**UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN**

**FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS**

# Técnicas de la Minería de Datos

Carla Mayela De la Garza Fernández

Matrícula: 1729734

Grupo 012

Minería de Datos

**Carla Mayela De la Garza Fernández** **1729734**

# Clustering

El clustering es una técnica dentro de la disciplina de Inteligencia Artificial, identifica de manera automática agrupaciones o clusters de acuerdo a una medida de similitud entre ellos. Su objetivo fundamental es identificar clústeres de elementos de tal manera que se tenga:

- Una similitud intra-clúster alta: La similitud media entre elementos del mismo clúster sea alta
- Una similitud inter-clúster baja: La similitud media entre elementos de distintos clústeres sea baja

El clustering se utiliza en diferentes campos como, por ejemplo:

- En el *marketing*
- En la *biología*
- En el *descubrimiento web*
- En la *detección de fraudes*

**Métricas de distancia**

Las funciones de distancia representan un método para calcular la cercanía entre dos elementos. Hay dos métricas de distancia las cuales son muy populares y se utilizan comúnmente en el clustering las cuales son:

- Distancia euclídea, su fórmula está dada por:

$$\text{Euclidean dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

- Distancia de Manhattan, su fórmula está dada por:

$$\text{Manhattan dist}((x, y), (a, b)) = |x - a| + |y - b|$$

**Tipos de clustering**

- **Jerárquico**: el cuál se divide a su vez en otros dos tipos

- **Clustering jerárquico aglomerativo**: Comienza con cada caso como cluster individual. Combina el par de clusters más cercano hasta que solo quede k clusters.
- **Clustering jerárquico divisivo**: Comienza con un único cluster que englobe todos los casos de nuestro conjunto de datos. En cada paso, se van partiendo el cluster hasta que queden k clusters.

- **De partición:** Se encarga de dividir un conjunto de datos en una pequeña cantidad de agrupaciones o particiones, basado en sus atributos. La técnica de clustering de partición entorno a centroides realiza una distribución de los elementos entre un número prefijado de grupos. Esta técnica recibe como dato de entrada el número de clústers a formar además de los elementos a clasificar y la matriz de similitudes.

Consiste en agrupar los elementos entorno a elementos centrales llamados *centroides* a cada clúster. Definimos el centroide de un clúster como aquel elemento que minimiza la suma de las similitudes al resto de los elementos del clúster.

El algoritmo más popular de clustering es el **K-Means**, en el cual **n** objetos se agrupan en **k** agrupaciones en función de características. La agrupación de objetos se realiza minimizando la suma de cuadrados de distancias. A continuación, se muestra la metodología que sigue el método de K-Means:

1. Seleccionar el número **k** de clusters aleatoriamente.
2. Escoger aleatoriamente **k** centroides.
3. Calcular las distancias de todos los puntos a los **k** centroides.
4. Formar **k** grupos, asignando cada punto al centroide más cercano.
5. Recalcular los **nuevos** centroides.
6. Se repiten los pasos **3, 4 y 5** hasta que los centroides no se muevan.

- **Método Density-based**
- **Método Grid-based**

**Carla Mayela De la Garza Fernández** **1729734**

# Outliers

Un outlier, también conocido como valor atípico, es una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente. Los tipos de outliers que existen son:

- Outliers que surgen por un error en el procedimiento, tales cómo entrada de datos errónea.
- Observación que ocurre como consecuencia de un evento extraordinario
- Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables
- Datos extraordinarios que no tiene explicación y por lo tanto se vuelve a realizar el análisis y averiguar el porqué de dichas observaciones

**Métodos de detección**

- Desviación estándar
- Boxplots: los bigotes inferiores y superiores pueden verse como los límites de la distribución de datos, cualquier punto que se muestre por abajo o encima de los bigotes, se considera como outlier.
- DBSCan Clustering: algoritmo de agrupación de clusters que utiliza datos agrupados.

**Carla Mayela De la Garza Fernández** **1729734**

# Patrones Secuenciales

La minería de patrones consiste en describir patrones interesantes, útiles e inesperados en una base de datos. Los tipos de patrones son:

- Itemsets frecuentes
- Asociaciones
- Subgrafos
- Reglas secuenciales
- Patrones periódicos

La tarea de minería de datos consiste en encontrar sub-secuencias interesantes dentro de un conjunto de secuencias. Estas sub-secuencias aparecen con frecuencia en una base de datos. Un ejemplo de cómo nos puede ayudar esto es a entender el comportamiento de clientes en cuanto a decisiones de mercado.

Para poder realizar patrones secuenciales se necesita una base de datos secuencial y especificar el umbral de soporte mínimo.

Otro ejemplo de patrones secuenciales es el análisis de textos, cuando, por ejemplo, el objetivo es encontrar las palabras más utilizadas en el texto y en este caso la base de datos secuencial sería el conjunto de frases del texto.

Algunos algoritmos para resolver problemas de patrones secuenciales son:

- PrefixSpan
- Spade
- SPAM
- GSP

**Carla Mayela De la Garza Fernández** **1729734**

# Regresión Lineal

En la regresión lineal se busca una variable aleatoria simple, el valor de esta variable está influenciado por los valores tomados por una o más variables. Esta variable aleatoria se denomina *variable independiente* y las variables influyentes se llaman *variables dependientes*.

Es importante mencionar que al realizar una predicción los regresores no se tratan como variables al azar.

La entrada o variable predictora a la hora de resolver un problema de regresión lineal nos ayuda a predecir el valor de la variable de salida. La variable de salida es la que queremos predecir.

El objetivo es minimizar la suma de los errores al cuadrado sobre todos los puntos del dataset: $X = \{(Xi, Yi)\}$

La ecuación de regresión lineal simple indica que el valor medio o valor esperado de y es una función lineal de x:

$$\boldsymbol{Ye} = \alpha + \beta * x$$

**Componente del error**

Siempre habrá un componente de error o residuo E que es una variable aleatoria con distribución normal.

$$\boldsymbol{Ye} = \alpha + \beta * x + E$$

**Carla Mayela De la Garza Fernández** **1729734**

# Reglas de Asociación

Los algoritmos de reglas de asociación tienen como objetivo encontrar relaciones dentro de un conjunto de transacciones o ítems que tienden a ocurrir en forma conjunta. Un ejemplo sencillo de esto es la cesta de compra de un supermercado.

A cada elemento que forma parte de una transacción se le llama item y a un conjunto de ítems se le llama itemset. Una transacción puede estar formada por uno o varios ítems, en el caso de ser varios, cada posible conjunto de ellos es un itemset distinto. Por ejemplo, la transacción {A, B, C} y sus posibles itemsets: {A}, {B}, {C}, {AB}, {AC}, {BC}, {ABC}.

Se puede representar una base de datos transaccional con las siguientes métricas de interés:

- Una *lista*: cada transacción como una fila
- Una *representación vertical*: ocupa solo dos columnas: una con el id y la otra con los ítems
- Una *representación horizontal*: se representa como una matriz binaria, cada fila es una transacción. Si el articulo está presente es 1 y si no, es 0.

**Soporte**

Dada la regla A → B, el soporte de esta se define como la frecuencia relativa con la que A y B aparecen juntos en una base de datos transaccional.

**Confianza**

Dada una regla si A → B, la confianza de esta regla es el cociente de soporte de la regla y soporte del antecedente solamente. La confianza mide la fortaleza de la regla.

$$\boldsymbol{Confianza}\ (\boldsymbol{A \rightarrow B}) \ = \ Soporte\ (A \rightarrow B)\ /\ Soporte\ (A)$$

**Lift**

Refleja la probabilidad. Si el lift < 1 es relación débil. Si el lift > 1 es una relación fuerte.

$$\boldsymbol{Lift}(\boldsymbol{A} \ \rightarrow \ \boldsymbol{B}) \ = \ Soporte(A \ \rightarrow \ B)/(Soporte\ (A) \ * \ Soporte(B))$$

**Algoritmo Apriori**

1. Identificar todos los itemsets que ocurren con una frecuencia por encima de un determinado límite (itemsets frecuentes).
2. Convertir esos itemsets frecuentes en reglas de asociación.

**Carla Mayela De la Garza Fernández** **1729734**

# Predicción

El análisis predictivo consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido del pasado, presente o futuro.

El análisis predictivo es muy popular actualmente debido a la importancia de la información que ha ganado últimamente ya que se busca que la información sea analizada en busca de tendencias.

Hoy en día, se considera a las personas como proveedores de datos, ya que simples cosas de la vida diaria como caminar, pagar con una tarjeta de crédito, ver una seria online, entre otras generan información que puede ser explotada.

**Modelo predictivo**

Se utiliza para predecir qué posibilidades hay de que una persona reaccione de una manera determinada a ciertos eventos.

**Técnicas aplicables al análisis de predictivo**

- Técnicas de regresión: regresión lineal, árboles de clasificación y regresión, curvas de regresión adaptativa multivariable.
- Técnicas de aprendizaje computacional: redes neuronales, máquinas de vectores de soporte, Naive Bayes, K-vecinos más cercanos.