MEASURING CLUSTER GOODNESS

Luis Alberto Vega Martínez

Justificación del cluster de medición de bondad

Toda técnica de modelado requiere una fase de evaluación múltiple.

Modelo de regresión múltiple

Error estándar de la estimación = $100000

¿$100000?

En el ámbito de la clasificación, esperaríamos que un modelo que prediga quién responderá a nuestra operación de marketing por correo directo arroje resultados más rentables que la línea de base "enviar-un-cupón-a-todos" o "enviar-no-cupones-en- -todos los modelos.

También es necesario evaluar los modelos de agrupación en clusters.

- ¿Mis clusters se corresponden realmente con la realidad o son simplemente artefactos de conveniencia matemática?

- No estoy seguro de cuántos grupos hay en los datos. ¿Cual es el óptimo

número de clusters para identificar?

- ¿Cómo mido si un conjunto de grupos es preferible a otro?

También examinamos un método para validar nuestros clústeres mediante validación cruzada con análisis gráfico y estadístico.

Cualquier medida de bondad o calidad de los clusters debe abordar los conceptos de separación de clusters y cohesión de clusters. La separación de grupos representa la distancia entre los grupos; la cohesión del clúster se refiere a cuán estrechamente relacionados los registros dentro de los grupos individuales son.

Método de silueta

La silueta es una característica de cada valor de datos y se define de la siguiente manera:

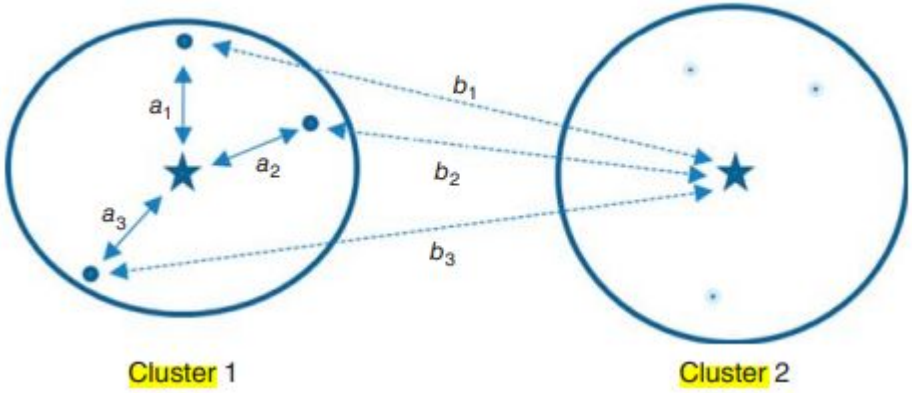
$$Silueta_i = s_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

Donde ai es la distancia entre el valor de los datos y su centro de grupo, y bi es la distancia entre el valor de los datos y el siguiente centro de conglomerado más cercano.
El valor de silueta se usa para medir qué tan buena es la asignación de clúster para ese punto particular.

Valor + : Indica que la asignación es buena, siendo mejores los valores más altos que los valores más bajos

Valor cercano a cero: se considera una asignación débil, ya que la observación podría haber sido asignada al siguiente conglomerado más cercano

Valor -: Está mal clasificado, ya que la asignación al siguiente grupo más cercano hubiera sido mejor.



INTERPRETACIÓN DEL VALOR PROMEDIO DE LA SILUETA

- 0.5 o mejor. Buena evidencia de la realidad de los clusters en los datos.
- 0,25–0,5. Alguna evidencia de la realidad de los clusters en los datos. Es de esperar que se pueda aplicar el conocimiento específico del dominio para respaldar la realidad de los clústeres.
- Menos de 0,25. Poca evidencia de la realidad del clúster

Ejemplo del método de silueta

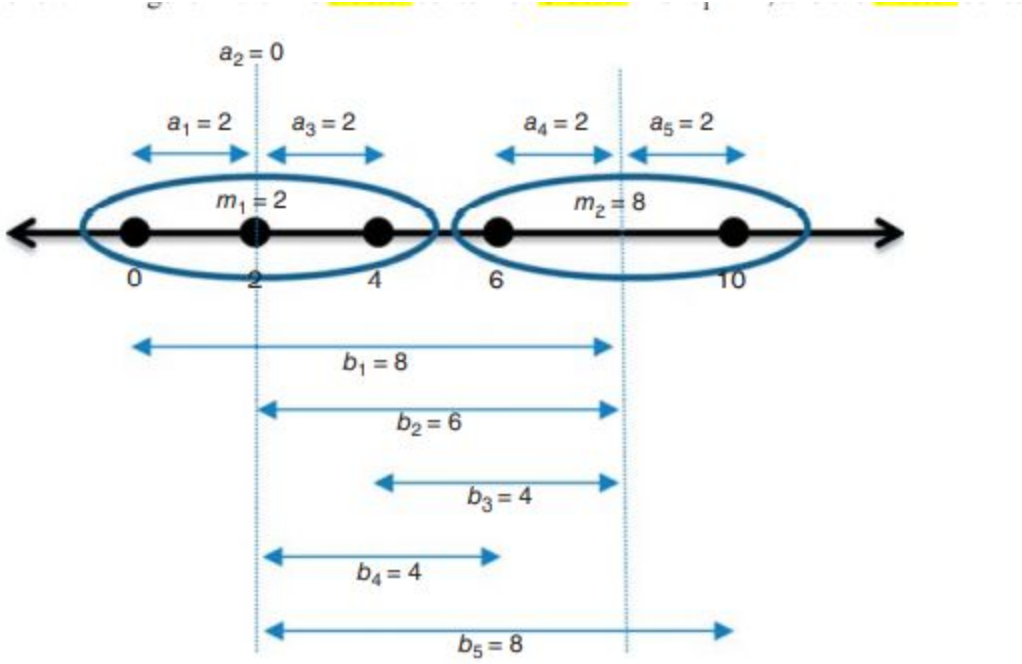
Supongamos que aplicamos la agrupación de k-medias al siguiente pequeño conjunto de datos unidimensionales:

$$x_1 = 0 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 6 \quad x_5 = 10$$

k-means asigna los primeros tres valores de datos al Cluster 1 y los dos últimos al Cluster 2, como

se muestra en la Figura 22.2.

Distances between the data values and the cluster centers



Utilizando nuestra regla de oro, la silueta media = 0,7 representa una buena evidencia de la realidad de los conglomerados en los datos. Tenga en cuenta que x2 está perfectamente clasificado como perteneciente al grupo 1, ya que se encuentra justo en el centro del grupo m1; por lo tanto, su valor de silueta es un 1,00 perfecto. Sin embargo, x3 está un poco más lejos de su propio centro de clúster y algo más cerca del otro centro de clúster; por lo tanto, su valor de silueta es menor, 0.50

TABLE 22.1 Calculations for individual data value silhouettes and mean silhouette

x_i	a_i	b_i	$\max(a_i, b_i)$	Silhouette$_i = s_i = \dfrac{b_i - a_i}{\max(b_i, a_i)}$
0	2	8	8	$\dfrac{8-2}{8} = 0.75$
2	0	6	6	$\dfrac{6-0}{6} = 1.00$
4	2	4	4	$\dfrac{4-2}{4} = 0.50$
6	2	4	4	$\dfrac{4-2}{4} = 0.50$
10	2	8	8	$\dfrac{8-2}{8} = 0.75$
				Mean silhouette = 0.7

La PSEUDO-F estadística

La estadística PSEUDO-F

-Supongamos que tenemos k clusters, con ni valores, por lo tanto $\Sigma ni = N$, es el

tamaño total de la muestra.

-Refiramonos a *xij* como el valor *j*th en el *i*th cluster.

-*mi* se refiere al centroide del cluster.

-*M* representa la gran media de todos los datos.

Entonces, definimos SSB, como la suma de los cuadrados entre los clusters:

$$\text{SSB} = \sum_{i=1}^{k} n_i \cdot \text{Distance}^2(m_i, M)$$

Definimos SSE, como la suma de los cuadrados sin los clusters:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_j} \text{Distance}^2(x_{ij}, m_i)$$

La estadística PSEUDO-F

Donde la distancia:

$$\text{Distance}(a,b) = \sqrt{\sum (a_i - b_i)^2}$$

La PSEUDO-F estadística **mide la relación de (i) la separación entre los clusters, medido por MSB, el cuadrado medio entre los grupos, a (ii) la propagación de los datos dentro de los clusters, medidos por el error del cuadrado medio, MSE.**

Luego la PSUEDO-F estadística queda como:

$$F = \frac{\text{MSB}}{\text{MSE}} = \frac{\text{SSB}/k-1}{\text{SSE}/N-k}$$

Se tienen dos hipótesis:

Ho: No hay clusters en los datos.

Ha: Existen K clusters en los datos.

Rechazar Ho por un valor p suficientemente pequeño, donde:

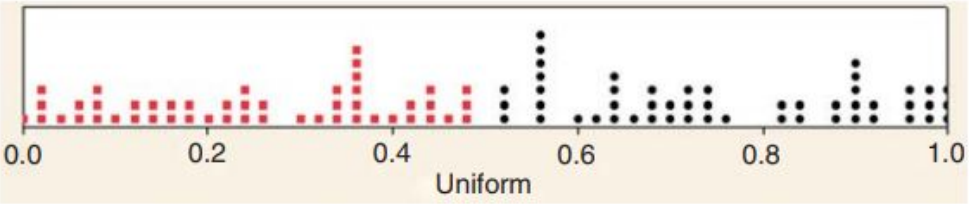
$$p\text{-value} = P(F_{k-1,n-k} > \text{pseudo-}F \text{ value})$$

La estadística PSEUDO-F

La razón por la que llamamos a esta estadística pseudo-F, es por que rechaza la hipótesis nula por mucho.

Por ejemplo, 100 valores aleatorios uniformes fueron dibujados y se le dijo a k-Means que encontrará k=2 agrupaciones dentro de los datos.

k-Means encontró debidamente los grupos como se muestra en la imagen, simplemente separando
los valores de datos superiores a 0,5 de los inferiores a 0,5.



El resultado de la PSEUDO-F estadística queda de la siguiente manera:

$$F = \frac{\mathrm{SSB}/k-1}{\mathrm{SSE}/n-k} = \frac{6.4606/1}{2.2725/98} = \frac{6.4606}{0.0232} = 278.61$$

La estadística PSEUDO-F

Si tenemos razones para creer que existen clusters en los datos, y no sé cuántos grupos hay, entonces el pseudo-F puede ser útil.

1. Utilice un algoritmo de agrupación en clústeres para desarrollar una solución de agrupación para una variedad de valores de k.

2. Calcule el PSEUDO-F estadístico y el valor p para cada candidato, y seleccione el candidato con el valor p más pequeño como la mejor solución de agrupamiento.

EJEMPLO PSEUDO-F

-Recordando que aplicamos la agrupación de k-means al siguiente conjunto de datos, y encontramos, para k = 2, k-means asigna los primeros tres valores de datos al cluster 1 y los dos últimos al cluster 2.

$x_1 = 0 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 6 \quad x_5 = 10$

Calculando la PSEUDO-F estatística para la desagrupación.

Tenemos k = 2 grupos, con n1 = 3 y n2 = 2 valores de datos, y N = 5. Los centros de los clusters son m1 = 2, m2=8 y la gran media es M = 4.4. Porque estamos en una dimensión, Distancia (mi, M) = | mi - M |, entonces:

$$\text{SSB} = \sum_{i=1}^{k} n_i \cdot \text{Distance}^2(m_i, M)$$
$$= 3 \cdot (2 - 4.4)^2 + 2 \cdot (8 - 4.4)^2 = 43.2$$

EJEMPLO PSEUDO-F

$x_1 = 0 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 6 \quad x_5 = 10$

Después.

$$\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_j} \text{Distance}^2(x_{ij}, m_i)$$

$$= (0-2)^2 + (2-2)^2 + (4-2)^2 + (6-8)^2 + (10-8)^2 = 16$$

Luego, la PSEUDO-F estadística queda como:

$$F = \frac{\text{MSB}}{\text{MSE}} = \frac{\text{SSB}/k-1}{\text{SSE}/N-k} = \frac{43.2/1}{16/3} = \frac{43.2}{5.33} = 8.1$$

La PSEUDO-F estadística aplicado al conjunto de Iris

A continuación, vemos qué valor de k es favorecido por la PSEUDO-F estatística para agrupar al conjunto Iris.

Para el vector **(longitud del sépalo, ancho del sépalo, largo del pétalo, ancho del pétalo)**, tienen N = 150 valores de datos, con la gran media.

$$M = (0.4287, \quad 0.4392, \quad 0.4676, \quad 0.4578)$$

Para k = 3 grupos, tenemos los siguientes recuentos y centros de grupos:

- Cluster 1: $n_1 = 50$ and $m_1 = (0.1961, \quad 0.5908, \quad 0.0786, \quad 0.06)$
- Cluster 2: $n_2 = 39$ and $m_2 = (0.7073, \quad 0.4509, \quad 0.7970, \quad 0.8248)$
- Cluster 3: $n_2 = 61$ and $m_3 = (0.4413, \quad 0.3074, \quad 0.5757, \quad 0.5492)$

La PSEUDO-F estadística aplicado al conjunto de Iris

Luego aplicando SSB a cada cluster:

- Cluster 1: $50 \times \{(0.1961 - 0.4287)^2 + (0.5908 - 0.4392)^2 + (0.0786 - 0.4676)^2 + (0.06 - 0.4578)^2\}$
- Cluster 2: $39 \times \{(0.7073 - 0.4287)^2 + (0.4509 - 0.4392)^2 + (0.7970 - 0.4676)^2 + (0.8248 - 0.4578)^2\}$
- Cluster 3:61 $\times \{(0.4413 - 0.4287)^2 + (0.3074 - 0.4392)^2 + (0.5757 - 0.4676)^2 + (0.5492 - 0.4578)^2\}$

Sumando los 3 clusters nos queda:

$$\text{SSB} = \sum_{i=1}^{k} n_i \cdot \text{Distance}^2(m_i, M) = \quad 34.1397.$$

La PSEUDO-F estadística aplicado al conjunto de Iris

Para saber SSE se obtiene el cuadrado de la distancia de cada observación y su centro y después sumando todas las entradas, obteniendo:

SSE=6.9891

Luego la fórmula queda:

$$F = \frac{\text{MSB}}{\text{MSE}} = \frac{SSB/k-1}{SSE/N-k} = \frac{34.1397/2}{6.9981/147} = 358.5$$

Validación de clúster

Validación de clusters

Como con cualquier otra técnica de modelado de minería de datos, la aplicación de un análisis

debe estar sujeto a validación cruzada, para asegurar que los grupos sean reales, y

no solo como resultado del ruido aleatorio en el conjunto de datos de entrenamiento.

Objetivo: Confirmar que los grupos encontrados en el conjunto de datos de prueba coincidan con los encontrados en el conjunto de datos de entrenamiento.

Metodología de validación de clusters

1. Aplique el análisis de clusters al conjunto de datos de entrenamiento.

2. Aplique el análisis de clusters al conjunto de datos de prueba.

3. Utilice gráficos y estadísticas para confirmar que los grupos del conjunto de datos de entrenamiento coinciden con los clústeres en el conjunto de datos de prueba.

Gracias por su atención.