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

$$Si\ A\ \rightarrow B$$

Donde:
A: antecedente
B: consecuencia

Nos permiten:

- Encontrar las combinaciones de artículos que ocurren con mayor frecuencia en una base de datos transaccional
- Medir la fuerza e importancia de estas combinaciones

Aplicaciones:

- Definir patrones de navegación dentro de la tienda
- Promociones de pares de productos
- Soporte para la toma de decisiones
- Análisis de información de ventas
- Distribución de mercancías en tiendas
- Segmentación de clientes con base en patrones de compra

Tipos de asociaciones:

- Asociación Cuantitativa
  - Cuantitativa
  - Booleana
- Asociación Multidimensional
  - Unidimensional
  - Multidimensional
- Asociación Multinivel
  - De un nivel
  - Multinivel

Métricas de interés:

- Soporte
  El número de veces con que A y B aparecen juntos en una base de datos de transacciones

$$Soporte(A => B) = P(A \cap B)$$

$$\frac{Frecuencia\ en\ que\ A\ y\ B\ aparecen\ en\ las\ transacciones}{Total\ de\ transacciones}$$

- Confianza
  Mide la fortaleza de la regla. Si existe baja confianza es probable que no exista relación entre antecedente y consecuente

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

- Lift
  Refleja el aumento de la probabilidad de que ocurra el consecuente cuando nos enteramos de que ocurrió el antecedente

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

  - Lift > 1 : Relación fuerte (complementos)
  - Lift = 1 : Relación al azar
  - Lift > 1 : Relación débil (sustitutos)

OUTLIERS

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

Aplicaciones:

- Aseguramiento de ingresos en las telecomunicaciones
- Detección de fraudes financieros (ya que se detecta como una actividad sospechosa y fuera de lo normal)
- Seguridad y la detección de fallas

Se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de algoritmos

Un método para detectar estos outliers, aplicable para las regresiones, es calcular los errores de cada observación, es decir, medir la distancia que existe entre los puntos observados y los puntos predichos.
Una vez teniendo esos errores se tienen que estandarizar utilizando el error cuadrático medio (el cual también se obtiene una vez hecha la regresión), después de ahí se buscará detectar aquellos valores lo cuales tengan un valor menor a -2 o mayor a 2.
Las observaciones que sobrepasen este límite son consideradas “atípicas)

Los outliers también se pueden detectar de manea gráfica a través de un diagrama de caja o boxplot.
Estas anomalías son todos los puntos que se encuentren fuera del bigote, o, en otras palabras, fueras del rango intercuartil

REGRESIÓN

Es una técnica de la minería de datos de la categoría predictiva. Predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos
Se encarga de analizar el vínculo entre una variable dependiente y una o varias variables independientes

Tipos:

- Regresión lineal simple
  Cuando sólo se trata de una variable regresora, tiene como modelo:
  $$y = \beta_0 + \beta_1 x + e$$
  Donde:
  e: variable aleatoria normalmente distribuida con E(e)=0 y Var(e)=σ^2

  La estimación de $y = \beta_0 + \beta_1 x$ debe ser una recta que proporcione un buen ajuste a los datos observados (el cual se representa por la R^2 ajustada)

- Regresión lineal múltiple
  Se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos. Se puede relacionar la respuesta "y" con los k regresores
  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

Es más cómodo manejar modelos de regresión múltiple cuando se expresan de forma matricial: $y = X\beta + e$, donde:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

Las ecuaciones normales de mínimos cuadrados quedan dadas por $X'X\hat{\beta} = X'y$
Y el estimador de mínimos cuadrados es $\hat{\beta} = (X'X)^{-1}X'y$

Aplicaciones:

- Medicina
- Informática
- Estadística
- Comportamiento humano
- Industria

PREDICCIÓN

Metodología de la partición de datos:

Para tener un modelo que se ajuste a todos los datos que se introduzcan y que no sólo sea aplicable a los datos con los que se construyó dicho modelo, es necesario partir la base de datos en 3 subconjuntos:

- Conjunto de entrenamiento: son los datos que van a servir de base para la construcción del modelo (deben de estar concentrados la mayor parte de los datos en este subconjunto)
- Conjunto de prueba: son con los cuales vamos a ir probando nuestro modelo y así poder hacer las modificaciones pertinentes
- Conjunto de validación: con estos se puede probar en última instancia si el modelo verdaderamente es eficaz

Árboles de decisión:

Modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable de respuesta. Para dividir el espacio muestral en subregiones es preciso aplicar una serie de reglas o decisiones para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones

- Estructura:
  - Primer nodo/nodo raíz
  - Nodos internos/intermedios
  - Nodos terminales/hojas
- Tipos de nodos:
  - Nodos de decisión: tienen una condición al principio y tienen más nodos debajo
  - Nodos de predicción (nodos hijo): no tienen ninguna condición ni nodos debajo
- Información de los nodos:
  - Condición
  - Gini (medida de impureza)
  - Samples
  - Value
  - Class

- Árboles de Clasificación:
  Consiste en hacer preguntas del tipo $¿x_k = nivel_j?$ Para las variables CUALITATIVAS, de esta forma todas las observaciones con las mismas características queden dentro de una misma región
- Árboles de Regresión:
  Cuando la variable dependiente es continua. En el caso de regresión, en lugar de usar Gini como medida de impureza, se usa el error cuadrático medio (MSE)

Bosques aleatorios:
Técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento de generalización. Esta mejora se consigue compensando los errores de las predicciones de los distintos árboles de decisión. Para asegurarnos que los árboles sean distintos, lo que se hace es que cada uno se entrena con una muestra aleatoria de los datos de entrenamiento

- Bagging: técnica que consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar resultados

Validación cruzada:
Se emplea para estimar el test error rate de un modelo y así evaluar su capacidad predictiva

- Validación simple
- Leave One Out Cross-Validation (LOOCV)
- K-Fold Cross Validation

Métricas de eficacia:
Para datos numéricos: Error cuadrático medio
Para datos categóricos: Curva ROC

CLUSTERING

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

Usos:

- Investigación de mercado
- Identificar comunidades
- Prevención del crimen
- Procesamiento de imágenes

Tipos de análisis

- Centroid Based Clustering:
  Cada cluster es representado por un centroide
  Los clusters se construyen basados en la DISTANCIA de punto de los datos hasta el centroide. Se realizan varias iteraciones hasta llegar al mejor resultado
- Connectivity Based Clustering
  Los clusters se definen agrupando a los datos más similares o cercanos.
  La carcaterística principal es que un cluster contiene a otros clusters
- Distribution Based Clustering
  Cada cluster pertenece a una distribución normal
  Los puntos son divididos con base en la PROBABILIDAD de pertenecer a la misma distribución normal
- Density Based Clustering
  Los clusters son divido por ÁREAS de concentración
  Se trata de conectar puntos cuya distancia entre sí en pequeña, se considera como irregular a las áreas esparcidas entre clusters

Método K-Medias:

Basado en centroides, K representa el número de clusters y es definido por el usuario

1. Se eligen k datos aleatorios que serán los centroides de cada cluster
2. Se analiza la distancia de cada dato hacia el centroide más cercano
3. Obtener media de cada cluster (será el nuevo centro)
4. Se repite hasta que los clusters no cambien

Varianza de los clusters:

Disminuye al aumentar k. Si sólo hay un elemento en el cluster la varianza es 0.
Entre menor sea la suma de las varianzas de los clusters, mejor es nuestro clustering

Métodod del Codo:

Consiste en graficar la reducción de la varianza total a medida que k aumenta. En un punto la reducción de la varianza no disminuirá de forma significativa, este punto es llamado elbow plot o codo (es la k a utilizar)

VISUALIZACIÓN DE DATOS

Representación gráfica de información y datos.
Esta herramienta proporciona una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos
Es esencial para analizar grandes cantidades de información

Tipos de visaulizaciones:

- Elementos básicos de representación de datos
  - Gráficas (barras, líneas, columnas, puntos, de pastel, etc)
  - Mapas (burbujas, de calor, de agregación, etc)
  - Tablas (con anidación, dinámicas, etc)
- Cuadros de mando
  Es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas
- Infografías
  No están destinadas al análisis de variables sino a la construcción de narrativas a partir de los datos

Importancia:
Para los profesionales es cada vez más valioso poder usar los datos para tomar decisiones y usar elementos visuales con los datos apara informar quién, qué, cuándo, dónde y cómo

Ejemplos:

- Gráfico de lineas
- Gráfico de barras
- Gráfico de dispersión
- Histograma
- Matriz de relaciones
- Boxplot
- Gráfica de violín

PATRONES SECUENCIALES

Se especializan en analizar datos y encontrar subsecuencias interesantes dentro de un grupo de secuencias, en este caso el orden de acontencimientos es considerado.
En otras palabras, son eventos que se enlazan con el paso del tiempo

Se trata de buscar asociaciones de la forma “si sucede el evento X en el instante t entonces sucederá el evento Y en el instante t+n”

Objetivo:
Poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos

Cracaterísticas:

- El orden importa
- Su objetivo es encontrar patrones de secuencia
- El tamaño de una secuencia es su cantidad de elementos
- La longitud de una secuencia es su cantidad de ítems
- El soporte es el porcentaje de secuencias que la contienen en un grupo de secuencias S
- Las secuencias frecuentes son las subsecuencias de una secuencia que tienen un soporte mínimo

Entornos de desarrollo

- Medicina
- Bilogía
- Web
- Análisis de mercado
- Deportes
- Finanzas

Resolución de problemas:

- Agrupamiento de patrones secuenciales
  Es separar en grupos a los datos, de manera que los miembros de un grupo sean muy similares entre sí y al miesmo tiempo tengan objetivos diferentes a los demás grupos
- Clasificación con datos secuenciales
  Expresan patrones de comportamiento secuenciales, es decir que se dan en instantes distintos (pero cercanos) en el tiempo
- Reglas de asociación con datos secuenciales
  Cuando los datos contiguos presentan algún tipo de relación

CLASIFICACIÓN

Organiza y mapea un conjunto de atributos por clase, dependiendo de sus características
Se entrena un modelo usando los datos recolectados para hacer predicciones futuras

Técnicas:

- Clasificación por inducción de árbol de decisión
  Serie de condiciones organizadas en forma jerárquica a modo de árbol. Útiles para problemas que mezclan datos categóricos y numéricos
- Clasificación Bayesiana
  Si tenemos una hipótesis H sustentada por una evidencia E

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

- Redes neuronales
  Trabajan directamente con números y en caso de que se desee trabar con datos nominales, estos deben enumerarse, consisten generalmente de tres capas:
  - Entrada
  - Oculta
  - Salida

- Support Vector Machines (SVM)
- Clasificación basada en asociaciones