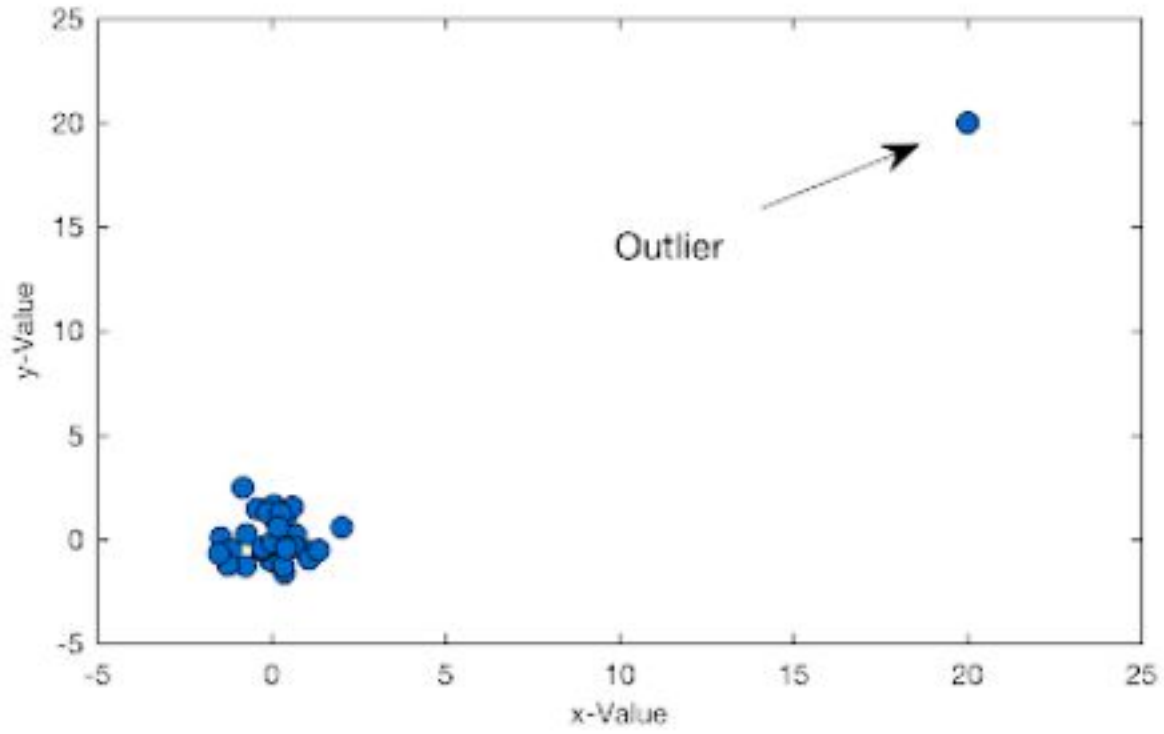


Detección de Outliers

18-09-2020

Lindsey Zugey Alejandro Castillo 1676950
Equipo #
Grupo 003
Minería de Datos

La detección de outliers (del inglés traducido como valores atípicos) pertenece a la categoría descriptiva de las tareas de la minería de datos. El objetivo es encontrar patrones que den un resumen de las relaciones ocultas dentro de los datos. De esta forma, la detección de valores atípicos estudia el comportamiento de valores extremos que difieren del patrón general de una muestra.

Pero ¿qué son los valores atípicos? son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Los datos atípicos son ocasionados por errores de entrada de datos y procedimiento, acontecimientos extraordinarios, valores extremos y/o faltantes, causas no conocidas. Los datos atípicos distorsionan los resultados de los análisis, y por esta razón hay que identificarlos y tratarlos de manera adecuada.

Y ¿cuándo un valor es atípico?, ¿cómo identificarlo?. Visualmente podemos apreciar en gráficas de dispersión, de líneas, diagramas de caja, distribuciones (normal, por ejemplo), de barras, etc., datos que son en extremo diferentes al promedio, datos que se ubican en lugares apartados de los demás, datos que son muy muy grandes y/o datos que son muy muy pequeños. Los valores observados en estas zonas de las gráficas corresponden a los outliers.

Existen distintos tipos de técnicas para detectarlos y se pueden dividir en dos categorías principales: métodos univariantes y métodos multivariantes. Una vez detectados los outliers se pueden eliminar o sustituir si se corrobora que se deben a un error de captura o en la medición de la variable.

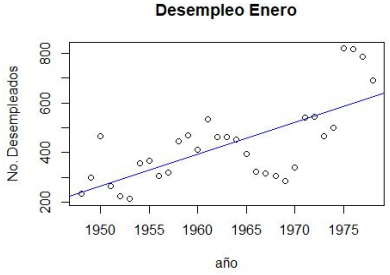
Podemos analizar distintos significados de outliers: error, límites, punto de interés. En el primero si tenemos un grupo de “edades de personas” y tenemos una persona con 160 años, seguramente sea un error de carga de datos. En este caso, la detección de outliers nos ayuda a detectar errores. El segundo se refiere a valores que se escapan del “grupo medio”, pero queremos mantener el dato modificado, para que no perjudique al aprendizaje del modelo de ML. Y el tercero Puede que sean los casos “anómalos” los que queremos detectar y que sean nuestro objetivo.

Ejercicio

Se tiene una base con datos de desempleo por cada mes de 1948 hasta 1978, encuentre en que meses hubo datos atípicos.

Año	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiem	Octubr	Noviem	Diciem
1948	235.1	280.7	264.6	240.7	201.4	240.8	241.1	223.8	206.1	174.7	203.3	220.5
1949	293.5	347.4	338.3	327.7	351.6	396.6	438.8	395.6	363.5	378.8	357	369
1950	464.8	479.1	431.3	368.5	326.3	355.1	331.6	261.3	249	205.5	235.6	240.9
1951	264.9	253.8	232.3	193.8	177	213.2	207.2	180.6	188.6	175.4	199	179.6
1952	225.8	234	200.2	183.6	178.2	203.2	208.5	191.8	172.8	148	159.4	154.5
1953	213.2	196.4	182.9	176.4	153.6	173.2	171	151.2	161.9	157.2	201.7	236.4
1954	356.1	398.3	403.7	384.6	365.8	368.1	367.9	347	343.3	292.9	311.5	300.9
1955	366.9	356.9	329.7	316.2	269	289.3	266.2	253.6	233.8	228.4	253.6	260.1
1956	306.6	309.2	309.5	271	279.9	317.9	298.4	246.7	227.3	209.1	259.9	266
1957	320.6	308.5	292.2	262.7	263.5	313.1	284.3	252.6	250.3	246.5	312.7	333.2
1958	446.4	511.6	515.5	506.4	483.2	522.3	509.8	460.7	405.8	375	378.5	406.8
1959	467.8	469.8	429.8	355.8	332.7	378	360.5	334.7	319.5	323.1	363.6	352.1
1960	411.9	388.6	416.4	360.7	338	417.2	388.4	371.1	331.5	353.7	396.7	447
1961	533.5	566.4	542.3	488.7	467.1	531.3	496.1	444	403.4	398.3	394.1	404.1
1962	462.1	448.1	432.3	386.3	395.2	421.9	382.9	384.2	345.5	323.4	372.6	376
1963	462.7	487	444.2	399.3	394.9	455.4	414	375.5	347	339.4	385.8	378.8
1964	451.8	446.1	422.5	383.1	352.8	445.3	367.5	355.1	326.2	319.8	331.8	340.9
1965	394.1	417.2	393.9	349.2	321.4	405.7	342.9	316.5	284.2	270.9	288.8	278.8
1966	324.4	310.9	299	273	279.3	359.2	305	282.1	250.3	246.5	257.9	266.5
1967	315.9	318.4	295.4	266.4	245.8	362.8	324.9	294.2	289.5	295.2	290.3	272
1968	307.4	328.7	292.9	249.1	230.4	361.5	321.7	277.2	260.7	251	257.6	241.8
1969	287.5	292.3	274.7	254.2	230	339	318.2	287	295.8	284	271	262.7
1970	340.6	379.4	373.3	355.2	338.4	466.9	451	422	429.2	425.9	460.7	463.6
1971	541.4	544.2	517.5	463.4	439.4	549	533	506.1	484	457	481.5	469.5
1972	544.7	541.2	521.5	469.7	434.4	542.6	517.3	485.7	465.8	447	426.6	411.6
1973	467.5	484.5	451.2	417.4	379.9	484.7	455	420.8	416.5	376.3	405.6	405.8
1974	500.8	514	475.5	430.1	414.4	538	526	488.5	520.2	504.4	568.5	610.6
1975	818	830.9	835.9	782	762.3	856.9	820.9	769.6	752.2	724.4	723.1	719.5
1976	817.4	803.3	752.5	689	630.4	765.5	757.7	732.2	702.6	683.3	709.5	702.2
1977	784.8	810.9	755.6	656.8	615.1	745.3	694.1	675.7	643.7	622.1	634.6	588
1978	689.7	673.9	647.9	568.8	545.7	632.6	643.8	593.1	579.7	546	562.9	572.5

Podemos observar los datos atípicos graficando la recta de regresión lineal sobre el gráfico de dispersión, dichos datos son los que se encuentran más alejados de la línea de regresión.



Dato Sospechoso: 856.4 (Junio)

Prueba de Grubbs:

Estadístico T=|(Xo-Xm)|/S(Xm= Media, Xo=Dato sospechoso, S=desviación std.) Según la prueba de Grubbs si T>Gtab entonces dicho dato es atípico, (Gtab: valor crítico en tablas para prueba de Grubbs)

Xm= 433.92

S=161.26

Xo=856.4

T=2.632

Gtab=2.76 (n=31, a= 0.05)

NO ES ATÍPICO

Resumen con Diagramas de caja