Análisis de Anomalías

Jose Fernando Zea y Fernando López-Torrijos

Abril de 2021

Prólogo

En español se define anomalía como "Desviación o discrepancia de una regla o de un uso". En inglés lo definen como "something that is unusual enough to be noticeable or seem strange". Según Aggarwal (2017) un outlier es un punto que difiere significativamente a los demás puntos (Aggarwal). Si el punto difiete del mecanismo generador de los datos.

Se toma la definición proveniente del inglés.

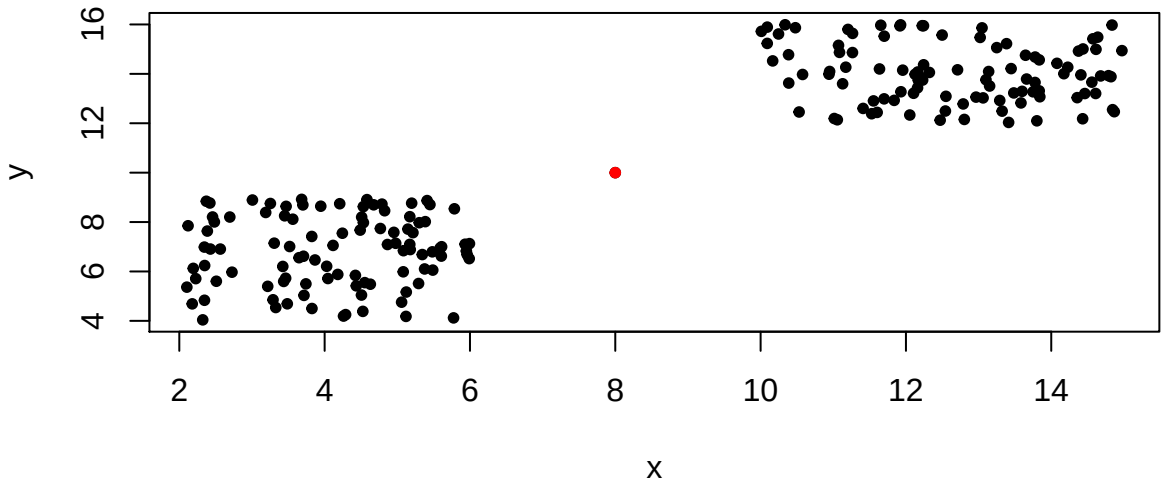
Podría hablarse de muchos tipos de anomalías dentro de un conjunto de datos:

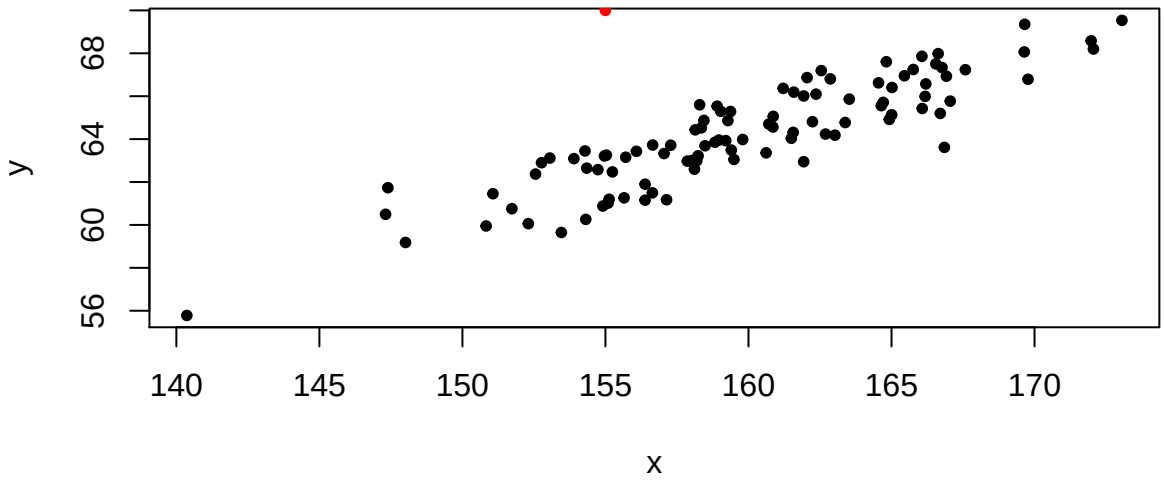
- Datos faltantes
- Valores no probables (¿error de digitación?)
- Valores atípicos
- No respuesta
- Formato incorrecto

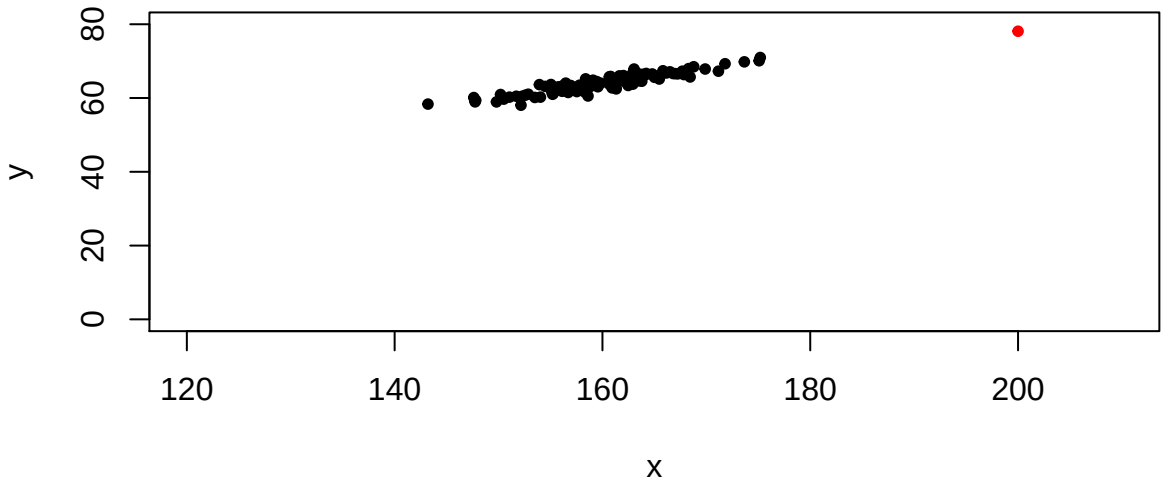
Tradicionalmente los datos atípicos son datos muy *grandes* o muy *pequeños* comparados con el grueso del conjunto de datos. Son observaciones con un comportamiento extraño porque toman valores que no se esperan. Pero esos datos es mejor denominarlos *valores extremos*.

Este documento se centra en las *anomalías*, distinguiendo éstas del ruido aleatorio, que también puede generar algunos **valores atípicos**. La separación entre ruido aleatorio y dato anómalo no es siempre clara.

¿Cómo lucen los valores atípicos?







Aplicaciones

- Sistema de detección de intrusiones: transacciones bancarías anómalas (demasiados movimientos en una cuenta), actividad inusual en una red de telefonía móvil.
- Fraude de tarjeta de crédito: localizaciones, movimientos inusuales. Eventos obtenidos a partir de sensores: comportamientos extraños que pueden estar asociados al mal funcionamiento de un dispositivo.
- Diagnóstico médico: imágenes de resonancia magnética, tomografías, electrocardiogramas pueden estar asociadas a detección de enfermedades. Incumplimiento de la ley: reclamaciones de seguros, actividad de trading, movimientos financieros.
- Ciencias de la tierra: anomalías ambientales, uso del suelo, condiciones ambientales anomalas.

Presentación de las técnicas tradicionales o más usuales.

Análisis variable a variable.

El primer caso que se trabaja es la definición de un dato anómalo cuando se trabaja una única variable a la vez.

El primer contexto en el que se encuentran es en el diagrama de caja o boxplot.

El Diagrama de caja se construye con base en los cuartiles (ver Figura 1):

Cuartil 1 (q1): valor a partir del cual el 25% de los datos tienen un valor menor a éste.

Ojo. El 25% hace referencia a la cantidad de datos, no a sus valores.

Cuartil 2 (q2): valor a partir del cual el 50% de los datos tienen un valor menor. Se denomina también mediana. Es un valor robusto frente a datos extremos, es decir, no se afecta por la presencia de datos extremos, sean datos atípicos o no.

Cuartil 3 (q3): valor a partir del cual el 75% de los datos tienen un valor menor a éste. O leído al revés, el valor a partir del cual el 25% de los datos tienen un valor mayor a éste.

Rango intercuartil (RI): La diferencia entre el cuartil 3 y el 1. $RI = q3 - q1$

Bajo las reglas de Tukey, un estadístico activo en las décadas del 40 al 80 del siglo XX, y quien popularizó los diagramas de caja, todo dato que está alejado más de 1.5 veces el RI del cuartil más cercano se dice que es un dato *atípico*. Un dato atípico lo denominan *extremo* si está ubicado a una distancia mayor de 3 veces el RI del cuartil más cercano y se llama *moderado* en caso contrario.

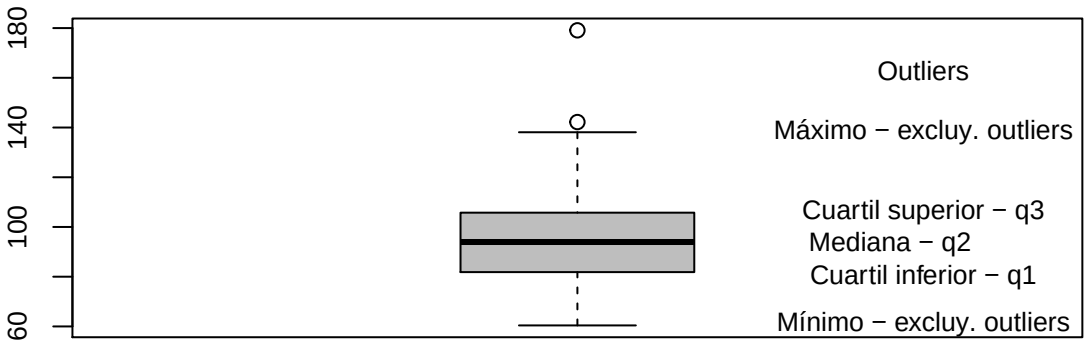


Figure 1: Elementos de un diagrama de caja

Sobre el conjunto de datos que generó el diagrama del ejemplo (ver Figura 1), hay dos círculos que reflejan los datos atípicos.

Atípico moderado: 142.2

Atípico extremo: 179.1

Los cuartiles de la Figura 1 son: 81.9, 93.9, 105.5

Regla empírica de Tukey

Tukey en los años 60 construyó una regla para identificar valores extremos en datos con distribuciones gausianas o normales, para esto se basó en el uso de dos cantidades conocidas como bisagras o bigotes (hinges/whiskers en la lieratura anglosajona).

El bigote inferior se calcula como $LW = q_1 - k \times iqr$\$Es decir, el bigote inferior (lower whisker) se calcula como el percentil 25 (o primer cuartil) menos k veces el rango intercuartílico (iqr). Tukey, Tukey propuso como valor de k = 1,5.

SI LW es menor al mínimo de los datos el valor que finalmente se deja como bigote inferior es el mínimo. Es decir,

$$LW = max(min(x), q_1 - k \times iqr)$$

El bigote superior se calcula como:

$$LW = min(max(x), q_3 + k \times iqr$$

En la expresión anterior q_3 corresponde al percentil 75. Bajo normalidad co un Valor de 1,5 observemos la probabilidad de que un dato sea considerado como un valor extremo:

```
## [1] 0.006976603
```

En otras palabras aproximadamente 7 de cada mil valores serán detectados como valores extremos

Filzmoser, Gussenbauer, and Templ (2016) propusieron realizar previamente una transformación de las distribuciones asimétricas a una distribución normal y posteriormente aplicar la regla de Tukey para detección de valores extremos.

$$y(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, \text{Sí } \lambda \neq 0 \\ log(\lambda), \text{Sí} \lambda = 0 \end{cases}$$

Ilustraremos la tranformación de Box - Cox con un ejemplo:

```
# importar modulos
import numpy as np
from scipy import stats

import seaborn as sns
import matplotlib.pyplot as plt
```

Simularemos una distribución asimétrica a la izquierda:

```
# Gnerar una distribución asimétrica (una exponencial en particular)
np.random.seed(0)

x = np.random.exponential(size = 1000)
np.mean(x), np.var(x)
```

```
## (1.003540208760709, 1.0590341339276639)
```

```
# Transofmración de box-Cox (Dupla: arreglo de valores transformados y lambda )
y, fitted_lambda = stats.boxcox(x)
```

Se compara la distribución orignal con la transformada mediante Box - Cox

```
# crear ejes para el gráfico
fig, ax = plt.subplots(1, 2)

# plotting the original data(non-normal) and
# fitted data (normal)
sns.distplot(x, hist = False, kde = True,
             kde_kws = {'shade': True, 'linewidth': 2},
             label = "Non-Normal", color ="green", ax = ax[0])

sns.distplot(y, hist = False, kde = True,
             kde_kws = {'shade': True, 'linewidth': 2},
```

```
            label = "Normal", color ="green", ax = ax[1])

# adding legends to the subplots
plt.legend(loc = "upper right")

# rescaling the subplots
fig.set_figheight(5)
fig.set_figwidth(10)
```

```
## Lambda value used for Transformation: 0.2420131978174143
```

Ejercicio: se ilustra la identificación de atípicos con la base de datos iris.

Se considera la variable Sepal.Length, dada la aparentemente normalidad de los datos se detectarán outliers con la regla de Tukey:

```
import numpy as np
def fivenum(x, range = 1.5, nan_remove = True):
    """DEvuelve los cinco números de Tukey (mínimo, bigote inferior, mediana,
    bigote superior, maximo) para una lista, arreglo univariado de numpy o serie de pandas"""

    if(isinstance(x, list)):
        x = np.array(x)
    try:
        np.sum(x)
    except TypeError:
        print('Error: you must provide a list or array of only numbers')
    if(nan_remove == True):
        y = x[~np.isnan(x)]
        q1 = np.percentile(y, 25)
        q3 = np.percentile(y,75)
        md = np.median(y)
        iqr = q3-q1
        lower_whisker = q1 - 1.5 * iqr
        upper_whisker = q3 + 1.5 * iqr
        lower_whisker = np.max([lower_whisker, np.min(y)])
        upper_whisker = np.min([upper_whisker, np.max(y)])

    else:
        q1 = np.percentile(x, 25)
        q3 = np.percentile(x,75)
        md = np.median(x)
        iqr = q3-q1
        lower_whisker = q1 - 1.5 * iqr
        upper_whisker = q3 + 1.5 * iqr
        lower_whisker = np.max([lower_whisker, np.min(x)])
        upper_whisker = np.min([upper_whisker, np.max(x)])

    salida = np.array([lower_whisker, q1, md, q3, upper_whisker])
    return salida
```

Se cargan los datos de ejemplo:

Se calculan los 5 números de Tukey:

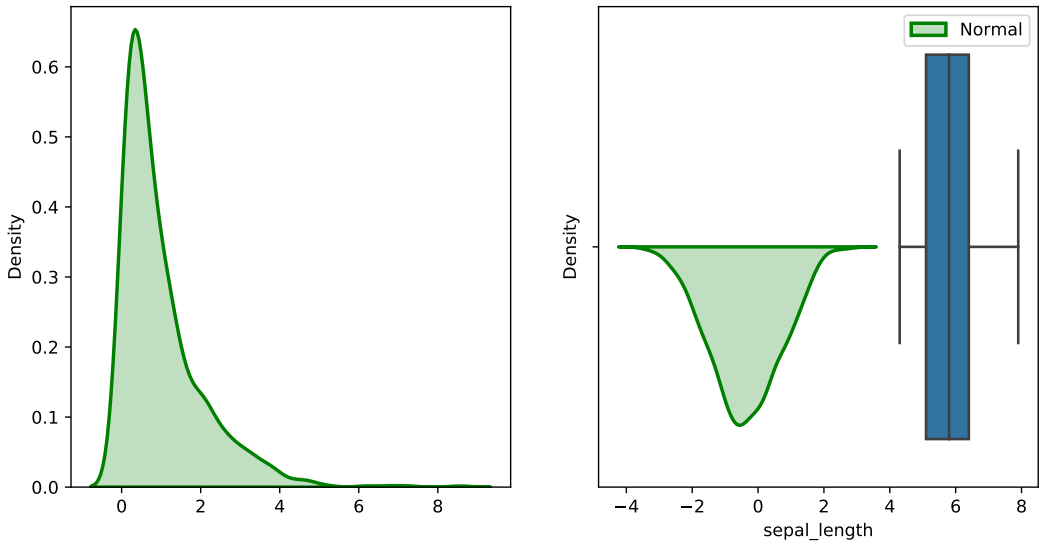
```
fivesnums = fivenum(iris_df.sepal_width)
fivesnums
```

```
## array([2.05, 2.8 , 3.  , 3.3 , 4.05])
```

```
iris_df['outlier_sepal_width'] = iris_df['sepal_width'].apply(lambda x: 'outlier' if (x < fivesnums[0]) 
iris_df.outlier_sepal_width.value_counts()
```

```
## no_outlier    146
## outlier         4
## Name: outlier_sepal_width, dtype: int64
```

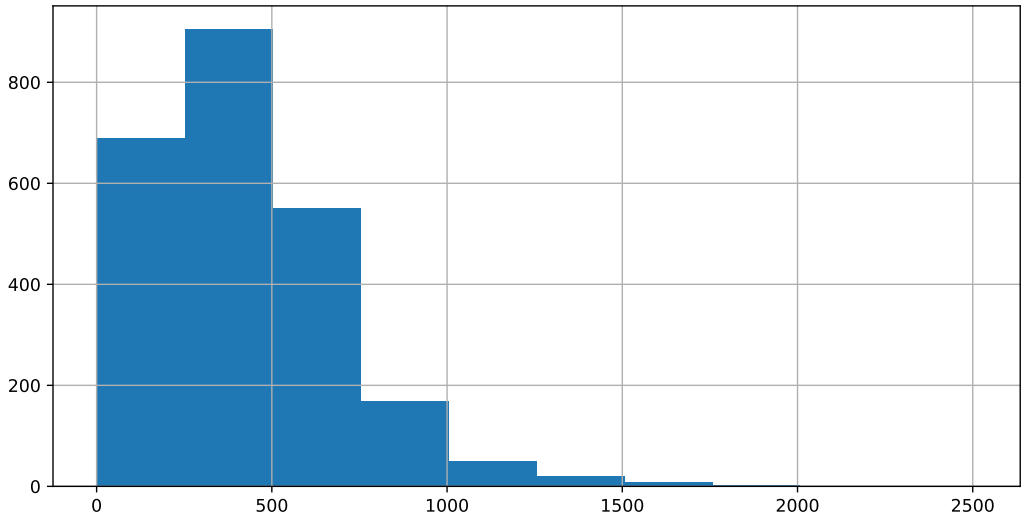
```
ax = sns.boxplot(x=iris_df.sepal_length)
ax
```



Ejemplo: Detectar los outliers de la variable Income de base de datos de empresas Lucy.

```
import pandas as pd
Lucy = pd.read_csv("Lucy.csv")
```

```
Lucy.Income.hist()
```



Se evidencia que la distribución del ingreso es muy asimétrica a la derecha, por lo tanto aplicaremos la transformación de Box-Cox:

```
##       ID Ubication  Level Zone  Income  Employees  Taxes SPAM  Income_bc
## 0  AB001       c1k1  Small    A     281         41    3.0   no  18.293605
## 1  AB002       c1k2  Small    A     329         19    4.0  yes  19.521134
## 2  AB003       c1k3  Small    A     405         68    7.0   no  21.248675
## 3  AB004       c1k4  Small    A     360         89    5.0   no  20.253834
## 4  AB005       c1k5  Small    A     391         91    7.0  yes  20.947118
```

El valor de lambda considerado es:

```
## 0.3589008867008163
```

Los cinco números de Tukey son:

```
fivesnums_income_bc = fivenum(Lucy.Income_bc)
fivesnums_income_bc
```

```
## array([ 5.34786971, 16.8315778 , 20.92531509, 24.48738319, 35.97109128])
```

```
Lucy['outlier_Income'] = Lucy['Income_bc'].apply(lambda x: 'outlier' if (x < fivesnums_income_bc[0]) |
Lucy.outlier_Income.value_counts()
```

```
## no_outlier    2379
## outlier         17
## Name: outlier_Income, dtype: int64
```

Ejercicio:: detectar los valores extremos de la variable Employees

Prueba de Grubbs

Unidimensionalmente los datos extremos son las anomalías. Una primera prueba para detección de datos extremos fue el test de Grubbs, denominada así por Frank E. Grubbs, quien la publicó en 1950. Se conoce también como prueba residual máxima normalizada o prueba de desviación extrema studentizada. Se utiliza para detectar valores atípicos en un conjunto de datos univariados que se supone que proviene de una población distribuida normalmente.

Obsérvese el supuesto. Es importante.

La prueba de Grubbs detecta un valor atípico a la vez. Este valor atípico se elimina del conjunto de datos y la prueba se repite hasta que no se detectan valores atípicos adicionales. Sin embargo, las probabilidades de detección cambian en cada iteración. La prueba no debe usarse para tamaños de muestra de seis datos o menos, ya que con frecuencia etiqueta la mayoría de los puntos como valores atípicos.

La hipótesis nula H_0 es que no hay datos atípicos. Y la estadística de prueba es:

$$G = \frac{max_{1,\ldots,n} \mid y_i - \bar{y} \mid}{s}$$

siendo $\bar{y}$ la media y s la desviación estándar.

Mide la máxima desviación y la divide entre la desviación estándar. El resultado lo compara con respecto a un valor de referencia sacado a partir de la distribución t de student. De ahí el origen de su nombre alternativo.

La hipótesis se rechaza si $G > \frac{n-1}{n}\sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}$

donde $t^2_{\alpha/(2n),n-2}$ denota el valor crítico después del cual se debe considerar un valor extremo.

Se puede definir el test con la estadística sólo hacia un lado (test de una cola): $G = \frac{\bar{y}-y_{min}}{s}$ ó $G = \frac{y_{max}-\bar{y}}{s}$

Para ejecutarlo en Python, se hace uso de la función smirnov_grubbs() del paquete outlier_utils, que usa la siguiente sintaxis:

```
smirnov_grubbs.test(datos, alfa = .05)
```

No vale la pena realizar una práctica acerca del cálculo. Se menciona como antecedente histórico de cómo en análisis unidimensional *datos extremos* y *datos anómalos* son equivalentes. Y se menciona porque tiene que ver con un método que se explicará más adelante. Pero también es un ejemplo de una de las formas de determinar anomalías. Si el valor de la estadística G es mayor al umbral especificado, es anómalo. Si no, es *normal*. Se trata de una asignación binaria. Sin ambiguedades. Se verán métodos que asignan una probabilidad, y es deber del analista determinar de una manera razonable el umbral de probabilidad a partir del cual lo clasificará como dato anómalo.

Análisis en dos o más dimensiones

El segundo contexto de análisis es el multidimensional. Se muestran ejemplos en dos dimensiones, pero se puede proyectar a un espacio de mayor cantidad de dimensiones.

En la Figura 2 se observa que el dato atípico no es un dato extremo ni para la variable X ni para la Y. Es un dato anómalo bidimensionalmente. No es un dato extremo. Cuando se está trabajando en dimensiones muy grandes, no es fácil identificar este tipo de datos anómalos.

Usualmente los algoritmos de búsqueda de datos atípicos cuantifican qué tan anómalo es un punto midiendo qué tan dispersa es la región de datos, la distancia al vecino(s) más cercano(s), o qué tan ajustado está al modelo de distribución subyacente.

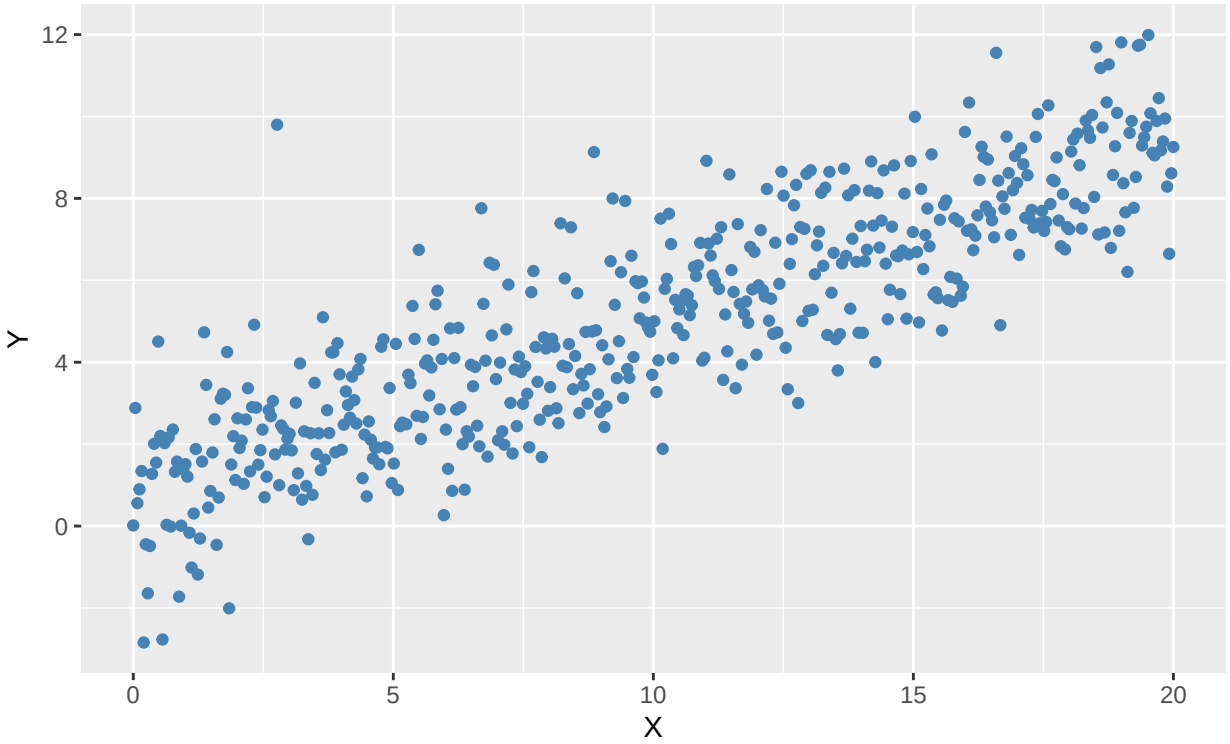


Figure 2: Identificación de datos anómalos en un diagrama de dispersión bidimensional

Hay modelos predictivos en donde una variable se predice en términos de otras. Por ejemplo, el valor de una vivienda se puede predecir en términos de diferentes atributos de la vivienda. En estos modelos se dispone de variables predictivas (X), que producen una variable respuesta (Y).

- Las **variables predictivas** se conocen también como variables independientes, explicativas o regresoras. En un lenguaje muy propio del Machine Learning, también se denominan atributos (features).

- La **variable objetivo** se conoce también como variable dependiente, explicada, respuesta o regresada.

Los modelos anteriormente descritos se conocen como modelos de aprendizaje supervisado. La palabra *supervisado* refleja el hecho de que la variable objetivo auxilia o supervisa el aprendizaje dado que se conoce cuál debe ser el resultado. Otra manera de nombrar el objetivo es decir que se conoce la *etiqueta* del resultado. Dados los datos, el algoritmo de aprendizaje optimiza una función para encontrar una combinación de las variables predictivas que esté lo más cercana al valor verdadero de la variable objetivo.

En el contexto de los modelos predictivos de aprendizaje automático (Machine Learning) que buscan datos anómalos, son preferibles los modelos supervisados ya que determinan de una manera clara qué se entiende por un dato anómalo. Como se observará, la definición de anomalía varía y es mejor ajustarse a la definición específica que requiere la investigación que se está adelantando.

No obstante su preferencia, la detección supervisada de valores atípicos es un caso difícil. Los datos normales suelen ser fáciles de recopilar y, por lo tanto, están disponibles en abundancia. Pero ejemplos de valores atípicos son escasos. En la literatura clásica sobre aprendizaje automático, este problema también se conoce como el problema de *detección de clases poco comunes*.

El desequilibrio en el número de etiquetas *anómalas* a menudo hace que el problema sea bastante difícil de resolver, porque muy pocas instancias de la clase *rara* pueden estar disponibles para los fines del modelado.

Esto también puede hacer que los modelos estándar sean propensos a sobre ajustarse a los datos de ejemplo y no ser muy buenos para predecir sobre nuevos datos.

Modelos estadísticos

Otra manera de hallar datos anómalos es comparar las observaciones con modelos de distribución estadística e identificar observaciones que no se adecuan al modelo subyacente. Es un buen método sujeto a que se elija un modelo adecuado.

Otra alternativa en donde se utiliza un modelo subyacente es la regresión lineal.

Suponga un modelo que se usa para pronosticar el precio de venta de una vivienda basados en sus características.

$$\text{Precio de venta} = \beta_0 + \beta_1 \times \text{Año construcción} + \beta_2 \times \text{Área del lote} + \beta_2 \times \text{Condición general} + \ldots$$

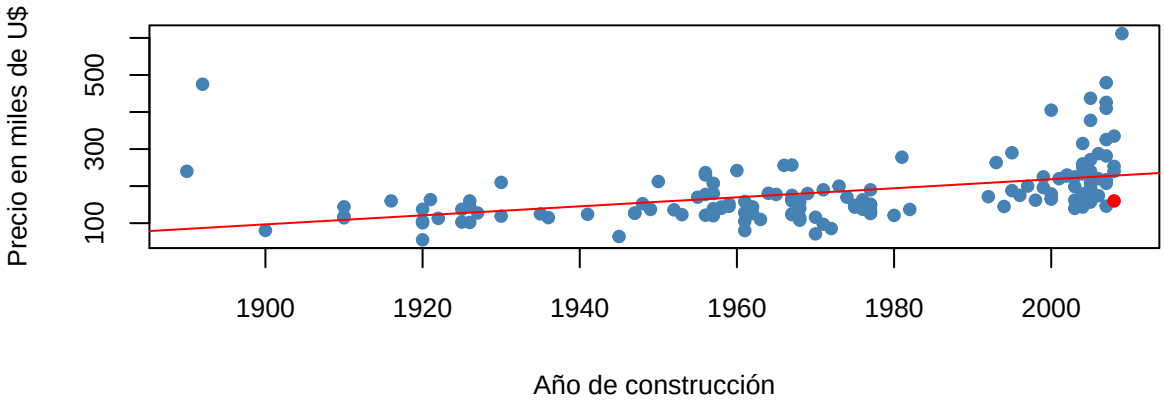


Figure 3: Identificación de viviendas infravaloradas mediante un modelo de regresión lineal

La vivienda señalada con el punto rojo en la Figura 3 es la que tiene el mayor residuo negativo. Se puede considerar un dato anómalo. Las firmas de finca raíz realizan análisis de este estilo para identificar viviendas con un precio muy por debajo del mercado, susceptibles de ser compradas directamente y negociadas en mejores condiciones posteriormente.

Los puntos de datos con grandes residuos y/o alto apalancamiento pueden distorsionar el resultado y la precisión de una regresión. La distancia de Cook mide el efecto de eliminar una observación determinada. Se considera que los puntos con una gran distancia de Cook merecen un examen más detenido en el análisis. Algunos son datos extremos. Otros tal vez sean datos anómalos.

Análisis de componentes principales

El análisis de componentes principales (ACP) es una técnica para reducción de dimensionalidad en donde usualmente interesan los primeros componentes. El ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas *componentes principales*. Por eso se utiliza para la reducción de dimensionalidad. Un uso alternativo es interesarse

en los últimos componentes y pronosticar datos con éstas. El dato que tenga un valor extremo respecto a estas componentes se puede considerar un dato anómalo multidimensional.

El estudio de datos anómalos suele requerir la interpretación de la razón por el cual se define como anómalo. El modelado por regresión lineal o por reducción de dimensionalidad tienen la desventaja de ser particularmente difícil de interpretar en términos de las variables originales, mayormente cuando la dimensionalidad de los datos subyacentes es alta. Esto se debe a que el nuevo subespacio se define como una combinación lineal de atributos con coeficientes positivos o negativos. Esto no suele poder interpretarse intuitivamente en términos de propiedades específicas de los atributos de los datos.

Técnicas más recientes

Nombradas las técnicas más comunes para la detección de datos anómalos diseñadas en el siglo XX, se procede a enunciar técnicas relacionadas con el Machine Learning que serán presentadas en algún detalle en capítulos propios: *Análisis del vecino más cercano* (KNN), *Factor de valor atípico local* e *Isolation Forest*

Métodos basados en la proximidad

La idea de los métodos basados en la proximidad es modelar los valores atípicos como puntos que están aislados de los datos restantes. Este modelado se puede realizar de tres formas: análisis de conglomerados, análisis basado en densidad o análisis del vecino más cercano (KNN). En análisis de conglomerados y otros métodos basados en la densidad, las regiones densas en los datos se identifican directamente y los valores atípicos se definen como aquellos puntos que no se encuentran en dichas regiones. La principal diferencia entre el análisis de conglomerados y los métodos basados en la densidad es que los primeros segmentan puntos, mientras que los otros segmentan el espacio.

Se presentará el método de análisis del vecino más cercano (KNN) más adelante.

Son métodos que proporcionan un alto nivel de interpretabilidad en cuanto que las regiones de datos dispersas se pueden presentar en términos de combinaciones de los atributos originales. Por ejemplo, los conjuntos de restricciones sobre los atributos originales se pueden presentar como criterios específicos para que los puntos de datos particulares se interpreten como valores atípicos.

Anomalías dentro de alta dimensionalidad

El caso de alta dimensionalidad es particularmente desafiante para la detección de valores atípicos. Esto se debe a que los datos se vuelven escasos por lo dispersos y todos los pares de puntos de datos se vuelven casi equidistantes entre sí. Desde una perspectiva de densidad, todas las regiones se vuelven casi igualmente escasas en plena dimensionalidad. La razón de este comportamiento es que muchas dimensiones pueden ser muy ruidosas y pueden mostrar un comportamiento similar por pares en términos de la suma de las distancias específicas de la dimensión. El comportamiento de escasez en alta dimensionalidad hace que todos los puntos se vean muy similares entre sí.

Hay alternativas de análisis. En alta dimensionalidad los verdaderos valores atípicos solo pueden descubrirse examinando la distribución de los datos en un subespacio local de menor dimensión. Los valores atípicos a menudo se ocultan en un inusual comportamiento local inusual de subespacios dimensionales, y este comportamiento desviado está enmascarado en el análisis dimensional completo.

El *factor de valor atípico local* (Local outlier factor LOF) se basa en un concepto de densidad local, donde la localidad viene dada por k vecinos más cercanos, cuya distancia se usa para estimar la densidad. Al comparar la densidad local de un objeto con las densidades locales de sus vecinos, se pueden identificar regiones de densidad similar y puntos que tienen una densidad sustancialmente más baja que sus vecinos. Estos se consideran valores atípicos.

La densidad local se estima mediante la distancia típica a la que se puede *llegar* a un punto desde sus vecinos. La definición de *distancia de accesibilidad* utilizada en LOF es una medida adicional para producir resultados más estables dentro de los conglomerados o clústeres.

Ensambles

En muchos problemas como los de agrupamiento y la clasificación, se utilizan una variedad de *ensambles* para mejorar la solidez de las soluciones. Por ejemplo, en el caso del problema de clasificación, se utilizan una variedad de métodos de conjunto como baggin, boosting y stacking para mejorar la solidez de la clasificación.

Por ejemplo, *Isolation Forest* es un método no supervisado para identificar anomalías cuando no se conoce la clasificación real (anomalía - no anomalía) de las observaciones.

Su funcionamiento está inspirado en el algoritmo de clasificación y regresión Random Forest del Machine Learning. Está formado por la combinación de múltiples árboles llamados *isolation trees*. Estos árboles se crean de forma similar a los de clasificación-regresión: las observaciones de entrenamiento se van separando de forma recursiva creando las ramas del árbol hasta que cada observación queda aislada en un nodo terminal. Sin embargo, en los *isolation trees*, la selección de los puntos de división se hace de forma aleatoria. Aquellas observaciones con características distintas al resto, quedarán aisladas a las pocas divisiones, por lo que el número de nodos necesarios para llegar a estas observación desde el inicio del árbol (profundidad) es menor que para el resto.

Series temporales

Las series temporales contienen un conjunto de valores que se generan mediante mediciones continuas a lo largo del tiempo. Se espera que los valores en momentos consecutivos no cambien de manera muy significativa o cambien de manera suave. En tales casos, los cambios repentinos en los registros de datos subyacentes pueden considerarse eventos anómalos. Por lo tanto, el descubrimiento de puntos anómalos en series de tiempo suele estar estrechamente relacionado con el problema de la detección de eventos anómalos, en forma de anomalías contextuales o colectivas sobre marcas de tiempo relacionadas. Por lo general, tales eventos son creados por un cambio repentino en el sistema subyacente y pueden ser de considerable interés para un analista.