Valores atípicos

Carlos Malanche

22 de febrero de 2018

Vamos primero a definir lo que es un valor atípico, pero vamos a hacerlo con una frase:

Un valor atípico es una observación que se aleja tanto de la mayoría de las otras observaciones que levanta la sospecha de no haber sido generada por el mismo mecanismo que el resto

Hay que ser muy cuidadosos con el manejo de valores atípicos, pues estos pueden tener distintas razones de fondo, por ejemplo:

- Error en la captura de la información
- Error en la transmisión de la información
- Error en el manejo de la información
- Errores experimentales
- Una *novedad*

Se le llama *novedad* a un valor atípico que se está considerando como tal por no prever su existencia, es decir, estos valores NO se deben tirar y se debe replantear la teoría sobre la que trabajamos para incluír su existencia.

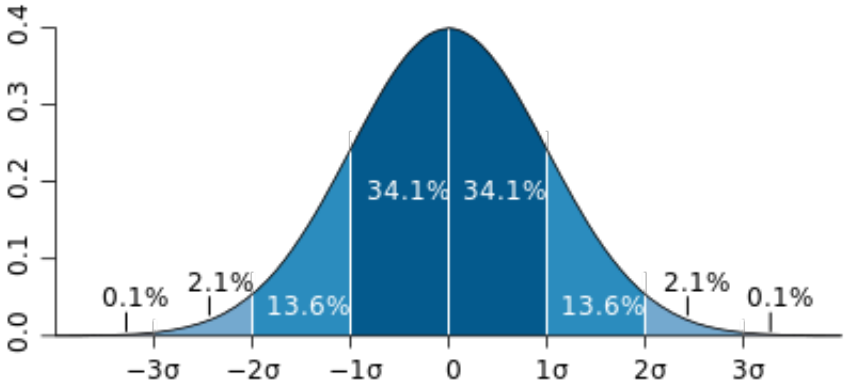
Existen diversos métodos para encontrar valores atípicos, pero nosotros nos vamos a enfocar por el momento en 3 pues 2 son los clásicos y uno algo nuevo.

1. Z-Score

Para utilizar este método, se asume que la serie de datos sigue una distribución Gaussiana, lo cual implica que la varianza es una parámetro estadístico pertinente. Dada la desviación estándar de la serie, se coloca un valor límite (usualmente un múltiplo de la desviación estándar) a partir de la cual se consideran valores atípicos. Un poco más formal, de la serie $S = \{s_i\}_{i=1}^n$ se deriva una serie de *Z-scores* $Z = \{z_i\}_{i=1}^n$ donde

$$z_i = \frac{s_i - \mu}{\sigma} \tag{1}$$

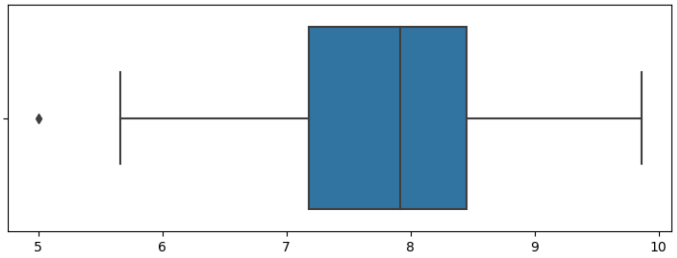
es el número de desviaciones estándar que el valor de la variable aleatoria s_i está alejado de la media.



En la figura anterior está la cantidad de información contenida bajo la curva de una distribución normal. En base a esta, valores límite populares para deshacernos de posibles valores atípicos son al menos 2.5 desviaciones estandar.

2. *Interquartile range* (IQR)

Como lo vimos para ver las gráficas de *caja* (en inglés *box with whiskers*), se utilizan los cuartiles para encontrar posibles outliers.



Bajo este esquema, se clasificará como valor atípico cualquier valor s_i de la serie que no esté contenido en el rango

$$(Q_1 - 1{,}5 * IQR, Q_3 + 1{,}5 * IQR) \tag{2}$$

En donde Q_1 y Q_3 son el primer y tercer cuartil respectivamente, y la región intercuartil IQR es la diferencia de los mismos ($IQR = Q_3 - Q_1$).

Si usted se pregunta *por qué exactamente 1.5 y no 1.123 ó $\pi/2$* la respuesta es muy sencilla: John Turkey, creador de la gráfica de caja, decidió arbitrariamente que 1.5 era un buen múltiplo para detectar valores atípicos. La gente lo siguió usando dado que dio *buenos resultados*, y entonces ya no se cuestiona realmente.

Sin embargo, asumiendo que se tiene una distribución normal, utilizando la función acumulativa de distribución se puede notar que $Q_3, Q_1 \approx \pm 0{,}68\sigma$, con eso obtenemos que $IQR \approx 1{,}36\sigma$, lo que nos dice que un valor atípico será todo aquel que esté más de $2{,}72\sigma$ lejos de la media, lo cual es masomenos el 1 % de la información bajo una curva de Bell.

3. DBSCAN

Las siglas significan ***D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise* [1], y es un algoritmo de clustering; así es, emociónense que nuestro primer algoritmo de machine learning ha llegado.

Aunque el propósito del algoritmo es encontrar clusters de información con un poco de ruido, ha adquirido popularidad al usarse para detectar valores atípicos. Piense en el siguiente escenario: Se hacen mediciones de un evento cuya distribución de probabilidad son dos gaussianas de mismos parámetros estadísticos, con sus medias separadas por 10σ. Un valor que se encontrara a la mitad no se podría detectar asumiento una distribución Gaussiana, y tampoco buscando un valor extremo.

La mentalidad del algoritmo es buscar la información que se encuentre *cercana* para agruparla. Si un elemento no parece tener un grupo, entonces quedará marcada como valor atípico.

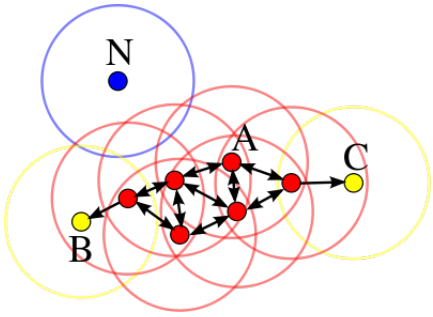
El algoritmo depende principalmente de una métrica establecida con la que se determina la distancia entre dos puntos de nuestros datos (puede ser una simple métrica euclidiana, o la de *Manhattan*), y de dos parámetros más: ϵ el radio de la vecindad conforme la métrica seleccionada y $minPts$, el número mínimo de vecinos para formar el núcleo de un *cluster*. A los datos se les asigna una métrica que los coloque espacialmente en un espacio. La métrica puede contemplar una de las variables que describen los puntos, un subconjunto del total, o todas.

Vamos a añadir las siguientes tres definiciones para el método:

- Punto núcleo: Un punto es un punto núcleo si en su vecindad de radio ϵ hay al menos tantos puntos como el parámetro $minPts$ lo indica.
- Punto frontera: Un punto es un punto frontera si tiene menos puntos en su vecindad de radio ϵ de lo que $minPts$ indica, pero en su vecindad contiene un punto núcleo al menos.

- Valor atípico: Los puntos que no caen en los primeros dos casos son valores atípicos.

Al calcular el número de elementos que vive en un cluster con respecto a un punto, se debe incluír el punto mismo en la cuenta.



En el ejemplo de arriba, se tiene que $minPts = 4$ (ϵ es gráfico nadamás). El algoritmo se corre en un orden específico:

- Primero se buscan todos los puntos núcleo, ignorando cualquiera que no califique.
- Con los puntos núcleo se hacen los clusters correspondientes (se ponen en un solo grupo los puntos que son vecinos).
- Los puntos restantes se separan en dos grupos: Los que se pueden anexar a un cluster, y los que son valores atípicos (*ruido*).

Al final, quedamos con un grupo de clusters que por el momento son indistinguibles. Los clusters de 1 elemento son marcados como valores atípicos. Como lo veremos en clases posteriores, el truco de este último método está en estimar correctamente los parámetros ϵ y $minPts$. $minPts$ tendrá el efecto de definir el número mínimo de elementos en un cluster, y ϵ qué tan *similar* queremos que la iformación sea (asumiendo que nuestra métrica mide similitud). Algunas estrategias comunes son calcular la interdistancia de todos los elementos. Dependiendo del histograma resultante de este cómputo, se utiliza una adivinanza de ϵ para contar cuántos vecinos tiene cada elemento y ver ahora qué porcentaje de la información será tirado en función de $minPts$.

Referencias

[1] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. *A density-based algorithm for discovering clusters in large spatial databases with noise*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.