UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Minería de Datos
Fase 1: Resúmenes

Gilberto Noé López Ávila

1812678

1 de octubre de 2020

Clustering

El **clustering** es una técnica de aprendizaje de máquina no supervisada que consiste en agrupar puntos de datos y de esta forma crear particiones basadas en similitudes.

Los tipos básicos de análisis son:

- **Centroid based clustering**: Cada cluster es representado por un centroide. Los clusters se construyen basados en la distancia del punto de los datos hasta el centroide. El algoritmo más usado es **k-medias**.
- **Connectivity based clustering:** Los clusters se definen agrupando los datos más similares o cercanos. La característica principal es que un cluster contiene a otros clusters. Un algoritmo de este tipo es **hierarchical clustering**.
- **Distribution based clustering:** En este método cada cluster pertenece a una distribución normal. Los puntos son divididos con base en la probabilidad de pertenecer a la misma distribución. Un algoritmo de este tipo es **Gaussian mixture models**.
- **Density based clustering:** Los clusters son definidos por áreas de concentración. Consiste en conectar puntos cuya distancia entre sí es relativamente pequeña.

El método **k-medias** es un algoritmo de clustering basado en centroides. El número K representa el número de clusters definido por el usuario. Una vez escogido el valor de K, se eligen K datos aleatorios que pasarán a ser los centroides representativos de cada cluster. Cada punto es asignado al cluster cuyo centroide sea más cercano al punto. Se define un nuevo centroide para cada cluster a partir de la media de los puntos del cluster, y se itera el proceso hasta que los clusters no cambien.

La varianza de cada cluster disminuye al aumentar K. Si solo hay un elemento en el cluster, la varianza es cero. Entre mejor sea la suma de varianzas del cluster, mejor será el clustering. El **método del codo** consiste en graficar la reducción de varianza total a medida que K aumenta. En un punto la varianza no disminuirá de forma significativa entre un valor de K y otro. Este punto es llamado **codo**, y representa el número K a utilizar.

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 tienden a ocurrir de forma conjunta.

Una regla de asociación se define como una implicación del tipo

$$A \Rightarrow B$$

donde A y B son ítems individuales. Las reglas de asociación nos permiten encontrar las combinaciones de ítems que ocurren con mayor frecuencia y medir la importancia de estas combinaciones.

Las reglas de asociación se pueden clasificar con base en los tipos de valores que manejan:

- **Asociación Booleana:** Asociaciones entre la presencia o ausencia de un ítem.
- **Asociación Cuantitativa:** Describe asociaciones entre ítems cuantitativos o atributos.

También se pueden clasificar con base en las dimensiones de datos que involucran:

- **Asociación Unidimensional:** Si los ítems de la regla se referencian en una sola dimensión.
- **Asociación Multidimensional:** Si los ítems de la regla se referencian en dos o más dimensiones.

El **soporte** de una regla de asociación $A \Rightarrow B$ se define como la frecuencia relativa con que A y B aparecen juntos en una base de datos de transacciones. En lenguaje de probabilidad,

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

El primer requisito para limitar el número de reglas es que tengan un soporte mínimo. Una regla de bajo soporte pudo haber aparecido por casualidad.

La **confianza** de una regla de asociación $A \Rightarrow B$ es el cociente del soporte de la regla y el soporte del antecedente. La confianza es la probabilidad condicional

$$Confianza(A \Rightarrow B) = \frac{Soporte(A \Rightarrow B)}{Soporte(A)} = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

El **lift** de una regla de asociación refleja el aumento de la probabilidad de que ocurra el consecuente cuando sabemos que ocurrió el antecedente.

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

Un lift mayor a 1 representa una relación fuerte y frecuencia mayor que el azar, mientras que un lift menor que 1 representa una relación débil y menor que el azar.

Detección de Outliers

Un **outlier** es una observación que se desvía mucho del resto, apareciendo como una observación sospechosa que pudo ser generada por mecanismos diferentes al resto de los datos. La detección de outliers, o datos atípicos, permite identificar comportamientos inusuales en los datos.

En la detección de outliers se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos. Uno de los algoritmos más conocidos para la detección de outliers es el algoritmo **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise).

La técnica de agrupación DBSCAN clasifica los puntos de las observaciones como puntos núcleo, puntos alcanzables o ruido, y utiliza los parámetros de distancia ϵ y número de puntos $MinPts$. Un punto p es un punto **núcleo** si al menos $MinPts$ puntos están a una distancia ϵ de él y estos son directamente alcanzables desde p. Un punto q es **alcanzable** desde p si existe una secuencia de puntos alcanzables desde p a q, donde todos los puntos son núcleos, excepto posiblemente q. Un punto que no sea alcanzable desde cualquier otro punto es considerado **ruido**.

Si p es un punto núcleo, este forma un cluster junto a otros puntos, núcleos o no, alcanzables desde p. Los puntos no núcleos alcanzables actúan como una barrera del cluster. Los puntos ruidosos representan outliers y no están asignados a otro cluster.

Visualización

La **visualización de datos** es la representación gráfica de información y datos. Al utilizar elementos visuales como cuadros, gráficos y mapas, las herramientas de visualización de datos proporcionan una manera accesible de ver y comprender tendencias, valores atípicos y patrones en los datos. La visualización es esencial para analizar grandes cantidades de información.

Según la complejidad y elaboración de la información, podemos tener los siguientes tipos de visualizaciones:

- **Elementos básicos de representación de datos:** En el caso más sencillo, se utilizan varios tipos de visualizaciones básicas.
    - **Gráficas:** Barras, líneas, columnas, puntos, treemaps, tarta, etc.
    - **Mapas:** Burbujas, mapa temático, mapa de calor, mapa de agregación, etc.
    - **Tablas:** Con anidación, dinámicas, de drill-down, de transiciones, etc.
- **Cuadros de mando:** Un cuadro de mando es una composición compleja de visualizaciones individuales que guardan una coherencia y una relación temática entre ellas. Son ampliamente utilizados para análisis de conjuntos de variables.
- **Infografías:** Están destinadas a la construcción de narrativas a partir de los datos. Esta narrativa se construye a través de la disposición de la información en la que las visualizaciones se combinan con otros elementos como símbolos, leyendas, dibujos, imágenes sintéticas, etc.

En los últimos años se han desarrollado estándares web para la creación de visualizaciones web de los datos:

- **HTML5:** Canvas, elemento HTML para dibujar gráficos 2D.
- **CSS3:** Permite diferenciar el contenido de las páginas web de la presentación de este contenido.
- **SCV:** Utilizado para crear gráficos 2D.
- **WebGL:** Gráficos 3D haciendo uso de Canvas.

Regresión

La **regresión** es una técnica de minería de datos que predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos. La regresión se encarga de analizar el vínculo entre una variable dependiente y una o varias variables independientes.

Cuando el análisis de regresión solo se trata de una variable regresora se llama **regresión lineal simple**. Esta regresión tiene como modelo

$$y = \beta_0 + \beta_1 x + e$$

La cantidad e es una variable aleatoria normal con media cero.

La estimación de $y = \hat{\beta}_0 + \hat{\beta}_1 x$ debe ser una recta que proporcione un buen ajuste a los datos observados. El modelo ajustado por mínimos cuadrados utiliza

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

Un modelo de **regresión lineal múltiple** relaciona la respuesta y con k regresores, o variables predictivas, bajo el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

Se requieren $p = k + 1$ ecuaciones para estimar cada uno de los coeficientes desconocidos de la regresión. La notación matricial del modelo es $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$, donde

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{p1} & \cdots & x_{pk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{pmatrix}$$

La estimación por mínimos cuadrados es

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

siempre y cuando existe la matriz inversa $(X^T X)$; es decir, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

Clasificación

La **clasificación** es la técnica de minería de datos más comúnmente aplicada, que organiza un conjunto de atributos por clase dependiendo sus características. Se entrena un modelo usando los datos recolectados para hacer predicciones futuras.

Algunas técnicas de clasificación son las siguientes:

- **Clasificación Bayesiana:** Si tenemos una hipótesis H, sustentada para una evidencia E,

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

  Donde $P(E)$ y $P(H)$ representan las probabilidades de los sucesos E y H respectivamente, y $P(E|H)$ representa la probabilidad de E condicionada al suceso H.
- **Redes Neuronales:** Trabajan directamente con números, y en caso de que se desee trabajar con datos nominales, estos deben enumerarse. Se utilizan en clasificación, agrupamiento, regresión, etc. Las redes neuronales consisten generalmente en tres capas: de entrada, oculta y de salida. Internamente pueden verse como una gráfica dirigida.
- **Árbol de Decisión:** Son una serie de condiciones organizadas en forma jerárquica. Son útiles para problemas que mezclen datos categóricos y numéricos. Entre algunos problemas con la inducción de reglas se encuentran no necesariamente formar un árbol, no poder cubrir todas las posibilidades o que las reglas entren en conflicto.

Entre otras técnicas de clasificación también se encuentran la **clasificación basada en asociaciones** y **Support Vector Machines** (SVM).

Patrones Secuenciales

Los **patrones secuenciales** se especializan en analizar datos y encontrar subsecuencias dentro de un grupo de secuencias. Estos patrones son eventos que se enlazan con el paso del tiempo.

El objetivo de los patrones secuenciales es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos. Se utilizan reglas de asociación secuenciales – reglas que expresan patrones que se dan en instantes distintos en el tiempo.

Una **secuencia** es una lista ordenada de itemsets; es decir, un conjunto de elementos para los cuales el orden importa. El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S. Las **secuencias frecuentes** (o patrones secuenciales) son las subsecuencias de una secuencia que tienen un soporte mínimo.

La agrupación de patrones secuenciales se define como la tarea de separar los datos en grupos, de manera que los miembros de un grupo sean muy similares entre sí, y al mismo tiempo sean diferentes a los objetivos de otros grupos. Para la creación de agrupamientos:

- Se selecciona arbitrariamente el centro del primer agrupamiento.
- Se procesan secuencialmente los demás patrones mediante cálculos de distancia.

Cada M patrones se mezclan agrupamientos. Estos pueden ser

- Mezcla por cercanía
- Mezcla por tamaño
- Mezcla forzada

Las reglas de asociación con datos secuenciales se presentan cuando los datos contiguos presentan algún tipo de relación. La clasificación con datos secuenciales expresa patrones de comportamiento en instantes distintos (pero cercanos) de tiempo.

Predicción

Para hacer un buen modelo de predicción, es necesario definir adecuadamente el problema, recopilar los datos a utilizar en el modelo, elegir una medida o indicador de éxito y preparar los datos, ya sea tratando los campos vacíos, convirtiendo variables categóricas, etc.

Es recomendable dividir el conjunto de datos en un conjunto de entrenamiento, utilizado para construir el modelo, un conjunto de pruebas, a partir del cual se realizarán ajustes al modelo, y un conjunto de validación, con el cual se verifica la exactitud de las predicciones. Es común dividir los datos en 70% para el conjunto de entrenamiento y 15% para tanto el conjunto de pruebas como el conjunto de validación.

Un **árbol de decisión** es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable de respuesta o dependiente. Los árboles se pueden clasificar en dos tipos:

- Árboles de regresión, en los cuales la variable respuesta es cuantitativa.
- Árboles de clasificación, en los cuales la variable respuesta es cualitativa.

Los árboles están formados por nodos y su lectura se realiza de arriba hacia abajo.

- **Primer nodo** o **nodo raíz:** En él se produce la primera división en función de la variable más importante.
- **Nodos internos** o **intermedios:** Vuelven a dividir el conjunto de datos en función de las variables.
- **Nodos terminales** u **hojas:** Su función es dar la clasificación definitiva.

Un **árbol de clasificación** consiste en hacer preguntas de tipo $x_k \leq c$ para las covariables cuantitativas y preguntas de tipo $x_k = nivel_j$ para las covariables cualitativas. De esta forma el espacio de las covariables es dividido en hiper-rectángulos, y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor grupo estimado.

Hay dos tipos de nodo:

- **Nodo de decisión:** Tienen una condición al principio y más nodos debajo de ellos.
- **Nodo de predicción:** No tienen ninguna condición ni nodos debajo de ellos. Al llegar a este nodo, se asigna una clase a la muestra.

Un **árbol de regresión** consiste en hacer preguntas de tipo $x_k \leq c$ para cada una de las covariables. De esta forma el espacio de las covariables es dividido en hiper-rectángulos, y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado $\bar{y}$.

Para realizar la partición del espacio se encuentra la covariable y el punto de corte sobre esta covariable que permita predecir mejor la variable respuesta. Estos pasos se repiten hasta alcanzar un criterio de parada.

Un **bosque aleatorio** es una técnica de aprendizaje automático basada en árboles de decisión. Su principal ventaja es que se obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar. Esta mejora en la generalización la consigue compensando los errores de las predicciones de los distintos árboles de decisión.

Una forma de mejorar el modelo es usando la técnica denominada **bagging**. Primero se seleccionan individuos al azar, usando muestreo con remplazo, para crear diferentes conjuntos de datos, y se crea un árbol de decisión con cada conjunto. Al crear los árboles se eligen variables al azar en cada nodo del árbol. El modelo predice los nuevos datos usando un “voto mayoritario”, donde la clasificación corresponde a la predicción de la mayoría de los árboles.

Algunas métricas de eficacia utilizadas son:

- Error cuadrático medio: Mide la diferencia del estimador y lo que se estima.
- Curva ROC: El área bajo la curva mide el rendimiento global de la prueba.