Lección 1. Métodos de detección de anomalías basados en estadística

UCO ONLINE

La detección de anomalía

Métodos de detección de anomalías basados en estadística

La detección de anomalías ha sido abordada desde una gran variedad de técnicas diferentes. El tipo de anomalías que se quieren detectar, el tipo de dato de los datos de entrada, si los datos están etiquetados o no y qué tipo de etiquetas usan, nos va a delimitar qué técnica podemos usar.

En esta sección veremos los métodos estadísticos, los cuales no necesitan que los datos estén etiquetados. La mayoría de ellos están basados en datos univariantes y aunque pueden extenderse fácilmente a trabajar con datos multivariantes, si existen dependencias entre ellos, no lo suelen tener en cuenta. Finalmente, la mayoría de ellos necesitan que los datos sean numéricos, aunque existen estadísticos que también trabajan con datos categóricos.

Estas técnicas son ampliamente utilizadas para realizar el análisis exploratorio de datos (EDA, *Explorative Data Analysis*). EDA resulta crucial cuando se trabaja en proyectos de ciencia de datos para poder aplicar las técnicas más adecuadas. Conocer los datos por dentro y por fuera puede simplificar la toma de decisiones con respecto a la selección de características, algoritmos e hiperparámetros. Una parte esencial de la EDA es la detección de valores atípicos o anomalías y determinar si son relevantes o deberían eliminarse.

1. Introducción

Los primeros métodos que se utilizaron para la detección anomalías estaban basados en modelos estadísticos. Estos métodos incluso fueron propuestos mucho antes de todos los avances que se han realizado en la ciencia de datos y en el cálculo de modelos con alto coste computacional y con complejas representaciones. Sin embargo, los modelos matemáticos subyacentes son muy útiles y eficientes, por lo que actualmente se siguen utilizando y se adaptan e hibridan con otros métodos.

Las técnicas estadísticas que se aplican para la detección de anomalías se basan en que una anomalía es una observación que se presupone parcial o totalmente irrelevante porque no ha sido generada por el modelo estocástico que se asume [1]. Por ello, estas técnicas estadísticas asumen que: "*las instancias normales están en regiones donde los modelos estadísticos determinan una alta probabilidad de ocurrencia, mientras que las anomalías ocurren en regiones donde los modelos estadísticos estiman una baja probabilidad de ocurrencia*".

Las técnicas estadísticas ajustan un modelo estadístico (generalmente para un comportamiento normal) para los datos de entrada y luego aplican un test de inferencia estadístico para determinar si una nueva instancia pertenece a este modelo o no. Las instancias que tienen una baja probabilidad de ser generadas a partir del modelo aprendido, basado en el test estadístico aplicado, se declaran como anomalías.

Dentro de los modelos estadísticos, podemos diferenciar técnicas estadísticas paramétricas y no paramétricas. Las técnicas paramétricas asumen que conocen la distribución subyacente que siguen los datos y presuponen una distribución normal. Las técnicas no paramétricas parten de que no conocen la distribución de los datos y son más generales. En esta lección, veremos varios ejemplos de cada una de ellas, y finalmente, se darán las ventajas e inconvenientes que presentan estos modelos.

Debemos recordar que la estadística solamente nos permite aceptar o rechazar que un dato se aleja significativamente de la distribución de los datos recogidos, pero una vez encontrado estos datos, debe ser otro el criterio que nos autorice científicamente para eliminarlo o añadirlo en nuestro estudio. Estos valores atípicos pueden sugerir errores experimentales, variabilidad en una medición dentro de la normalidad o una anomalía que debe tenerse en cuenta. Por ejemplo, la edad de una persona puede registrarse erróneamente como 200 en lugar de 20 años. Este valor atípico definitivamente debe descartarse del conjunto de datos. En otros casos, puede no ser un error e indican que realmente es un dato almacenado correctamente, pero es significativamente diferentes del resto. Por ejemplo, puede indicar un fraude bancario o una enfermedad rara.

De este modo, resulta muy relevante encontrar los datos que se consideran anormales y estudiarlos para ver qué se debe hacer con ellos. Podemos diferenciar tres importantes dominios por los que deben detectarse:

- Los valores atípicos afectan gravemente la desviación estándar y media del conjunto de datos. Estos pueden dar resultados estadísticamente erróneos.
- La mayoría de los algoritmos de aprendizaje automático no funcionan bien en presencia de valores atípicos. Por tanto, es deseable detectar y eliminar estos valores si realmente no deberían ser considerados.
- Los valores atípicos son muy útiles en la detección de anomalías, como la detección de fraudes, donde las transacciones de fraude presentan características diferentes de las transacciones normales.

2. Técnicas paramétricas

Las técnicas paramétricas parten de que los datos de entrada siguen una distribución paramétrica con parámetros θ y función de densidad de probabilidad $f(x, \theta)$, donde x es una observación o instancia. El *score* de anomalía de una nueva instancia x es la inversa de la función de densidad de probabilidad, $f(x, \theta)$. Los parámetros θ se estiman a partir de los datos de entrada.

Estas técnicas se basan en un test de hipótesis estadístico (también conocido como test de discordancia en el área de detección de anomalías). La hipótesis nula (H0) es que la instancia x ha sido generada utilizando la distribución estimada (con parámetros θ). Si el test estadístico rechaza H_0, x se define como una anomalía.

Los métodos que vemos aquí asumen que los datos siguen una distribución normal o gaussiana. La distribución normal es una distribución con forma de campana donde las desviaciones estándar sucesivas con respecto a la media establecen valores de referencia para estimar el porcentaje de observaciones de los datos (figura 1). Estos valores de referencia son la base de muchos tests de hipótesis. Aunque los métodos que veremos se basan en que los datos siguen una distribución normal, veremos en los ejemplos que aplicar un test para ver la distribución de los datos en datos con anomalías no nos van a dar ningunas garantías de la distribución; debido a que, si los datos tienen anomalías, aunque sus datos sigan una distribución normal, si se analizan junto con las anomalías, los distintos estadísticos nos determinarán que no se puede asumir una distribución normal.

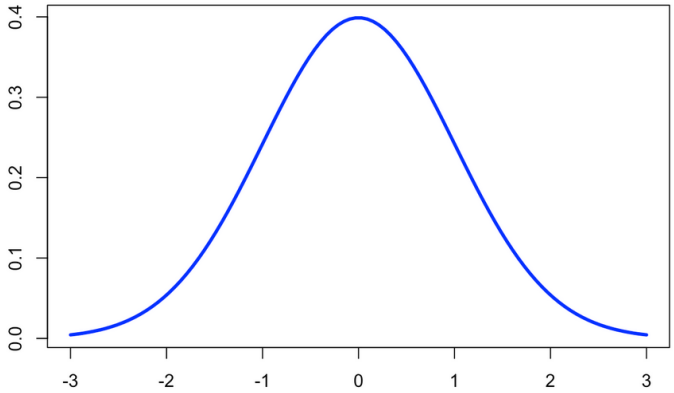

Figura 1. Función de distribución gaussiana

Las diferentes técnicas estadísticas se basan en estimaciones de máxima verosimilitud. El método de máxima verosimilitud nos dice que escogeremos como valor estimado de un parámetro aquél que tiene mayor probabilidad de ocurrir según lo que hemos observado, es decir aquél que es más compatible con los datos observados, siempre suponiendo que es correcto el modelo matemático postulado. Así, la distancia de una instancia de datos a la media estimada estima el *score* de anomalía para esa instancia. Se aplica un umbral a los *scores* de anomalías asignados para determinar cuáles se consideran verdaderamente una anomalía. Como veremos, las diferentes técnicas en esta categoría calculan la distancia a la media y el umbral de diferentes maneras [1,2].

2.1 Z-score

Esta técnica se basa en declarar como anómalas todas las instancias de datos que están a una distancia mayor de 3σ de la media de distribución μ, donde σ es la desviación estándar de la distribución. Esto se debe a que la región $\mu \pm 3$ contiene el 99.7% de las instancias de datos. Con lo que fuera de esa región, los datos se pueden considerar atípicos.

Si observamos la figura 2, vemos que un z-score nos revela dónde se encuentra la puntuación en una curva de dispersión típica. Una puntuación z de cero te revela que las cualidades son realmente normales mientras que una puntuación de +3 te revela que el valor es mucho más alto de lo normal o de -3 mucho más bajo.

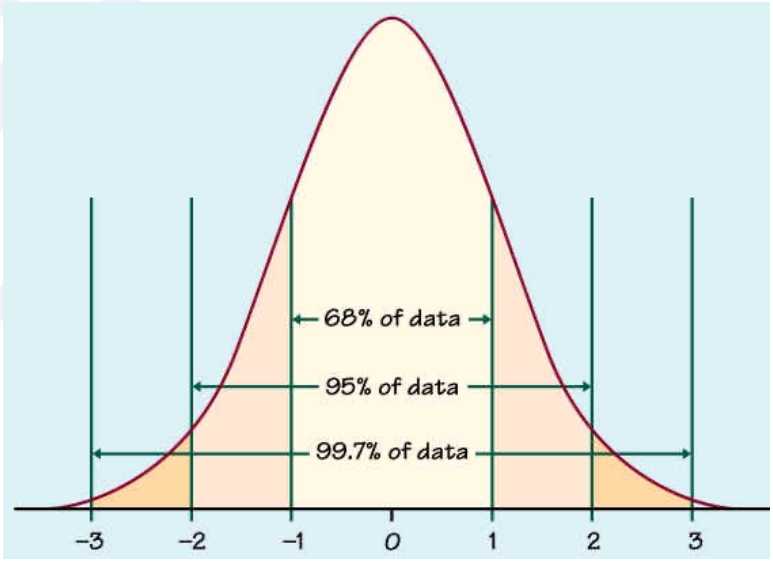

Figura 2. Relación de z-score en una distribución normal

Para utilizar un z-score, hay que conocer la media (μ) y además la desviación estándar (σ) de la población. Los resultados de las pruebas o estudios tienen un gran número de posibles resultados y unidades. Por ejemplo, darse cuenta de que el peso de alguien es de 150 gramos puede ser fácil, pero contrastarlo con el peso del individuo normal analizando millones de registros puede ser costoso si se hace forma manual. Z-score puede revelar fácilmente aquellas mediciones que se desvían considerablemente de la media normal de la población.

La ecuación de Z-score es:

$$z = (x - \mu)/\sigma$$

Por ejemplo, si estamos analizando los pesos de una población y tenemos un valor de 190 gramos y la media de la población (μ) es 90 con una desviación estándar (σ) de 25. El valor que nos daría el Z-score sería:

$$z = (x - \mu)/\sigma = (190 - 90)/25 = 4$$

Este valor de 4 nos releva el número de desviaciones estándar con respecto a la media. En este caso diríamos que se trata de un valor anómalo, ya que supera el valor de 3.

En el momento en que se tienen numerosos ejemplos y se necesita representar la desviación estándar que esos ejemplos implican (el error típico), se utiliza la ecuación:

$$z = (x - \mu)/(\sigma/\sqrt{N})$$

donde N es el tamaño de la población.

2.2 Test de Tukey

Este test también se conoce como el cálculo del rango intercuartílico (IQR, *InterQuartil Range*). Se usa para medir la variabilidad dividiendo un conjunto de datos en cuartiles. Los datos se clasifican en orden ascendente y se dividen en 4 partes: Q1, Q2 y Q3 (llamados primer, segundo y tercer cuartil) son los valores que separan las 4 partes.

- Q1 representa el percentil 25 de los datos.
- Q2 representa el percentil 50 de los datos.
- Q3 representa el percentil 75 de los datos.

IQR es el rango entre el primer y el tercer cuartil: IQR = Q3 - Q1. Los puntos de datos que caen por debajo de [Q1 - 1.5 * IQR] o por encima de [Q3 + 1.5 * IQR] se consideran valores atípicos o anormales.

Estos datos pueden verse muy bien reflejados en los diagramas de cajas (*bloxplot*). Estos diagramas se han aplicado desde hace mucho tiempo para detectar anomalías univariantes y multivariantes, ya que reflejan visualmente los datos que salen del IQR. Un *boxplot* representa gráficamente los datos usando atributos de resumen como la observación más pequeña que no se considera anómala (*min*), el cuartil inferior (Q1), la mediana, el cuartil superior (Q3) y el mayor valor de la observación que tampoco se considera anómalo (*max*).

Los *bloxplots* también indican los límites más allá de los cuales cualquier observación será tratada como una anomalía. La instancia de datos que se encuentra [1.5 × IQR] más baja que Q1 o [1.5 × IQR] más alta que Q3 se definen como anómalas. La región entre (Q1 - 1.5 × IQR) y (Q3 + 1.5 × IQR) contiene el 99.3% de las observaciones y, por lo tanto, la elección del límite 1.5 × IQR hace esta regla similar a la z-score vista anteriormente.

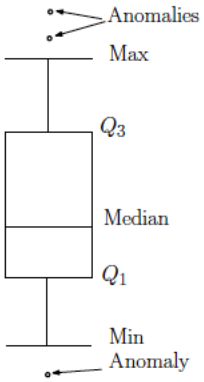

Figura 3.Boxplot para una variable

2.3 Test de Grubb

El test de Grubb (también conocido como prueba residual máxima normada) se utiliza para detectar anomalías en un conjunto de datos univariantes bajo el supuesto de que los datos son generados por una distribución normal. Para cada instancia x, se calcula el valor G, de forma similar a como se ha comentado con zscore:

$$G = (x - \mu)/\sigma$$

donde μ es la media y σ es la desviación estándar de la muestra de datos.

Una instancia se considera anómala si G es mayor que:

$$\frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

donde N es el tamaño de los datos y $t_{\alpha/(2N,\ N-2)}$ es el umbral utilizado para declarar una instancia anormal o normal. Este umbral es el valor que toma una distribución-t con un nivel significativo de $\alpha/2N$. El nivel de significación refleja la confianza asociada con el umbral e indirectamente controla el número de instancias declaradas como anómalas.

La prueba de Grubb se utiliza para identificar la presencia de valores atípicos en un conjunto de datos. Para utilizar esta prueba, el conjunto de datos debe tener una distribución aproximadamente normal.

2.4 Desviación absoluta mediana

La desviación absoluta mediana (MAD, *Median Absolute Desviation*) es la diferencia entre cada observación y la mediana de esas observaciones. Una observación que se desvía más del resto de la observación se considera una anomalía.

El valor MAD se calcula con la siguiente fórmula:

$$MAD = \frac{\sum_{x \in X} x - median}{N}$$

donde N es el tamaño de los datos, x la instancia de datos observada del conjunto de datos X y *median* la mediana de la población.

2.5 Determinante de Covarianza Mínima

El determinante de covarianza mínima (MCD, *Minimal Covariance Determinant*) [3]. Como el resto de los métodos comentados en esta sección, se basa en que los datos considerados como normales han sido generados por una distribución normal. De esta manera, partiendo de esta distribución de nuestros datos, nos resultará sencillo detectar las anomalías, ya que únicamente tendremos que comprobar cómo de probable es que un punto haya sido generado por esta distribución o no.

MCD trata de buscar las h observaciones de la muestra cuya matriz de covarianzas tiene el menor determinante. Para decidir qué puntos van a ser incluidos, este método utiliza la distancia de Mahalanobis.

La distancia de Mahalanobis, en lugar de medir la separación en el plano, como hace la distancia euclídea, busca cuantificar cómo de parecidos son dos puntos. En este caso, dados dos puntos p y q, ambos en el espacio R^n, la distancia de Mahalanobis entre ambos es:

$$d(p,q) = d(q,p) = \sqrt{(p-q)^T \sum\nolimits_{-1} (p-q)}$$

Donde Σ es la matriz de covarianza. Esta matriz debe ser definida positiva e invertible. Si Σ es la matriz de identidad, entonces la distancia de Mahalanobis es igual a la distancia Euclídea.

Otras técnicas como fast-mcd o el test t-student también son ampliamente utilizadas dentro de las técnicas paramétricas.

3. Técnicas no paramétricas

Las técnicas de detección de anomalías en esta categoría utilizan técnicas estadísticas no paramétricas. Tales técnicas no presuponen una distribución de probabilidad para los datos, por ello se conocen también como de distribución libre.

3.1 Histogramas

Esta técnica se basa en la frecuencia de los valores. Las técnicas basadas en histogramas son particularmente populares en la comunidad de detección de intrusos donde a partir de los datos se generan perfiles utilizando los histogramas.

Una técnica de detección de anomalías basada en histogramas se puede aplicar en dos pasos. El primer paso consiste en construir un histograma basado en los diferentes valores tomados por esa característica en los datos de entrada. En el segundo paso, se comprueba si una nueva instancia está dentro de alguno de las barras del histograma. Si lo hace, la nueva instancia se considera normal, de lo contrario es anormal. En otros análisis se utiliza una variante donde a cada instancia nueva se le asigna un *score* en función de la altura (frecuencia) de la barra en el que cae (figura 4).

Figura 4. Histograma para detección de anomalías (una variable)

El tamaño de la barra utilizada al construir el histograma es clave para la detección de anomalías. Si las barras son pequeñas, muchas instancias nuevas normales caerán en barras vacías o raras resultando en una alta tasa de falsas alarmas (falsos positivos). Si las barras son grandes, muchas instancias anómalas caerán en barras frecuentes, lo que dará como resultado una alta tasa de falsos datos normales (falsos negativos). Así el desafío en estas técnicas es determinar un tamaño óptimo de las frecuencias para construir el histograma que mantiene un bajo índice de falsas positivos y un bajo índice de falsos negativos.

Las técnicas basadas en histogramas requieren datos normales para construir los histogramas. Algunas técnicas construyen histogramas para las anomalías si hay disponibles instancias anómalas que están etiquetadas. Para datos multivariados, se construyen histogramas para cada atributo (figura 5). Durante el test, para cada instancia nueva, se asigna un s*core* de anomalía para cada valor de atributo de la instancia como la altura de la barra que contiene el valor del atributo. Los *scores* de anomalías por atributo se agregan para obtener un s*core* de anomalía para la instancia.

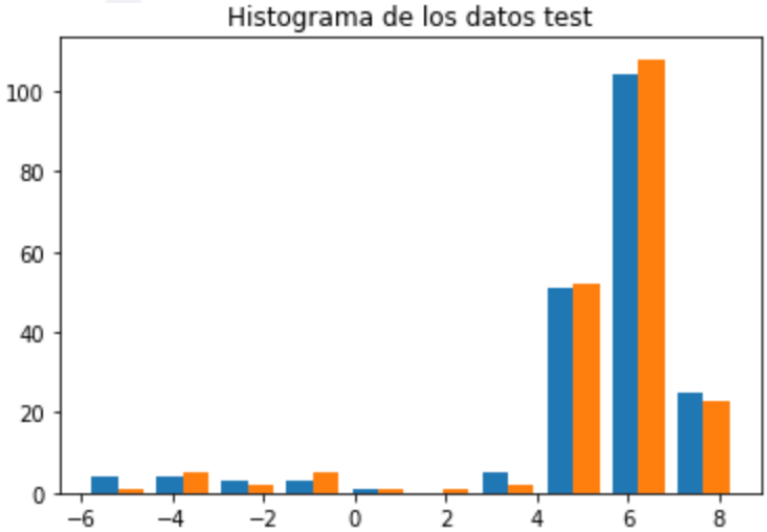

Figura 5. Histograma para detección de anomalías (dos variables)

3.2 Puntaje de valores atípicos basado en histogramas

El puntaje de valores atípicos basados en histogramas (HBOS, *Histogram-based Outlier Score*) [4] se basa en la idea anterior y se puede utilizar con datos multivariantes. Su principio se basa en histogramas. Es un algoritmo rápido y bastante eficiente. El parámetro crítico para el correcto funcionamiento es el número de barras (k). Una regla de uso frecuente para la obtención del valor k está fijando k a la raíz cuadrada del número de instancias.

Su funcionamiento se basa en el siguiente procedimiento. Para cada característica individual (dimensión) del conjunto de datos, primero se construye un histograma de variable única. Si la característica está comprendida por datos categóricos, se realiza el recuento de los valores de cada categoría y se computa la frecuencia relativa (altura del histograma). Los valores sucesivos se agrupan en una sola barra donde N es el número total de instancias y k el número de barras. Ya que el área de una barra en un histograma representa el número de observaciones, es el mismo para todos los contenedores en nuestro caso. Debido a que el ancho del contenedor está denotado por el primer y último valor, y es el mismo para todos, la altura de cada barra individual puede ser calculada. Esto significa que las barras que cubren un intervalo mayor del rango de valores tienen menos altura y representan así una menor densidad. Después, para cada dimensión d, se calcula un histograma individual donde la altura de cada una de las barras representa una estimación de la densidad.

Los histogramas se normalizan de tal manera que la altura máxima es de 1.0. Esto asegura un peso igual de cada dimensión para la anomalía. Finalmente, HBOS da a cada instancia x un valor que es una multiplicación del inverso de las densidades estimadas asumiendo independencia de las características y que se calcula utilizando la altura correspondiente de las barras donde se encuentra la instancia:

$$HBOS(x) = \sum_{i=0}^{d} \log\left(\frac{1}{hist\_i(x)}\right)$$

Otras técnicas no paramétricas pueden ser usadas para detección anomalías, por ejemplo, el test de $\chi 2$ [5] es también ampliamente utilizado o técnicas probabilísticas que también se clasifican con estos modelos habitualmente.

4. Coste computacional

El coste computacional de las técnicas estadísticas de detección de anomalías depende de la naturaleza del modelo estadístico que se requiere ajustar a los datos. El ajuste de las distribuciones paramétricas individuales de la familia exponencial, como la gaussiana, suele ser lineal con respecto al tamaño de datos y el número de atributos.

El ajuste de distribuciones complejas usando técnicas de estimación iterativas como la maximización de la esperanza, son lineales con respecto a las iteraciones, aunque pueden ser lentas en la convergencia dependiendo del problema y/o criterio de convergencia.

5. Ventajas y desventajas de las técnicas estadísticas

Entre las ventajas de las técnicas estadísticas podemos encontrar:

(1) Si las suposiciones con respecto a la distribución de datos son verdaderas, estadísticamente estas técnicas proporcionan una solución significativa estadísticamente para la detección de anomalías. Sus resultados son fáciles de explicar a los expertos del dominio.

(2) El *score* de la anomalía proporcionada por una técnica estadística está asociado con un intervalo de confianza, que se puede utilizar como información adicional al realizar una decisión con respecto a cualquier instancia nueva.

(3) Si el paso de la estimación de la distribución es robusto a las anomalías en los datos, las técnicas estadísticas pueden operar en un entorno no supervisado sin necesidad de etiquetas en los datos de entrada.

Entre las desventajas de las técnicas estadísticas podemos encontrar:

(1) La principal desventaja de las técnicas estadísticas es que se basan de la hipótesis que los datos se generan a partir de una distribución particular. Esta suposición a menudo no es cierta, especialmente para conjuntos de datos reales de alta dimensión.

(2) Incluso cuando la hipótesis puede justificarse razonablemente, hay distintas hipótesis de los test estadísticos que se pueden aplicar para detectar anomalías, elegir el mejor estadístico, a menudo, no es una tarea sencilla. En particular, trabajar con distribuciones complejas y alta dimensionalidad de los datos no es nada trivial con estas técnicas.

(3) Son relativamente simples de implementar, pero un problema para estas técnicas con los datos multivariados es que la mayoría no son capaces de capturar las interacciones entre diferentes atributos. Una anomalía podría tener valores de atributo que individualmente son muy frecuentes, pero su combinación es muy rara, y estas técnicas no lo detectarían.

Referencias

[1] V. Chandola, A. Banerjee, V. Kumar. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58. 2009.

[2] C.C. Aggarwal. “Outlier analysis second edition”. Springer International Publishing, 2º edición, 465 páginas. 2016.

[3] P. J. Rousseeuw, K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3):212 223, 1999.

[4] M. Goldstein and A. Dengel. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. Annual German Conference on Artificial Intelligence (KI-2012), pages 59–63, 2012.

[5] M. Gol, A. Abur. A modified Chi-Squares test for improved bad data detection. IEEE PowerTech, (1): 1-5. 2015