Data Mining

Jhon Jairo Padilla Aguilar, PhD.

(Extraido del libro Data Mining: Practical Machine Learning Tools and Techniques. Autores: Ian H. Witten and Eibe Frank)

Definición

- Se define como el proceso de descubrir patrones en los datos.
- El proceso debe ser automático ó también puede ser semi-automático.
- Los patrones descubiertos deben ser útiles y deben proveer alguna ventaja que suele ser económica.
- Los datos suelen estar presentes en grandes cantidades.

Cómo se expresan los patrones?

- Los patrones son útiles porque nos permiten hacer predicciones no triviales sobre nuevos datos.
- Hay dos extremos para la expresión de un patrón:
  - Caja Negra: No se comprende la estructura de los datos
  - Caja transparente: Los datos revelan la estructura del patrón
- Los patrones que permiten analizar su estructura se denominan Estructurales, porque capturan la estructura de decisión en una forma explicita. Es decir, ayudan a explicar algo de los datos.

Objetivo del curso

- Este curso se enfoca en las técnicas para encontrar y describir patrones estructurales en los datos.
- La mayoría de las técnicas que se cubren han sido desarrolladas dentro de un campo conocido como Machine Learning.

Descripción de Patrones estructurales

Necesidad

- Estamos interesados en técnicas para encontrar y describir patrones estructurales en los datos como una herramienta para ayudar a explicar esos datos y hacer predicciones a partir de ellos.

Data Mining: Entradas y Salidas

- Los datos tomarán la forma de un conjunto de ejemplos.
- Por ejemplo, clientes que cambian sus preferencias, o situaciones en las cuales cierta clase de lentes de contacto pueden ser formuladas.
- La salida toma la forma de predicciones sobre nuevos ejemplos:
  - una predicción de si un cliente particular cambiará
  - o una predicción de qué clase de lentes deberán prescribirse bajo ciertas circunstancias.

Data Mining: Salidas

- La salida debería incluir también una descripción real de una estructura que pueda ser usada para clasificar ejemplos desconocidos para explicar la decisión
- Esto también es útil para suplir una representación explicita del conocimiento que es adquirido.
- Esto refleja ambas definiciones de aprendizaje consideradas previamente: adquisición de conocimiento y capacidad para usarlo.

Data Mining: representación del conocimiento (salidas)

- Muchas técnicas de aprendizaje buscan descripciones estructurales de lo que se aprendió Estas descripciones pueden ser muy complejas y son expresadas típicamente como conjuntos de reglas o arboles de decisión

- Debido a que estas pueden ser entendidas por la gente, estas descripciones sirven para explicar lo que ha sido aprendido y explicar la base para nuevas predicciones.

- La experiencia muestra que las estructuras explicitas de conocimiento que se adquieren, las descripciones estructurales, son al menos tan importantes, y a menudo mas importantes, que la capacidad de aplicarlas en nuevos ejemplos.

- La gente frecuentemente usa Data Mining para ganar conocimiento, no solo para predicciones.

Ejemplo: Problema del Clima

- Este problema supone que debe haber unas condiciones para jugar un juego cualquiera.
- En general, las instancias en un conjunto de datos están caracterizadas por los valores de ciertas características, o atributos, que miden diferentes aspectos de la instancia.
- En este caso hay 4 atributos: panorama, temperatura, humedad, viento.
- La salida es si se jugara o no.

Ejemplo: Problema del Clima

- En su forma simple, todos los cuatro atributos tienen valores que son categorías simbólicas en lugar de numeros.
- El panorama puede ser: soleado, nublado, o lluvioso;
- La temperatura puede ser: caliente, leve, o fresca;
- La humedad puede ser: alta, normal;
- El viento puede ser: verdadero, falso.
- Esto crea 36 posibles combinaciones (3 * 3 *2 * 2 = 36), de las cuales 14 están presentes en el conjunto de ejemplos de entradas.

Conjunto de reglas obtenido

Table 1.2 The weather data.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true    then play = no
If outlook = overcast                  then play = yes
If humidity = normal                   then play = yes
If none of the above                   then play = yes
```

1. Listas de Decisión

- Estas reglas se deben interpretar en orden: la primera, si no aplica, entonces la segunda, y así sucesivamente.
- Un conjunto de reglas que son establecidas e interpretadas en secuencia se conocen como una **Lista de decisión**
- Interpretadas como una lista de decisión, las reglas clasifican correctamente todos los ejemplos en la tabla, mientras que si son tomadas individualmente, fuera de contexto, algunas reglas son incorrectas.
- Por ejemplo, la regla: “if humidity = normal then play = yes”, da un ejemplo erróneo (chequear cuál)
- El significado de un conjunto de reglas depende de cómo son interpretadas (no es sorprendente!)

1. Listas de Decisión (Reglas de Clasificación)

- Un problema un poco mas complejo se muestra en la tabla siguiente.
- Dos de los atributos (temperatura, humedad) tienen valores numéricos
- Esto significa que cualquier método de aprendizaje debe crear desigualdades que envuelven estos atributos en lugar de simples pruebas de igualdad
- Esto se llama el **problema de atributo numérico**, en este caso un problema de atributos mezclados porque no todos los atributos son numéricos
- Por tanto, la primera regla dada antes podría tomar la forma:

```
If outlook = sunny and humidity > 83 then play = no
```

Datos del Clima

Table 1.3 Weather data with some numeric attributes.

Outlook	Temperature	Humidity	Windy	Play
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

2. Reglas de Asociación

- Las reglas vistas son reglas de Clasificación: ellas predicen la clasificación de los datos en términos de si se juega o no.
- Es igualmente posible ser indiferente a la clasificación y solo buscar reglas que asocien fuertemente diferentes valores de atributos. Estas son llamadas **Reglas de Asociación.** Muchas reglas de asociación pueden ser derivadas de los datos del clima.

Reglas de Asociación

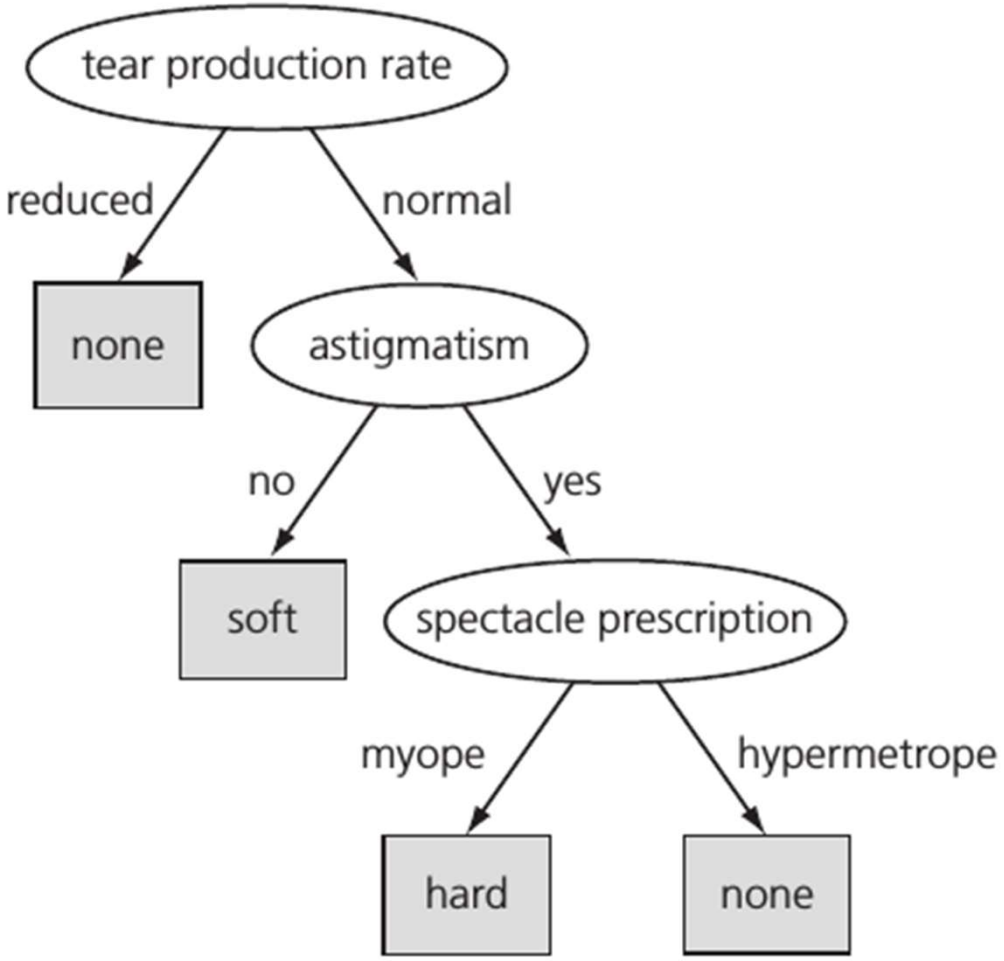
- Algunas reglas buenas pueden ser:

```
If temperature = cool                          then humidity = normal
If humidity = normal and windy = false         then play = yes
If outlook = sunny and play = no               then humidity = high
If windy = false and play = no                 then outlook = sunny
                                                    and humidity = high.
```

Reglas de Asociación

- Todas estas reglas son 100% correctas sobre los datos dados, no hacen predicciones falsas.
- Las dos primeras aplican para cuatro ejemplos en el conjunto de datos, la tercera a tres ejemplos, y la cuarta a dos ejemplos.
- Hay muchas otras reglas: de hecho, pueden encontrarse unas 60 reglas de asociación que aplican a dos o mas ejemplos de los datos del clima y son completamente correctas con estos datos.
- Si busca reglas que son menos del 100% correctas, entonces podrá encontrar muchas mas.
- Hay muchas porque a diferencia de las reglas de clasificación, las reglas de asociación pueden predecir cualquiera de los atributos, no solo una clase especificada, y pueden predecir mas de una cosa.
- Por ejemplo, la cuarta regla predice que ambos, el panorama será soleado y que la humedad será alta.

3. Árboles de decisión



Árbol de decisión para datos de negociaciones laborales

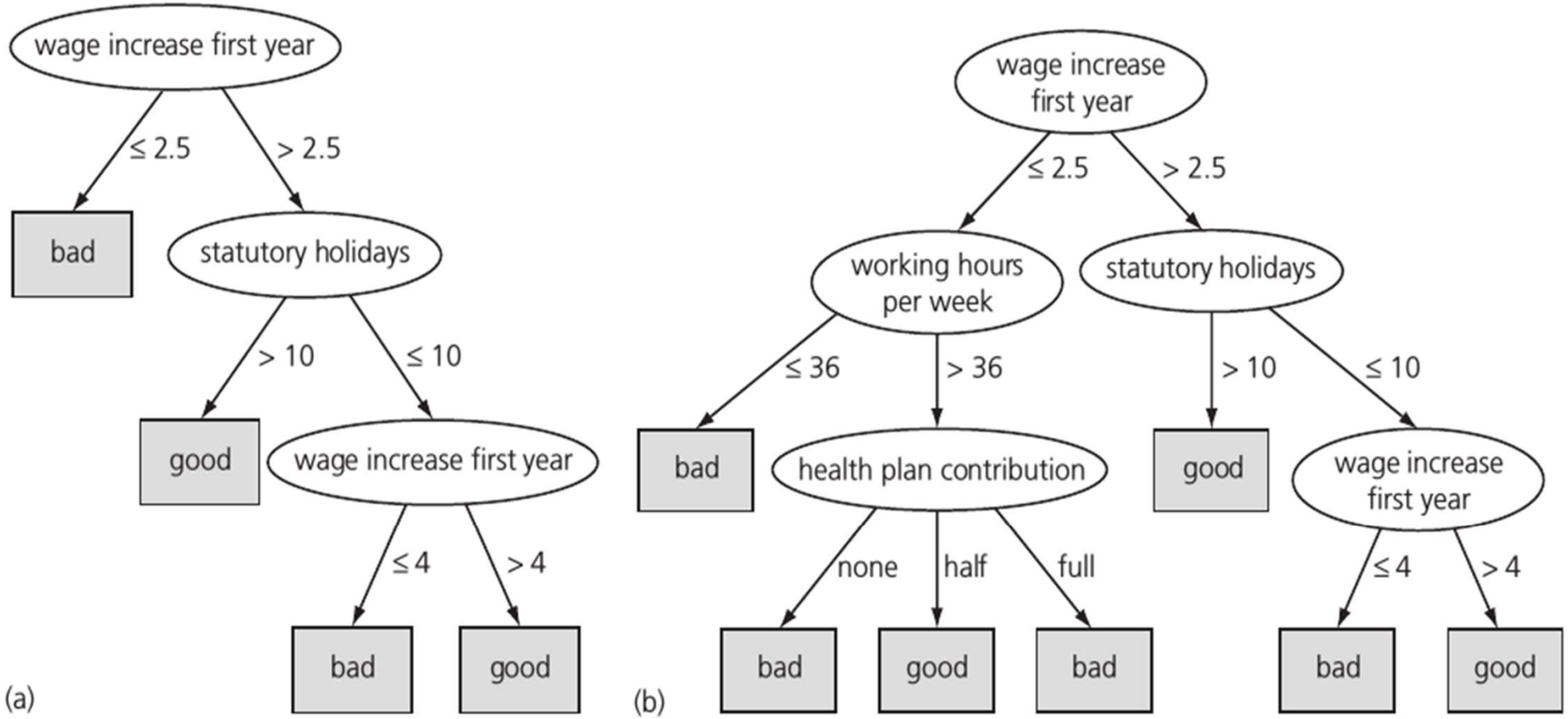


Figure 1.3 Decision trees for the labor negotiations data.

```
If    [leaf condition is normal and
       stem condition is abnormal and
       stem cankers is below soil line and
       canker lesion color is brown]
then
       diagnosis is rhizoctonia root rot

If    [leaf malformation is absent and
       stem condition is abnormal and
       stem cankers is below soil line and
       canker lesion color is brown]
then
       diagnosis is rhizoctonia root rot
```

Table 1.7 The soybean data.

	Attribute	Number of values	Sample value
Environment	time of occurrence	7	July
	precipitation	3	above normal
	temperature	3	normal
	cropping history	4	same as last year
	hail damage	2	yes
	damaged area	4	scattered
	severity	3	severe
	plant height	2	normal
	plant growth	2	abnormal
	seed treatment	3	fungicide
	germination	3	less than 80%
Seed	condition	2	normal
	mold growth	2	absent
	discoloration	2	absent
	size	2	normal
	shriveling	2	absent
Fruit	condition of fruit pods	3	normal
	fruit spots	5	—
Leaf	condition	2	abnormal
	leaf spot size	3	—
	yellow leaf spot halo	3	absent
	leaf spot margins	3	—
	shredding	2	absent
	leaf malformation	2	absent
	leaf mildew growth	3	absent
Stem	condition	2	abnormal
	stem lodging	2	yes
	stem cankers	4	above soil line
	canker lesion color	3	—
	fruiting bodies on stems	2	present
	external decay of stem	3	firm and dry
	mycelium on stem	2	absent
	internal discoloration	3	none
	sclerotia	2	absent
Root	condition	3	normal
Diagnosis		19	diaporthe stem canker