Cómo ajustar y evaluar algoritmos basados en densidad

Los algoritmos de aprendizaje tienen parámetros que debemos ajustar a los datos para poder obtener buenas soluciones. En el caso de DBSCAN, estos son:

- El radio de la vecindad (épsilon, Eps), el cual define el tamaño de la región a considerar en torno a un dato.
- El mínimo número de puntos dentro de la vecindad (minPtos), que es utilizado para decidir cuándo un punto se puede considerar como núcleo.

Es importante entender el impacto de los valores de los parámetros de un algoritmo en el resultado. Por ejemplo, para DBSCAN, un valor bajo de minPtos puede generar muchos grupos (al considerar más datos como núcleos) y que puntos aislados generen grupos dispersos, mientras que un valor alto puede generar menos grupos, al exigir una densidad muy alta. Por su parte, un valor demasiado pequeño para Eps puede conducir a que muchos puntos sean considerados como ruido o atípicos, mientras que uno demasiado grande puede hacer que los posibles grupos se fusionen en uno solo. Así, la elección de la mejor combinación (minPtos, Eps) dependerá de la distribución de los datos.

En general, hallar una buena combinación de parámetros requiere experimentación. Una manera de sistematizar esta "búsqueda" en un contexto no supervisado es utilizando una métrica de evaluación a partir de la cual podamos juzgar la calidad del modelo generado. Se puede entonces construir una "grilla" de posibles valores de minPtos y Eps y generar una agrupación para cada combinación, con su respectivo valor de la métrica de evaluación. Al final, se seleccionaría la combinación que arroje el mejor valor para esta (Liu *et.al.*, 2010). El proceso es similar al método del codo o el de la silueta para determinar el valor de K para un algoritmo como K-medias y asimismo es aplicable a otros algoritmos basados en densidad.

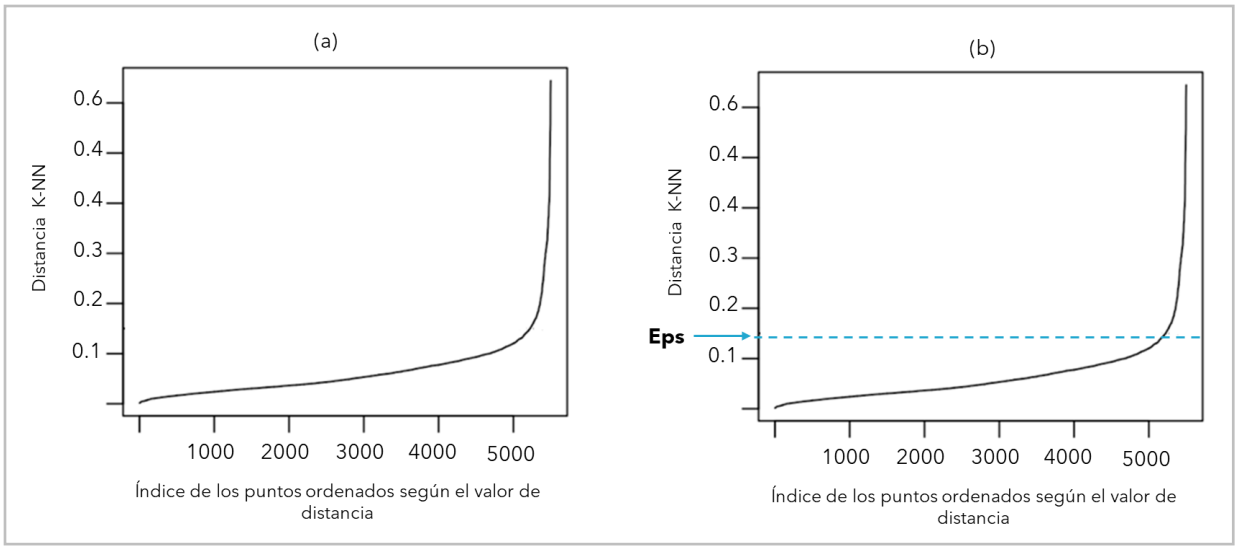
También se han propuesto heurísticas para determinar los valores de estos parámetros. Acá te describo algunas para el caso de DBSCAN:

Para MinPtos. Lo recomendable es que se asigne un valor utilizando conocimiento del dominio de aplicación. Sin embargo, muchas veces esto no es posible, por lo que se pueden probar algunas reglas empíricas que derivan este valor a partir de la dimensionalidad del conjunto de datos. Por ejemplo, $minPtos \geq d + 1$ (donde d es el número de características); o si el conjunto de datos es muy ruidoso, $minPts = 2 * d$.

Para Eps. Una forma de determinar este parámetro es mediante el método de k-distancias, donde k = minPtos. El procedimiento es el siguiente:

1. Calcular la distancia promedio de cada punto a sus k-vecinos más cercanos.
2. Ordenar los puntos en orden ascendente con base en las distancias calculadas en el paso anterior.
3. Graficar las k-distancias como se muestra en la Fig. 1.a, donde el eje x representa los puntos ordenados y el eje y representa la distancia.
4. El valor de Eps será la distancia correspondiente al punto de "codo" en el gráfico (ver Fig. 1.b), el cual generalmente marca el umbral a partir del cual las distancias comienzan a crecer rápidamente.

Fig. 1. Método de las K-distancias para estimar el parámetro Eps.



Métricas de evaluación para algoritmos basados en densidad

Para evaluar este tipo de modelos se pueden utilizar tanto métodos intrínsecos como extrínsecos. Por ejemplo, una métrica ampliamente utilizada en este contexto es el coeficiente de silueta, pero se pueden utilizar otras, como el índice de Davies-Bouldin. En caso de contar con datos etiquetados, se podría utilizar cualquier métrica que considere el *ground truth*, como el índice Rand.

También se han propuesto métricas específicas para algoritmos basados en densidad. Una de ellas es el índice DBCV (*Density-based Clustering Validation*), el cual evalúa una agrupación con base en la densidad intragrupo e intergrupo (Moulavi *et al*., 2014). Otra que podemos mencionar es el índice CDbw (*Composing Density Between and Within Cluster*), que al igual que el anterior. mide la compacidad y separabilidad evaluando la densidad de la distribución dentro y entre grupos. En ambos, valores cercanos a 1 sugieren grupos de buena calidad.

Para finalizar, te dejo estas preguntas relacionadas con esta lectura:

- La métrica de distancia utilizada por el algoritmo ¿puede ser considerada como un hiperparámetro? ¿Crees que puede tener impacto en la calidad de los resultados con base en las características del conjunto de datos?
- ¿Es posible utilizar la inercia para evaluar este tipo de modelos?

Bibliografía

Halkidi, M., Vazirgiannis, M. (2008). *A density-based cluster validity approach using multi-representatives*. Pattern Recognition Letters. 29 (6):773-786.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J. (2010). *Understanding of Internal Clustering Validation Measures*. IEEE International Conference on Data Mining. DOI: 10.1109/ICDM.2010.35.

Moulavi, D., Jaskowiak, P., Campello, R., Zimek, A. Sander, J. (2014). *Density-Based Clustering Validation*. Proceedings of the 14th SIAM International Conference on Data Mining (SDM). 839-847.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.