupo (cohesión) en comparación con otros grupos (separación).

Método de las Siluetas

Observaciones:

- La silueta varía de -1 a $+1$: un valor alto indica que el objeto se corresponde bien con su propio grupo y no con los grupos vecinos.
- Si la mayoría de los objetos tienen un valor alto, entonces la configuración de agrupamiento es apropiada. Si muchos puntos tienen un valor bajo o negativo, entonces la configuración de la agrupación en clústeres puede tener demasiados o muy pocos clústeres.
- La silueta se puede calcular con cualquier métrica de distancia, como la distancia euclidiana o la distancia de Manhattan.

Método de las Siluetas

Suponga que los datos se han agrupado en k grupos.
Para cada punto $\mathbf{x}_i$, en el grupo $C_i = g(i)$, asociamos la distancia media entre $\mathbf{x}_i$ y todos los demás puntos del mismo grupo

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

Podemos interpretar $a(i)$ como una medida de qué tan bien $\mathbf{x}_i$ está asignado a su grupo (cuanto menor sea el valor, mejor será la asignación).

Luego, definimos la distancia media del punto $\mathbf{x}_i$ con algún grupo C_j, como la media de la distancia desde $\mathbf{x}_i$ a todos los puntos en C_j. Definimos

$$b(i) = \min_{j: j \neq g(i)} \frac{1}{|C_j|} \sum_{r: g(r) = j} d(\mathbf{x}_i, \mathbf{x}_j),$$

Método de las Siluetas

como la menor de las distancias medias desde $\mathbf{x}_i$ a todos los puntos en cualquier otro grupo. Se dice que el grupo con esta menor distancia media es el **grupo vecino** de $\mathbf{x}_i$, ya que es el grupo de mejor ajuste para dicho punto.

Ahora definimos el valor de la **silueta** para el punto $\mathbf{x}_i$:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{si } |C_i| > 1,$$

y $s(i) = 0$, si $|C_i| = 1$. También podemos escribirla como

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{si } a(i) < b(i); \\ 0, & \text{si } a(i) = b(i); \\ \frac{b(i)}{a(i)} - 1, & \text{si } a(i) > b(i). \end{cases}$$

Método de las Siluetas

Observaciones:

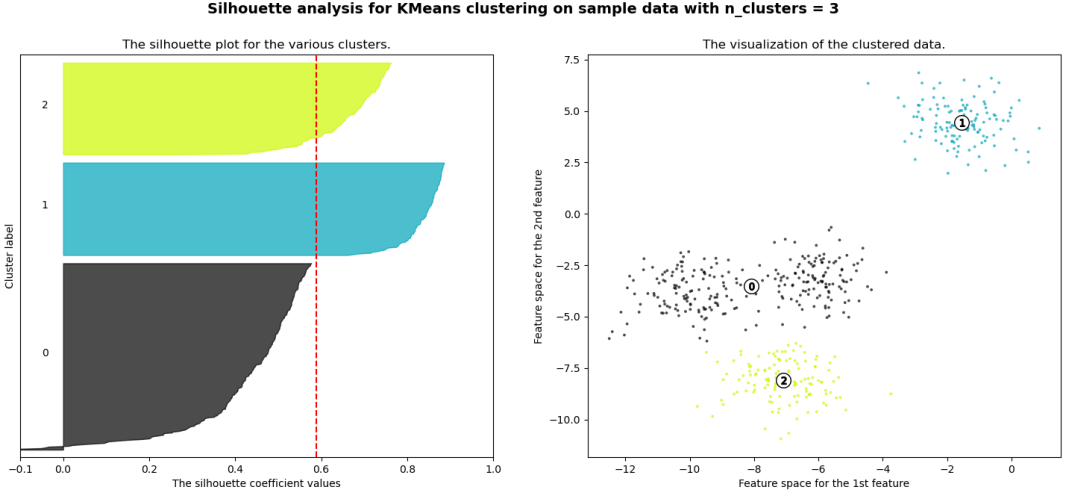
- De la definición anterior queda claro que $-1 \leq s(i) \leq 1$.
- $a(i)$ no está claramente definido para clústeres con tamaño 1, en cuyo caso establecemos $a(i) = 0$. Esta elección es arbitraria, pero neutral en el sentido que es el punto medio de los límites, -1 y $+1$.
- Para que $s(i)$ esté cerca de 1, necesitamos $a(i) << b(i)$. Como $a(i)$ es una medida de lo diferente que es $\mathbf{x}_i$ con su propio grupo, un valor pequeño significa que está bien adaptado. Además, un $b(i)$ grande implica que $\mathbf{x}_i$ no coincide con su grupo vecino. Por lo tanto, $s(i)$ cercano a 1 significa que los datos están agrupados de manera apropiada.
- Si $s(i)$ está cerca de -1, entonces $\mathbf{x}_i$ sería más apropiado si estuviera agrupado en su grupo vecino.

Método de las Siluetas

- Un $s(i)$ cercano a cero significa que el dato está en el borde de dos conglomerados naturales.
- La media de $s(i)$ es una medida de cuán estrechamente agrupados están todos los puntos.
- Si hay demasiados o muy pocos conglomerados (*e.g* mala elección de k), algunos de los conglomerados normalmente mostrarán siluetas mucho más estrechas que el resto. Por lo tanto, se pueden usar gráficos de silueta y medios para determinar el número natural de conglomerados dentro de un conjunto de datos.
- Kaufman y Rousseeuw (1990). definen el término **coeficiente de silueta** para el valor máximo de la media $s(i)$ por grupos:

$$SC = \max_{1 \leq j \leq k} \widetilde{s}(j), \quad \text{con } \widetilde{s}(j) = \frac{1}{|C_j|} \sum_{i:g(i)=j} s(i).$$

Método de las Siluetas



UFM UNIVERSIDAD FRANCISCO MARROQUÍN

Método de las Siluetas