

PROGRAMA DE TUTORIAS

PREGUNTAS METODOLOGÍA DE LA INVESTIGACIÓN – IMPLEMENTACIÓN DE PROYECTOS DE INVESTIGACIÓN

2

¿Existe un criterio definido para el tratamiento de los valores extremos?. Algunos toman la media mas dos o tres desvios y los valores fuera de estos parametros son excluidos y otros manejan indices para descartar valores.

Valores atípicos ("Outliers")

Conceptualmente los valores atípicos son observaciones cuyo valor es numéricamente distante del resto de los datos. El problema de esta definición es establecer un significado preciso de "distante", es decir de cuan distante debe ser un dato para ser considerado un valor inusual.

Algunos valores extremos, pueden ser valores genuinos poco frecuentes que son significativamente diferentes del resto. El reconocimiento de valores extremos y su significado puede ser la clave de descubrimientos mayores, especialmente en campos como medicina y física, de tal forma que se requiere ser muy cuidadoso antes de simplificarmente eliminarlos o ajustarlos a valores "normales". *Priciples of data mining. Max Bramer.*

Un valor atípico puede ser un dato genuino, es decir reflejar fielmente la magnitud que se intenta medir o ser la consecuencia de un error en la medición o la transcripción del dato.

La aproximación analítica para con los valores atípicos comprende dos aspectos I) su identificación y II) su tratamiento.

La identificación de un valor atípico implica obtener una medida estadística de su distancia del resto de las observaciones. Una de las sugeridas se obtiene el siguiente procedimiento: Siendo Q_1 el percentilo 25, Q_3 el percentilo 75, y $\Delta Q_3 Q_1$ el intervalo intercuartílico decimos que x es un valor extremo cuando: $x > Q_3 + k(\Delta Q_3 Q_1)$ ó $x < Q_1 - k(\Delta Q_3 Q_1)$, donde usualmente se toma $k=1.5$. Con este mismo criterio de análisis, cuando x cumple esta premisa con $K=3$ se lo denomina valor atípico extremo.

Otras medidas de distncia identifican como valores atípicos a aquellos datos, que bajo el paradigma de una distribución de densidad normal, se encuentran a una distancia mayor a 2 o 3 desvíos estándar de la media, con justificación en el hecho que valores con estas características pertenecen a dos subconjuntos de la distribución con una probabilidad asociada < 0.05 y < 0.003 respectivamente. Este procedimiento, al igual que el anterior basado en el intervalo intercuartílico, no provee ninguna información respecto de la pertenencia o no de estos valores a todo el conjunto, dado que justamente solo está reflejando una de las propiedades fundamentales de la distribución de densidad normal, a saber que entre dos y tres desvíos estándar se encuentran el 95 y 99.7 % respectivamente de los valores de la distribución, es decir en toda distribución normal existen subconjuntos de valores pertenecientes a la misma que presentan una probabilidad < 0.05 y < 0.003 .

Una vez identificados por cualquiera de los métodos descriptos, y bajo la premisa que ninguno de dichos métodos indica si los valores pertenecen o no al conjunto de los datos, el tratamiento de los valores atípicos depende, como se indicó previamente, de la postura que se tenga respecto de los mismos, si se los considera secundarios a error, obviamente serán eliminados, si en cambio son valores genuinos poco frecuentes, deben considerarse otros aspectos. Por lo tanto, como se comprenderá, la postura respecto del significado de un valor atípico, una vez identificado, es enteramente dependiente del proceso que originó el dato, y es ahí donde deberá indagarse por la explicación.

Si el valor atípico es considerado genuino, en primer lugar debe considerarse su poder explicativo en la teoría de la que forman parte. Ahora si los valores extremos forman parte de una variable determinada cuyo efecto sobre otra variable de interés desea evaluarse, deberá

investigarse la influencia de los mismos sobre los resultados del modelo. Si su influencia es marcada, es decir si los resultados del modelo dependen de la inclusión de los valores extremos, deberá considerarse la utilización de métodos robustos para la evaluación de la relación entre dichas variables.