Ben Chams - Fotolia

UNIVERSIDAD NACIONAL DE COLOMBIA

Diplomado
BIG DATA ANALYTICS
Análisis Predictivo

Por
Ing. Fabio González,
Ph.D.

Agenda

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Análisis Predictivo
  - ❑ Clasificación
  - ❑ Evaluación de Algoritmos de Clasificación
  - ❑ Generalización y Sobre-ajuste
  - ❑ Clasificación Sensible al Costo
  - ❑ Regresión y Series de Tiempo

Agenda

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

1. Análisis Predictivo de Datos
   - ❑ Clasificación
   - ❑ Evaluación de Algoritmos de Clasificación
   - ❑ Generalización y Sobre-ajuste
   - ❑ Clasificación Sensible al Costo
   - ❑ Regresión y Series de Tiempo

Análisis Predictivo de Datos

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

Uso de algunas variables para predecir valores desconocidos o futuros de otras variable

Clasificación

- Dada una colección de registros (***conjunto de entrenamiento***)
  - Cada registro contiene un conjunto de ***atributos***, uno de ellos es la ***clase***

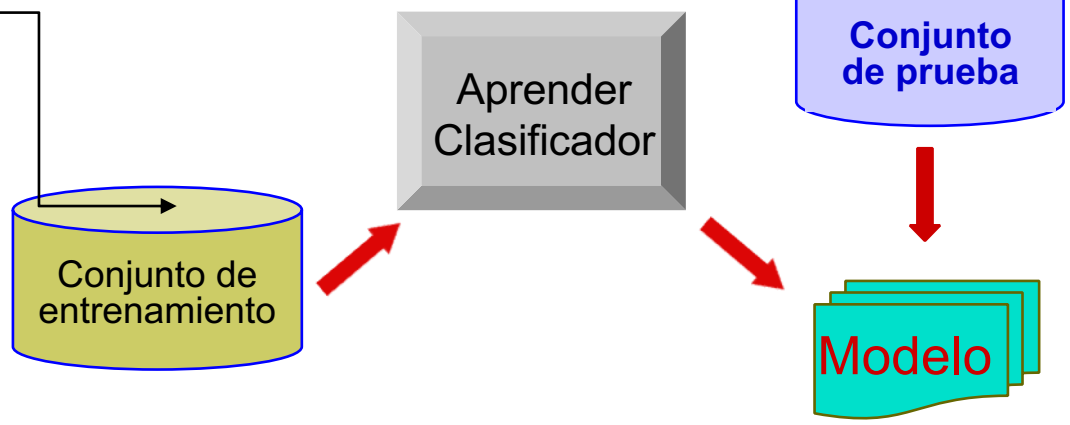
- Encontrar un modelo para el atributo de clase como una función de los valores de los atributos

- Objetivo: asignar una clase lo más preciso posible a los registros que no se han visto antes
  - Un ***conjunto de prueba*** es usado para determinar la exactitud del modelo. A menudo se divide el conjunto de datos en conjuntos de entrenamiento (construcción del modelo) y de prueba (validación)

Clasificación

Tid	Reem-bolso (categórico)	Estado Civil (categórico)	Ingreso (continuo)	Evade
1	Sí	Soltero	125K	No
2	No	Casado	100K	No
3	No	Soltero	70K	No
4	Sí	Casado	120K	No
5	No	Divorciado	95K	Sí
6	No	Casado	60K	No
7	Sí	Divorciado	220K	No
8	No	Soltero	85K	Sí
9	No	Casado	75K	No
10	No	Soltero	90K	Sí

Reem-bolso	Estado Civil	Ingreso	Evade
No	Soltero	75K	?
Sí	Casado	50K	?
No	Casado	150K	?
Sí	Divorciado	90K	?
No	Soltero	40K	?
No	Casado	80K	?



Clasificación: Aplicación 1

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Marketing Directo
  - Objetivo: Reducir el costo de envío de correo al *enfocarse* en los clientes que probablemente compren el nuevo producto
  - Enfoque:
    - Usar datos de un producto similar que se haya lanzado antes
    - Sabemos cuáles clientes decidieron comprar y cuáles no. Esta decisión *{compra, no compra}* forma el *atributo de clase*
    - Reunir información demográfica, del estilo de vida o económica de estos compradores
      - Tipo de negocio, donde viven, cuanto ganan, etc
    - Usar esta información como variables de entrada para construir el modelo de clasificación

Clasificación: Aplicación 2

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Detección de Fraudes
  - Objetivo: Predecir casos fraudulentos en transacciones de tarjeta de crédito.
  - Enfoque:
    - Usar las transacciones y la información sobre los titulares de la cuenta como los atributos
      - Cuándo compra un cliente, qué compra, qué tan a menudo paga a tiempo, etc
    - Etiquetar las transacciones pasadas como fraude o correcta. Con esto se crea el atributo de clase.
    - Construir un modelo para la clase de transacción Tipo de negocio, donde viven, cuanto ganan, etc
    - Usar el modelo para detectar fraudes observando las transacciones de una tarjeta de crédto

Clasificación: Aplicación 3

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Pérdida/Cancelación de Clientes:
  - Objetivo: Predecir si es probable que se pierda un cliente frente a un competidor
  - Enfoque:
    - Usar registros detallados de transacciones de clientes antiguos y actuales para hallar atributos
      - Qué tan frecuente llama el cliente, dónde llama, a qué hora hace la mayoría de llamadas, estado financiero, estado civil, etc.
    - Etiquetar los clientes como leal o desleal
    - Encontrar un modelo para la lealtad

Clasificación: Aplicación 4

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

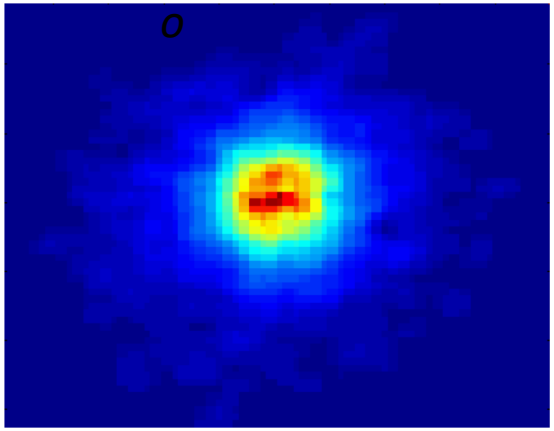
- Pérdida/Cancelación de Clientes
  - Objetivo: Predecir la clase (estrella o galaxia) de los objetos del cielo, especialmente aquellos apenas visibles, con base en las imágenes de telescopio (desde el Observatorio Palomar).
    - 3000 imágenes con 23,040 x 23,040 píxeles por imagen
  - Enfoque:
    - Usar registros detallados de transacciones de clientes antiguos y actuales para hallar atributo
    - Segmentar la imagen
    - Medir los atributos de la imagen (características) - 40 por objeto
    - Modelar la clase con base en estas características
    - Historia de éxito: Se encontraron 16 nuevos cuásares con alto desplazamiento al rojo, unos de los objetos más lejanos que son difíciles de encontrar!

Clasificación: Aplicación 5

- Clasificando galaxias

Tempran

o



Clase:

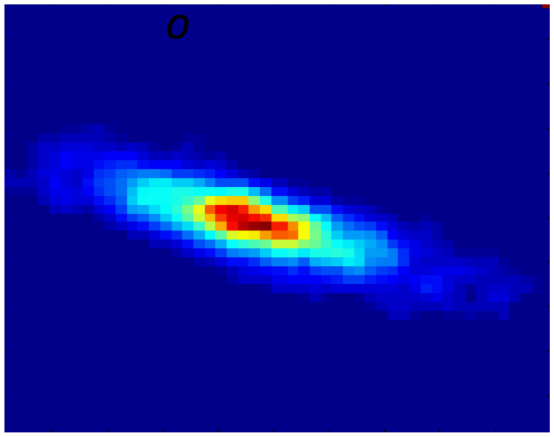
- Estados de Formación

Atributos:

- Características de la Imagen,
- Características de ondas de luz recibidas, etc.

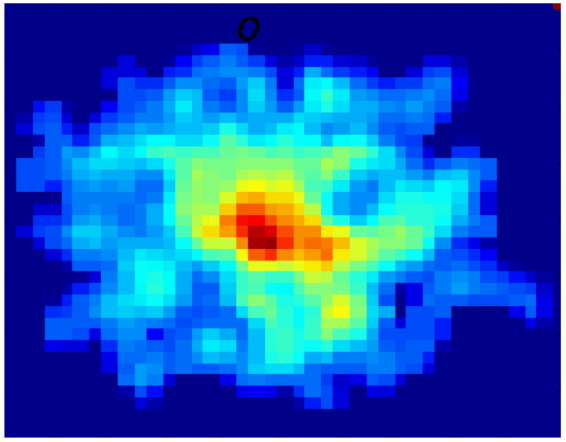
Intermedi

o



Tardí

o



Tamaño de los datos:

- 72 millones de estrellas, 20 millones de galaxias
- Catálogo de Objetos: 9 GB
- Base de Datos de Imágenes: 150 GB

Clasificación: Aplicación 6

- Detectar desviaciones significativas del comportamiento normal

- Aplicaciones:
  - Detección de fraude en tarjetas de crédito
  - Detección de intrusos en redes de computadores

Agenda

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

1. Análisis Predictivo de Datos
   - ❑ Clasificación
   - ❑ Evaluación de Algoritmos de Clasificación
   - ❑ Generalización y Sobre-ajuste
   - ❑ Clasificación Sensible al Costo
   - ❑ Regresión y Series de Tiempo

Evaluación de Algoritmos

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Criterios (dependiendo de la aplicación):
  - Error de clasificación o riesgo
  - Complejidad espacio/temporal del entrenamiento
  - Complejidad espacio/temporal de la aplicación
  - Interpretabilidad
  - Programación sencilla

Matrices de Confusión

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Usadas cuando más de dos clases están involucradas

	Clase Predicha		
Clase verdadera	Setosa	Virginica	Versicolor
Setosa	50	0	0
Virginica	0	48	2
Versicolor	0	1	49

Medidas de Error

	Clase Predicha	
Clase verdadera	+	-
+	True Positive	False Negative
-	False Positive	True Negative

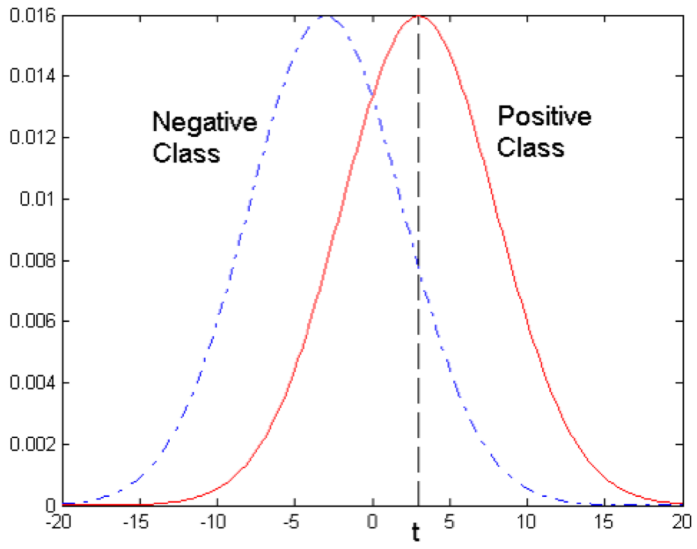
- %**error** = #errores/#instancias = (FN+FP)/N
- **sensitividad** = #+ encontrados/#+ = TP/(TP+FN) (recall, hit rate)
- **precisión**= #+ encontrados/ #total de encontrados = TP/(TP+FP)
- **especificidad** = TN/(TN+FP)
- %**falsas alarmas** = FP/(FP+TN)

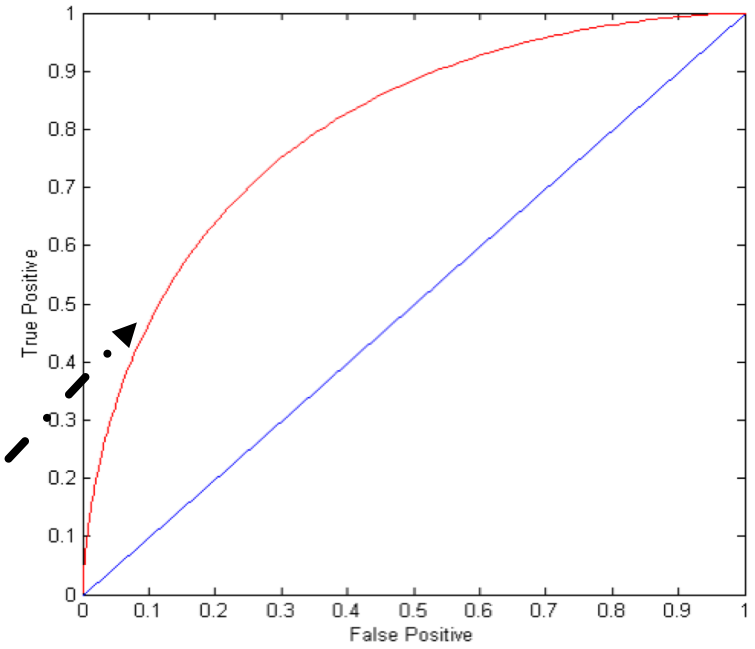
Curvas ROC

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Caracteriza el umbral entre hits positivos y falsas alarmas
- La curva ROC modela los TP en el eje Y, contra los FP en el eje X
- El comportamiento de cada clasificador es representado como un punto en la curva ROC

Curvas ROC



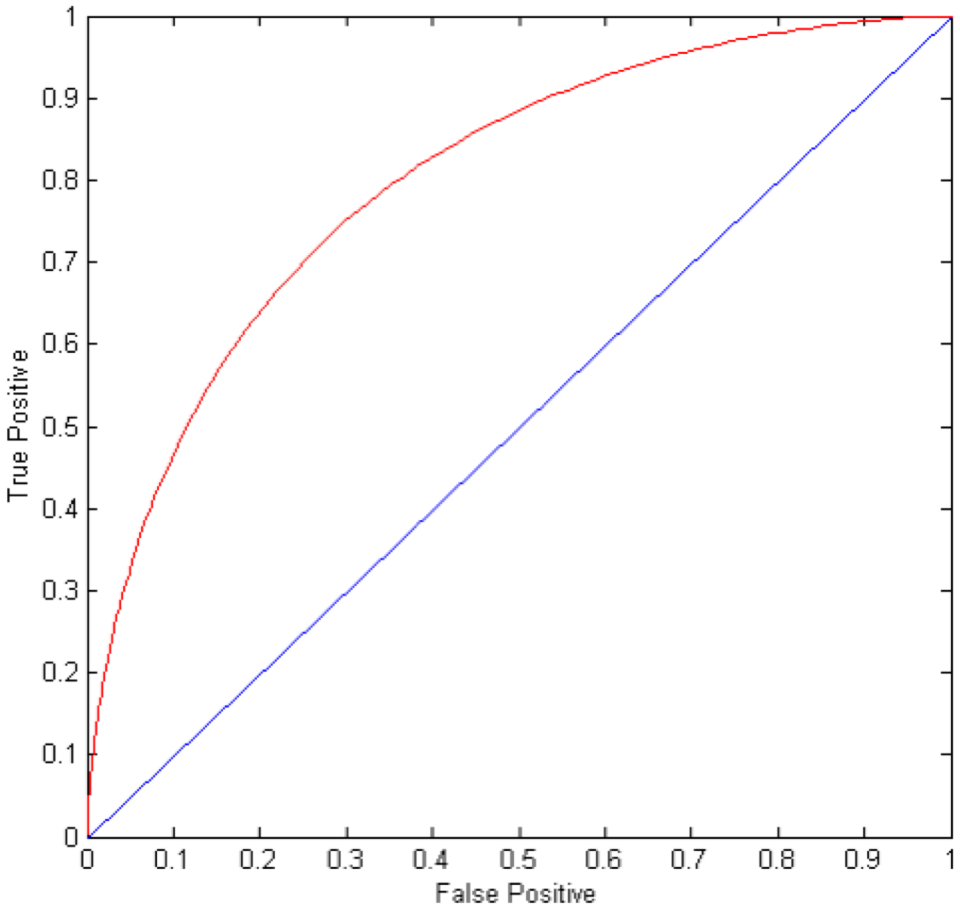


At threshold t:
TP=0.5, FN=0.5, FP=0.12, FN=0.88

Curvas ROC

(TP,FP):

- (0,0): Declara todo para ser clase negativa
- (1,1): Declara todo para ser clase positiva
- (1,0): Ideal
- Línea diagonal: Suposición



Métodos de Estimación

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- **Holdout**
  - Mantener un porcentaje de instancias (2/3) para entrenamiento y el resto (1/3) para pruebas
  - Se sugiere que el de entrenamiento sea mayor que el de pruebas

- **Muestreo Aleatorio**
  - Repetir varias veces holdout y calcular estadísticos sobre dicho proceso
  - Se sugiere repetir como mínimo 30 veces

- **Validación Cruzada (Cross validation)**
  - Partir el conjunto de datos en k subgrupos disjuntos
  - k-fold: entrenar con k-1 subgrupos, validar con el restante. Repetir usando cada grupo en validación
  - Dejar uno afuera (Leave-one-out): k=n

- **Muestreo Estratificado**
  - sobremuestreo vs submuestreo

- **Bootstrap**
  - Muestreo con repetición

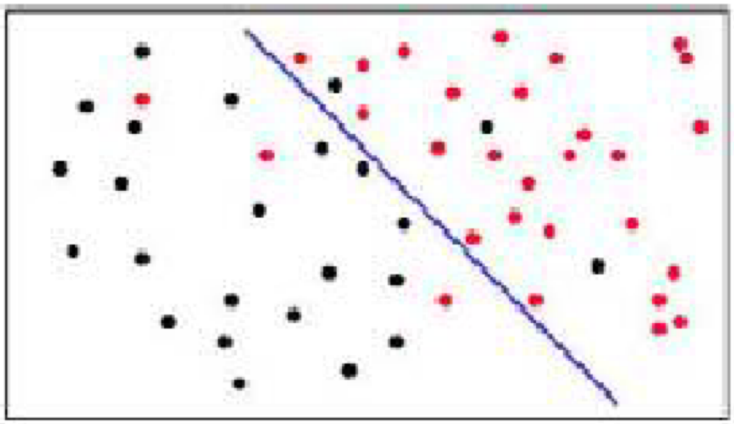
Agenda

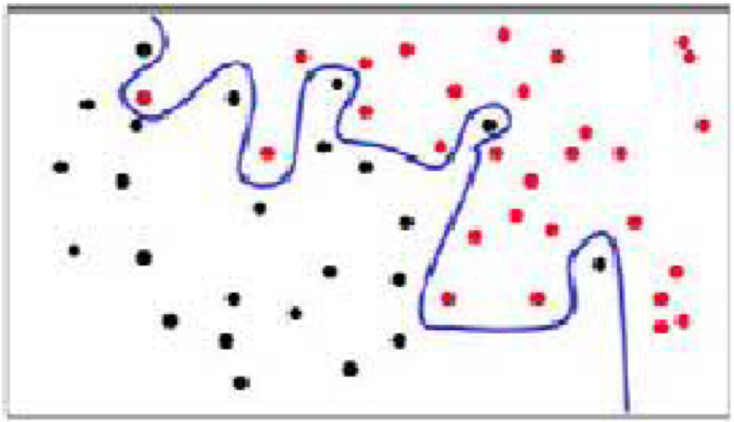
UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

1. Análisis Predictivo de Datos
   - ❑ Clasificación
   - ❑ Evaluación de Algoritmos de Clasificación
   - ❑ Generalización y Sobre-ajuste
   - ❑ Clasificación Sensible al Costo
   - ❑ Regresión y Series de Tiempo

Sobre-ajuste (Overfitting)

- Cuando el algoritmo de aprendizaje se ajusta tanto a los datos de entrada que pierde su capacidad de generalizar



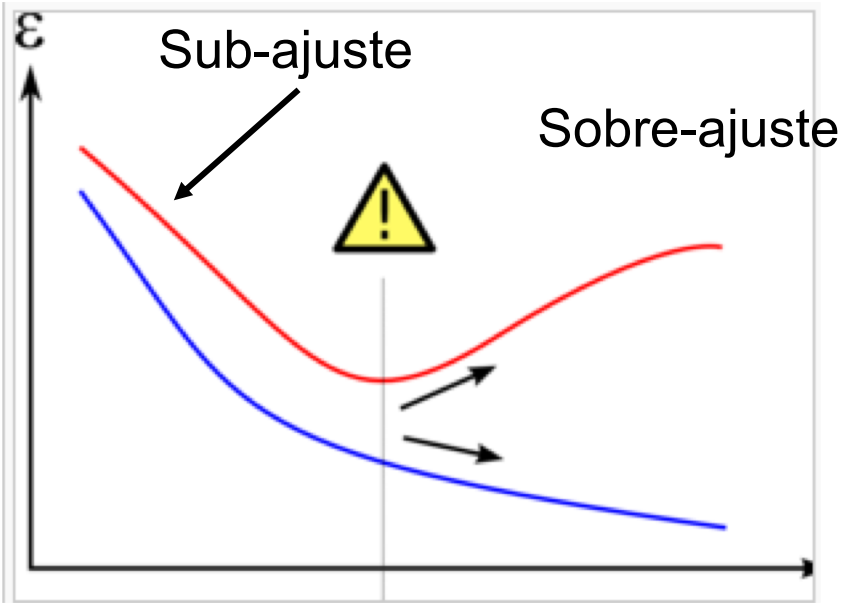


- El error de cálculo de los ejemplos futuros será alto

Sobre-ajuste (Overfitting)

- **Sobre-ajuste:** Bajo error de entrenamiento pero error de generalización alto.

- **Sub-ajuste (underfitting):** Errores de entrenamiento y generalización altos



Sobre-ajuste (Overfitting)

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Ejemplo

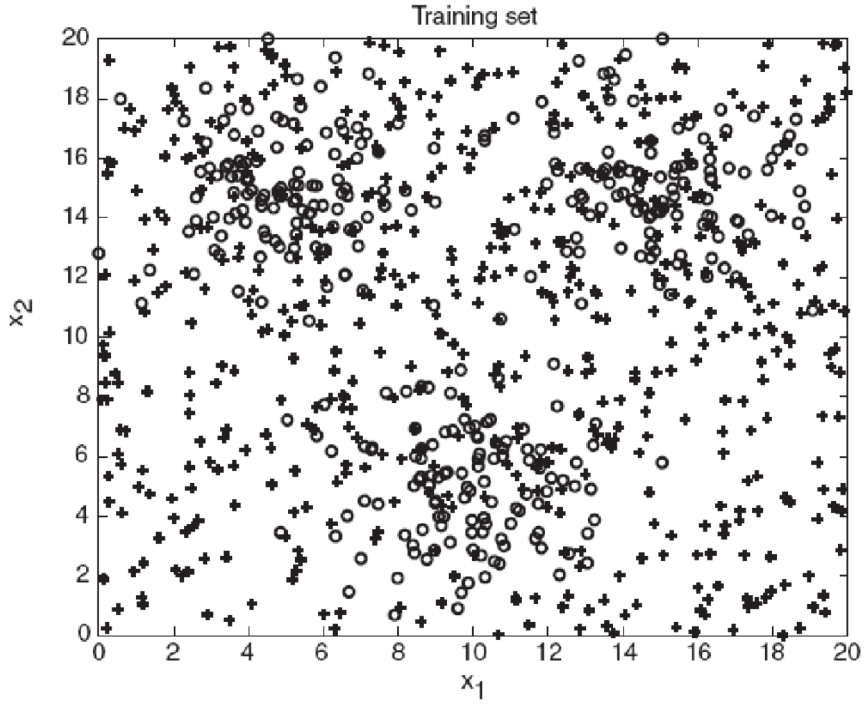


Figure 4.22. Example of a data set with binary classes.

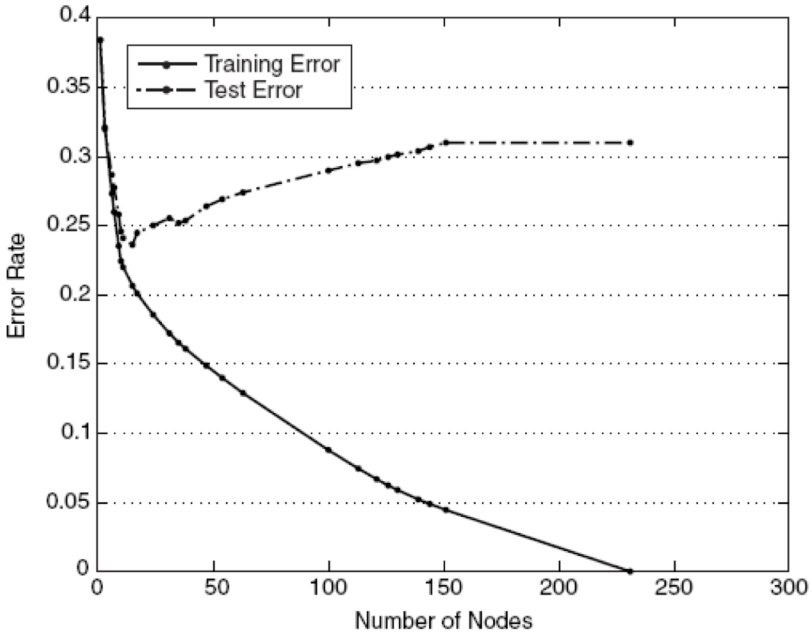
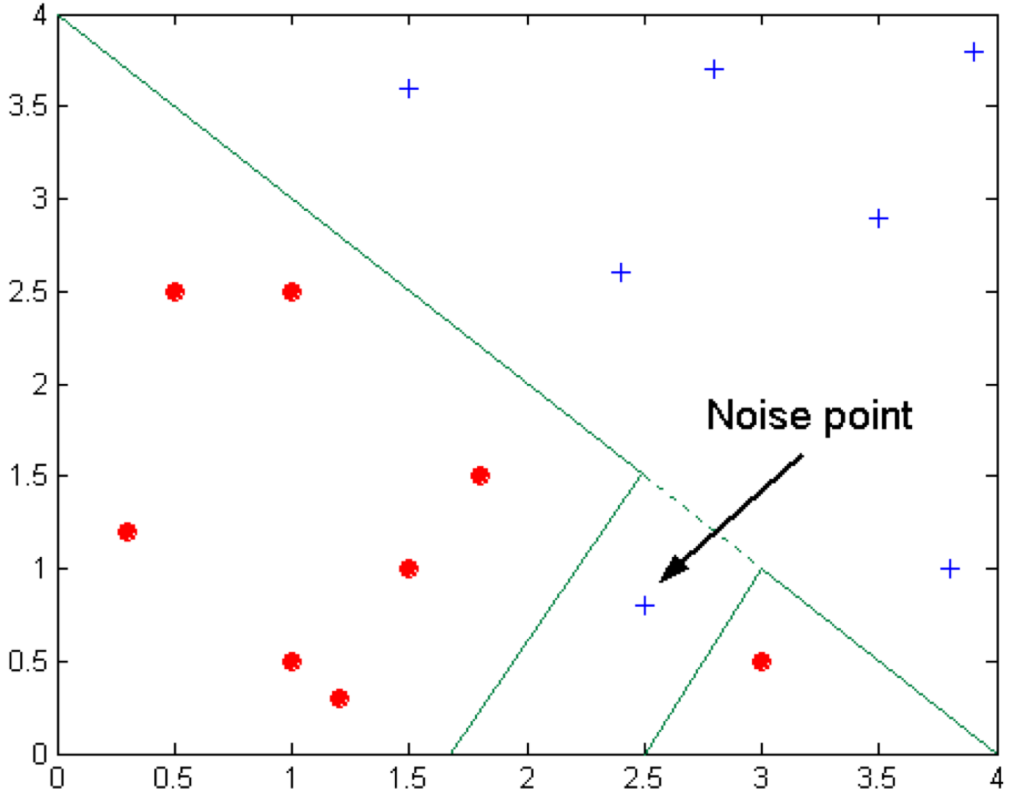


Figure 4.23. Training and test error rates.

Sobre-ajuste (Overfitting)

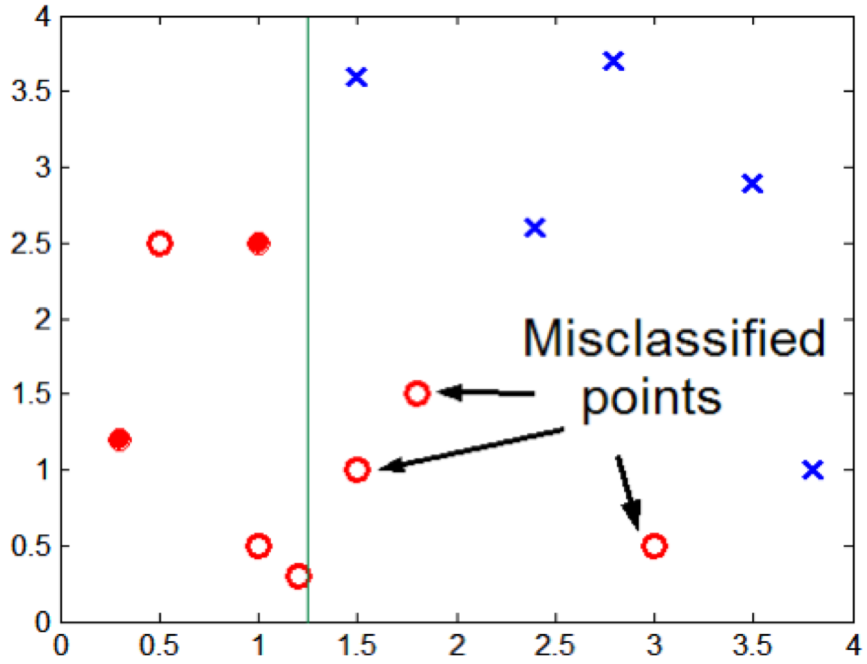
- Presencia de ruido



La frontera de decisión es distorsionada por el ruido

Sobre-ajuste (Overfitting)

- Falta de ejemplos representativos



La falta de ejemplos en la parte inferior del diagrama hace difícil que el modelo realice una predicción acertada en esta región

Error de Clasificación

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Error de entrenamiento:
  - **e(modelo, datos)**
  - Número de ejemplos de entrenamiento clasificados incorrectamente
  - Conocido como error de re-substitución o error aparente

- Error de generalización:
  - **e'(modelo, datos)**
  - Error esperado del modelo en ejemplos no usados en el entrenamiento

- Un buen modelo debe tener errores de entrenamiento y generalización bajos

Estimación del Error de Generalización

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Estimación optimista: Usando re-substitución

e'(modelo, datos) = e(modelo, datos)

- Incorporando la complejidad del modelo – Cuchilla de Occam

e'(modelo, datos) = e(modelo, datos)+costo(modelo, datos)

  - Estimación pesimista
  - Principio MDL (Descripción de mínima longitud)

Complejidad del Modelo

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Parámetro que controla lo complejo del modelo

- En árboles de decisión (tamaño)
  - Pre-podado
  - Post-podado

- En redes neuronales
  - Número neuronas ocultas y/o conexiones
  - Tipo de red neuronal

Agenda

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

1. Análisis Predictivo de Datos
   - ❑ Clasificación
   - ❑ Evaluación de Algoritmos de Clasificación
   - ❑ Generalización y Sobre-ajuste
   - ❑ Clasificación Sensible al Costo
   - ❑ Regresión y Series de Tiempo

Costo

Área	Ejemplo
Marketing	Comprador / no Comprador
Medicina	Enfermo / no Enfermo
Finanzas	Prestar / no Prestar
Spam	Spam / no Spam

Costo

- Suponer que los errores son igualmente costosos puede llevar a malas decisiones

Examples	
Marketing	El costo de hacerle una oferta a un no comprador es pequeña comparada con no contactar un comprador
Finance	El costo de un mal prestamo es mayor que negarle un prestamo aun buen cliente
Spam	Rechazar correo que no sea Spam es más costoso que aceptar correo Spam

Matriz de Costos

Predicted \ Actual	Sunny	Snowy	Rainy
Sunny	0	10	15
Snowy	1	1	11
Rainy	2	2	2

Matriz de Costos

- Costos dependientes. Fraude con tarjeta de crédito

Real

Predicho	Fraude	No fraude
Rechazo	20	- 20
Aprobar	$-x$	$(0.2)x$

x = *valor transación*

Aprendizaje Sensitivo al Costo

- Aprendizaje no sensitivo al costo:

$$\max_{C_i} P(C_j \mid A_1, \ldots, A_n)$$

- Aprendizaje sensitivo al costo:
  - Escoger acción que minimice el costo esperado

$$\min_{C_i} \sum_{C_j \neq C_i} P(C_j \mid A_1, \ldots, A_n) Costo(C_j, C_i)$$

  - Costo(C_j,C_i) = costo de clasificar como C_i cuando realmente es C_j

- Los dos enfoques son equivalentes cuándo los costos son iguales para todos los errores

Metacost

- Es una algoritmo que permite volver cualquier clasificador sensitivo al costo

- Se debe especificar una matriz de costos

- El algoritmo reetiqueta los ejemplos de entrenamiento de manera que el costo esperado se minimice

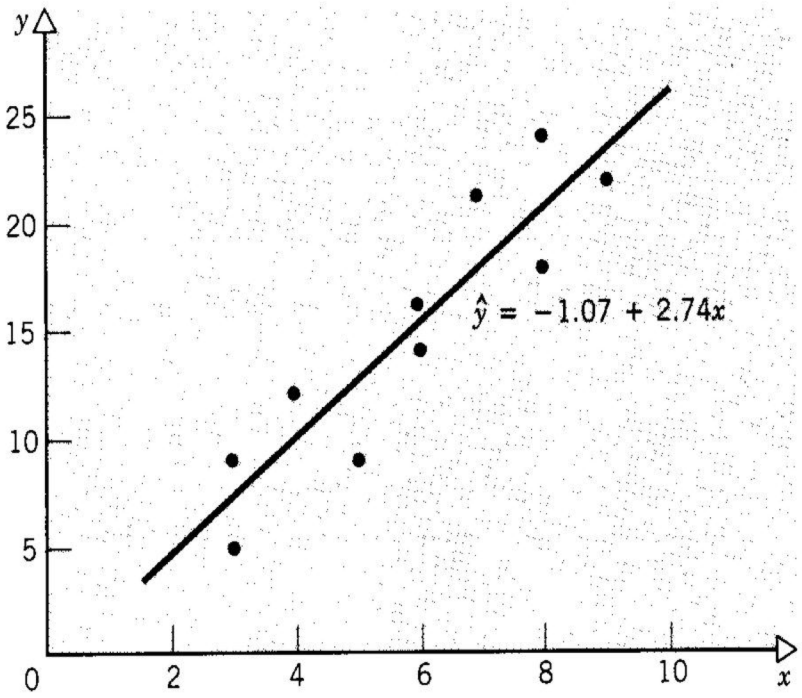
Agenda

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

1. Análisis Predictivo de Datos
   - ❑ Clasificación
   - ❑ Evaluación de Algoritmos de Clasificación
   - ❑ Generalización y Sobre-ajuste
   - ❑ Clasificación Sensible al Costo
   - ❑ Regresión y Series de Tiempo

Regresión

- Similar al problema de clasificación pero el atributo de clase es continua
- Problema: Encontrar una relación funcional entre una variable dependiente y uno o varias variables independientes
- Tipos:
  - Regresión lineal
  - Regresión no lineal
  - Otros: regresión logísitica, árboles de decisión



Series de Tiempo

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ D.C.

- Base de Datos de Series de Tiempo
  - Consiste en secuencias de valores o eventos que cambian con el tiempo
  - Los datos son almacenados en **intervalos regulares**
  - Componentes principales de una serie de tiempo:
    - Tendencia, ciclo, estacional, irregular

- Aplicaciones
  - Finanzas: inventarios, precio, inflación
  - Industria: consumo de energía
  - Ciencia: resultados de experimentos
  - Metereología: precipitación

Referencias

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2005, Introduction to Data Mining, Addison-Wesley

- Stuart Russell and Peter Norvig “Artificial Intelligence: A Modern Approach”, Second Edition.

- Domingos. *MetaCost: A General Method for Making Classifiers Cost-Sensitive*. In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99). 1999.

- Alan Abrahams, An Introduction to Cost-Sensitive Learning , Lecture Slides, **http://opim.wharton.upenn.edu/~asa28/opim_410_672_spring05/opim_410_guest_lecture_dan_fleder_cost_sensitive_learning.ppt**

¿Preguntas?

fagonzalezo@unal.edu.co

http://www.mindlaboratory.org