Validación

Nicolás Kossacoff

Octubre 2024

1. Introducción

Hay varios métodos de validación que se pueden aplicar sobre los resultados de nuestro algoritmo:

- **Relativa.** Evalúa la estructura de los clusters modificando los distintos hiperparámetros utilizados para el entrenamiento. Se suele utilizar para determinar el número óptimo de clusters para el problema.
- **Externa.** Compara los resultados de nuestro modelo con información externa disponible. Usualmente, esta información proviene de un experto en el problema en cuestión.
- **Interna.** Utiliza los resultados del modelo para medir la bondad de la estructura de nuestros clusters.
- **Estabilidad.** Genera pequeñas modificaciones a nuestro conjunto de datos con el objetivo de evaluar la consistencia de los clusters (i.e., que tan similares son en distintos contextos).

2. Validación externa

Existen muchos tipos de información externa que podemos tener a disposición:

- La información puede venir de un experto en el problema en cuestión.
- Podemos tener variables que, a priori, sabemos que pueden estar relacionadas con la partición (e.g., cuando trabajamos con países, una partición muy común es la de economías desarrolladas, economías en vías de desarrollo o economías sub-desarrolladas).
- Podemos contar con las etiquetas de antemano. En estos casos , el problema de clustering ya se encuentra resuelto (i.e., ya conocemos a que grupo pertenece cada observación). Sin embargo, estos conjuntos de datos suelen ser muy útiles como **benchmarks** para comparar el rendimiento de distintos métodos de clustering.

Algunos de los índices que se suelen utilizar para medir el rendimiento de los algoritmos son:

- Tasa de clasificación correcta (CCR).
- Rand Index.
- Adjusted Rand Index.

3. Estabilidad

Generalmente, cuando evaluamos la estabilidad de nuestro algoritmo buscamos:

1. Generar nuevos conjuntos de datos a partir del conjunto original.
2. Para cada uno de los nuevos conjuntos de datos, ajustar el modelo y calcular un estadístico que mida la similaridad entre los nuevos clusters con los originales.
3. Finalmente, determinamos si los nuevos clusters son similares a los originales. Esto nos dará una idea de la estabilidad del algoritmo: si ante pequeñas variaciones de nuestro conjunto de datos los clusters resultantes difieren significativamente, entonces eso implica que el algoritmo es poco estable.

Es importante tener en cuenta que, si bien tener estabilidad es una buena propiedad, no nos garantiza que nuestros clusters sean 'buenos'. Es decir, podemos tener un algoritmo que devuelve clusters estables pero que clasifican incorrectamente las observaciones.

3.1. Algoritmo de Hening

3.1.1. Índice de Jaccard

El índice de Jaccard mide la similaridad entre dos conjuntos, comparando la cantidad de elementos en la intersección con respecto a la cantidad de elementos en la unión. Si ambos conjuntos son similares, entonces tienen muchos elementos en común, y el índice de Jaccard toma un valor cercano a uno:

$$\gamma\left(C, D\right) = \frac{|C \cap D|}{|C \cup D|}$$

3.1.2. Algoritmo:

El algoritmo de Hening se resume a continuación.

1. Generar B muestras bootstrap a partir del conjunto de datos original. Se pueden utilizar otros métodos de muestreo, pero la interpretación de los resultados puede cambiar.
2. Para cada muestra $b = \{1, \ldots, B\}$:
    - Ajustamos el algoritmo elegido.

- Para cada $C \in \mathcal{C}$ (i.e., para cada cluster en la partición) calculamos el índice de Jaccard.

  Ahora, como las etiquetas asignadas a los clusters originales pueden no ser las mismas etiquetas asignadas a los nuevos clusters, calculamos el índice de Jaccard entre el nuevo cluster y cada uno de los clusters originales. Finalmente, nos quedamos con el índice de Jaccard más grande para el cluster C:

$$m_{b,C} = \max \gamma(C, D)$$

3. Calculamos el valor promedio del índice de Jaccard para cada C:

$$\bar{\gamma} = \frac{1}{|B|} \sum_{b \in B} m_{b,C}$$

Finalmente, y dado un corte de estabilidad (Hening propone 0.5), tomamos una decisión:

- Si el índice de Jaccard promedio para el cluster C es menor al corte de estabilidad, $C < ths$, decimos que el cluster no es estable.
- Si el índice de Jaccard promedio para el cluster C es mayor al corte de estabilidad, $C \gg ths$, decimos que el cluster es estable.

4. Validación interna

Este tipo de validación utiliza la información disponible (e.g., la información que obtenemos de los modelos) para validar nuestros resultados.

Muchos de los índices utilizados para la validación interna se construyen con las siguientes medidas:

- **Cohesión.** Mide que tan cercanas son las observaciones contenidas en cada cluster. La idea es que, si las observaciones son cercanas entre sí, entonces pertenecen al mismo cluster.
- **Separación.** Mide que tan separados están los clusters entre sí[1]. La idea es que si las observaciones pertenecen a distintos clusters, entonces deben ser lejanas.
- **Conexión.** Mide la relación entre una observación y las observaciones cercanas. Define un entorno alrededor de la observación.

4.1. Silhouette

El índice de Silhouette compara una medida de cohesión con una medida de separación. Entonces, dados nuestros datos $D = \{x_1, \ldots, x_n\}$ y considerando una partición $\mathcal{C} = \{C_1, \ldots, C_K\}$, para cada observación $i \in \{1, \ldots, n\}$ calculamos:

[1] Hay muchas maneras de calcular la distancia intra-clusters. Se puede calcular a partir de la distancia entre centroides o calculando las distancias entre observaciones de distintos clusters.

1. La disimilaridad media con el resto de las observaciones en el cluster (i.e., que tan distinta es esa observación del resto de las observaciones del cluster). Esta es una medida de cohesión.

$$a(i) = \frac{1}{|C| - 1} \sum_{x_j \in C, x_j \neq x_i} d(x_i, x_j)$$

2. La disimilaridad media con el resto de los clusters (i.e., que tan distinta es esa observación respecto de las observaciones en el resto de los clusters). Nos quedamos con la disimilaridad media entre la observación y el cluster más cercano. Esta es una medida de separación.

$$b(i) = \min_{B \in \mathcal{C}: B \cap C = \emptyset} \frac{1}{|B|} \sum_{z \in B} d(x_i, z)$$

Definimos el índice de Silhouette para la observación i como:

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

donde $s(i) \in [-1, 1]$. Este índice nos da una idea de que tan bien clasificada se encuentra la observación i.

Luego:

- Si $s(i) \to 1$, entonces la cohesión es muy cercana a cero (i.e., la observación es muy parecida al resto de las observaciones del cluster y, por lo tanto, las disimilaridad es cercana a cero) y la separación es muy cercana a uno (i.e., la observación se encuentra muy separada del resto de los clusters y su disimilaridad con las observaciones del cluster más cercano es cercana a uno).
- Si $s(i) \to -1$, entonces la cohesión es muy cercana a uno (i.e., la observación es muy distinta al resto de las observaciones del cluster y, por lo tanto, las disimilaridad es cercana a uno) y la separación es muy cercana a cero (i.e., la observación se encuentra muy cerca del resto de los clusters y su disimilaridad con las observaciones del cluster más cercano es cercana a cero).

También podemos calcular el índice de Silhouette medio para cada cluster en la partición. De esta manera podemos tener una idea de que tan bien clasificadas están las observaciones en el cluster.

4.2. Dunn-Index

Al igual que antes, el índice de Dunn compara la separación de los clusters con su cohesión. La fórmula general del índice es la siguiente[2]:

$$DI(C) = \frac{\min_{1 \leq i < j \leq K} Dist(C_j, C_i)}{\max_{1 \leq i \leq K} \text{diam}(C_i)}$$

[2]El valor final del índice depende de como calculamos la distancia entre clusters.

donde diam(C_i) se calcula como la distancia máxima entre las observaciones del cluster (nos da una idea de que tan compactos son) y $Dist(C_j, C_i)$ puede ser cualquier tipo de distancia entre clusters (como, por ejemplo, single-linkage o complete-linkage). Luego:

- Valores grandes del índice de Dunn implican clusters bien separados y compactos (i.e., separación y cohesión alta).
- Valores chicos del índice de Dunn implican clusters poco separados y poco compactos (i.e., separación y cohesión baja).

5. Cantidad óptima de clusters

Hay varios métodos que nos permiten encontrar la cantidad óptima de clusters, aunque no existe una forma cerrada para la elección. También es válido utilizar información externa o la opinión de un experto en estos casos.

5.1. Elbow Method

Es un método heurístico que se basa en graficar la varianza intra-grupo para distintas particiones. Como la varianza intra-grupo es decreciente en el número de clusters, K, este método propone quedarse con el valor de K a partir del cual el decrecimiento comienza a ser más lento. Gráficamente:

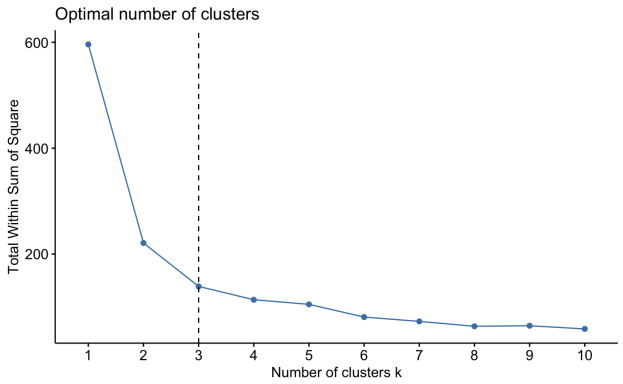

Figura 1: Elbow method con $K = 3$ clusters.

5.2. Max-Mean Silhouette

Para distintos tamaños de la partición, calculamos el índice de Silhouette medio, el cual nos da una idea de que tan cohesionadas se encuentran las observaciones en los clusters:

$$S_K = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

Luego, la partición óptima es aquella que nos devuelve los clusters más compactos y separados. Para eso nos quedamos con el máximo índice de Silhouette medio:

$$K_{opt} = \arg \max_{K \geq 2} S_K$$

La Figura 2 muestra el número óptimo de clusters calculados con el índice de Silhoutte medio. También se pueden utilizar otros índices, como el índice de Dunn, siguiendo el mismo criterio.

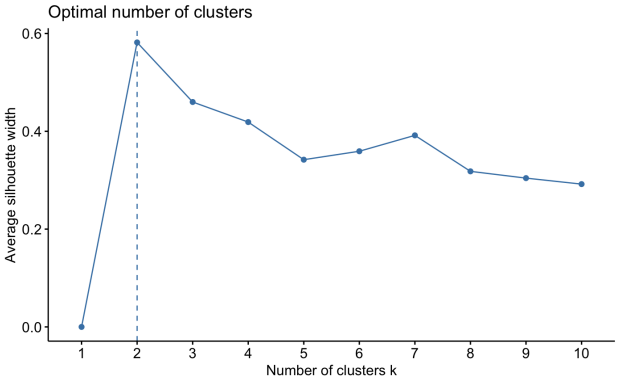

Figura 2: Max-Mean Silhouette con $K = 2$ clusters.

5.3. Gap Statistic

La idea de este método es comparar la distribución de nuestros datos con una distribución que, de antemano, sabemos que no tiene ningún grupo. Por lo general se utiliza una distribución uniforme.

El método compara la curva $\log(W_K)$, donde W_K es la dispersión intra-grupo para la partición de tamaño K, con la misma curva para la distribución uniforme. El número óptimo de clusters es aquel en donde ambas curvas se encuentran más separadas. Graficamente:

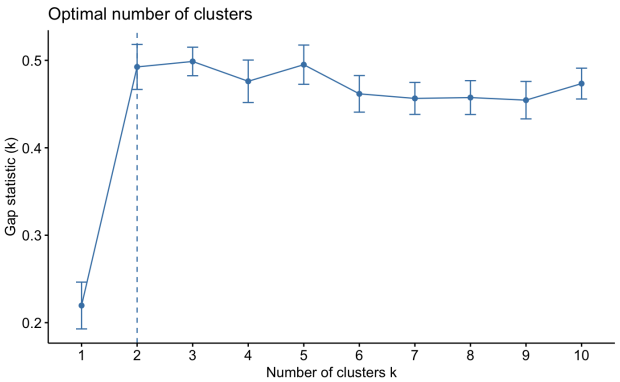

Figura 3: Gap Statistic con $K = 3$ clusters.