TÉCNICAS DE MINERÍA DE DATOS

PREDICCIÓN

Para hacer una predicción de datos es útil usar árboles de decisión. Estos dividen el espacio de los predictores agrupando observaciones con valores similares para la variable de respuesta; se divide con reglas o decisiones para que cada subregión contenga la mayor proporción posible de dividuos en una de las poblaciones. Si una subregión contiene dos datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en subregiones menores que integran datos de la misma clase.

Los árboles se pueden clasificar en:

- Árboles de regresión. La variable respuesta es cuantitativa.
- Árboles de clasificación. La variable de respuesta es cualitativa.

Los árboles de decisión están formados por nodos y su lectura se realiza de arriba hacia abajo. Se distinguen diferentes tipos de nodos:

- Primer nodo o raíz, en él se produce la primera división en función de la variable más importante.
- Nodos internos o intermedios, se encuentran tras la primera división, vuelven a dividir el conjunto de datos en función de las variables.
- Nodos terminales u hojas, se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva.

La profundidad de un árbol está dada por el número máximo de nodos de una rama.

Árbol de clasificación: Consiste en hacer preguntas del tipo $¿x_k \leq c?$ para las covariables cuantitativas o preguntas del tipo $¿x_k$=nivel? para las covariables cualitativas, de forma que el espacio de las covariables es dividido en hiper-rectángulos y todas las observaciones quedan dentro de un hiper-rectángulo con el mismo valor del grupo estimado.

Hay dos tipos de nodo:

- De decisión: Condición al principio y tienen más nodos debajo de ellos.
- De predicción: No tienen ninguna condición ni nodos debajo de ellos. También denominados “nodos hijo”.

La información de cada nodo es:

- Condición: Si es un nodo donde se toma alguna decisión
- Gini: Medida de impureza.
- Samples: Número de muestras que satisfacen las condiciones para llegar a este nodo.
- Value: Cuántas muestras de cada clase llegan a este nodo.
- Class: Qué clase se les asigna a las muestras que llegan a este nodo.

La medida de limpieza “Gini” se refiere a qué tan mezcladas están las clases de cada nodo, por lo que si vale 0, el nodo es totalmente puro.

$$gini = 1 - \sum_{k=1}^{n} p_c^2$$

donde p_c es la probabilidad de cada clase.

Árbol de regresión: Consiste en hacer preguntas de tipo ¿$x_k \leq c$? para cada una de las covariables, de forma que el espacio de las covariables es dividido en hiperectángulos y todas las observaciones dentro de un hiper-rectángulo tendrán el mismo valor estimado y^.

Bosques aleatorios: Técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento para la generalización de un rendimiento durante entrenamiento similar. La mejora de la generalización la consigue compensando los errores de las predicciones de los distintos árboles de decisión, que para asegurarse sean distintos, cada uno se entrena con una muestra aleatoria de datos de entrenamiento. Esta estrategia se denomina bagging.

REGLAS DE ASOCIACIÓN

Las reglas de asociación se derivan de un tipo de análisis que extrae información por coincidencias con el objetivo de encontrar relaciones dentro de un conjunto de transacciones, en concreto, ítems o atributos que tienen a ocurrir de forma conjunta.

Una regla de asociación se define como una implicación del tipo “Si A (antecedente) => B (consecuencia)” donde A y B son ítems individuales. Nos permite encontrar combinaciones de ítems y la fuerza e importancia de estas combinaciones.

Se pueden aplicar para definir patrones de navegación dentro de una tienda, promociones de pares de productos, soporte para la toma de decisiones, análisis de información de ventas, distribución de mercancías y segmentación de clientes con base en patrones de compra.

Algunos tipos de asociación son:

- Asociación cuantitativa: Incluyen Booleana (ausencia o presencia de item), Cuantitativa (ítems cuantitativos).
- Asociación multidimensional: Incluyen unidimensional (ítems referenciados a una sola dimensión) y multidimensional (referenciados en dos o más dimensiones)
- Asociación multinivel: Involucra de las asociaciones de un nivel (único nivel) y multinivel.

Las métricas de interés son el Soporte, frecuencia en que A y B aparecen juntos.

$$Soporte(A => B) = P(A \cap B) = \frac{Frecuencia\ en\ que\ A \cap B\ aparece\ en\ las\ transacciones}{Total\ de\ transacciones}$$

La confianza, que es el cociente del soporte de la regla y el soporte del antecedente solamente. Mide la fortaleza de la regla, si es baja es probable que no exista relación entre antecedente-consecuente.

$$Confinza(A => B) = \frac{Soporte(A => B)}{Soporte\ (A)} = P(B|A) = \frac{P(A \cap B)}{P(A)}$$

El Lift, refleja el aumento de probabilidad de que ocurra el consecuente cuando nos enteramos de que ha ocurrido el antecedente. Si es >1 hay una relación fuerte y frecuencia mayor que el azar, si es ≈1 representa relación del azar y si es <1 representa relación débil y frecuencia menor que el azar.

$$Lift(A => B) = \frac{Soporte\ (A => B)}{Soporte(A) * Soporte(B)} = \frac{P(A \cap B)}{P(A) * P(B)}$$

CLUSTERING

Es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basándonos en similitudes.

Tiene usos en la investigación del mercado, identificar comunidades, prevención crímenes y procesamiento de imágenes.

Los tipos básicos de análisis son:

- Centroid Based Clustering: Cada cluster tiene un centroide. Los clusters se construyen basados en la distancia de un punto de los datos hasta el centroide. El algoritmo más usado para este tipo es el K-medias.
- Connectivity Based Clustering: Se definen agrupando los datos más similares o cercanos. La característica principal es que un cluster contiene otros clusters (representando una jerarquía), el algoritmo más usado es el Hierarchical clustering.
- Distribution Based Clustering: Cada cluster pertenece a una distribución normal. La idea es que los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución normal. El algoritmo más usado es el Gaussian mixture model.
- Density Based Clustering: Los clusters son definidos por áreas de concentración. Trata de conectar puntos cuya distancia entre sí sea pequeña. Un cluster contiene a todos los puntos relacionados dentro de una distancia delimitada y se considera irregular a las áreas esparcidas entre clusters.

El método de K medias es el más usado si de Clustering se trata. Debemos considerar la varianza de los clusters, que disminuye conforme la cantidad de clusters aumenta (por lo que si tenemos n datos y n clusters la varianza es 0). Un método para encontrar el mejor número de clusters es el método del codo.

VISUALIZACIÓN

Es la representación gráfica de información y datos para ser más accesibles de ver, comprender tendencias, valores atípicos y patrones; es esencial para análisis de grandes cantidades de información y la toma de decisiones.

Hay múltiples técnicas y aproximaciones para la visualización de datos según su naturaleza. Se clasifican en:

- Elementos básicos de representación de datos: Incluyen gráficas (barras, líneas, tarta), mapas (burbujas, de calor, de agregación) y tablas (anidación, dinámicas).
- Cuadros de mando: Es una composición compleja de visualizaciones individuales que tienen coherencia y relación temática entre ellas. Se usan principalmente para análisis de conjuntos de variables y toma de decisiones.
- Infografías: Destinadas para la construcción de narrativas a partir de los datos, cuentan "historias". Se combinan con otros elementos como símbolos, leyendas, dibujos, etc.

Estos son los estándares web desarrollados en los últimos años para la evolución de las aplicaciones web, que son base fundamental para creación de visualizaciones web basadas en datos: HTML5, CSS3, SCV y WebGL.

La visualización de datos se encuentra justo en el centro del análisis y la narración visual.

REGRESIÓN

La primera forma de regresión lineal documentada es la de mínimos cuadrados por Legendre (1805). Gauss lo desarrolló de manera más profunda y en 1889 el término de "regresión" fue introducido por Francis Galton.

La regresión es una técnica de minería de datos de la categoría predictiva. Predice el valor de un atributo con base en datos recolectados. Analiza el vínculo entre una variable dependiente y una(s) independiente(s), encontrando una relación matemática.

- Regresión lineal simple: Sólo una variable regresora. $y = \beta_o + \beta_1 x + e$ donde el modelo ajustado por mínimos cuadrados utiliza: $\widehat{\beta_O} = \bar{y} - \widehat{\beta_1} x$
- Regresión lineal múltiple. Se dice lineal porque la ecuación del modelo es una función lineal con parámetros desconocidos.

  $$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + e$$

  donde la estimación por mínimos cuadrados es

  $$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i}^{n} (y_i - \beta_o - \sum_{j=1}^{k} \beta_j x_{ij})\text{^}2$$

Estos modelos se pueden manejar más fácil expresados de forma matricial. La notación matricial del modelo es $y = X\beta + e$ en donde "y" y "e" son vectores yi y ei respectivamente con i de 1 a n, B es un vector de Bi con i de 0 a k y X es una matriz de (n,k) con columna 1 = 1. De forma en que las ecuaciones de mínimos cuadrados quedan dadas por $X'X\hat{\beta} = X'y$ y el estimador de mínimos cuadrados es $\hat{\beta} = (X'X)^{-1}X'y$ siempre y cuando exista $(X'X)^{-1}$.

Los modelos de regresión tienen grandes aplicaciones en la medicina, informática, estadística, comportamiento humano y en la industria.

CLASIFICACIÓN

La clasificación es la técnica de minería de datos más común, organiza o mapea un conjunto de atributos por clase dependiendo de sus características. Algunos ejemplos de aplicación son la calificación de crédito, reconocimiento de imágenes y patrones, detección de fallos industriales, clasificar tendencias de mercados financieros, etc.

Para esto, se estima un modelo usando los datos recolectados para hacer predicciones futuras. Algunas de las técnicas usadas para este fin son:

- Clasificación por inducción de árbol de decisión
- Clasificación Bayesiana
- Redes neuronales
- Support Vector Machines (SVM)
- Clasificación basada en asociaciones.

Regla de Bayes: Si tenemos una hipótesis H para una evidencia E, entonces $P(H|E) = (P(E|H) * P(H))/P(E)$ donde P(H) representa la probabilidad del suceso y P(H|E) la probabilidad de que ocurra H dado que ya ocurrió E.

Redes neuronales: Trabajan directamente con números, por lo que si se desea trabajar con datos nominales hay que enumerarlos. Consisten generalmente en tres capas: de entrada, oculta y de salida. Formalmente podemos definir a una red neuronal como un conjunto de elementos de procesamiento de la información altamente interconectados que son capaces de aprender con la información que se les alimenta; pueden aplicarse a un gran número de problemas desde complejos hasta modelos teóricos sofisticados.

Support Vector Machines(SV): Son un conjunto de algoritmos de aprendizaje supervisado. Dado un conjunto de ejemplos de entrenamiento podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra.

PATRONES SECUENCIALES

Los patrones secuenciales se especializan en analizar datos y encontrar subsecuencias interesadas dentro de un grupo de secuencias. Es una clase especial de dependencia en las que el orden de acontecimiento es considerado. Describe modelos de compras, por ejemplo. Son eventos que se enlazan con el paso del tiempo.

Hay que buscar asociaciones de la forma “Si sucede el evento X en el instante del tiempo t entonces sucederá el evento Y en el instante t+n”. El objetivo es describir de forma concisa las relaciones temporales que existen entre los valores de los atributos y el conjunto de empleos.

Algunas de las características principales es que el orden importa, una secuencia es una lista ordenada de itemsets, el tamaño de una secuencia es su cantidad de elementos (itemsets), la longitud de una secuencia es su cantidad de ítems, el soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto S, las secuencias frecuentes son subsecuencias de una secuencia con soporte mínimo.

Tiene usos en la medicina, análisis de mercado, finanzas, seguros y deportes así como bases de datos temporales, documentales y relacionales.

Resuelven problemas mediante:

- Agrupamiento de patrones secuenciales: Separar los datos en grupos de manera en que los elementos del grupo sean muy similares entre sí y diferentes a los de otros grupos.
- Reglas de asociación con datos secuenciales: Se da cuando los datos contiguos presentan algún tipo de relación.
- Clasificación con datos secuenciales: Estos expresan patrones de comportamiento secuenciales, que se dan en instantes distintos pero cercanos en el tiempo.

Algunos de los métodos representativos son: GSP, SPADE, AprioriAll, ISM, ISE.

OUTLIERS

Datos atípicos: Observación que se desvía mucho del resto de las observaciones apareciendo como una observación sospechosa que pudo haber sido generada por mecanismos diferentes al resto de los datos.

Se puede usar en el aseguramiento de ingresos en telecomunicaciones, detección de fraudes financieros, seguridad y detección de fallas.

Para detectarlos se realizan pruebas estadísticas no paramétricas y así se comparan los resultados basados en la capacidad de detección de los algoritmos.

El método más impartido académicamente por su sencillez y resultados es el test de Tukey, que toma como referencia la diferencia entre el primer cuartil y el tercer cuartil o rango intercuartílico. En un diagrama de caja se considera un valor atípico el que se encuentra 1,5 veces esa distancia de uno de esos cuartiles (atípico leve) o a 3 veces esa distancia (atípico extremo).

Si queremos arreglar nuestro problema de datos atípicos, eliminarlos no siempre es la solución ya que puede no deberse a algún error y podría modificar las inferencias realizadas. La mejor opción es quitarle peso a estas observaciones atípicas mediante técnicas robustas, que son técnicas modernas pensadas precisamente en estos problemas.

Algunos de estos métodos es la comparación de medias, por ejemplo, la prueba robusta de Yuen que utiliza las medias recortadas, es capaz de detectar diferencias significativas entre dos grupos.

Hay casos en que los métodos robustos no son la solución, pues suponen que la distribución subyacente de los datos es más o menos normal (unimodal y simétrica) pero perturbada por valores extremos. Por lo tanto, no son demasiado útiles si se aplican a datos que presentan una marcada distribución multimodal o sesgada.