Outliers

Un outlier, o parte aislada, es una observación o punto que se aísla mucho del resto. Esto puede ocurrir debido a errores en el diseño/implementación del modelo o por alguna circunstancia extraordinaria. Se puede trabajar con estos datos siempre y cuando sean verídicos. Cuando se trata de datos extraordinarios que no pueden ser comprobados o explicados por el investigador puede optarse por efectuar una comparación haciendo un análisis sin esos datos a fin de conocer su influencia sobre los resultados. En los outliers existen dos tipos de detección: univariados y multivariados. Dentro de los tipos anteriormente mencionados, los multivariados son más difíciles de identificar debido a que son observaciones extrañas por el conjunto que forman. Existen varios métodos de detección, entre ellos se encuentran la desviación estándar, boxplots y DBScan Clustering. En el método de desviación estándar se considera que los valores están dentro de un rango normal si estos se encuentran dentro de máximo tres desviaciones estándar, en caso de que sean más ya son considerados valores anómalos. En el caso de los boxplots, los datos son acomodados en graficas de bigotes y cualquier dato que se encuentre fuera de los limites ya es considerado anómalo o atípico. Y por último, en el caso de DBScan Clustering los datos son agrupados en grupos y se consideran anómalos los datos que están fuera de estas agrupaciones.

Reglas de asociación

Estos algoritmos tienen como objetivo encontrar relaciones entre un conjunto de transacciones, también conocidos como ítems. Estos algoritmos tienen como requisito que los datos tengan alguna relación entre ellos y se sugiere que no sea un conjunto demasiado extenso para evitar que pierda su efectividad. Las bases de datos transaccionales pueden ser representadas de varias maneras, entre ellas están las listas, las representaciones verticales y también las horizontales. En el caso de las listas son representadas como una fila, que representa los artículos comprados por el consumidor. La representación vertical es la más eficiente de ellas ya que solo requiere dos columnas para distribuir los datos. Por último, las horizontales son representadas como una matriz binaria, donde el 1 representa la presencia de un artículo y el 0 la ausencia. Suponiendo que se tuviera una regla de asociación, el soporte se refiere a la frecuencia relativa con la que estos elementos se repiten. La confianza, por otro lado, se define calculando el cociente del soporte de la regla en cuestión y el soporte del antecedente. A priori es uno de los primeros algoritmos que se desarrollaron para la búsqueda de reglas de asociación. Este consiste en repetir un par de pasos hasta encontrar las reglas. El primer paso es identificar todos los itemset frecuentes y convertirlos en reglas. Para este último, se debe estipular al inicio un nivel de confianza aceptable para hacer las iteraciones correctas.

Regresión lineal

El objetivo de la regresión lineal consiste en buscar una variable aleatoria, y esta última está influenciada por más valores. La regresión lineal se suele representar haciendo uso de graficas de dispersión. Los datos se visualizan y una línea indica los valores más constantes. Para poder predecir un valor es necesario tener una variable de salida, que es la que se quiere predecir, y una variable de entrada o predictora. El componente del error se agrega a la formula $Ye=\alpha+\beta*x$ debido a que es casi imposible que exista un modelo perfectamente lineal en el mundo real. Se debe establecer un nivel de significancia valido y si el valor P es menor se rechaza la hipótesis nula y se acepta que existe una relación entre x e y. El error estándar residual son los datos que el modelo no puede explicar debido a falta de información y sirve para reducir el error cuando se agregan más variables.

Clustering

Se trata de una técnica utilizada en la Inteligencia artificial. Forma agrupaciones de manera automática de acuerdo a similitudes entre los datos. El clustering tiene dos características a considerar: la primera es que la similitud media entre cada elemento debe ser alta y la segunda es que la similitud media entre elementos de distintos clústeres debe ser baja. Esta disciplina tiene diversas aplicaciones, entre las cuales se encuentran el marketing (porque ayuda a encontrar grupos distintivos entre los clientes), la biología (porque sirve para clasificaciones y ayuda a identificar genes con similitudes), el descubrimiento web (en la clasificación de documentos) y la detección de fraudes (como para detectar fraudes en tarjetas de crédito). El clustering debe poseer ciertas características como la escalabilidad, el manejo de diferentes atributos, independencia del orden de datos e interpretabilidad. En la escalabilidad se refiere a que debe poder funcionar con conjuntos pequeños de datos y también con grandes. Con el manejar diferentes atributos se refiere a que debe poder trabajar tanto con datos binarios como cualitativos o numéricos. Y por último la interpretabilidad, que se refiere a que los datos deben ser lógicos y utilizables. Las métricas de distancia sirven para especificar la distancia entre los elementos de un conjunto de números reales no negativos, y dicta que dos elementos son iguales si la distancia entre ellos es cero. Entre estas funciones se encuentra la distancia euclídea y la de manhattan. Existen varios tipos de clustering pero los que más se utilizan son el jerárquico y el de partición aunque también existe el método Density-based y Grid-based.

Análisis predictivo

Para llevar a cabo este análisis es indispensable disponer de una gran cantidad de datos, tanto actuales como pasados, para poder establecer un patrón y generar un análisis. Sin los datos sería imposible obtener variables ni la relación entre ellas. Una parte fundamental sería el correcto almacenamiento de los datos, ya que podrían arrojar una predicción incorrecta y generar pérdidas. Las tecnologías de la información han ido innovándose a lo largo del tiempo, a tal grado que los humanos pasaron a ser proveedores de datos. Por ejemplo, plataformas como amazon o google que almacenan las búsquedas del usuario y en base a eso efectúan predicciones o sugerencias de algo que podría ser del interés del usuario. Los modelos predictivos se utilizan para predecir las probabilidades de que alguien, en función a sus datos, compre algo (por mencionar un ejemplo). Existen varias técnicas que pueden ser aplicadas al análisis predictivo, entre ellas se encuentran la regresión lineal, los arboles de clasificación y regresión, redes neuronales, máquinas de vectores, entre otros. El modelo de regresión lineal analiza la relación existente entre una variable dependiente y un conjunto de variables independientes. Los árboles de clasificación y regresión se consideran una técnica de aprendizaje que produce árboles de regresión dependiendo de si las variables son categóricas o numéricas. Las redes neuronales se utilizan cuando no se conoce una relación exacta entre los valores de entrada y salida. Esta última tiene una característica clave, aprenden las relaciones entre los valores gracias al entrenamiento. Las máquinas de vectores de soporte son usadas para detectar patrones complejos. Son máquinas de aprendizaje que se utilizan para realizar clasificaciones binarias y regresiones. Hay algunas aplicaciones en particular en las que puede ser aplicado el análisis predictivo como en la segmentación de cliente, personalización de oferta, detectar riesgos de que un cliente abandone, entre otros.