Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software

T.2 Integración y Adaptación de Modelos

Cèsar Ferri Ramírez

Objetivos

Aprender cómo evaluar modelos de aprendizaje automático

Descubrir técnicas para adaptar modelos en contexto con coste asociado

Conocer métodos de combinación de modelos

Temario

2.1. Técnicas y Medidas de Evaluación.
2.2. Análisis ROC.
2.3. Combinación de Modelos.

Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

- El **aumento del volumen y variedad de información** que se encuentra informatizada en bases de datos digitales ha crecido espectacularmente en la última década.
- Gran parte de esta **información es histórica**, es decir, representa transacciones o situaciones que se han producido.
- Aparte de su función de "memoria de la organización", la información histórica es útil **para predecir la información futura**.

Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

- La mayoría de ***decisiones*** de empresas, organizaciones e instituciones se basan también en información de experiencias pasadas extraídas de fuentes muy diversas.
- las **decisiones colectivas** suelen tener consecuencias mucho más graves, especialmente económicas, y, recientemente, se deben basar en **volúmenes de datos que desbordan la capacidad humana**.

El área de la extracción (semi-)automática de conocimiento de bases de datos ha adquirido recientemente una importancia científica y económica inusual

Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

- "Descubrimiento de Conocimiento a partir de Bases de Datos" (KDD, del inglés *Knowledge Discovery from Databases*).

  "*proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos*". Fayyad et al. 1996

- Diferencia clara con métodos estadísticos: la estadística se utiliza para validar o parametrizar un *modelo sugerido y preexistente*, no para generarlo.
- Diferencia sutil "Análisis Inteligente de Datos" (IDA, del inglés Intelligent Data Analysis) que correspondía con el uso de técnicas de inteligencia artificial en el análisis de los datos.

Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

- Además el resultado de KDD debe ser COMPRENSIBLE.
- Se excluyen, a priori, por tanto, muchos métodos de aprendizaje automático (redes neuronales, CBR, k-NN, Radial Basis Functions, Bayes Classifiers...).

- Cambia la Manera de Extraer el Conocimiento:
  - Eficiente.
  - Entornos de Descubrimiento ('Navegación').
  - Consultas Inductivas.

Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

- KDD nace como interfaz y se nutre de diferentes disciplinas:
  - estadística.
  - sistemas de información / bases de datos.
  - aprendizaje automático / IA.
  - visualización de datos.
  - computación paralela / distribuida.
  - interfaces de lenguaje natural a bases de datos.

Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

- Datos poco habituales para algoritmos clásicos:
  - número de registros (ejemplos) muy largo (10^8-10^{12} bytes).
  - datos altamente dimensionales (nº de columnas/atributos): 10^2-10^4.
- El usuario final no es un experto en ML ni en estadística.
- El usuario no se puede perder más tiempo analizando los datos:
  - industria: ventajas competitivas, decisiones más efectivas.
  - ciencia: datos nunca analizados, bancos no cruzados, etc.
  - personal: "information overload"...

Los sistemas clásicos de estadística son difíciles de usar y no escalan al número de datos típicos en bases de datos.

Nuevas Necesidades del Análisis de Grandes Volúmenes de Datos

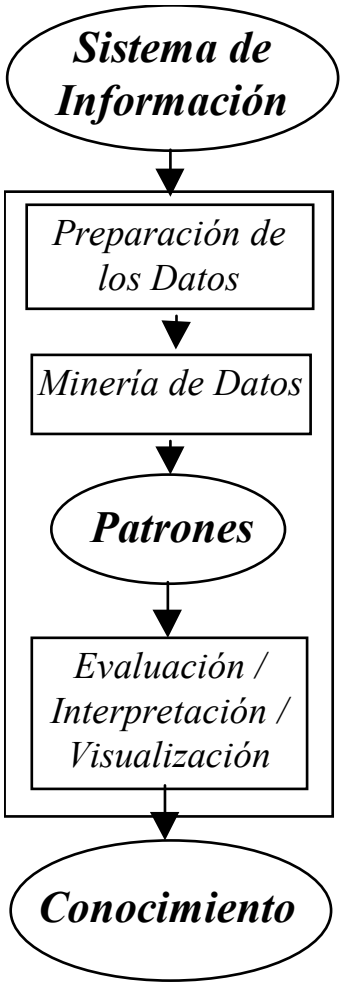
Evaluación del 'conocimiento':

- válido?
- útil?
- inteligible?
- novedoso?
- interesante?

Uso del 'conocimiento' obtenido:

- hacer predicciones sobre nuevos datos.
- explicar los datos existentes
- resumir una base de datos masiva para facilitar la toma de decisiones.
- visualizar datos altamente dimensionales, extrayendo estructura *local* simplificada.

FASES DEL KDD

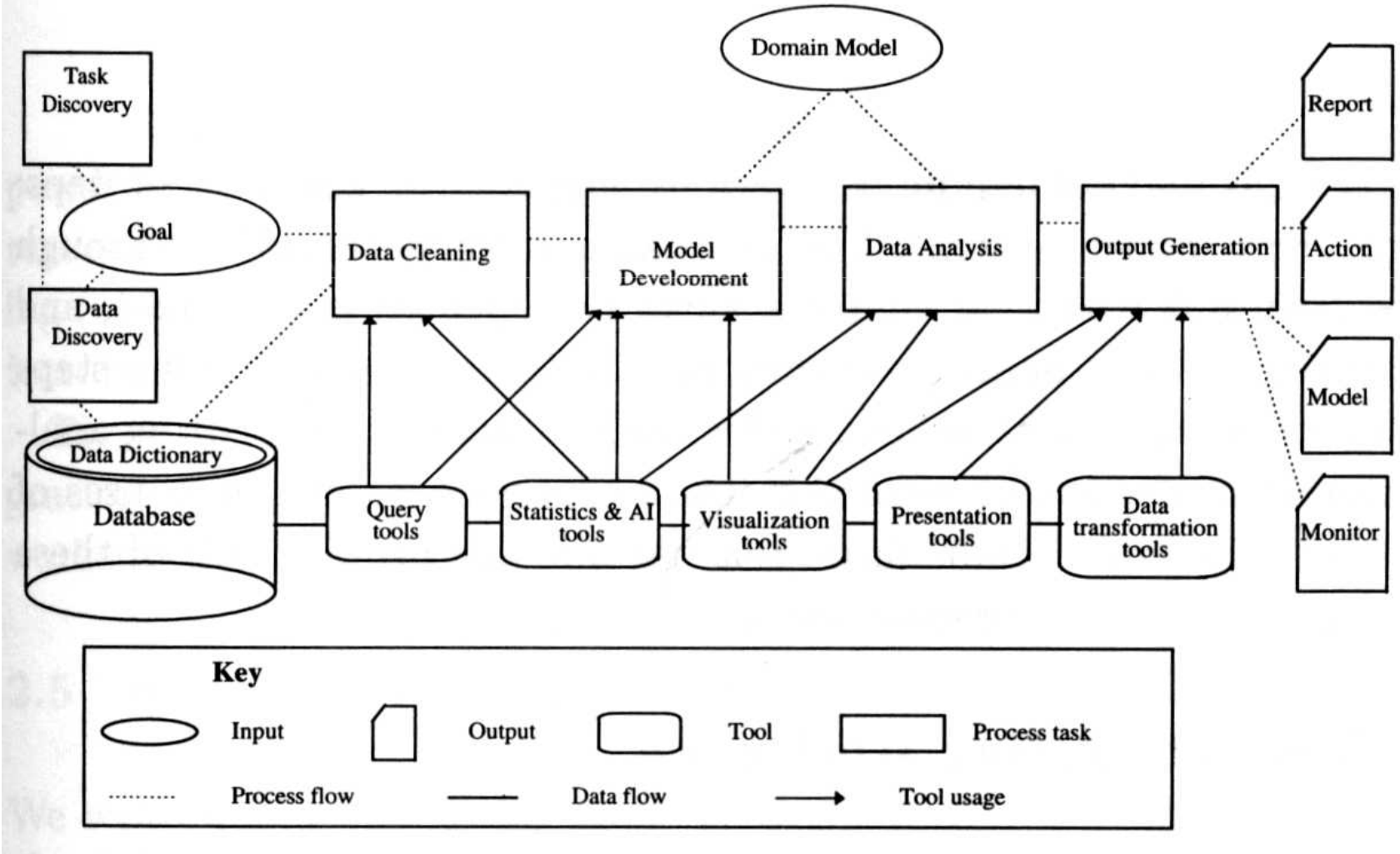


KDD

1. Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.
2. Diseñar el esquema de un almacén de datos (Data Warehouse) que consiga unificar de manera operativa toda la información recogida.
3. Implantación del almacén de datos que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.
4. Selección, limpieza y transformación de los datos que se van a analizar. La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).
5. Seleccionar y aplicar el método de minería de datos apropiado.
6. Interpretación, transformación y representación de los patrones extraídos.
7. Difusión y uso del nuevo conocimiento.

Fases y Técnicas del KDD

Las distintas técnicas de distintas disciplinas se utilizan en distintas fases:



Evaluación

- ¿Cómo se validan/descartan los modelos?
- ¿Cómo se elige entre varios modelos?
- ¿Cuánto afecta el número de ejemplos?
- ¿Cómo afecta la presencia de ruido?
- ¿Cómo se comportará mi modelo en el futuro?

Evaluación

- La evaluación de modelos depende del tipo de tarea:
  - Predictivas: evaluación más sencilla y general
  - Descriptivas: evaluación más dependiente de la técnica utilizada.

Evaluación

- **Evaluación de modelos predictivos:**

¿Qué medida usamos para comparar el valor correcto “f” del valor estimado “h” ?

  - Clasificación:
    - %Acierto o, inversamente, %Error
    - Alcance y precisión (recall & precision).
    - Área bajo la curva ROC.
    - …
  - Regresión:
    - Error cuadrático medio.
    - Error absoluto medio.
    - …

Evaluación

- **Evaluación de modelos predictivos:**
  - Dado un conjunto S de *n* datos, el error se define:
    - Clasificación: Error

$$error_S(h) = \frac{1}{n}\sum_{x \in S} \partial(f(x), h(x))$$

donde δ(a,b)=0 si a=b y 1 en caso contrario.

Clase predicha (h(x))	Clase real (f(x))	**Error**
Compra	Compra1	**No**
No Compra	Compra	**Sí**
Compra	No Compra	**Sí**
Compra	Compra	**No**
No Compra	No Compra	**No**
No Compra	Compra	**Sí**
No Compra	No Compra	**No**
Compra	Compra	**No**
Compra	Compra	**No**
No Compra	No Compra	**No**

Fallos / Total

Error = 3/10 = 0.3

Evaluación

- **Evaluación de modelos predictivos:**
  - Dado un conjunto S de *n* datos, el error se define:
    - Regresión: Error Cuadrático Medio

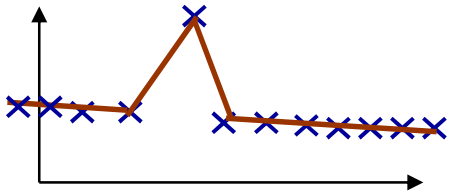
$$error_S(h) = \frac{1}{n}\sum_{x \in S}(f(x) - h(x))^2$$

Valor predicho (h(x))	Valor real (f(x))	**Error**	**Error²**
100 mill. €	102 mill. €	**2**	**4**
102 mill. €	110 mill. €	**8**	**64**
105 mill. €	95 mill. €	**10**	**100**
95 mill. €	75 mill. €	**20**	**400**
101 mill. €	103 mill. €	**2**	**4**
105 mill. €	110 mill. €	**5**	**25**
105 mill. €	98 mill. €	**7**	**49**
40 mill. €	32 mill. €	**8**	**64**
220 mill. €	215 mill. €	**5**	**25**
100 mill. €	103 mill. €	**3**	**9**

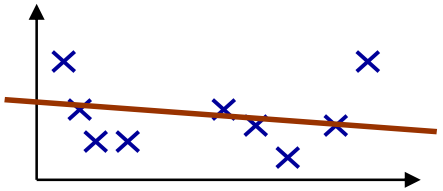
Error = 744/10 = 74,4

Evaluación

- **Evaluación de modelos predictivos.**
  - ¿Qué muestra S usamos para evaluar las medidas anteriores?
    - Si usamos todos los datos para entrenar los modelos y esos mismos datos para evaluar, tendremos:
      - sobreajuste (over-fitting).

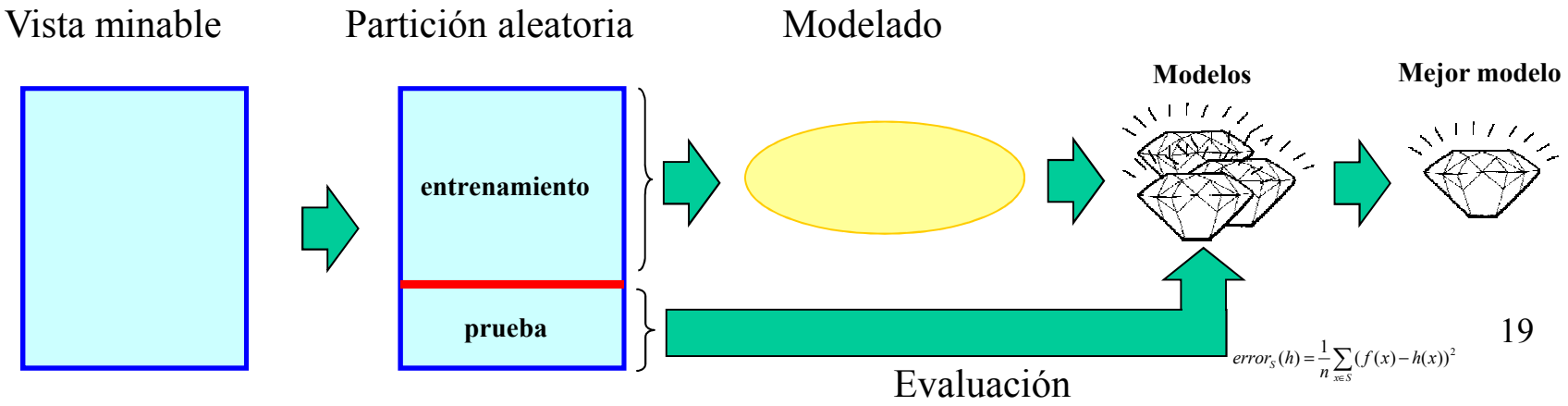


    - Si intentamos evitar el sobreajuste generalizando los modelos sin una referencia externa, podemos tener.
      - subajuste (under-fitting)



Evaluación

- **Evaluación de modelos predictivos.**
  - PARTICIÓN DE LOS DATOS:
    - Separación de los datos en:
      - Conjunto de entrenamiento (train).
        - Los modelos se entrenan con estos datos
      - Conjunto de prueba (test).
        - Los modelos se evalúan con estos datos.



Evaluación

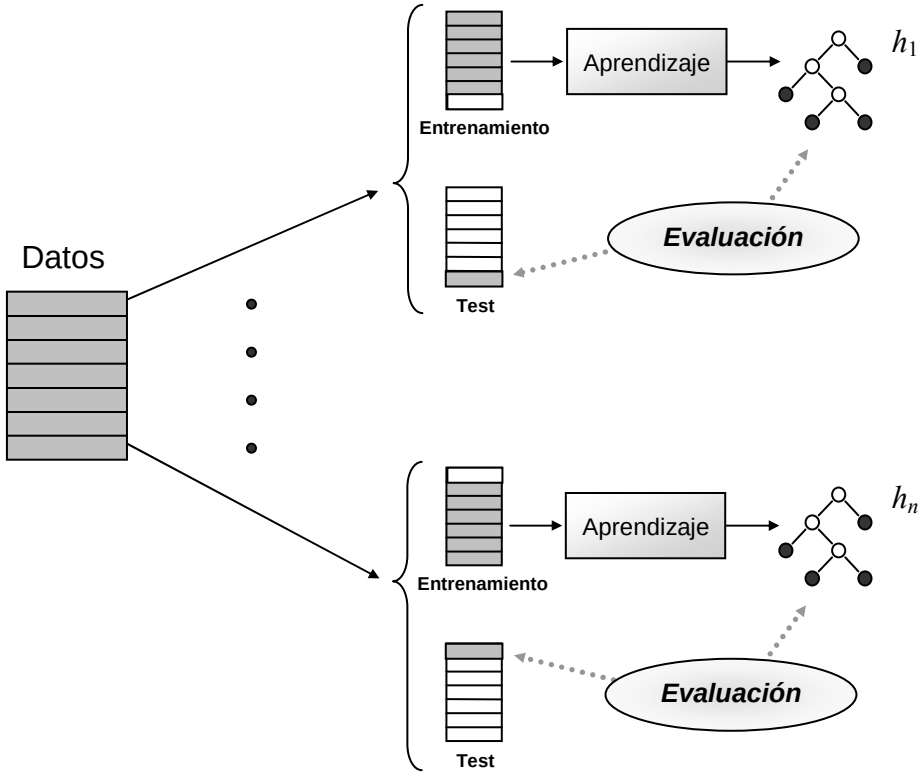
- **Evaluación de modelos predictivos.**
  - Particiones más elaboradas:
    - Validación cruzada: Se parten los datos aleatoriamente en n pliegues de igual tamaño.
    - Bootstrap: se realizan n muestras con repetición de los datos iniciales.


Especialmente útiles si hay pocos datos.


Proceso mucho más lento

Evaluación

- **Evaluación de modelos predictivos.**
  - Validación cruzada (detalle):



- Se realizan n particiones, incluyendo n-1 pliegues para entrenamiento y 1 para evaluación.
- El error medio se calcula promediando las 10 veces.
- Se reentrena un último modelo con todos los datos.

Evaluación

- **Evaluación de Modelos Descriptivos:**
  - Reglas de asociación:
    - Evaluación sencilla:
      - *soporte*
      - *confianza*
    - Se ordenan usando funciones que combinan ambos indicadores.
    - No se suele utilizar conjunto de prueba.
    - Las medidas se estiman sobre el conjunto total.

Evaluación

- **Evaluación de Modelos Descriptivos:**
  - Agrupamiento: mucho más compleja

    Concepto de error más difícil de definir

    - En los métodos basados en distancia se puede mirar:
      - GRUPOS:
        - distancia entre bordes de los clusters
        - distancia entre centros (de haberlos)
        - radio y densidad (desv. típica de la dist.) de los clusters.
      - Para cada ejemplo a agrupar se comprueba su distancia con el centro o con el borde de cada cluster.
    - Se pueden hacer diferentes agrupamientos con distintas técnicas y comparar los grupos formados (matriz de confusión)

Evaluación

- **Evaluación con sesgos o desequilibrios.**
  - Desequilibrios**:**
    - En clasificación puede haber muchos ejemplos de una clase y muy pocos del resto.
      - Problema: la clase escasa se puede tomar como ruido y ser ignorada por la teoría.
      - Ejemplo: si un problema binario (*sí* / *no*) sólo hay un 1% de ejemplos de la clase *no*, el modelo "todo es de la clase sí" tendría un 99% de acierto.

Este modelo es inútil

    - Soluciones:
      - Utilizar sobremuestro...
      - Macromedia,
      - Análisis ROC

Evaluación

- **Evaluación con sesgos o desequilibrios.**
  - Desequilibrios: Solución → SOBREMUESTREO**:**
    - Intenta equilibrar el porcentaje de clases.
      - Ejemplo: en el caso anterior "repite" los ejemplos de la clase *no* 99 veces. Ahora, la proporción pasa a ser del 50% para cada clase.
  - ¿Cuándo se debe usar sobremuestreo?
    - Cuando una clase es muy extraña: p.ej. predecir fallos de máquinas, anomalías, excepciones, etc.
    - Cuando todas las clases (especialmente las escasas) deben ser validadas. P.ej. si la clase escasa es la de los clientes que compran el producto.
  - Inconvenientes: cuidado a la hora de interpretar y evaluar los modelos (la proporción original ha sido modificada).

Evaluación

- **Evaluación con sesgos o desequilibrios.**
  - Desequilibrios: Solución → MACROMEDIA**:**
    - Otra solución es evaluar usando la macromedia, en vez del porcentaje de aciertos.

$$macromedia(h) = \frac{\frac{aciertos_{\text{clase 1}}}{total_{\text{clase 1}}} + \frac{aciertos_{\text{clase 2}}}{total_{\text{clase 2}}} + \ldots + \frac{aciertos_{\text{clase m}}}{total_{\text{clase m}}}}{m}$$

      - Ejemplo anterior:
        - Acierto global: 99%
        - Macromedia:

          Acierto para la clase sí: 100%

          Acierto para la clase no: 0%

          MACROMEDIA: 50%

Se ve claramente que el modelo es inútil

Estimadores de Probabilidad

• La mayoría de los modelos, dado un nuevo caso, estiman la probabilidad de pertenencia a cada clase, asignándole la clase con mayor probabilidad (clasificadores suaves).

• Sin embargo, algunos ámbitos requieren que esta asignación venga acompañada con alguna información sobre la fiabilidad de la clasificación.

•Los clasificadores suaves son también útiles en las aplicaciones donde nos interesa ordenar ejemplos: Mailings, predicciones de apuestas…

Estimadores de Probabilidad

•Un árbol de decisión, lo podemos convertir en un clasificador suave o PET (*Probabilistic Estimator Tree*) si utilizamos la distribución de los ejemplos en la hojas

•Si una hoja tiene las siguientes frecuencias absolutas n_1, n_2, ..., n_c (obtenidas a partir del conjunto de entrenamiento) las probabilidades estimadas pueden calcularse como: $p_i=n_i/\sum n$

•Podemos mejorar las estimaciones aplicando correcciones a la estimación de probabilidad: Laplace, M-estimate..

$$p_i = \frac{n_i + 1}{\left(\sum_{i \in C} n_i\right) + c}$$

Estimadores de Probabilidad

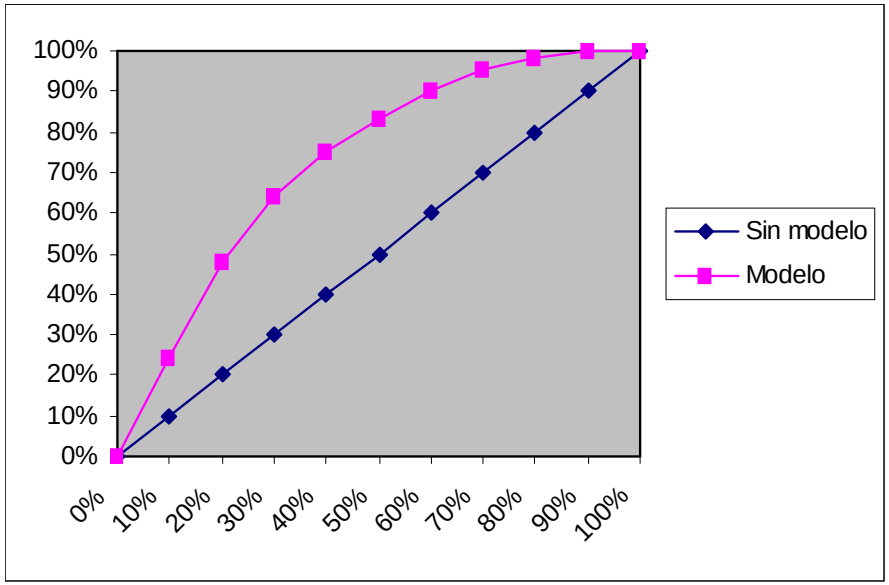
- Mailings:
  - Existen técnicas específicas para evaluar la conveniencia de campañas de 'mailings' (propaganda por correo selectiva):
    - EJEMPLO: Supongamos que una empresa de venta de productos informáticos por catálogo posee una base de datos de clientes. Esta empresa desea promocionar la venta de un nuevo producto: un mando de piloto para ser utilizado en programas de simulación de vuelo.
  - Podríamos enviar propaganda a todos sus clientes:
    - Solución poco rentable
  - Podemos utilizar técnicas de aprendizaje automático para poder predecir la respuesta de un determinado cliente al envío de la propaganda y utilizar esta información para optimizar el diseño de la campaña.

Estimadores de Probabilidad

- Mailings:
  1. Selección de una muestra aleatoria y suficientemente numerosa de clientes
  2. Se realiza el envío de la propaganda a los clientes seleccionados
  3. Una vez pasado un tiempo prudencial etiquetamos a los clientes de la muestra: 1 ha comprado el producto, 0 no ha comprado el producto
  4. Con la muestra etiqueta aprendemos un clasificador probabilístico
     - Asigna a cada ejemplo (cliente) no la clase predicha, sino una estimación de la probabilidad de respuesta de ese cliente

Estimadores de Probabilidad

- Mailings:
  - Con el clasificador probabilístico podemos ordenar a los clientes según su interés y dibujar un gráfico de respuesta acumulada

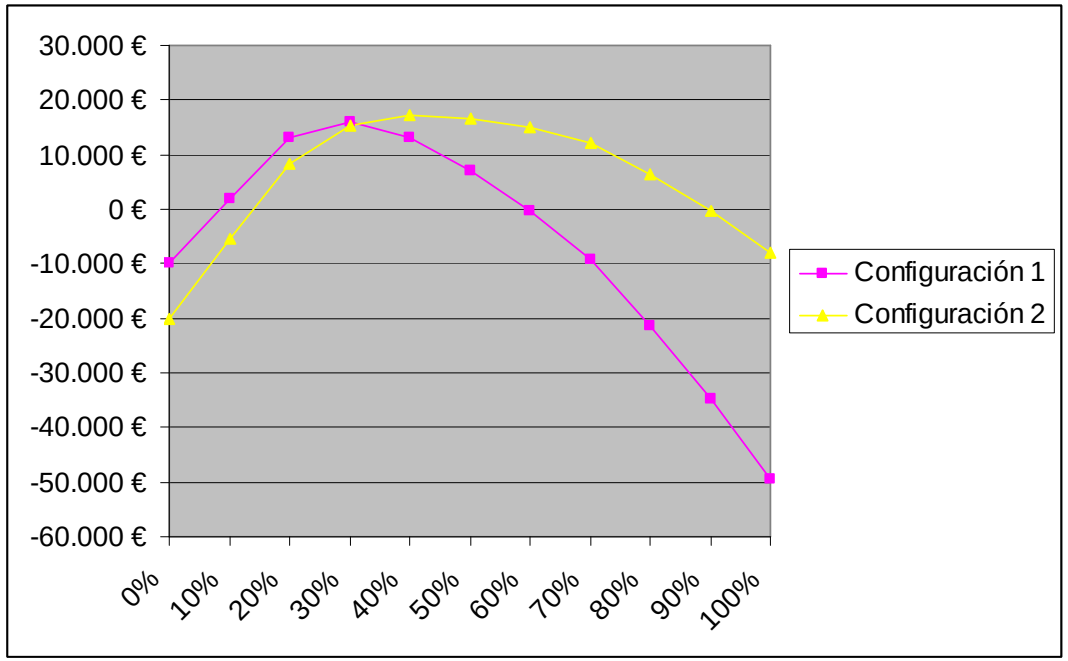


  - Nos indican qué porcentaje de las posibles respuestas vamos a obtener dependiendo del porcentaje de envíos que realicemos sobre la población total

Estimadores de Probabilidad

Además si estimamos la matriz de coste, podemos conocer la configuración optima mediante los gráficos de beneficio

- Configuración 1: Coste inicial de la campaña 10.000€, coste de envío de cada folleto 1,5€. Por cada producto vendido ganamos 3€
- Configuración 2: Coste inicial de la campaña 20.000€, coste de envío de cada folleto 0,8€. Por cada producto vendido ganamos 2,5€



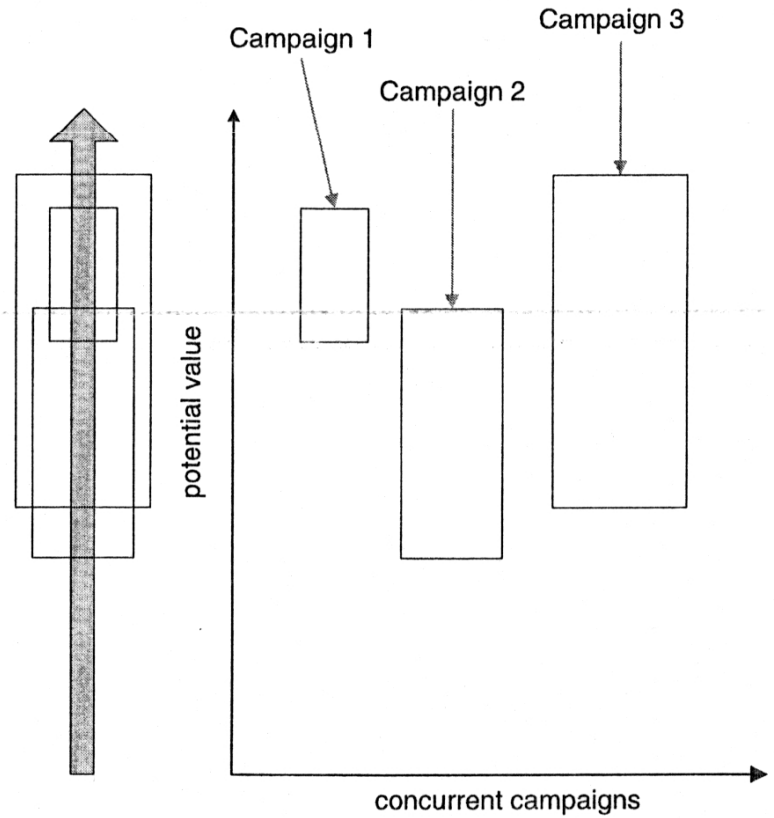
Estimadores de Probabilidad

Secuenciación de Mailings:

No sobrecargar los clientes con demasiados mensajes de márketing...

O bien acabarán ignorándolos o bien se cambiarán de compañía.

El mismo pequeño grupo de gente se elige una y otra vez y otros no se eligen nunca.

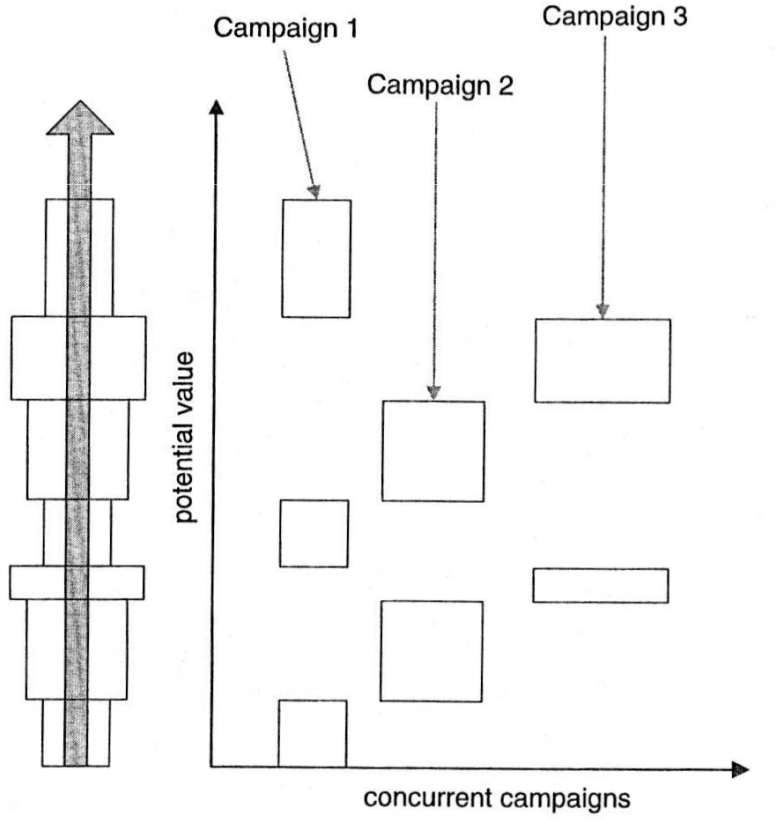


Estimadores de Probabilidad

Secuenciación de Mailings:

- Hay que intentar abarcar mejor los clientes:

Ahora todos los clientes participan en una campaña.



Aprendizaje Sensible al Coste

• **Contexto**: Una manera sencilla de definir un contexto es mediante dos aspectos:

- La distribución del valor de salida:
  - o Clasificación: distribución de las clases.
  - o Regresión: distribución de la salida.

- El coste de cada error:
  - o Clasificación: matriz de costes.
  - o Regresión: función de coste.

Aprendizaje Sensible al Coste

•Clasificación: matriz de costes.

•Ejemplo: Dejar cerrada una válvula en una central nuclear cuando es necesario abrirla, puede provocar una explosión, mientras que abrir una válvula cuando puede mantenerse cerrada, puede provocar una parada.

•Matriz de costes:

Predicho \ Real	abrir	cerrar
Abrir	0	100€
cerrar	2000€	0

•Lo importante no es obtener un "clasificador" que yerre lo menos posible sino que tenga un coste menor.

•A partir de la matriz se calcula el coste de un clasificador.

•Los clasificadores se evalúan con dichos costes.

•Se selecciona el clasificador de menos coste.

Aprendizaje Sensible al Coste

- Regresión: función de costes.
  - Ejemplo: un modelo de predicción de stocks debe penalizar más un error por exceso (al predecir mucho stock nos hemos quedado sin algún producto) que por defecto (el almacén estará más lleno pero se servirá el producto).
  - El modelo que esté "centrado" no será un buen modelo.
- Función de coste:

$$Coste = 1 - e^{\alpha \cdot (\hat{y} - y)}$$

  - Con un $\alpha = 0{,}01$:
    - Si el error es -200 el Coste= 0,86
    - Si el error es +200 el Coste= 6,3
  - De modo similar, se elige el modelo que minimice la función de coste.

Aprendizaje Sensible al Coste

Real

Pred

c_1	abrir	cerrar
ABRIR	300	500
CERRAR	200	99000

Real

c_2	abrir	cerrar
ABRIR	0	0
CERRAR	500	99500

Real

c_3	abrir	cerrar
ABRIR	400	5400
CERRAR	100	94100

Real

Predicho

	abrir	cerrar
ABRIR	0	100€
CERRAR	2000€	0

Matriz de coste

Matrices resultado

c_1	abrir	cerrar
ABRIR	0€	50.000€
CERRAR	400.000€	0€

COSTE TOTAL: 450.000€

c_2	abrir	cerrar
ABRIR	0€	0€
CERRAR	1.000.000€	0€

COSTE TOTAL: 1.000.000€

c_3	abrir	cerrar
ABRIR	0€	540.000€
CERRAR	200.000€	0€

COSTE TOTAL: 740.000€

Aprendizaje Sensible al Coste

- ¿De qué depende el coste final?
  - Para dos clases. Depende de un **contexto**:
    - El **coste** de los falsos positivos y falsos negativos: FPcost y FNcost
    - **Distribución de clases**: % de ejemplos de la clase negativa respecto de los de la clase positiva. (*Neg / Pos*).
    - Se calcula: (para el ejemplo anterior)

$$\frac{FPcost}{FNcost} = \frac{100}{2000} = \frac{1}{20} \qquad \frac{Neg}{Pos} = \frac{99500}{500} = 199 \qquad slope = \frac{1}{20} \cdot 199 = 9{,}95$$

- Para dos clases, el valor "slope" es suficiente para determinar qué clasificador será mejor.

Clasifi. 1: FNR= 40%, FPR= 0,5%
Coste Unitario =
1 x 0,40 + 9,95 x 0,005 = 0,45

Clasifi. 2: FNR= 100%, FPR= 0%
Coste Unitario =
1 x 1 + 9,95 x 0 = 1

Clasifi. 3: FNR= 20%, FPR= 5,4%
Coste Unitario =
1 x 0,20 + 9,95 x 0,054 = 0,74

Aprendizaje Sensible al Coste

- Adaptación de métodos para contextos con coste
  - ❑ Muchos métodos devuelven la probabilidad de pertenencia a la clase para cada ejemplo.
  - ❑ En estos casos en vez de asignar la clase con mayor probabilidad, se asigna la clase que minimice el coste.
  - ❑Ejemplo: un árbol de decisión retorna {0.4,0.6} a una instancia *t* con la siguiente matriz de coste:

	Real	
Predicho	c_1	+	-
	+	-20	200
	-	500	-10

  - ❑Teóricamente deberíamos asignar la clase - a *t*, sin embargo, dada la matriz de costes, es más sensato asignar +, ya que

Coste(+)=0.6*200+0.4*(-20)=112

Coste(-)=0.4*500+0.6*(-10)=194

Aprendizaje Sensible al Coste

- Adaptación de métodos para contextos con coste

  - Otra opción es modificar la distribución de los ejemplos de acuerdo a la matriz de costes (*Stratification*):
    - *Undersampling*: Eliminar instancias de las clases a reducir
    - *Oversampling*: Duplicar instancias de las clases a significar

  - Una solución más elegante es modificar los pesos los ejemplos de cada clase de acuerdo a la matriz de coste, siempre que los métodos lo permitan

Análisis ROC

- Problema

  - En muchas aplicaciones, hasta el momento de aplicación, no se conoce la distribución de clases y/o es difícil estimar la matriz de costes. P.ej. un clasificador de spam.

  - Pero los modelos se aprenden antes generalmente.

  - SOLUCIÓN:

Análisis ROC

(*Receiver Operating Characteristic*)

Análisis ROC

- El espacio ROC
  - Se normaliza la matriz de confusión por columnas: TPR, FNR TNR, FPR.

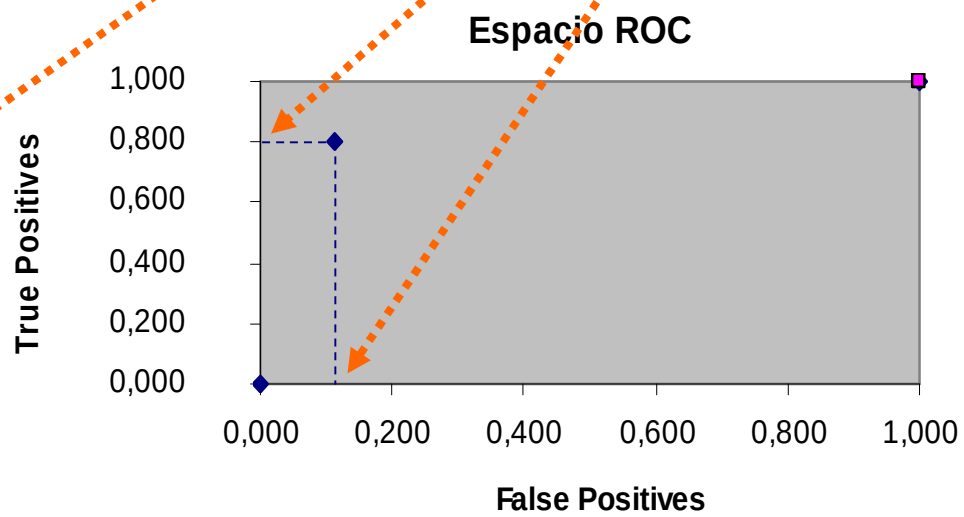
Real

Pred	abrir	cerrar
ABRIR	400	12000
CERRAR	100	87500

Real

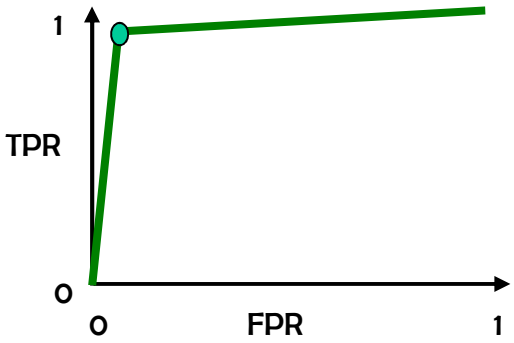
Pred	abrir	cerrar
ABRIR	0,8	0,121
CERRAR	0,2	0,879

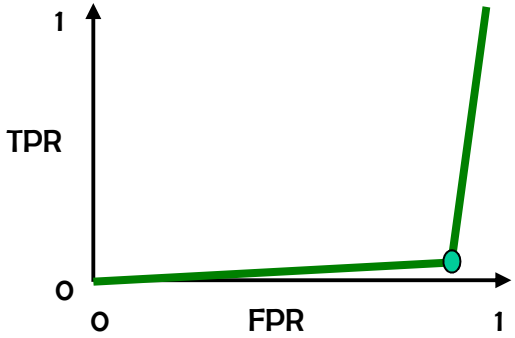
TPR= 400 / 500 = 80%
FNR= 100 / 500 = 20%
TNR= 87500 / 99500 = 87,9%
FPR= 12000 / 99500 = 12,1%

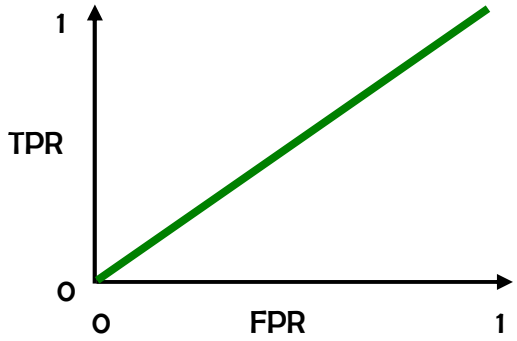


Análisis ROC

- Espacio ROC: buenos y malos clasificadores.



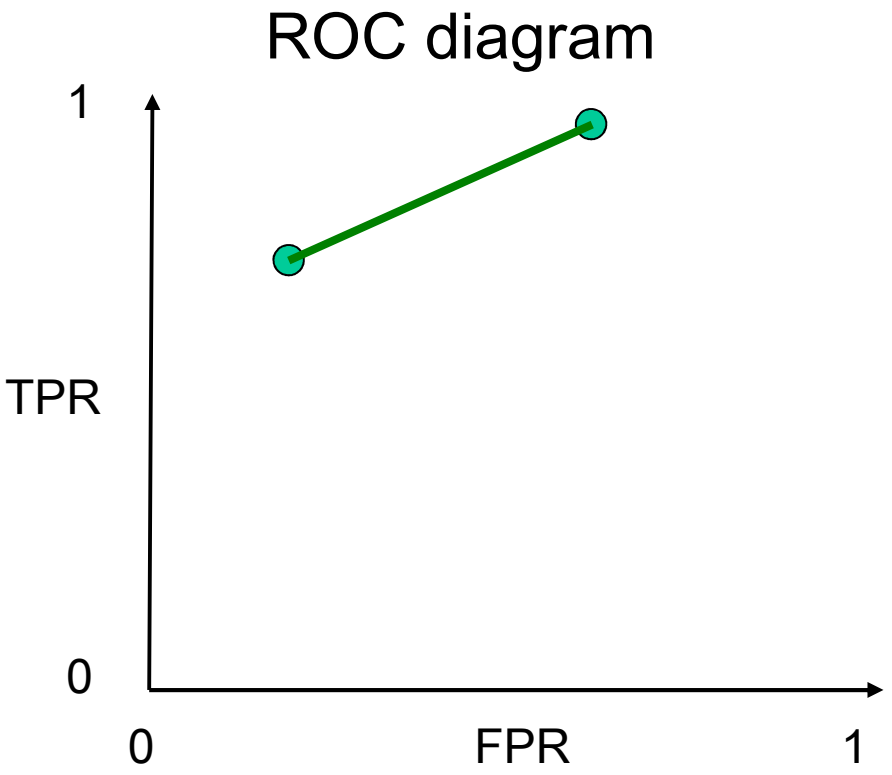




- Buen clasificador.
  - Alto TPR.
  - Bajo FPR.
- Mal clasificador.
  - Bajo TPR.
  - Alto FPR.
- Mal clasificador (en realidad).

Análisis ROC

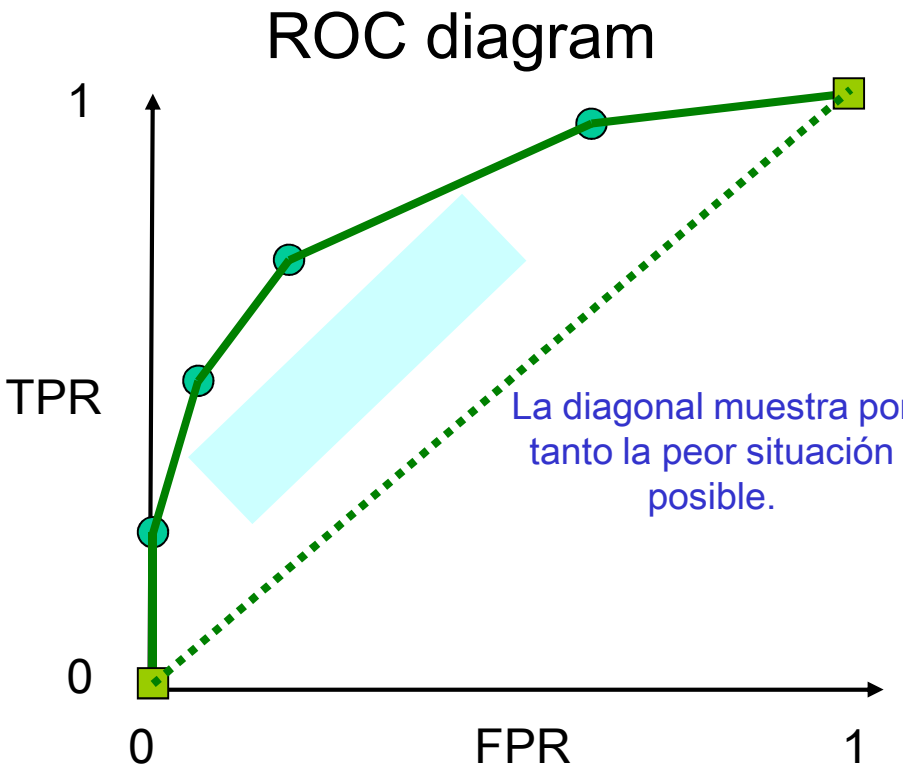
- La Curva ROC. “Continuidad”.



- Podemos construir cualquier clasificador “intermedio” ponderando aleatoramiente los dos clasificadores (con más peso a uno u otro).
- Esto en realidad crea un “continuo” de clasificadores entre cualesquiera dos clasificadores.

Análisis ROC

- La Curva ROC. Construcción.

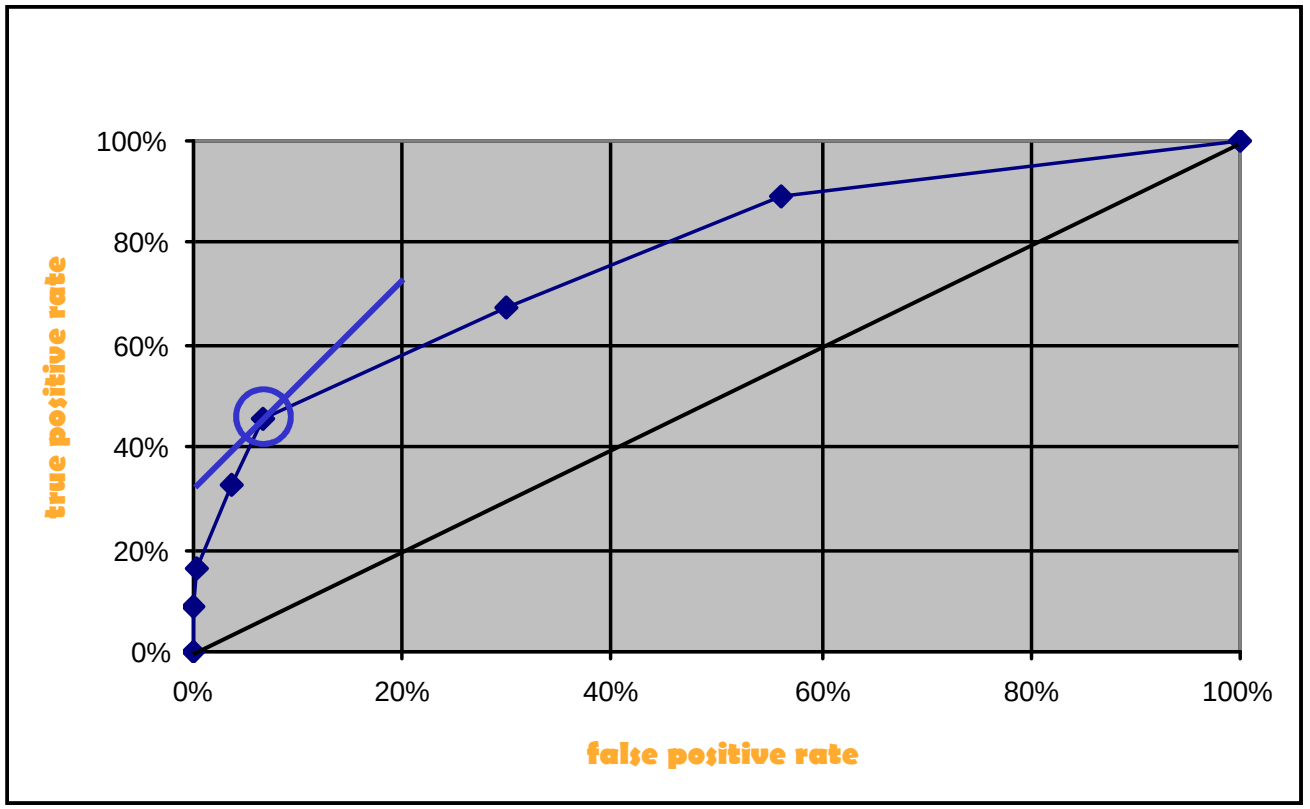


- Dados varios clasificadores:
  - Construimos el "casco convexo" (convex hull) de sus puntos (FPR,TPR) además de los dos clasificadores triviales (0,0) y (1,1).
  - Los clasificadores que caen debajo de la curva ROC se descartan.
  - El mejor clasificador de los que quedan se seleccionará en el momento de aplicación…

Podemos descartar los que están por debajo porque no hay ninguna combinación de distribución de clases / matriz de costes para la cual puedan ser óptimos.

Análisis ROC

- En el **contexto de aplicación**, elegimos el clasificador óptimo entre los mantenidos. Ejemplo 1:



Contexto:

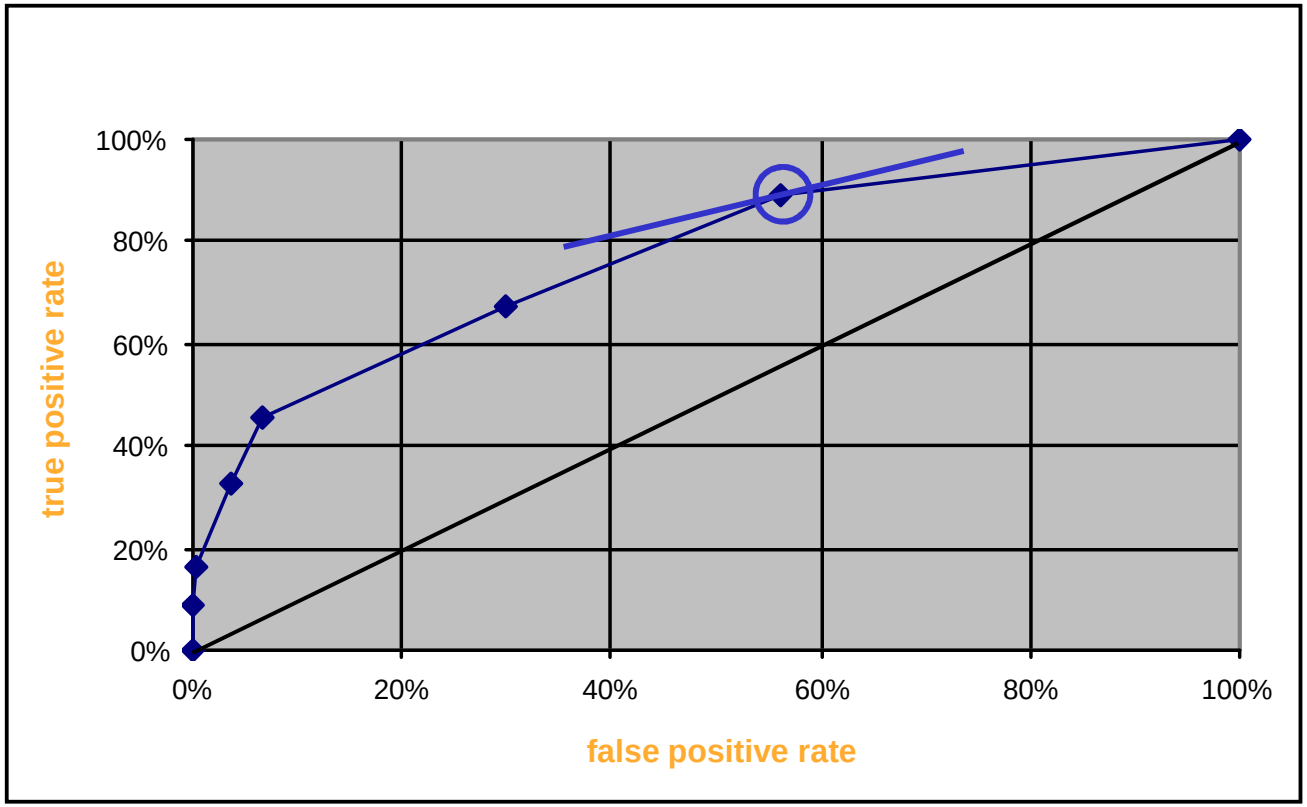
$$\frac{FPcost}{FNcost} = 1/2$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/2 = 2$$

Análisis ROC

- En el **contexto de aplicación**, elegimos el clasificador óptimo entre los mantenidos. Ejemplo 2:



Contexto:

$$\frac{FPcost}{FNcost} = 1/8$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/8 = .5$$

Análisis ROC

- ¿Qué hemos aprendido?
  - La optimalidad de un clasificador depende de la distribución de clases y de los costes de los errores.
  - A partir de este **contexto** se puede calcular una inclinación (o "skew") característica del contexto.
    - Si sabemos este contexto, podemos seleccionar el mejor clasificador, multiplicando la matriz de confusión por la matriz de coste.
    - Si desconocemos el contexto de aplicación en el momento de generación, usando el análisis ROC podemos elegir un subconjunto de clasificadores, entre los cuales seguro estará el clasificador óptimo para cualquier contexto posible, cuando éste se conozca.

¿Podemos ir más allá?

Análisis ROC

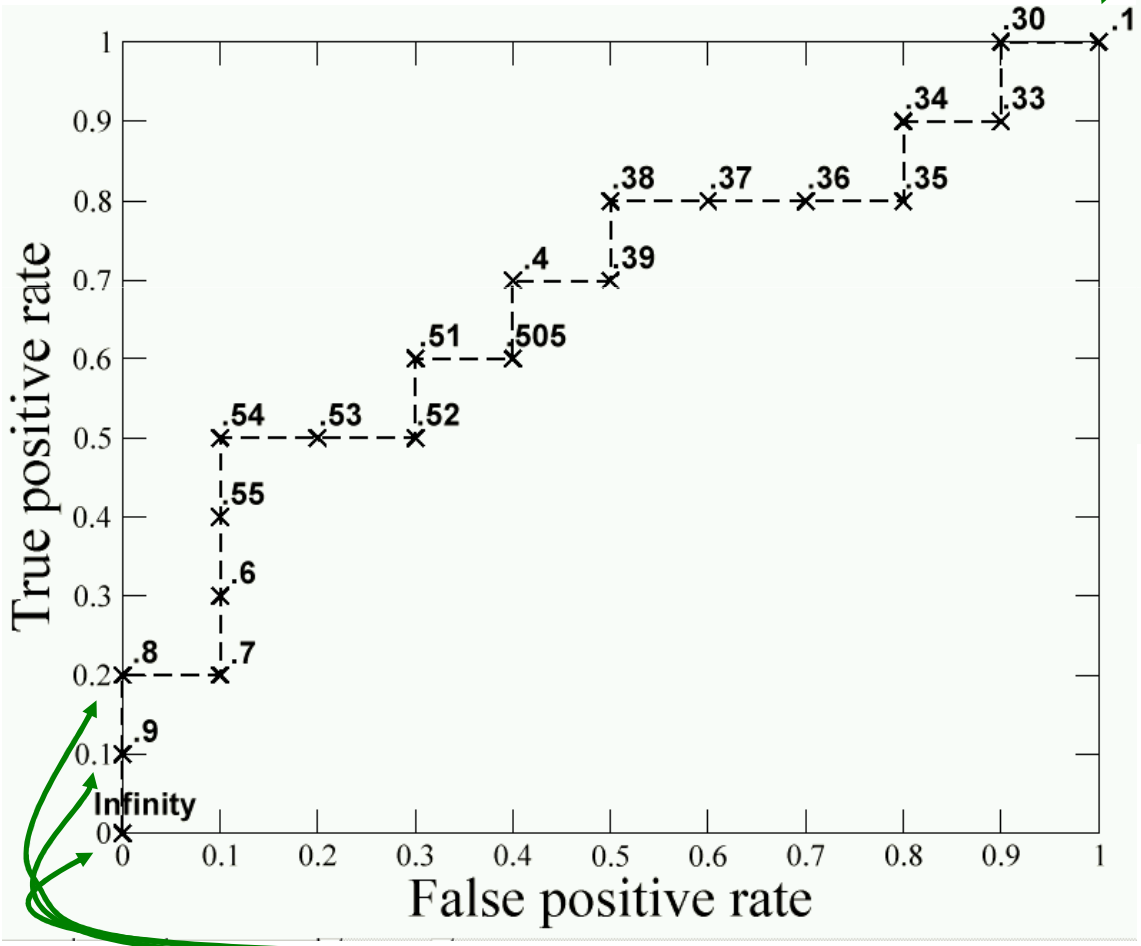
- Curva ROC de un Clasificador Probabilístico:
  - Un clasificador probabilístico (soft) se puede convertir en un clasificador discreto con un umbral.
    - Ejemplo: "si score > 0.7 entonces clase A, si no clase B".
    - Con distintos umbrales, tenemos distintos clasificadores, que les dan más o menos importancia a cada una de las clases (sin necesidad de sobremuestreo o submuestreo).
  - Podemos considerar cada umbral como un clasificador diferente y dibujarlos en el espacio ROC. Esto genera una curva...

Tenemos una "curva" para un solo clasificador "soft"

- Esta curva es escalonada (no se suele realizar el "convex hull").

Análisis ROC

- Curva ROC de un Clasificador “soft”:
  - Ejemplo:



Clase Real

Clase Predicha

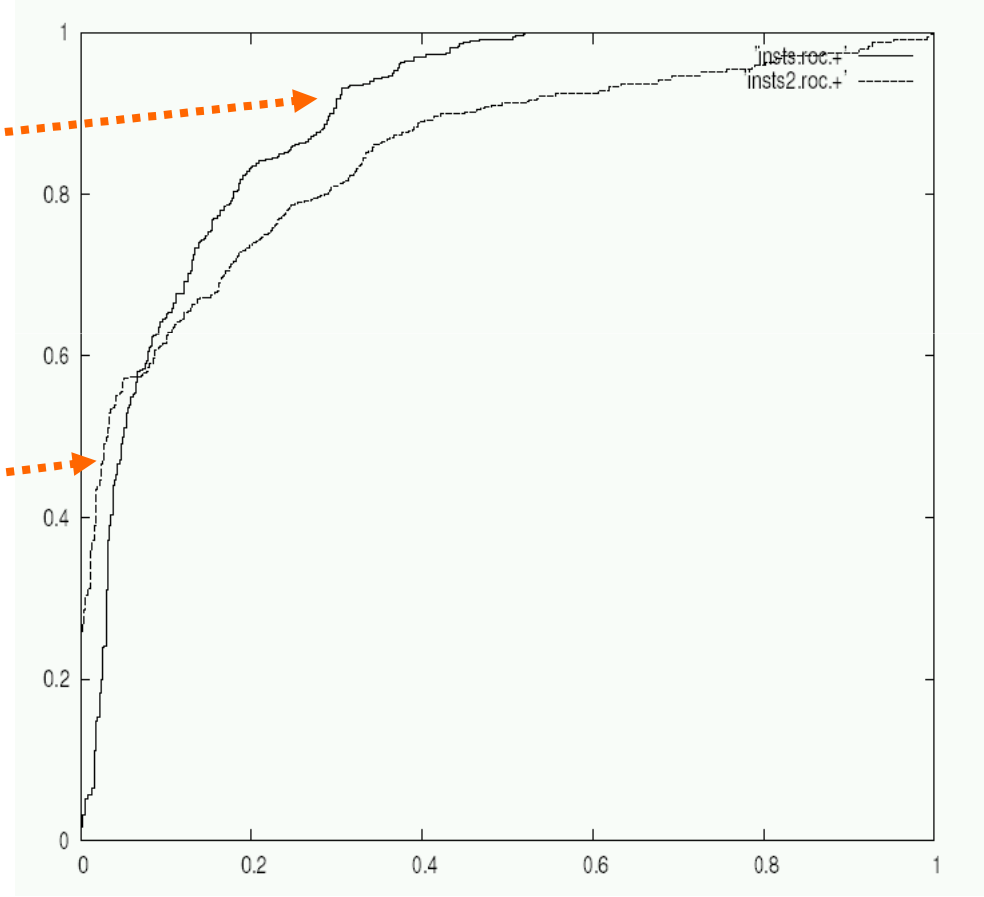
Inst#	Class	Score				
1	p	.9	n	p	p	p
2	p	.8	n	n	p	p
3	n	.7	n	n	n	p
4	p	.6	n	n	n	p
5	p	.55	n	n	n	p
6	p	.54	n	n	n	p
7	n	.53	n	n	n	p
8	n	.52	n	n	n	p
9	p	.51	n	n	n	p
10	n	.505	n	n	n	... p
11	p	.4	n	n	n	p
12	n	.39	n	n	n	p
13	p	.38	n	n	n	p
14	n	.37	n	n	n	p
15	n	.36	n	n	n	p
16	n	.35	n	n	n	p
17	p	.34	n	n	n	p
18	n	.33	n	n	n	p
19	p	.30	n	n	n	p
20	n	.1	n	n	n	p

Análisis ROC

- Análisis ROC de varios clasificadores “soft”:

En esta zona es mejor el clasificador “insts”

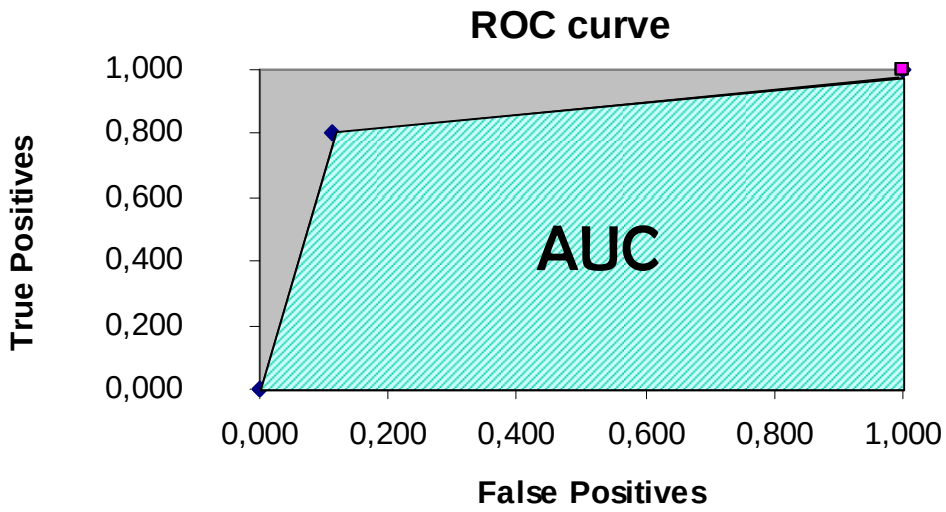
En esta zona es mejor el clasificador “insts2”



- Debemos mantener los clasificadores que tengan al menos una “zona mejor” y después actuar igual que en el caso de los clasificadores discretos.

Análisis ROC

- ¿Para seleccionar un solo clasificador discreto?
  - Se selecciona el que tiene mayor área bajo la curva ROC (AUC, *Area Under the ROC Curve*).

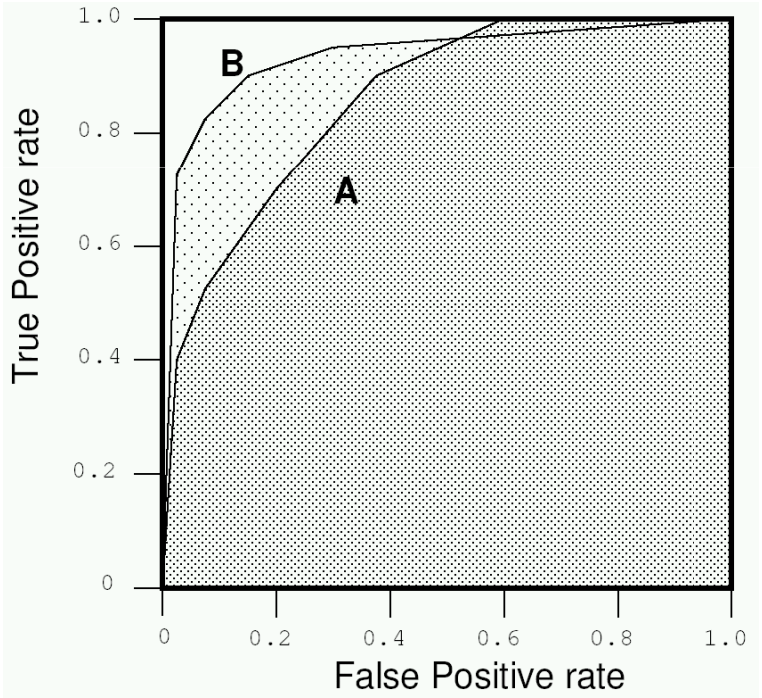


Alternativa al error para evaluar clasificadores

- Un método de aprendizaje / MD será mejor si genera clasificadores con alta AUC.

Análisis ROC

- ¿Para seleccionar un solo clasificador probabilístico?
  - Se selecciona el que tiene mayor área bajo la curva ROC (AUC, *Area Under the ROC Curve*).

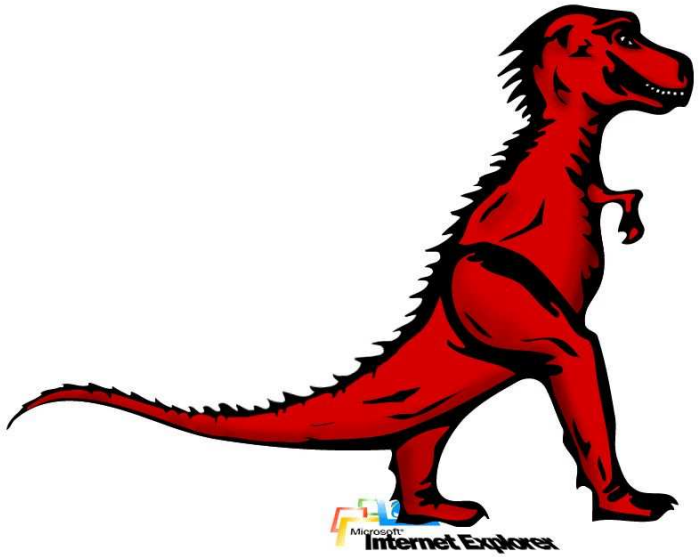


En este caso seleccionamos el B.

  - Evalúa cuán bien un clasificador realiza un ránking de sus predicciones.

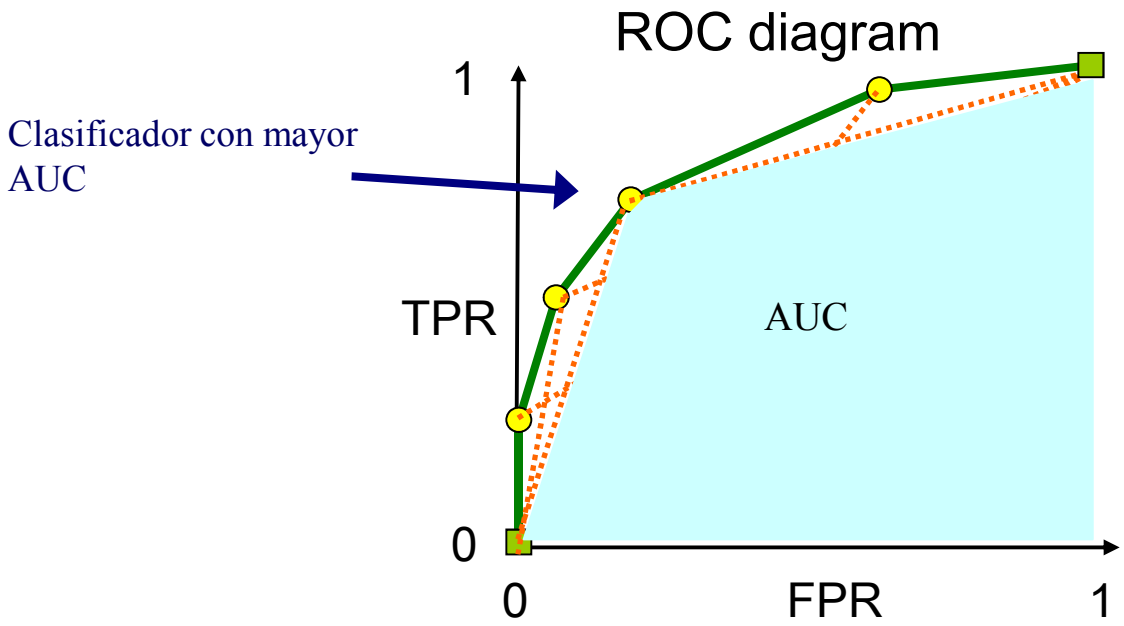
Análisis ROC

•Ejemplo: Detección Spam



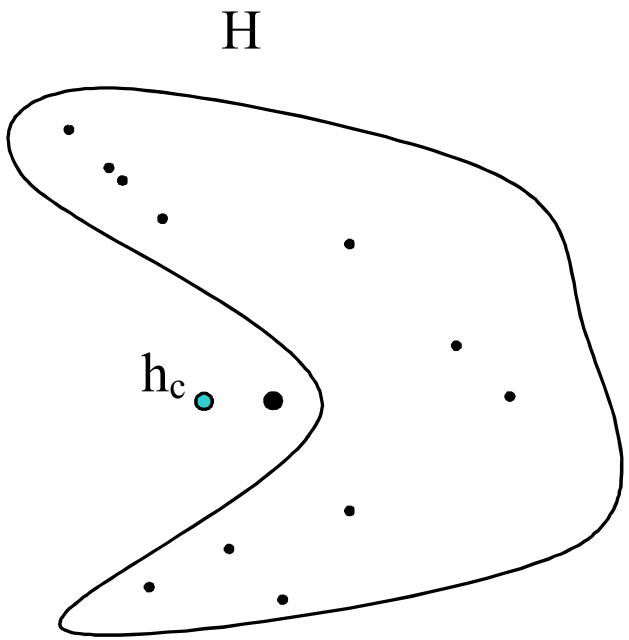
Análisis ROC

•Ejemplo: Detección Spam



Multi-clasificadores

•Una manera de mejorar las predicciones es combinar varios modelos



Multi-clasificadores

• Para obtener buenos resultados en la clasificación es necesario tener un conjunto de modelos (*ensemble*):

- Precisión alta
- Diferentes

•Dados 3 modelos {h1, h2, h3}, considere un nuevo dato x a ser clasificado:

- Si los tres clasificadores son similares, entonces cuando h1(x) sea erróneo, probablemente h2(x) y h3(x) también lo serán.

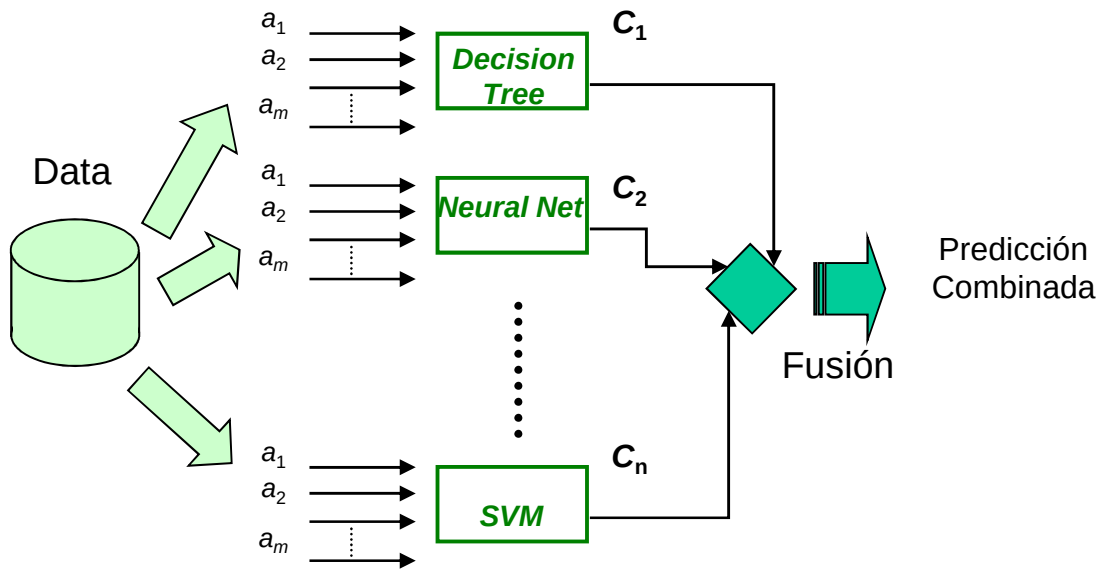
- Si los clasificadores son lo bastante diversos, cuando h1(x) sea erróneo, h2(x) y h3(x) podrían ser correctos, y entonces el conjunto combinado clasificaría correctamente el dato x.

Multi-clasificadores

- Métodos para generar *ensembles*:
  - Manipulación de los datos de entrenamiento
  - Manipulación de los atributos
  - Manipulación de las clases
  - Métodos aleatorios

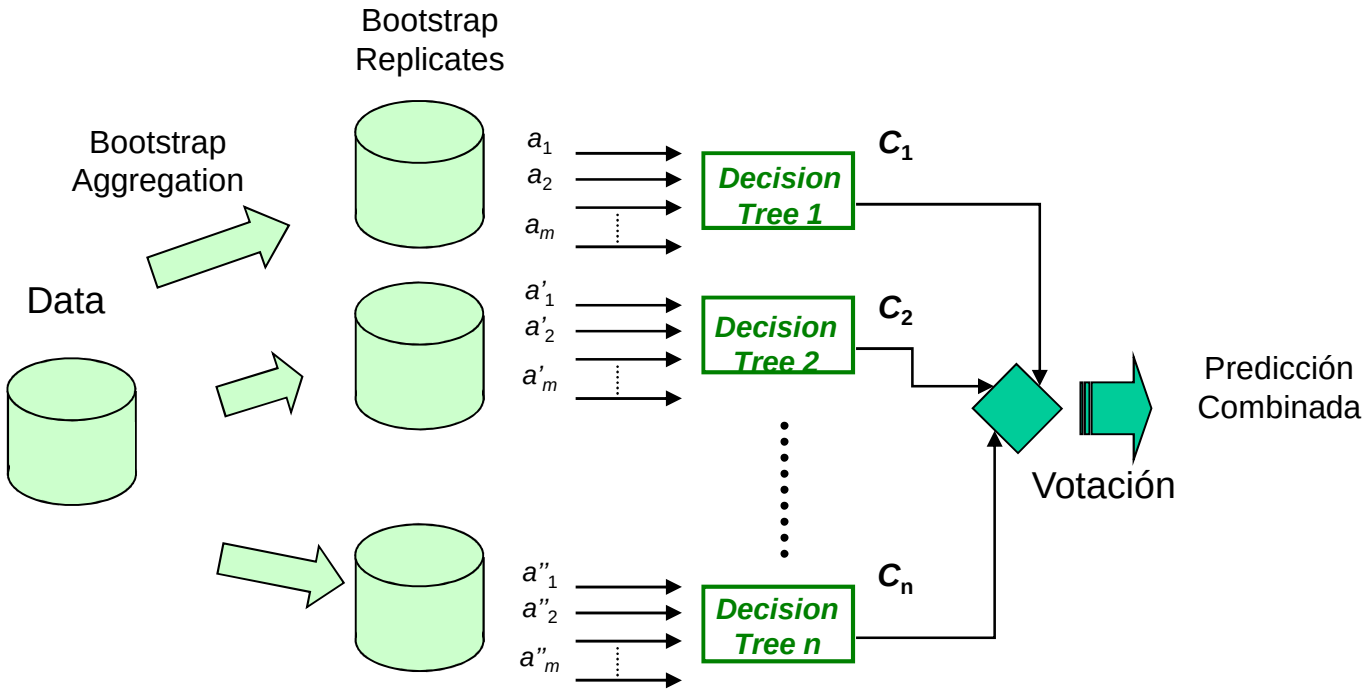
Multi-clasificadores

- Combinación simple (*voting*):



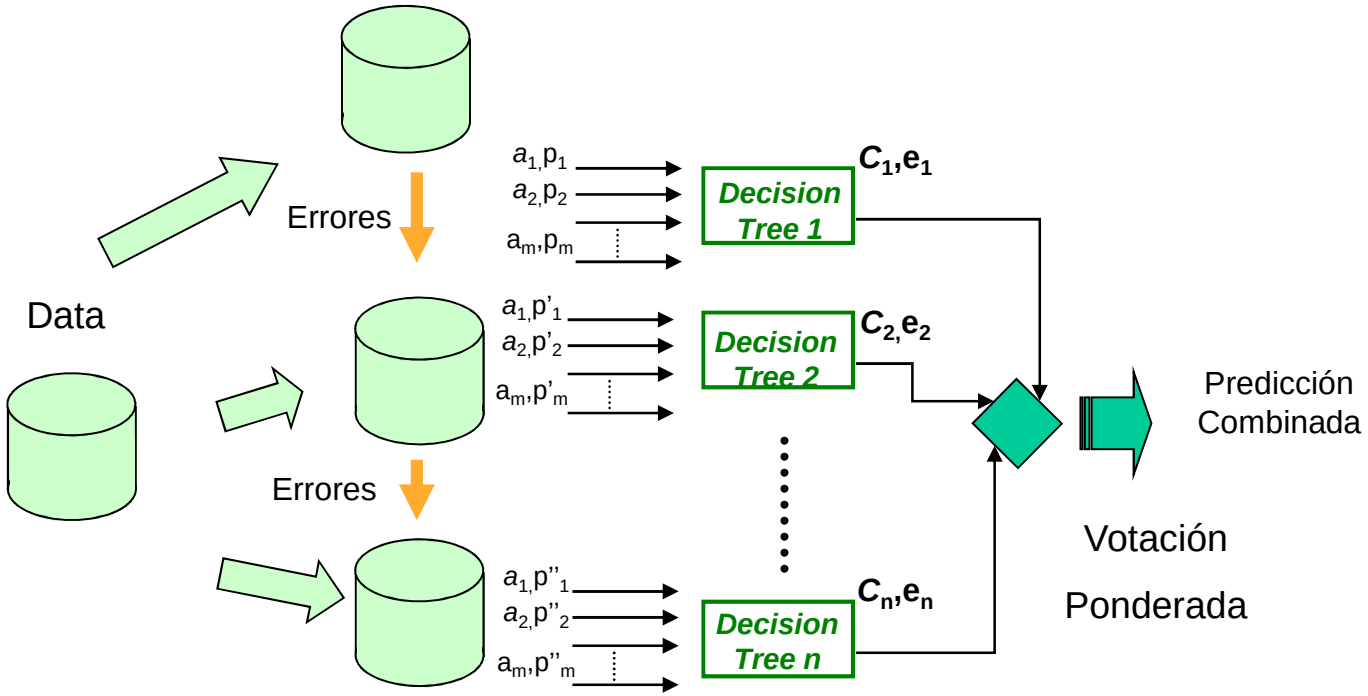
Multi-clasificadores

- Bagging (*Bootstrapt Aggregation*):



Multi-clasificadores

- Boosting



Multi-clasificadores

- Varios trabajos han comparado Boosting y Bagging
  - Boosting obtiene mejor precisión en general
  - En problemas con ruido Bagging es más robusto

Multi-clasificadores

- Stacking