Data Mining basado en Teoría de la Información

Marcelo Ferreyra

Temas a tratar

- Etapas de un proyecto de Data Mining
- Introducción a la Teoría de la Información
- La Teoría de la Información como base de DM
- Ventajas del enfoque de la TI
- Ejemplos
- Preguntas

Etapas de un proyecto

1. Definir el problema a resolver
2. Reunir los datos y darle forma de tabla
3. Data Mining
   I. Transformar (preparar) los datos
   II. Seleccionar las mejores variables
   III. Realizar los modelos
   IV. Probarlos con datos nuevos
4. Entregar el modelo a sistemas (si hace falta)
5. Actuar en base a lo descubierto
6. Medir los resultados

Etapas de un proyecto

1. Definir el problema a resolver
2. Reunir los datos y darle forma de tabla
3. Data Mining
   I. Transformar (preparar) los datos
   II. Seleccionar las mejores variables
   III. Realizar los modelos
   IV. Probarlos con datos nuevos
4. Entregar el modelo a sistemas (si hace falta)
5. Actuar en base a lo descubierto
6. Medir los resultados

Transformar los datos

Es un paso crítico

Si no se transforman las variables, el modelo resultará de inferior calidad, incluso inservible

Se necesita conocimiento estadístico

No es simple transformar variables. Hay un libro de más de 500 páginas dedicado exclusivamente a este tema

Lleva mucho tiempo

Dependiendo de la cantidad de variables, podría llevar días, incluso semanas

Transformar los datos

Es un paso crítico

Un modelo realizado con datos preparados

- ✓ Es más preciso
- ✓ Puede ser desarrollado con herramientas más simples
- ✓ Es más confiable
- ✓ Es más fácil de entender

Transformar los datos

Se necesita conocimiento estadístico

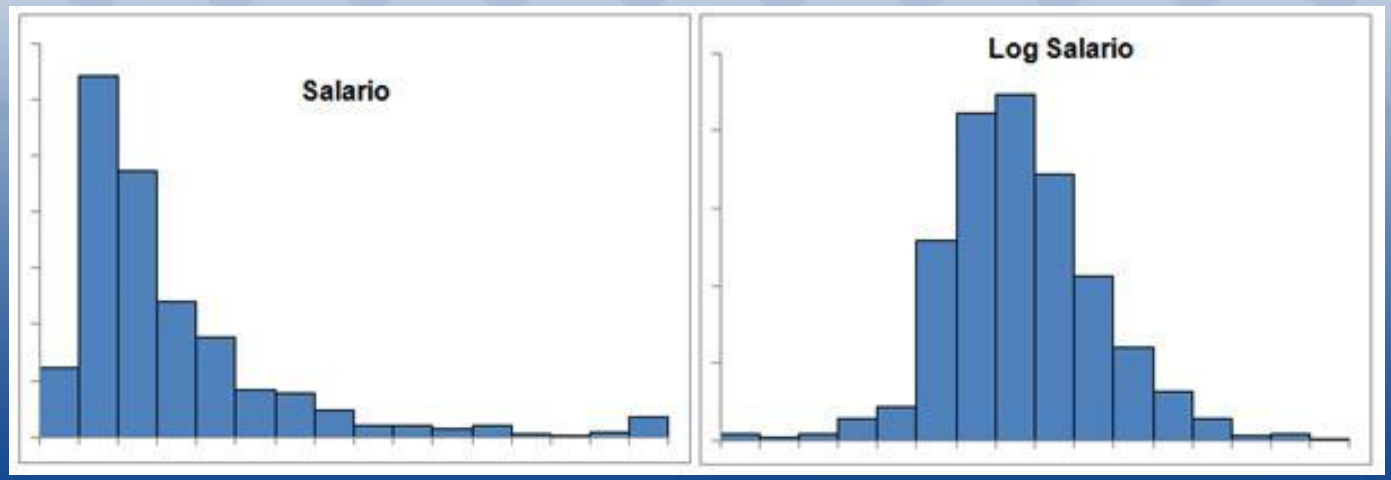
Una vez ensamblados los datos en una tabla, recién empieza el trabajo de preparación de las variables:

- ✓ Hay que decidir que hacer con los outliers (valores muy lejos del promedio)
- ✓ Hay que decidir cómo tratar los valores nulos
- ✓ Hay que modificar las distribuciones de las variables numéricas
- ✓ Dependiendo de la herramienta de modelado, hay que numerar las variables categóricas

Transformar los datos

Cambio en la distribución. Ejemplo

Aplicar un logaritmo a una variable que representa dinero, mejora el acceso a la información



Selección de Variables

Una base de datos podría contener cientos de variables

- ✓ Cuantas más variables mejor
- ✓ Es una práctica común derivar nuevas variables. Esto incrementa aún más el número de variables.

Un modelo debe realizarse con pocas variables

- ✓ Para que un modelo sea fácil de entender debe estar basado en muy pocas variables, usualmente de 5 a 10
- ✓ Un modelo con muchas variables corre el riesgo de no ser confiable

Es necesario reducir la cantidad de variables

Selección de Variables

Reducción de la dimensionalidad

- ✓ Utilizar métodos como PCA no es buena opción.
- ✓ Elegir las variables utilizando un criterio estrictamente de negocio, no es una buena práctica.
- ✓ Además de seleccionar qué variables utilizar, se debe decidir cuántas.
  - Si se eligen pocas, el modelo perderá precisión.
  - Si se eligen muchas, el modelo se torna complejo y tiene mayores chances de no funcionar correctamente cuando se lo pone en producción (técnicamente se dice que el modelo está sobre-entrenado)

Selección de Variables

Se necesita un algoritmo capaz de seleccionar las variables

Pero la selección es una tarea compleja. Por ejemplo, existen más de 1.000 millones de combinaciones para elegir 6 variables entre 100 posibles.

Modelo

Un modelo debe ser confiable y preciso

Es *confiable* si funciona bien aún con datos nuevos

Es *preciso* si sus predicciones son correctas

Los dos pasos anteriores son importantes para un buen modelo

- ✓ Una buena transformación y selección de variables favorecen enormemente la creación de un modelo confiable y preciso.
- ✓ Si la transformación de variables está bien hecha, la elección de la herramienta de modelado es secundaria

Pruebas

Un modelo debe probarse con datos nuevos

- ✓ Para tener la seguridad de que el modelo funcionará bien en producción, debe probarse con datos nuevos
- ✓ Es un error común saltear esta prueba o utilizar datos aparentemente nuevos pero que en realidad no lo son
- ✓ Utilizar un modelo incorrecto puede hacernos perder mucho dinero, ya que se descubrirá su falla sólo después de usarlo con datos reales

Temas a tratar

- Etapas de un proyecto de Data Mining
- Introducción a la Teoría de la Información
- La Teoría de la Información como base de DM
- Ventajas del enfoque de la TI
- Ejemplos
- Preguntas

Información

Vivimos en una sociedad de Información

Las bases de datos contienen Información

Hablamos de tecnología de la Información

Pero ¿qué es la Información?

¿Podemos medirla?

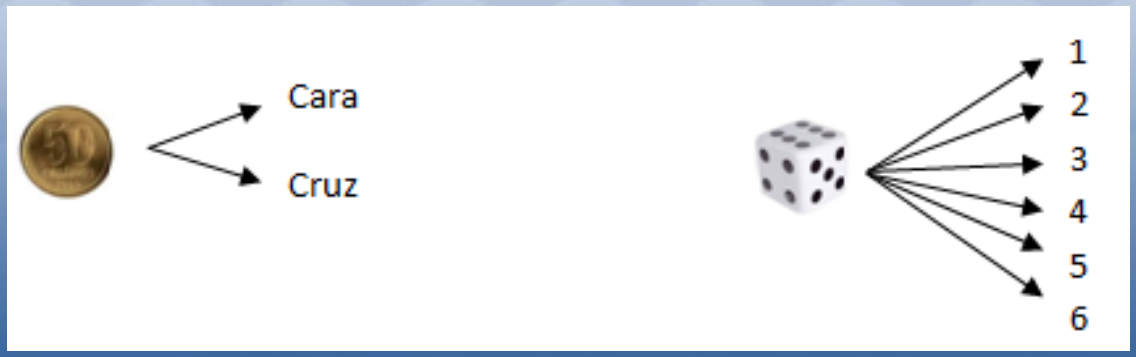
¿Qué es el Ruido? ¿Cómo se relaciona con la Información?

Información e Incertidumbre

Una fuente de información se comunica por medio de *mensajes*

Una moneda contiene dos *posibles* mensajes para comunicar

Un dado contiene 6 *posibles* mensajes



Hay mayor grado de incertidumbre en el mensaje que puede comunicar un dado, que una moneda, entonces decimos que el dado contiene mayor información que la moneda

Información y Entropía

La Entropía es una medida de la cantidad de información que contiene una fuente y su unidad es el bit

Shannon especificó cómo se mide la información por medio de la siguiente ecuación

$$H = -\sum_{i}^{n} p_i * \log_2(p_i)$$

En el caso de la moneda

Lanzamiento de una moneda			
Mensaje	p	$-\log_2(p)$	$-p*\log_2(p)$
Cara	0.5	1	0.5
Cruz	0.5	1	0.5
		H	1

Información y Entropía

¿Qué pasa si la moneda está cargada?

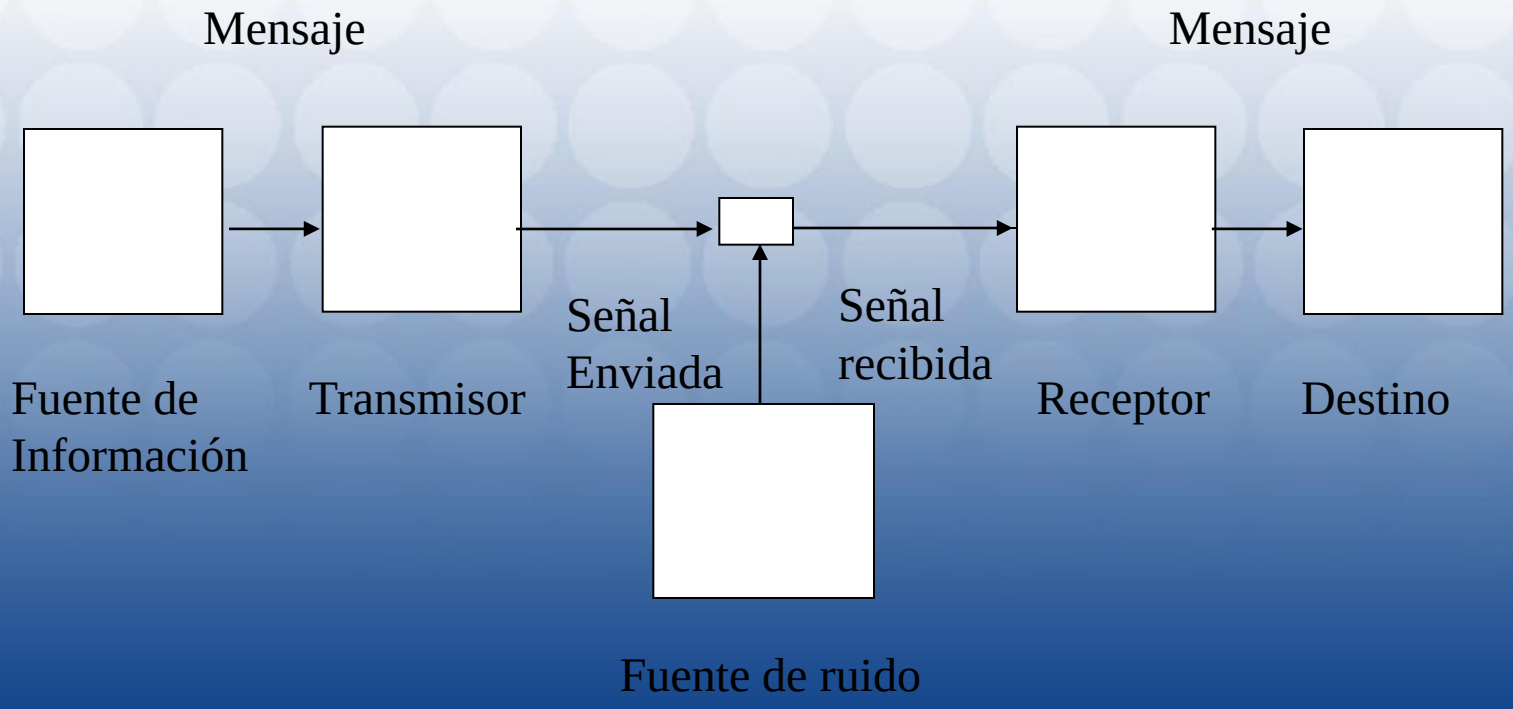
Si las probabilidades no son iguales la entropía es menor

$$H = -\sum_{i}^{n} p_i * \log_2(p_i)$$

Lanzamiento de una moneda			
Mensaje	p	$-\log_2(p)$	$-p*\log_2(p)$
Cara	0.3	1.737	0.521
Cruz	0.7	0.515	0.36
		H	0.881

Teoría de la Información

Sistema de comunicación planteado por Shannon



The Mathematical Theory of Communication, Shannon & Weaver

Entropía de una Variable

Para medir la entropía de una variable se utiliza:

$$H = -\sum_{i}^{n} p_i * \log_2(p_i)$$

Variables Categóricas

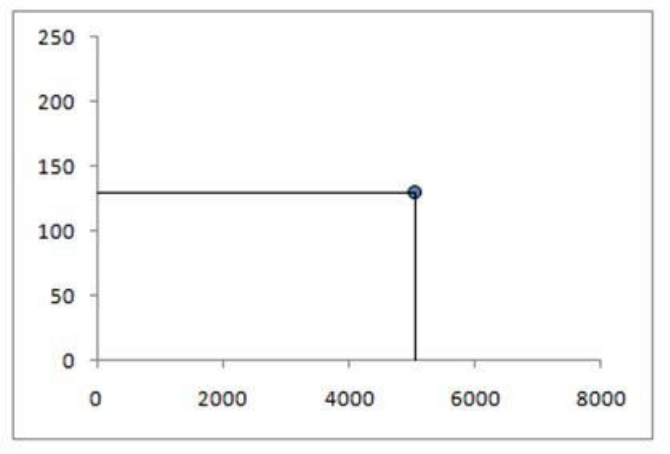
Las probabilidades se calculan en base a la ocurrencia de cada categoría

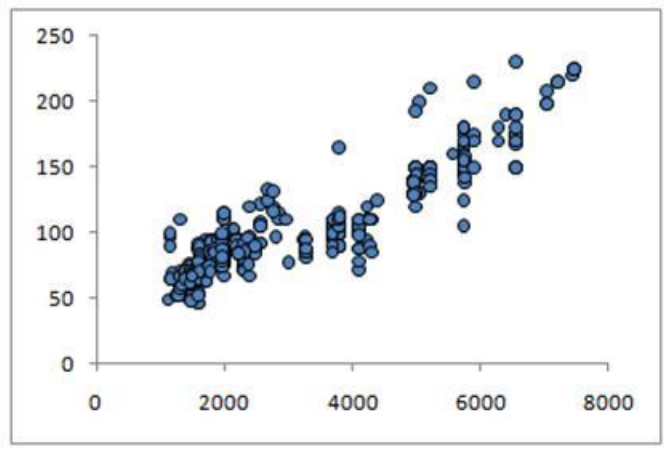
Variables Numéricas

Se convierten en categóricas mediante algún algoritmo apropiado, y se procede como en el caso anterior

Entropía de varias variables

Un conjunto de variables forman un espacio de estados

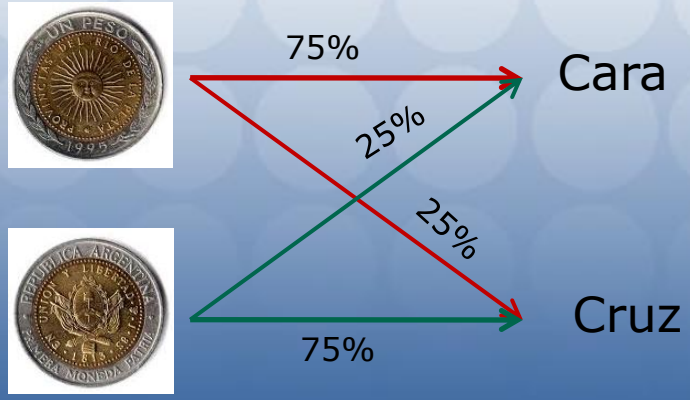




Cada estado representa un posible mensaje de la fuente de información, que se usará para calcular la entropía

Entropía relativa

Cuando existe ruido, la señal recibida no siempre es igual a la transmitida.



Ruido y Equivocación

Ruido

Conocemos el mensaje enviado, pero tenemos incertidumbre respecto del recibido.

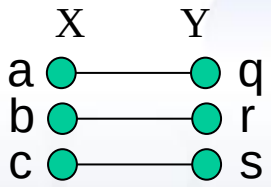
¿Cuál es la probabilidad de que el mensaje recibido haya sido Cara, si el mensaje enviado fue Cara?

Equivocación

Conocemos el mensaje recibido, pero tenemos incertidumbre respecto del enviado.

¿Cuál es la probabilidad de que el mensaje enviado haya sido Cara, si el mensaje recibido fue Cara?

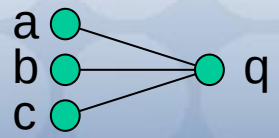
Ruido y Equivocación

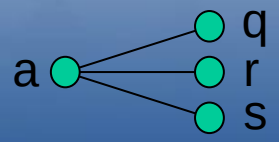


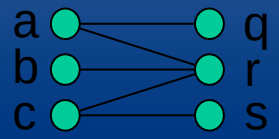
Ideal

El caso más simple es cuando hay una relación biunívoca entre las señales de X e Y

En este caso la información se transmite sin interferencias

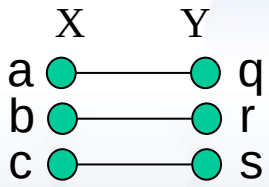






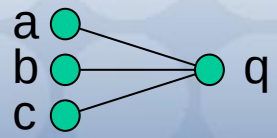
X \ Y	q	r	s
a	1.0		
b		1.0	
c			1.0

Ruido y Equivocación

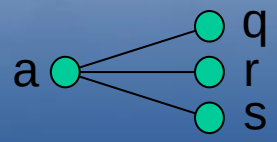


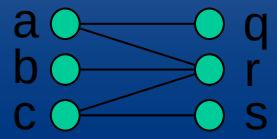
El segundo caso es cuando varias señales distintas de entrada apuntan a una sola señal de salida.

Muchas voces distintas están diciendo lo mismo.



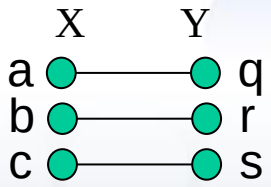
Equivocación

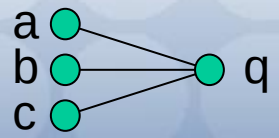


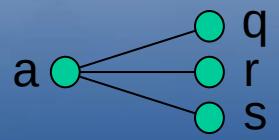


		Y: q	Y: r	Y: s
X	a	1.0		
	b	1.0		
	c	1.0		

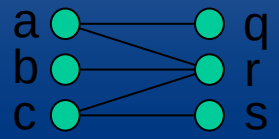
Ruido y Equivocación







Ruido

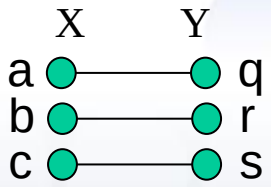


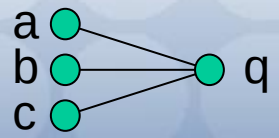
Cuando la relación contiene Ruido, una misma señal de entrada apunta a distintas señales de salida.

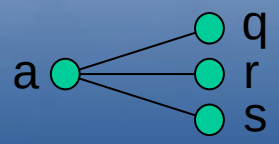
La señal de salida es incierta para una determinada señal de entrada

X \ Y	q	r	s
a	0.3	0.3	0.3
b			
c			

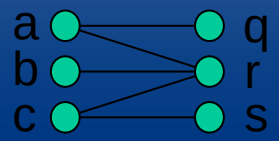
Ruido y Equivocación







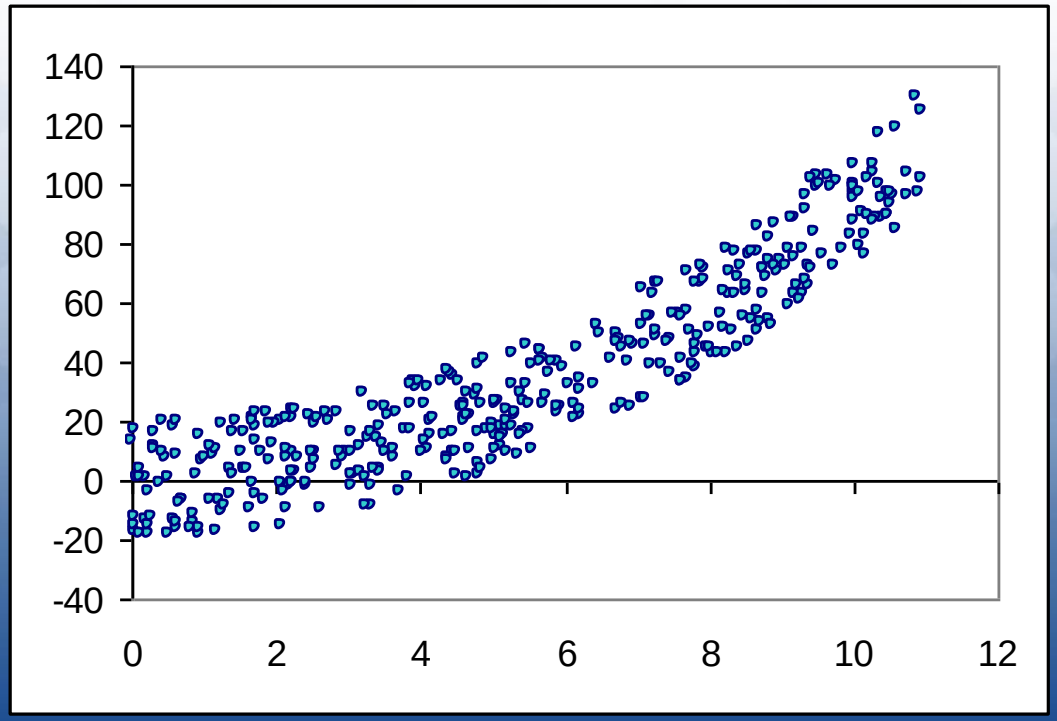
Ruido y Equivocación



Con datos reales lo normal es que exista una mezcla de información, ruido y equivocación.

X \ Y	q	r	s
a	0.5	0.5	
b		1.0	
c		0.5	0.5

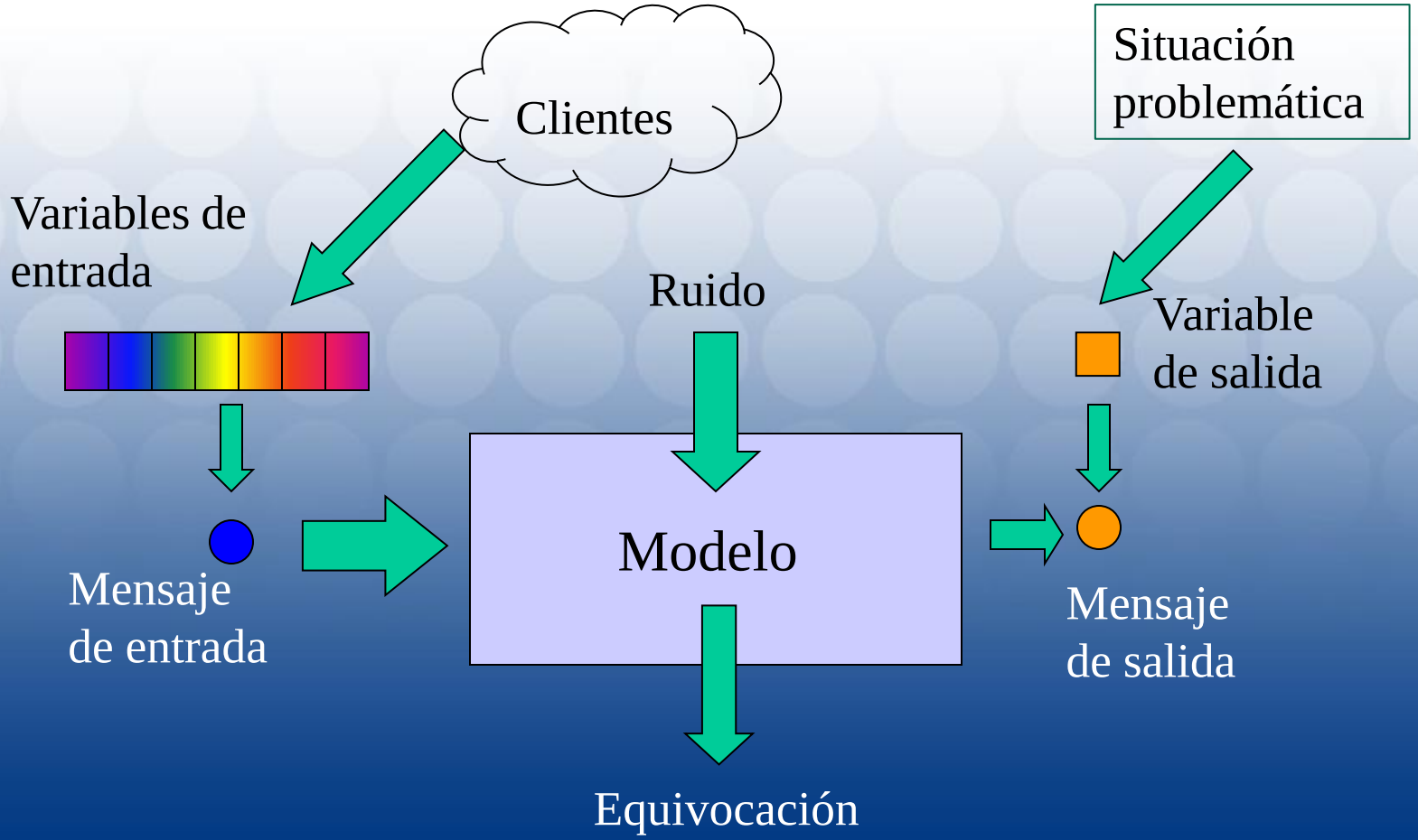
Una relación con ruido



Temas a tratar

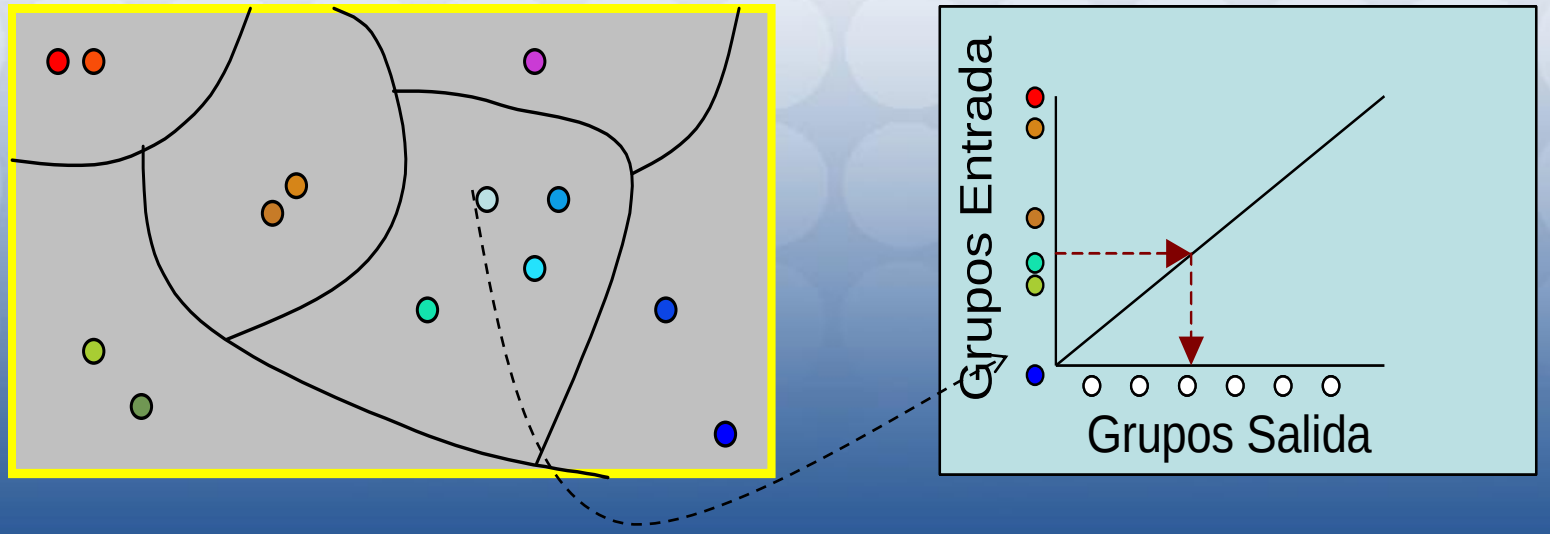
- Etapas de un proyecto de Data Mining
- Introducción a la Teoría de la Información
- La Teoría de la Información como base de DM
- Ventajas del enfoque de la TI
- Ejemplos
- Preguntas

TI como base del DM



Modelos de Predicción

Para predecir la señal de salida se usa un mapa que relaciona las señales de entrada con las de la salida



Mapeo de la Información usando una función de Transferencia lineal

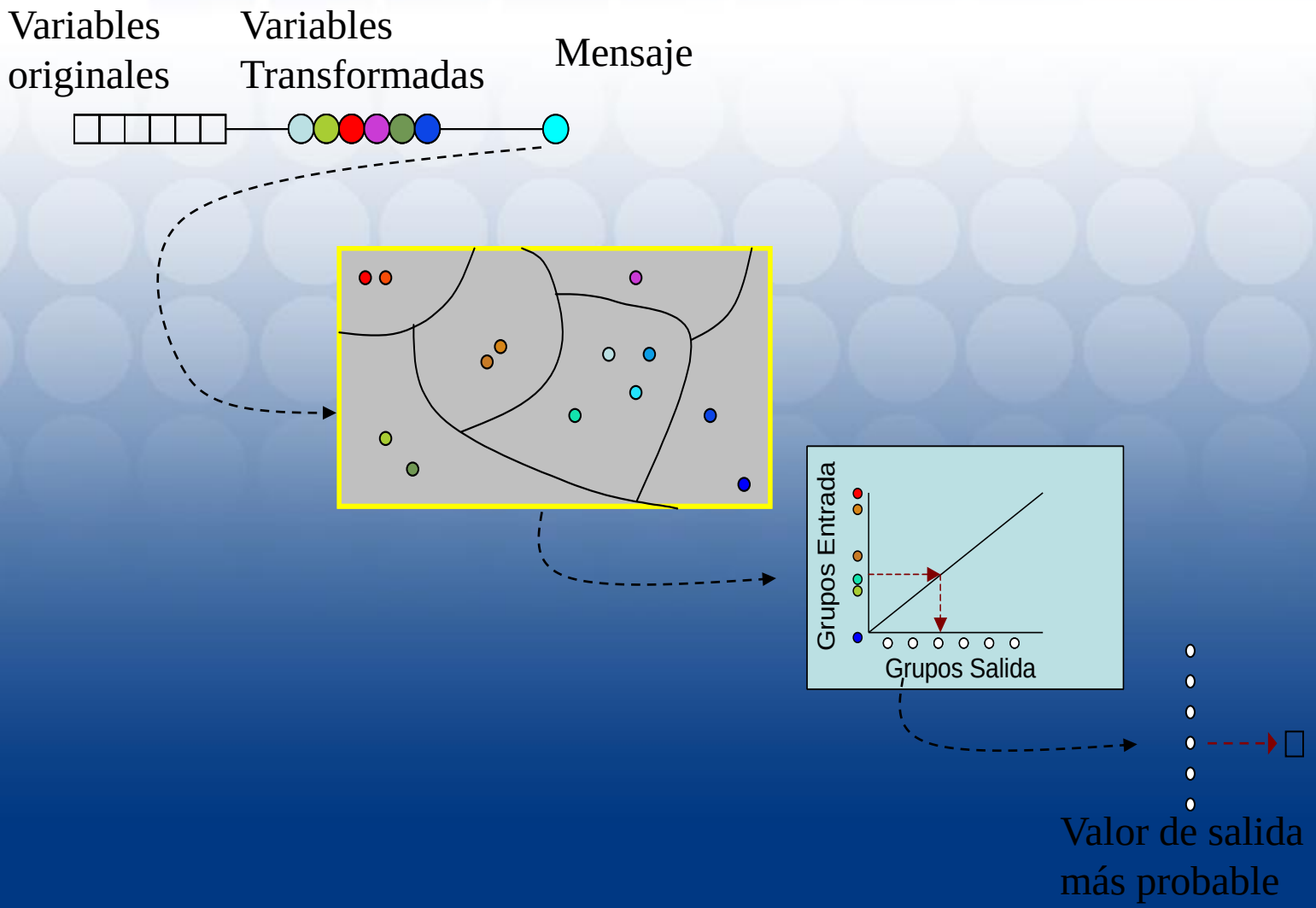
Modelos de Predicción

Para predecir la señal de salida se usa un mapa que relaciona las señales de entrada con las de la salida

Si v1 es aprox. 35 y
v2 es aprox. 189 y
v3 es Femenino y
Entonces
Resultado es 1

Mapeo de la Información usando Reglas

El proceso completo



Temas a tratar

- Etapas de un proyecto de Data Mining
- Introducción a la Teoría de la Información
- La Teoría de la Información como base de DM
- Ventajas del enfoque de la TI
- Ejemplos
- Preguntas

Ventajas

La Teoría de la Información se convierte en una herramienta muy poderosa:

1. Permite medir la información y el ruido en los datos
2. Facilita la creación de un algoritmo simple para seleccionar variables
3. Brinda una referencia objetiva para comparar los modelos
4. Posibilita la creación de algoritmos para la preparación de datos en forma automática
5. Habilita el uso de funciones de transferencia lineales, que se traduce en modelos claros y confiables

Selección de Variables

1. Seleccionar la variable que mayor información contenga acerca de la variable a predecir
2. Seleccionar la siguiente variable con mayor información *adicional* acerca de la variable a predecir
3. Continuar con el paso 2 hasta que la cantidad de información que aporte la variable no justifique la pérdida de representatividad

Temas a tratar

- Etapas de un proyecto de Data Mining
- Introducción a la Teoría de la Información
- La Teoría de la Información como base de DM
- Ventajas del enfoque de la TI
- Ejemplos
- Preguntas

Ejemplos

1. ¿Cuándo jugar Tenis? Mediremos información en los datos
2. El problema XOR
3. Precio de los diamantes. Interacción entre variables
4. Reconociendo un tipo de onda. Rechazo del ruido
5. Identificando riesgo de mora.

Preguntas

¿?

Gracias

www.dataxplore.com.ar

www.powerhousedm.blogspot.com