Minería de Datos

Clasificación

Introducción, conceptos básicos, modelos de evaluación

Por

Elizabeth León Guzmán

Introducción

Asignar objetos a una de muchas categorías predefinidas

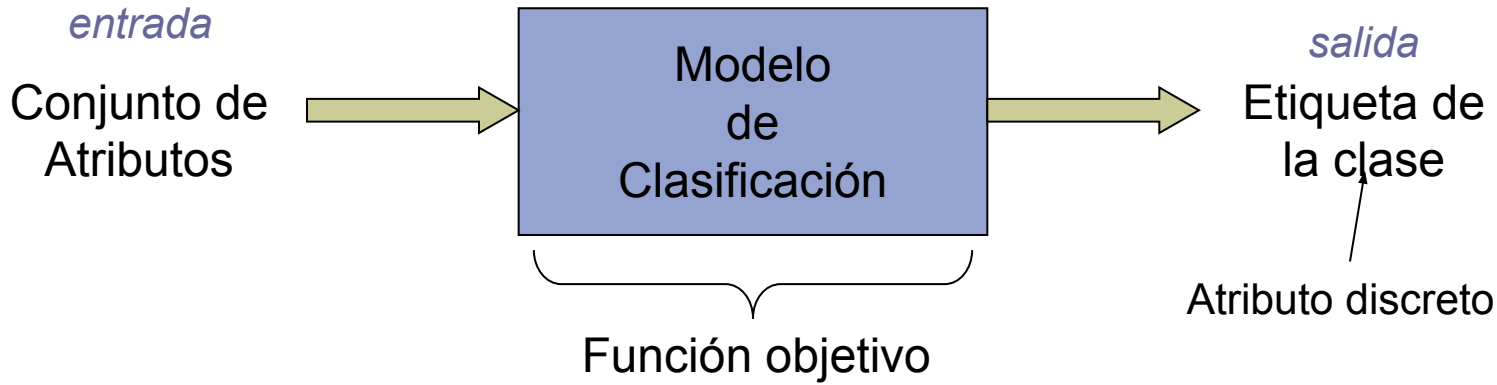
Introducción

Aplicaciones

- Diagnostico médico (detección de anomalías) basado en las características de los síntomas
- Detección de intrusos en redes de computadores basado en el comportamiento normal de la red y de los ataques conocidos
- Detección “spam email” basado en el encabezado y contenido del mensaje
- Clasificación de galaxias basado en su forma
- Detección de fraude

Introducción

Clasificación es la tarea de aprender una función objetivo F que asigne un conjunto de atributos a una clase predefinida



Introducción

Datos de entrada: Conjunto de registros/atributos

Cada **registro** (instancia, atributo o ejemplo) es caracterizado por una tupla (**x**,**y**)

x es el conjunto de atributos

y es un atributo especial (variable objetivo, etiqueta o label)

Introducción

Ejemplo

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

clase (etiqueta)
Variable objetivo

Modelo Predictivo

- Predecir la clase de un registro u objeto desconocido.

- Es una caja negra (automáticamente asigna la etiqueta de la clase)

Aprendizaje Supervisado

¡Supervisión!

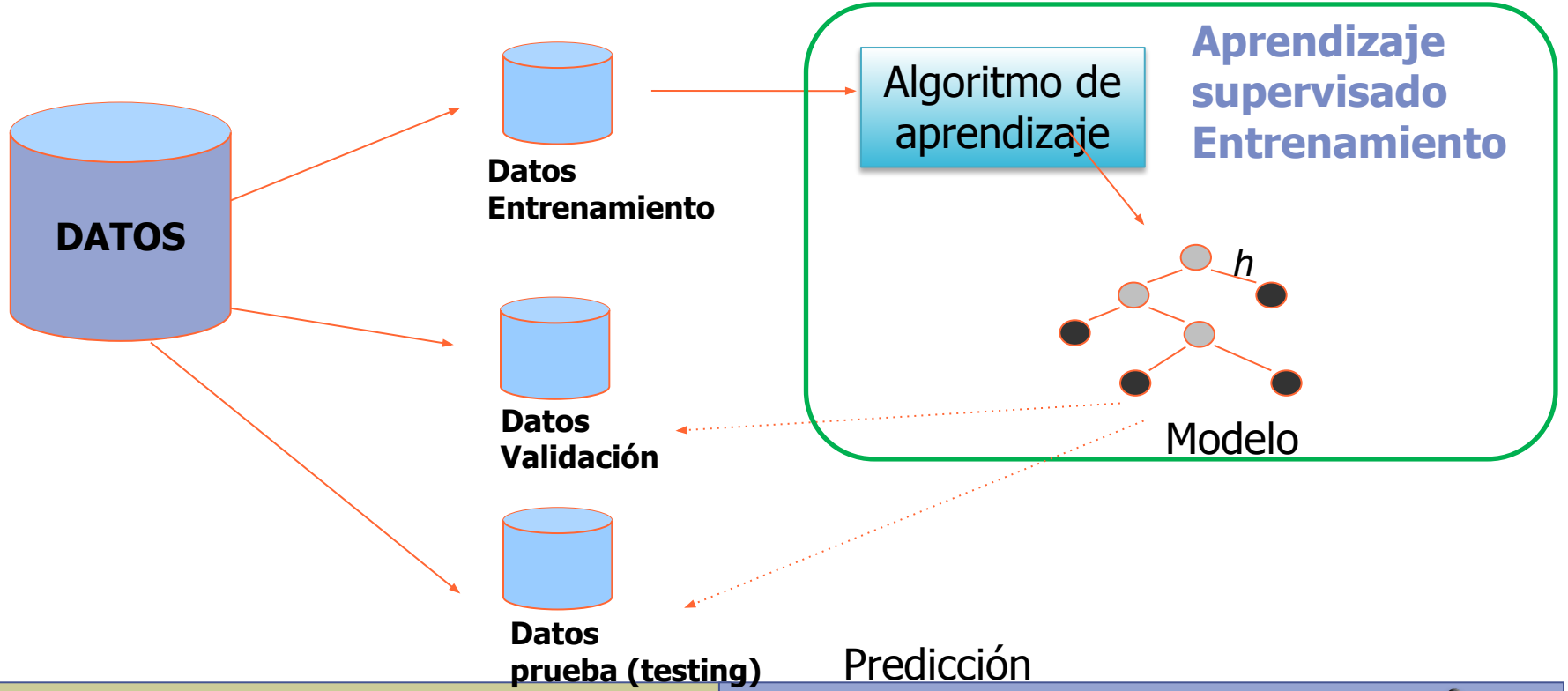
El aprendizaje es hecho usando las salidas de los datos, es decir, las etiquetas.

El conjunto de entrenamiento (observaciones, medidas, etc.) están acompañados de las etiquetas que indican la clase a la que pertenecen.

Aprendizaje Supervisado vs No Supervisado

- **Aprendizaje Supervisado (clasificación)**
  - Supervisión: El conjunto de entrenamiento están acompañados de las etiquetas que indican la clase a la que pertenecen
  - Nuevos datos son clasificados basados en el conjunto de entrenamiento
- **Aprendizaje No Supervisado (clustering)**
  - La etiqueta de la clase del conjunto de entrenamiento es desconocida
  - Dados el conjunto de medidas u observaciones sew intenta

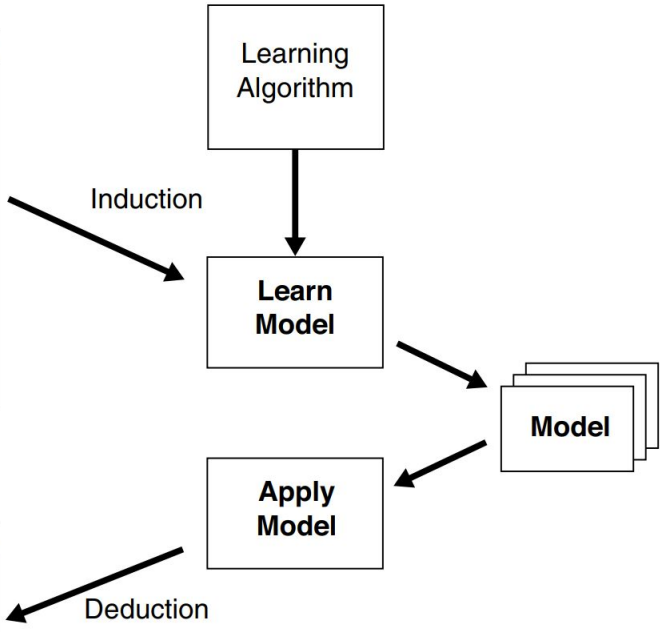
Modelo de Clasificación
Construcción



Modelo de Clasificación
Construcción

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes



Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

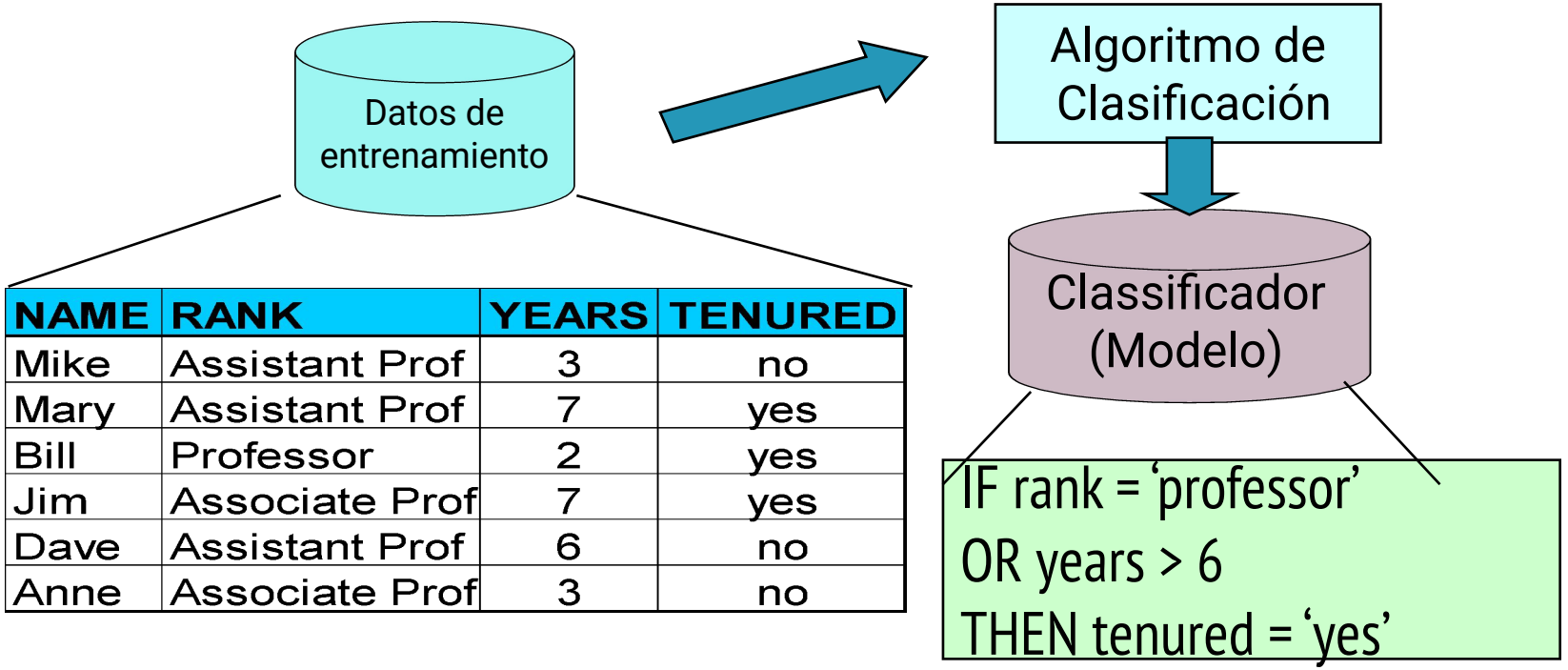
Tomado de libro Kumar

Clasificación
Proceso de 2 pasos

- **Paso 1: Construcción del Modelo (Inducción):** describir un conjunto de determinadas clases

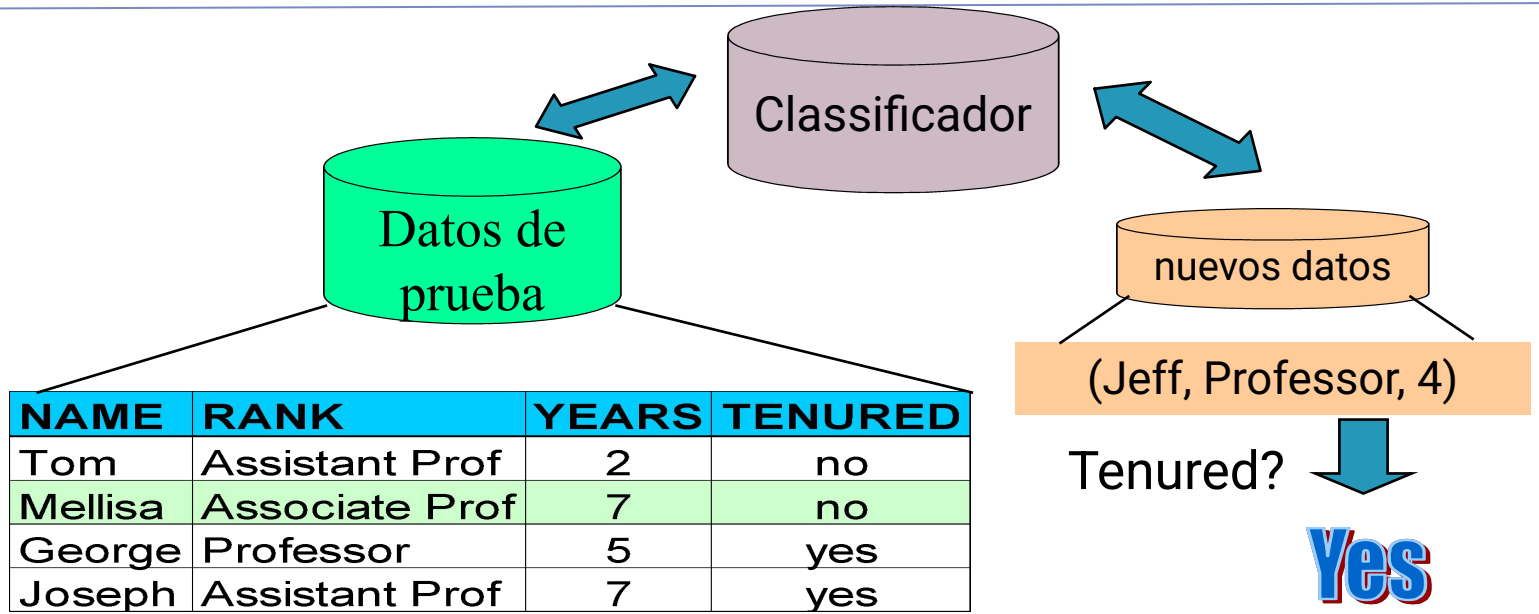
- **Paso 2: Validación y uso del Modelo (deducción):** Para clasificar futuros y desconocidos objetos

Paso 1: *Construcción del modelo* (Problema de 2-clases)



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

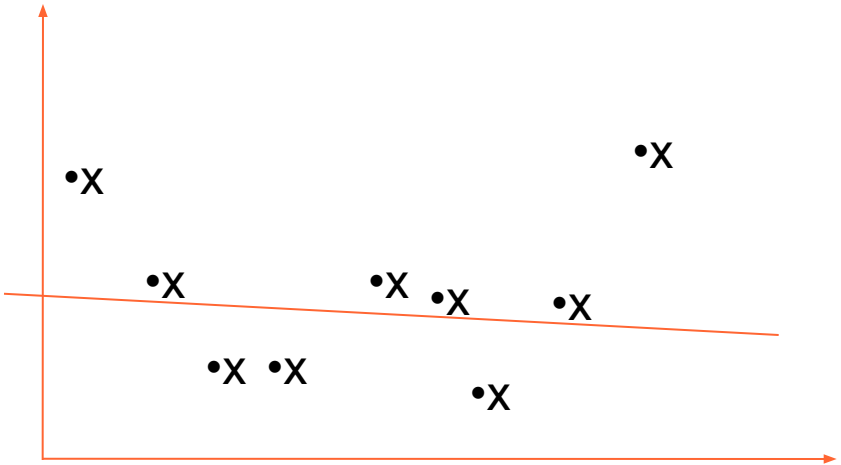
Paso 2: *Usar el modelo para predecir*

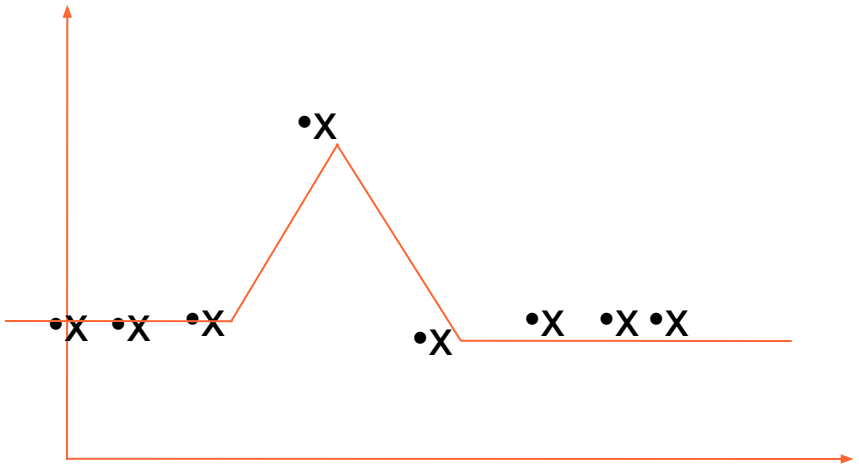


NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Mellisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

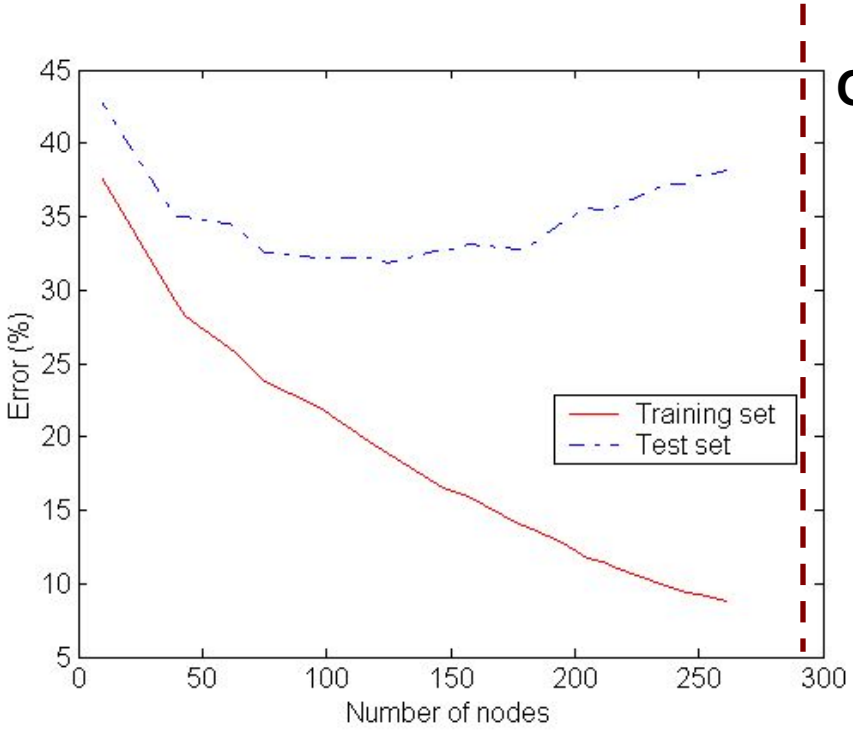
Los datos de prueba debe ser independientes del conjunto de entrenamiento, Overfitting puede ocurrir (el modelo funciona extremadamente bien con los datos de entrenamiento, pero pobremente con nuevos datos.

Subajuste y Sobreajuste



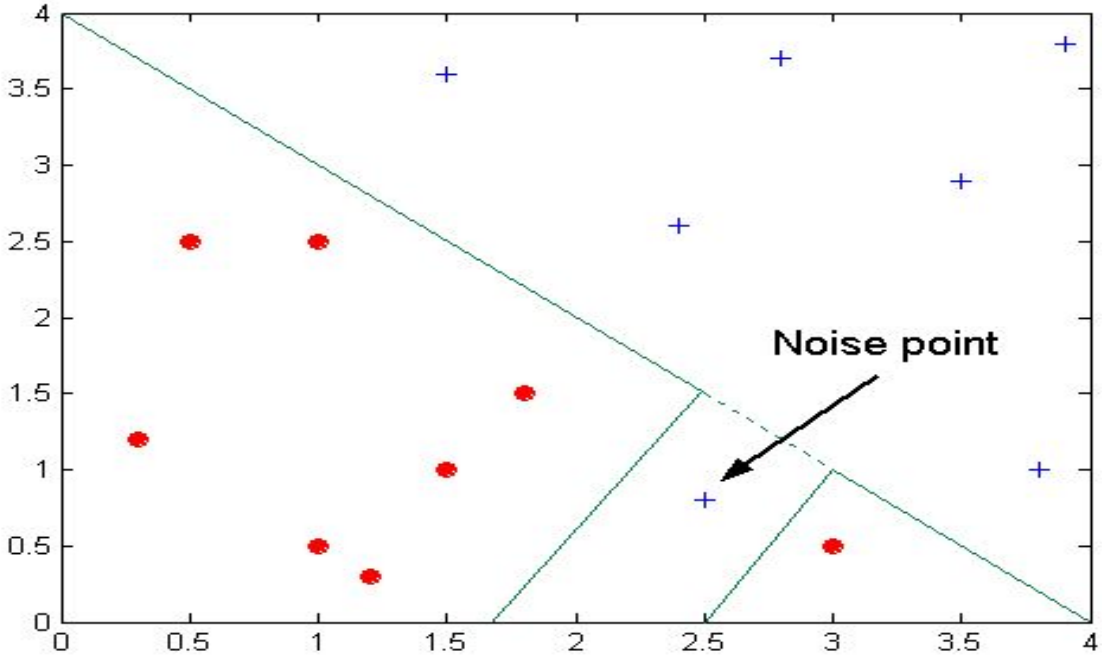


Sobre-entrenamiento o sobreajuste (overfitting)



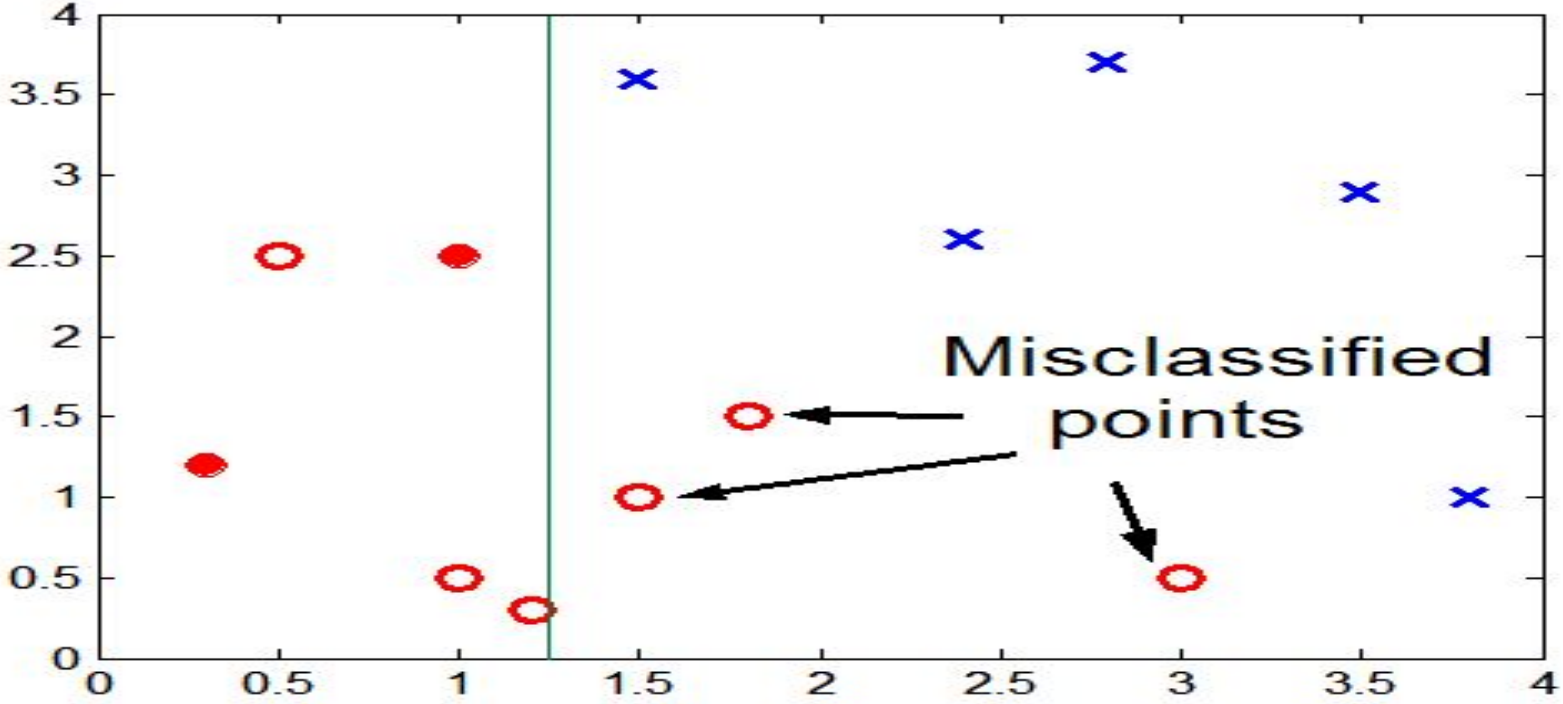
Sub entrenamiento: el modelo es muy simple, los errores de entrenamiento y prueba son grandes.

Sobre-entrenamiento o sobreajuste (overfitting)



El borde de decisión se distorsiona por un punto de ruido

Sobre-entrenamiento por falta de ejemplos



Evaluación de un clasificador

Métodos

“Holdout Method”

- Conjunto particionado en 2 conjuntos disjuntos (entrenamiento y prueba)
- Evaluación con respecto al conjunto de prueba
- El modelo puede tener dependencia de la composición de la partición
  - Pequeño tamaño del conjunto de entrenamiento varianza alta

Evaluación de un clasificador

Métodos

“Random subsampling”

- Repite el método *holdout* muchas veces
- La exactitud del modelo es dada por el promedio
- Mantiene algunos problemas, ya que no utiliza suficientes datos para entrenamiento
- No hay control sobre los ejemplos que ya han sido usados para entrenamiento

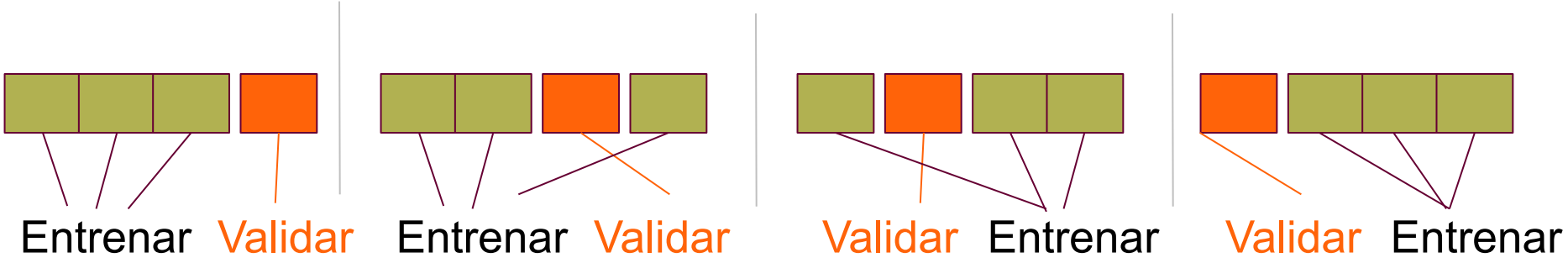
Evaluación de un clasificador

Métodos

“K Fold Cross-validation” (validación cruzada)

K=4

datos divididos en 4



Medidas de validación:Media de los cuatro modelos

Evaluación de un clasificador

Métodos

“K fold Cross-validation” (validación cruzada)

- Sampling aleatorio
- Cada ejemplo es usado el mismo número para entrenamiento y una vez para pruebas.
  - 10 fold cross-validation (la más usada)
  - Leave one out

Evaluación de un clasificador

Métodos

“**Bootstrap**”

- Los ejemplos para entrenamiento pueden estar repetidos(sampling with replacement)
- Muchas variaciones de bootstrap

Evaluación del modelo

Métricas

- Basado en el número de registros que fueron clasificados correcta e incorrectamente.
- "Confusion Matrix"

		Clase predicción	
		Clase =1	Clase =2
Clase	Clase =1	f_{11}	f_{10}
	Clase =2	f_{01}	f_{00}

Matriz de confusión para dos clases

Evaluación del modelo

Métricas

- Basado en el número de registros que fueron clasificados correcta e incorrectamente.
- "Confusion Matrix"

		Clase predicción	
		Clase =1	Clase =2
Clase	Clase =1	f_{11} **TP**	f_{10} **FN**
	Clase =2	f_{01} **FP**	f_{00} **TN**

$$Accuracy = \frac{Número \quad correctos}{Total} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$Error \quad rate = \frac{Número \quad incorrectos}{Total} = \frac{f_{10} + f_{10}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Evaluación del modelo

Métricas

$$Precision\ (p) = \frac{f_{11}}{f_{11} + f_{01}}$$

$$Recall(r) = \frac{f_{11}}{f_{11} + f_{10}}$$

$$F - measure(F) = \frac{2rp}{r + p} = \frac{2f_{11}}{2f_{11} + f_{10} + f_{01}}$$

		Clase predicción	
		Clase =1	Clase =2
Clase real	Clase =1	f_{11}	f_{10}
	Clase =2	f_{01}	f_{00}

f_{11}: TP (true positive)

f_{10}: FN (false negative)

f_{01}: FP (false positive)

f_{00}: TN (true negative)

Limitante del “Accuracy”

Se tiene un problema de 2-clases:

Número de ejemplos de la Clase 0 = 9990
Número de ejemplos de la Clase 1 = 10

Si el modelo predice todos los ejemplos de prueba como clase 0, el **accuracy** es 9990/10000 = **99.9 %**

En este caso el Ac[illegible]iente para validar ya que el modelo no detecta la clase 1

Desbalance de clases

Matrices de costo

		Clase predicción	
Clase real	C(i|j)	**Clase=Si**	**Clase=No**
	Clase=Si	C(Si|Si)	C(No|Si)
	Clase=No	C(Si|No)	C(No|No)

C(i|j): Costo de clasificar mal el ejemplo de la clase j como ejemplo de la clase i

Matrices de costo

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

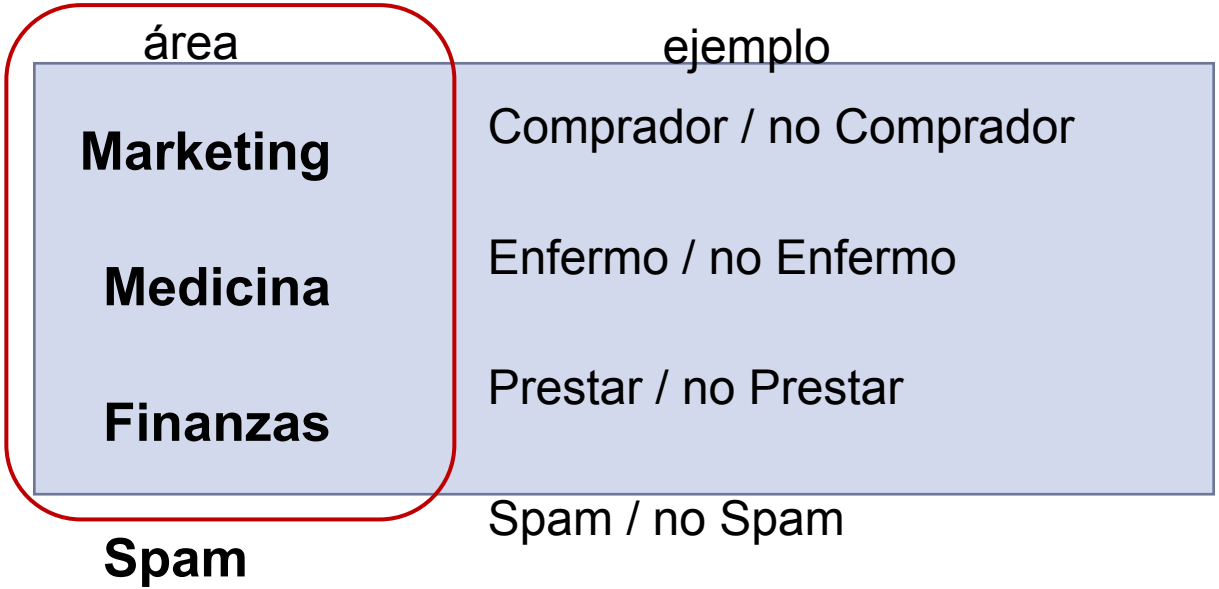
Accuracy = 80%

Cost = 3910

Accuracy = 90%

Cost = 4255

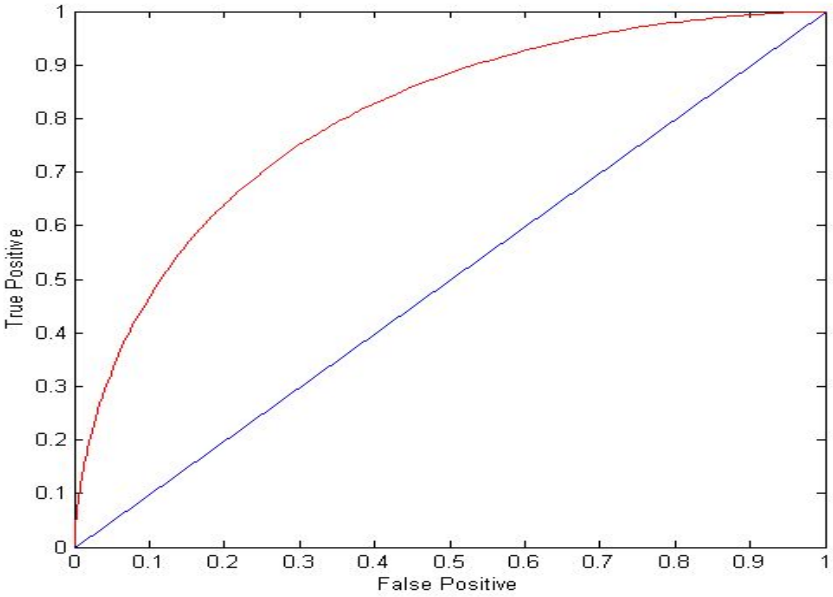
Matrices de costo



En cada ejemplo, ¿cuál métrica castigaría más, FP o FN?

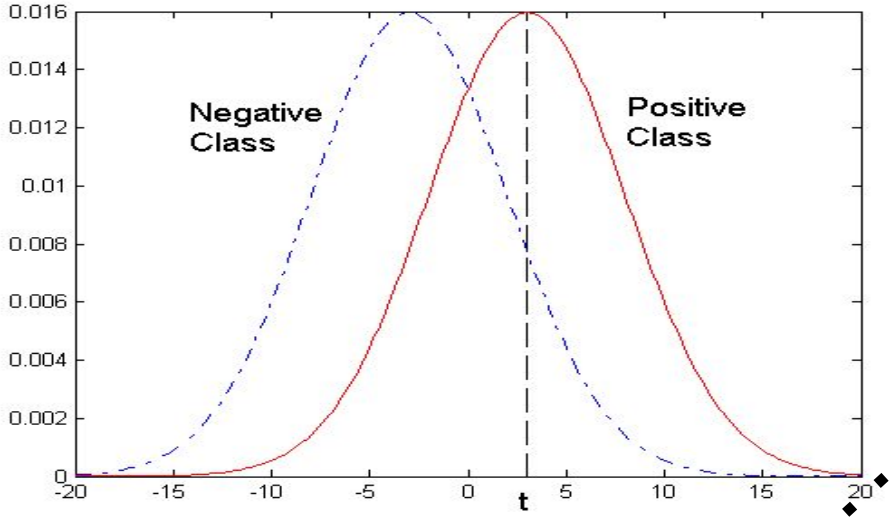
ROC (Receiver Operating Characteristic)

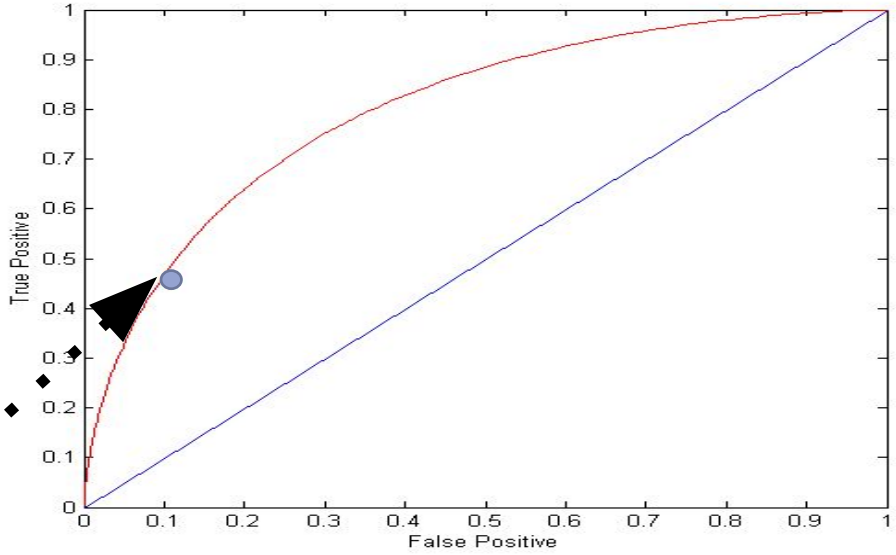
- Desarrollada en 1950s para analizar ruido en señales
- Caracteriza el trade-off entre detección y falsas alarmas
- La curva ROC curve dibuja TP (en eje y) vs FP (en eje x)
- El rendimiento de cada clasificador representado como un punto en la curva ROC cambiando el umbral del algoritmo, sampling, etc.



ROC (Receiver Operating Characteristic)

- Data set de una dimensión que contiene dos clases (positivo y negativo)
- Cualquier punto x, x > t es clasificado como positivo





t umbral TP=0.5, FN=0.5, FP=0.12, TN=0.88

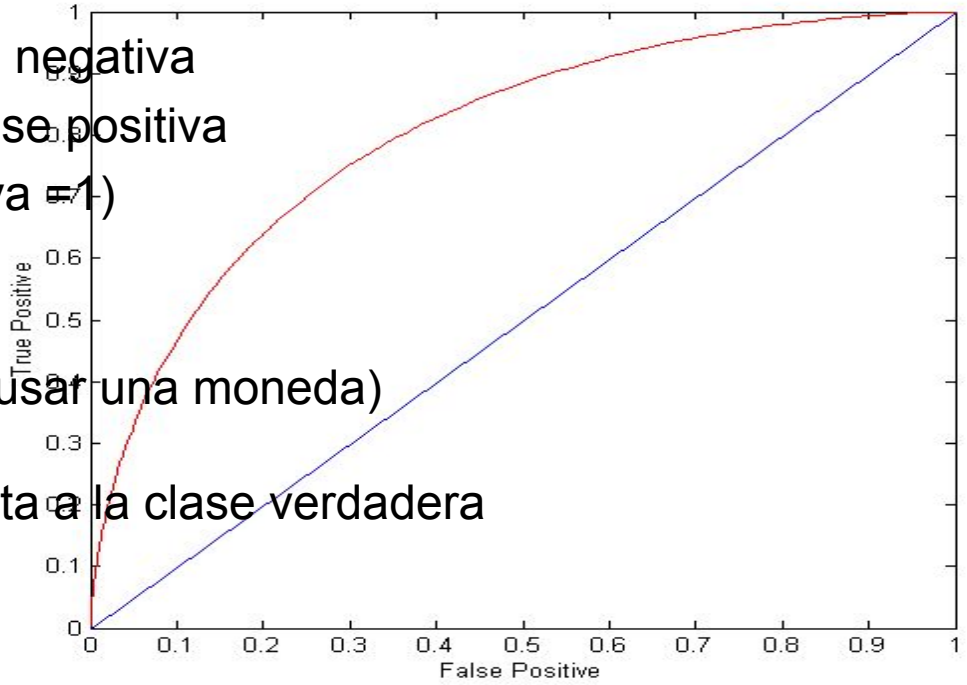
ROC (Receiver Operating Characteristic)

(TP,FP):

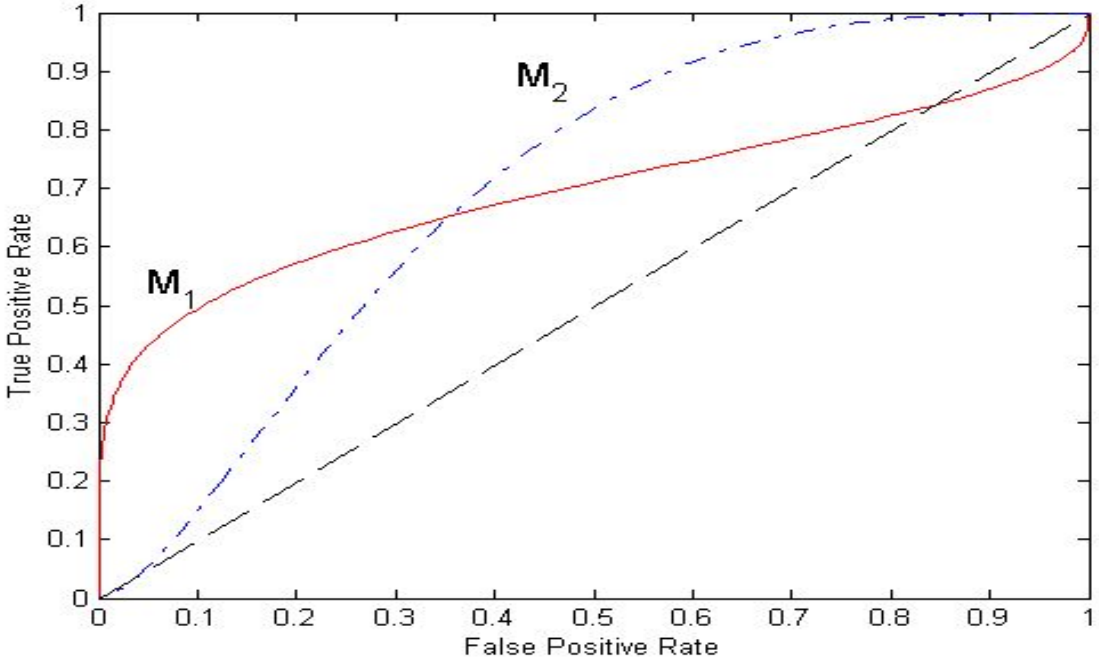
- (0,0): Todo pertenece a clase negativa
- (1,1): Todo pertenece a la clase positiva
- (1,0): ideal (área bajo la curva =1)

Diagonal line:

- Adivinar aleatoriamente (usar una moneda)
- Bajo la línea diagonal:
  - predicción es opuesta a la clase verdadera



Usando curva ROC para comparar Modelos



- Ninguno de los dos modelos es mejor
  - M_1 es mejor para valores bajos de FPR
  - M_2 es mejor para valores altos de FPR
- Área bajo la curva ROC
  - Ideal:
    - Area = 1
  - Aleatorio:
    - Area = 0.5

Construcción de la curva ROC

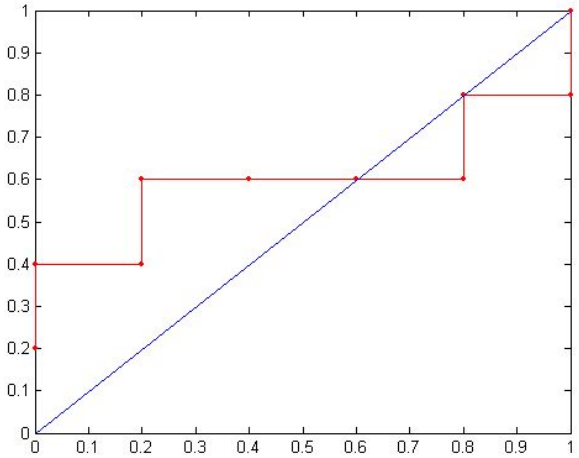
Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Se usa un clasificador que produce la probabilidad posterior para cada instancia de test P(+|A)
- Se ordenan la instancias decrecientemente de acuerdo a P(+|A)
- Se aplica el umbral a cada único valor de P(+|A)
- Se calcula TP, FP, TN, FN en cada umbral
- TP rate, TPR = TP/(TP+FN)
- FP rate, FPR = FP/(FP + TN)

Construcción de la curva ROC

	Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=		0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
	TP	5	4	4	3	3	3	3	2	2	1	0
	FP	5	5	4	4	3	2	1	1	0	0	0
	TN	0	0	1	1	2	3	4	4	5	5	5
	FN	0	1	1	2	2	2	2	3	3	4	5
→	TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
→	FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC Curve:

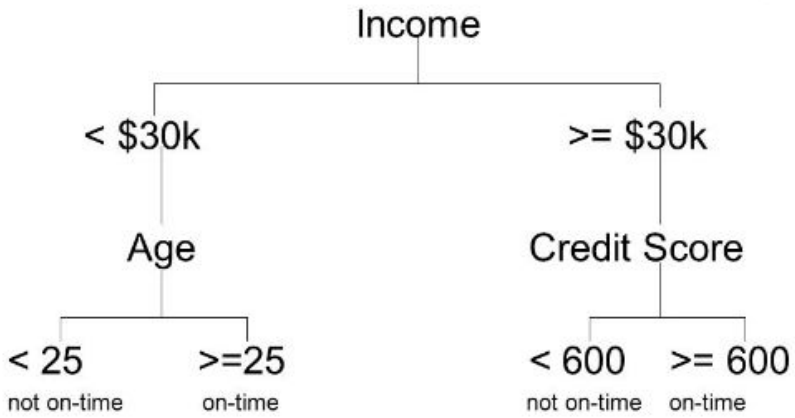


Aplicaciones

Crédito

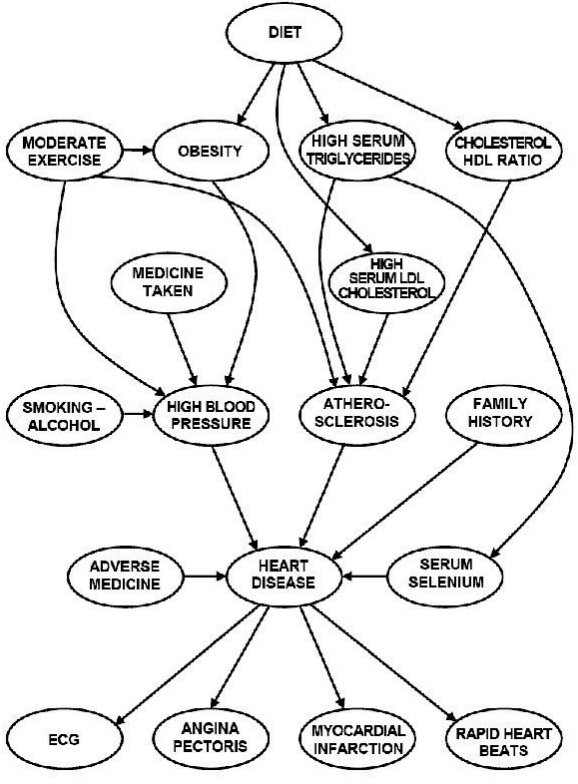
La entidad bancaria cuenta con los datos correspondientes a los créditos, concedidos con anter uración en años....) y otros datos d, etc.

Árbol de Decisión



Aplicaciones

- Diagnostico médico
  - Red Bayesiana



Aplicaciones

- Minería de Texto
  - Asignar "clases" a los documentos de acuerdo al conteni
  - Spam filtering

Aplicaciones

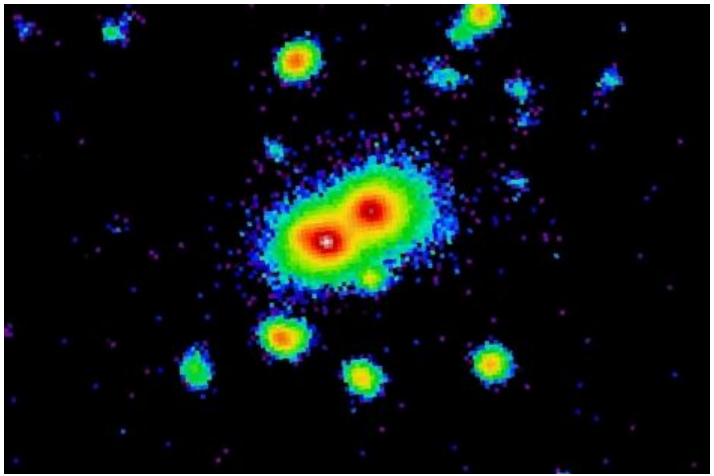
- Minería de Texto
  - Marketing o politics (opinion, sentimental analysis)

Aplicaciones

- Detección de intrusos en redes de computadores basado en el comportamiento normal de la red y de los ataques conocidos

Aplicaciones

- Clasificación de galaxias basado en su forma



Aplicaciones

Anális de crédito bancario

Un banco por Internet desea obtener reglas, para predecir qué personas de las que solicitan un crédito no lo devuelven. La entidad bancaria cuenta con los datos correspondientes a los créditos, concedidos con anterioridad a sus clientes (cuantía del crédito, duración en años....) y otros datos personales como el salario del cliente, si posee casa propia, etc.

Clase

IDC	D - Credito (Años)	C – Credito (Euros)	Salario (Euros)	Casa propia	Cuentas Morosas	…	Devuelve-credito
101	15	60.000	2.200	Sí	2	...	No
102	2	30.000	3.500	Sí	0	...	Sí
103	9	9.000	1.700	Sí	1	...	No
104	15	18.000	1.900	No	0	...	Sí
105	10	24.000	2.100	No	0	...	No
...	...	...	...	...	...	...	...

Referencias

- Tan, Steinbach, Kumar. Introduction to Data Mining