2-Las entradas del proceso de Data Mining

Jhon J. Padilla A.

Las entradas

- La entrada toma la forma de conceptos, instancias y atributos.
- Concepto: La Cosa que es aprendida
- La idea de un concepto es lo que estamos tratando de encontrar, el resultado de un proceso de aprendizaje.
- Un concepto es una descripción que es entendible, discutible y disputable, y es operacional o puede ser aplicado a ejemplos reales.

Qué es un concepto?

- En las aplicaciones de Minería de Datos hay cuatro diferentes estilos de aprendizaje.
- En el **aprendizaje por clasificación**, el esquema de aprendizaje es presentado con un conjunto de ejemplos desde los cuales se espera aprender una manera de clasificar ejemplos no vistos antes.
- En el **aprendizaje por asociación**, se busca cualquier asociación entre características, no solo uno que prediga un valor de clase particular.
- En el **Clustering**, se buscan grupos de ejemplos que están juntos. En predicción numérica, la salida a predecirse no es una clase discreta, sino una cantidad numérica
- Sin importar el tipo de aprendizaje involucrado, llamamos a la cosa a ser aprendida el concepto y a la salida producida por un esquema de aprendizaje la descripción del concepto.

Qué es un ejemplo?

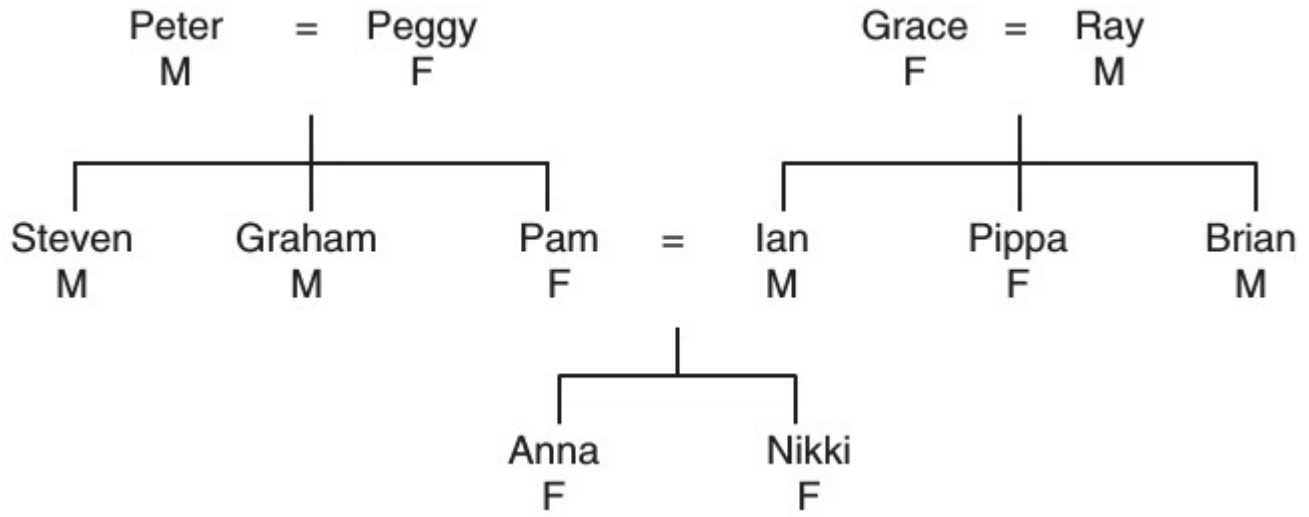
- La entrada a un esquema de aprendizaje de maquina es un conjunto de instancias.
- Las instancias son las cosas que serán clasificadas, asociadas o agrupadas
- Aunque hasta ahora las hemos llamado ejemplos, a partir de ahora utilizaremos el termino instancia para referirnos a la entrada.
- Cada instancia es un ejemplo independiente e individual del concepto a ser aprendido. Adicionalmente, cada uno es caracterizado por los valores de un conjunto de atributos predeterminados.
- Esto es el caso de los ejemplos estudiados hasta el momento (clima, lentes de contacto, iris, problemas de negociación laboral).
- Cada conjunto de datos esta representado como una matriz de instancias versus los atributos, lo cual es una relación simple en términos de bases de datos, o un archivo plano.

- Expresar los datos de entrada como un conjunto de instancias independientes es la más común de las situaciones para la Minería de datos en la práctica.
- Sin embargo, esta es una manera restrictiva de formular los problemas.
- Los problemas suelen involucrar relaciones entre objetos en lugar de tener instancias separadas e independientes.

Ejemplo

- Suponga que se tiene un árbol familiar y queremos aprender el concepto de hermano. Imagine su propio árbol familiar, con sus parientes y sus géneros ubicados en los nodos.
  Este árbol es la entrada al proceso de aprendizaje, junto con una larga lista de pares de personas y una identificación de si son hermanos o no.

Árbol familiar



Otra forma de expresar relaciones de familia

first person	second person	sister of?
Peter	Peggy	no
Peter	Steven	no
...	.....	
Steven	Peter	no
Steven	Graham	no
Steven	Pam	yes
Steven	Grace	no
...	.....	
Ian	Pippa	yes
...	.....	
Anna	Nikki	yes
...	.....	
Nikki	Anna	yes

first person	second person	sister of?
Steven	Pam	yes
Graham	Pam	yes
Ian	Pippa	yes
Brian	Pippa	yes
Anna	Nikki	yes
Nikki	Anna	yes
All the rest		no

Figure 2.1 A family tree and two ways of expressing the sister-of relation.

- Hay una gran cantidad de NO en la tercera columna de la tabla de la izquierda; hay 12*12=144 posibles combinaciones, y muchas de ellas no cumplen con la condición de hermandad.
- La tabla de la derecha, da la misma información, pero solo registra las instancias positivas y asume que todas las demás son negativas. La idea de especificar solo los ejemplos positivos y adoptar un supuesto de que el resto son negativas es llamado el **supuesto de mundo cerrado**. Esto se asume frecuentemente en estudios teóricos, pero en la practica estos supuestos no son muy útiles

- La tabla de la Figura 2.1 no es de ninguna utilidad sin el árbol familiar.
- El árbol puede expresarse en la forma de una tabla, parte de la cual se expresa en la tabla 2.3. Ahora, el problema es expresado en términos de dos relaciones. Pero estas tablas no contienen conjuntos independientes de instancias porque los valores en las columnas Name, Parent1, and Parent2 de la relación de hermandad se refieren a la relación del árbol familiar.
- Podemos construirla en un conjunto único de instancias, fusionando las dos tablas en una única tabla como la Tabla 2.4.

Table 2.3 **Family tree represented as a table.**

Name	Gender	Parent1	Parent2
Peter	male	?	?
Peggy	female	?	?
Steven	male	Peter	Peggy
Graham	male	Peter	Peggy
Pam	female	Peter	Peggy
Ian	male	Grace	Ray
. . .			

Table 2.4 **The sister-of relation represented in a table.**

First person				Second person				
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	Sister of?
Steven	male	Peter	Peggy	Pam	female	Peter	Peggy	yes
Graham	male	Peter	Peggy	Pam	female	Peter	Peggy	yes
Ian	male	Grace	Ray	Pippa	female	Grace	Ray	yes
Brian	male	Grace	Ray	Pippa	female	Grace	Ray	yes
Anna	female	Pam	Ian	Nikki	female	Pam	Ian	yes
Nikki	female	Pam	Ian	Anna	female	Pam	Ian	yes
				all the rest				no

Qué es un atributo?

- Una **instancia** es una fila de las tablas que contienen los datos de entrada.
- Cada instancia a su vez tiene diferentes campos o **Atributos**.

Restricciones de la representación por tablas

- Esta representación impone una estructura fija a los atributos e impide describir situaciones más flexibles o dinámicas que se presentan en Minería de Datos.
- Ejemplo de datos flexibles:
  - Si hay diferentes tipos de medios de transporte: camiones, barcos, aviones, etc. Un atributo de un camión podría ser número de ruedas, pero no aplica para un barco. El barco puede variar en el número de mástiles, pero el carro no los posee...

- Una solución para esta situación es crear atributos que tengan un valor “irrelevante” para indicar que el atributo no está disponible para un caso particular (p.ej. “N/A”-No Aplica)
- Otro ejemplo de flexibilidad es el apellido de una mujer si es casada o soltera. Si es casada, depende del apellido del esposo.

Tipos de Atributos

- Hay dos tipos:
  - **Numérico:** Contienen números enteros o reales. También llamados *contínuos* (aunque los números enteros no lo son).

    También pueden darse en diferentes formas como: valor nominal, ordinal, intervalo, razón.
  - **Nomina**l: También llamados *categóricos*. Los valores sirven como etiquetas o nombres.

Otros tipos de Atributos

- Algunos atributos pueden tomar valores de diferentes tipos:
  - Nominal
  - Ordinal
  - Intervalo
  - Razón

Atributo Nominal

- Ejemplo:
  - El panorama puede estar:
    - Soleado
    - Nublado
    - LLuvioso

Atributo Ordinal

- Son aquellas que pueden ser ordenadas.
- Puede ser que no sean numéricas pero tienen una noción de orden.
- Ejemplo: La temperatura puede ser:
  - Caliente
  - Templada
  - Fría
- Estos valores pueden ser ordenados: Caliente>Templado>Frío
- Pero no pueden ser restados o sumados.
- Se pueden aplicar reglas que usen comparaciones: “Si (Temperatura = Caliente) Entonces Encienda ventilador”

Atributo: Razón

- Se usa para medir con respecto a un valor o punto de referencia.
- Ejemplo:
  - Distancia al centro de la ciudad. El centro de la ciudad se toma como punto de referencia
- Se tratan como números y se pueden hacer operaciones matemáticas con ellas (sumas, multiplicaciones, etc).
- El punto de referencia es relativo a la situación que se quiere modelar.

Atributo: Intervalo

- Puede referirse a dos límites de una medición.
- P.ej:
  - del año 1939 a 1945, se tiene un intervalo de 6 años
  - Del año 1949 a 1955 se tiene igualmente un intervalo de 6 años.
- No tiene mucho sentido sumar o multiplicar los años.

Preparando las Entradas

- Los datos que se usan para hacer minería de datos suelen venir desordenados y sin estructura.
- Se requiere darles cierta estructura y orden antes de ingresarlos al proceso de minería de datos.
- Los datos de entrada deben estar en un formato apropiado para el programa que hace la minería de datos.
- P.ej: El programa Weka requiere un formato para sus archivos de entrada que tienen una extensión ARFF.

Preparando las Entradas

- La organización de los datos originales en un orden mínimo para el procesamiento se llama **Limpieza de los datos**.
- Este proceso ahorra grandes cantidades de trabajo en el proceso de minería de datos.
- La limpieza de los datos suele ser un proceso que conlleva un buen esfuerzo.

Juntando datos de diferentes fuentes

- En la práctica, suelen recibirse datos de diferentes fuentes.
- P.ej: datos de un estudio de mercadeo se pueden tomar del departamento de ventas, del departamento de servicio al cliente y del departamento de facturación.

Juntando datos de diferentes fuentes

- Integrar los datos en una sola base de datos a partir de diferentes fuentes, suele tener sus desafíos:
  - Diferentes estilos de llevar los registros
  - Diferentes convenciones
  - Diferentes períodos de tiempo
  - Diferentes grados de agregación de datos
  - Diferentes claves primarias de búsqueda
  - Diferentes márgenes de error

Data Warehousing

- Limpieza de los datos: Requiere ensamblaje, integración y limpieza.
- A la idea de una base de datos integrados de toda una empresa se le conoce como Data Warehousing.
- Data Warehouse ofrece un único punto de acceso a todos los datos organizacionales, trascendiendo los departamentos.
- Data Warehousing toma datos parciales de cada departamento y los integra en una base de datos que permite tomar decisiones más acertadas con base en los datos completos y no solo una parte de ellos.
- **Overlay data:** Datos que no han sido recogidos en el Data Warehouse, y que son necesarios para una situación particular. Estos deben ser preparados para el proceso de Datamining.

Agregación de datos

- Para muchas situaciones se requiere agregar los datos en intervalos.
- Por ejemplo: las ventas de un mes, de un trimestre, de un cuatrimestre.
- Pero cuál nivel de agregación es mejor?
- La selección adecuada del nivel de agregación suele ser clave para el análisis.

Concluyendo...

- El ensamblaje, la integración, la limpieza, la agregación de los datos, suelen tomar su tiempo para hacerse si se desea obtener unos buenos resultados en la minería de datos.

Formato ARFF

Def. Atributos

Comentario

instancias

```
% ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play? { yes, no }

@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
rainy, 75, 80, false, yes
sunny, 75, 70, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rainy, 71, 91, true, no
```

Formato ARFF

- Sirve para describir las tablas o relaciones.
- Se define una relación llamada weather mediante:

  @relation weather

- Las especificaciones de los atributos (definidas con @attribute) permiten al programa lector del archivo ARFF determinar si están correctos los valores asignados para cada instancia.

Formato ARFF

- Además de los atributos del ejemplo, se pueden definir dos tipos más:
  - String (ej: @attribute description string)
  - Date (ej: @attribute today date)
- String:
  - La cadena de caracteres debe estar descrita entre comillas.
  - La cadena podría ser un documento entero, por lo que se debe poder manipular y hacer tareas como por ejemplo buscar una palabra y colocar el número de veces que aparece.

Formato ARFF

- Date:
  - Formato: yyyy-MM-dd-THH:mm:ss
  - La T indica que a continuación viene la hora
  - Un parámetro tipo fecha es una cadena de caracteres que se convierte a números en el programa lector.
- Ejemplo: 2004-04-03T12:00:00
- Estos parámetros pueden ser muy útiles para encontrar patrones de comportamiento en el tiempo.

Tablas con datos dispersos

```
0, 26, 0,  0, 0, 0, 63, 0, 0, 0, "class A"
0,  0, 0, 42, 0, 0,  0, 0, 0, 0, "class B"
```

```
{1 26, 6 63, 10 "class A"}
{3 42, 10 "class B"}
```

Columna de ubicación
(inicia en cero)

Valor
(los que no se indican aquí son tomados como cero)
?: indica un valor desconocido