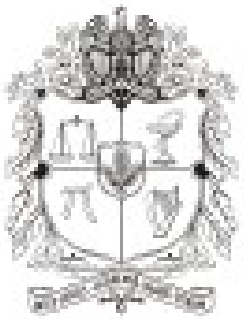
Diplomado en Inteligencia de Negocios

Módulo

Minería de Datos



UNIVERSIDAD NACIONAL DE COLOMBIA

Análisis Supervisado: Técnicas Alternativas

UNIVERSIDAD NACIONAL DE COLOMBIA

Agenda

- Clasificadores Basados en Ejemplos
- Redes Neuronales Artificiales
- Métodos de Ensamble
- Generalización y Sobreajuste

UNIVERSIDAD NACIONAL DE COLOMBIA

Agenda

- Clasificadores Basados en Ejemplos
- Redes Neuronales Artificiales
- Métodos de Ensamble
- Generalización y Sobreajuste

Clasificadores basados en ejemplos

UNIVERSIDAD NACIONAL DE COLOMBIA

Conjunto de ejemplos

Atr1	……	AtrN	Clase
			A
			B
			B
			C
			A
			C
			B

- Almacenar los registros de entrenamiento
- Usar los registros de entrenamiento para predecir la etiqueta de clase de los casos desconocidos

Casos-desconocidos

Atr1	……	AtrN

Clasificadores basados en ejemplos

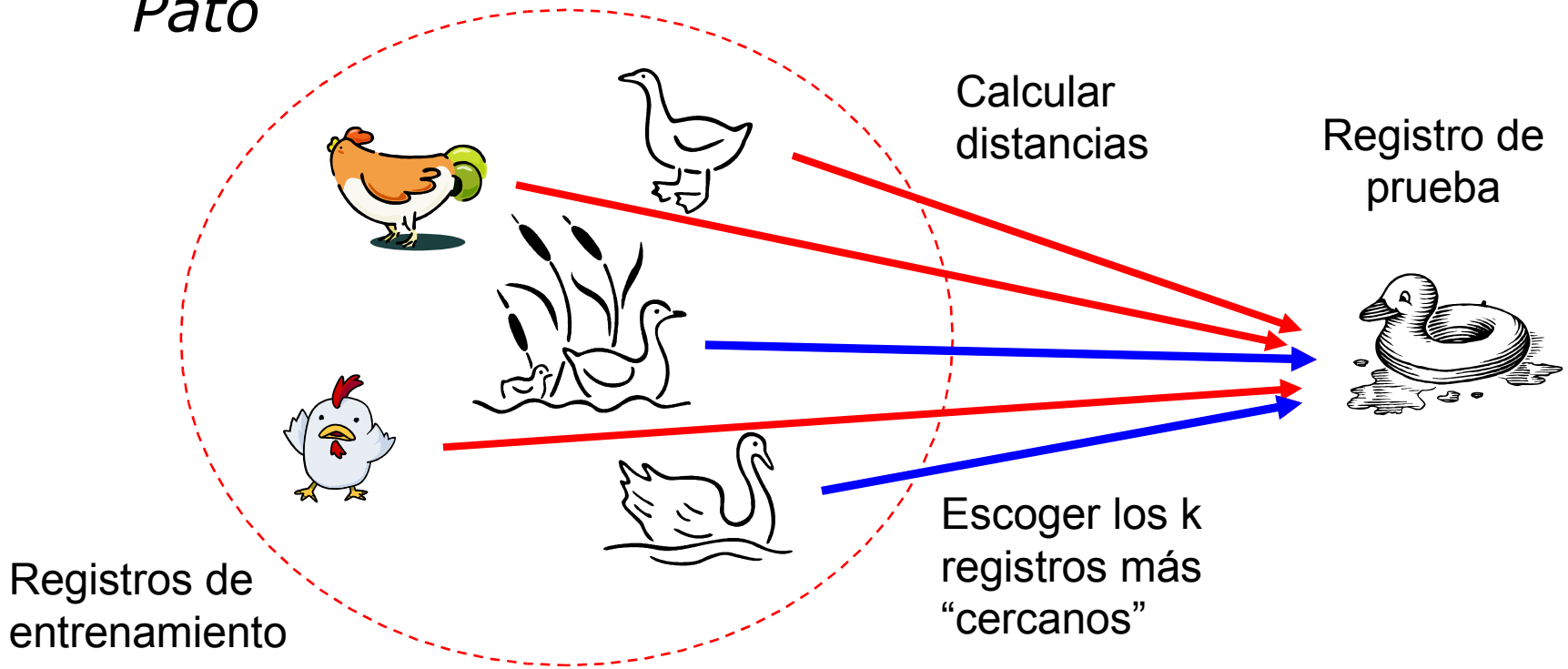
UNIVERSIDAD NACIONAL DE COLOMBIA

- Ejemplos:
  - Rote-learner
    - Memoriza los datos de entrenamiento y realiza la clasificación sólo si los atributos del registro coincide exactamente con uno de los ejemplos de entrenamiento
  - Vecino más cercano (Nearest Neighbor)
    - Utiliza los k puntos "más cercano"(vecinos más cercanos) para realizar la clasificación

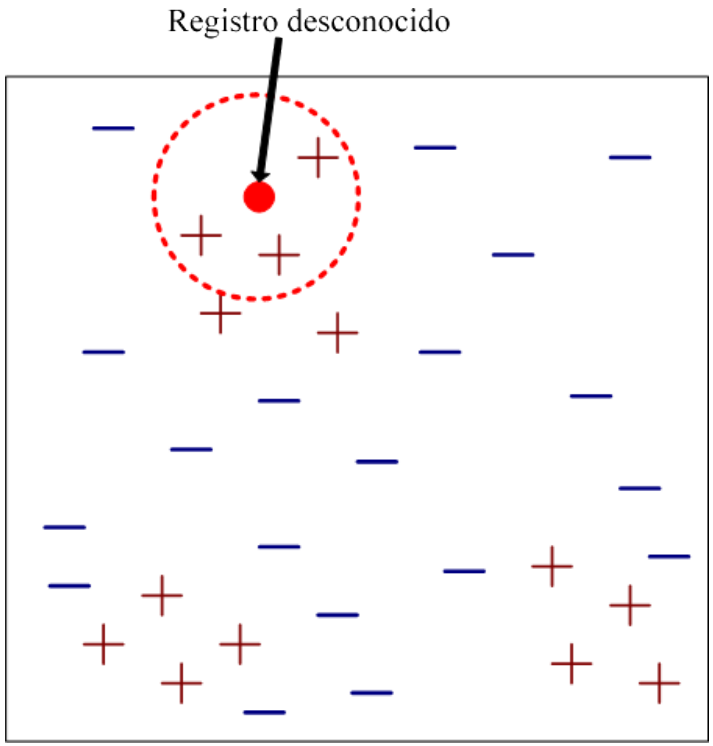
Clasificadores de vecino más cercano

UNIVERSIDAD NACIONAL DE COLOMBIA

- Idea básica
  - *si camina como Pato, hace como Pato y se parece a un Pato probablemente se trata de un Pato*

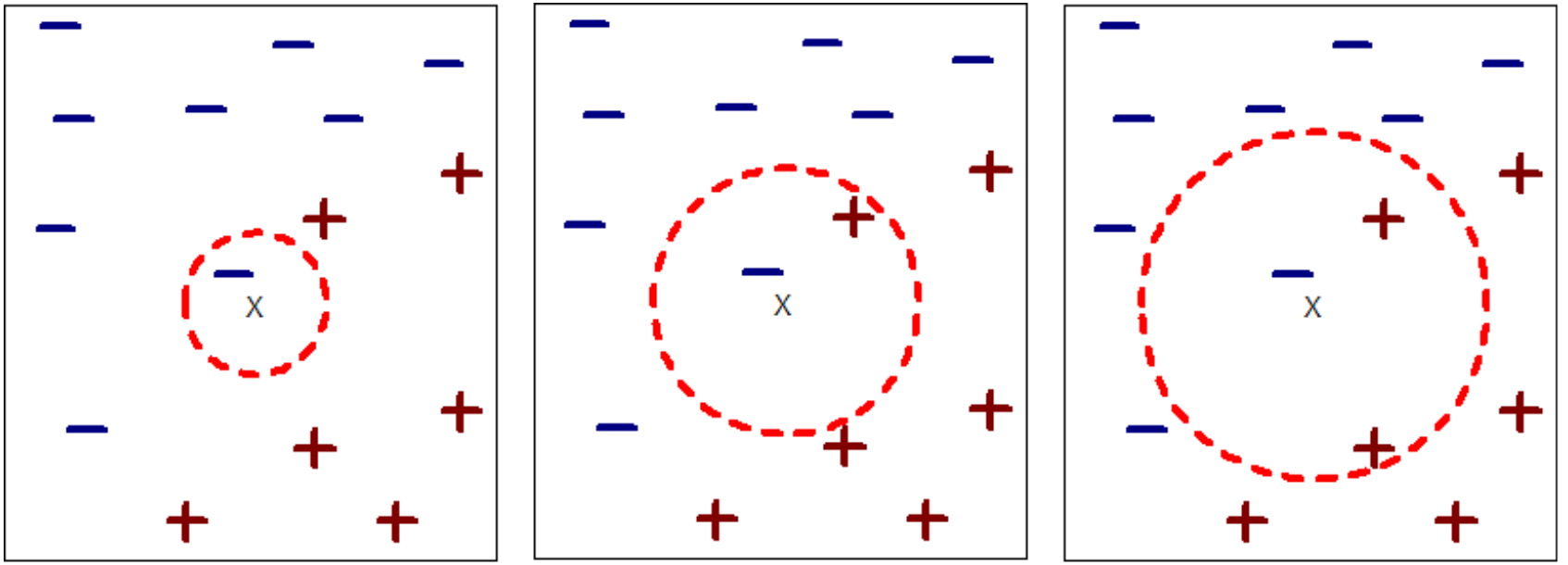


Clasificadores de vecino más cercano



- Exige tres cosas
  - Conjunto de registros almacenados
  - Métrica de Distancia para calcular la distancia entre los registros
  - Valor de k, el número de vecinos más cercanos para recuperar
- Para clasificar un registro desconocido:
  - Calcular la distancia a otros registros de entrenamiento
  - Identificar los k vecinos más cercanos
  - Uso de etiquetas de clase de los vecinos más cercanos para determinar la etiqueta de clase del registro desconocido (e.g., tomando el voto de la mayoría)

Definición de vecino más cercano



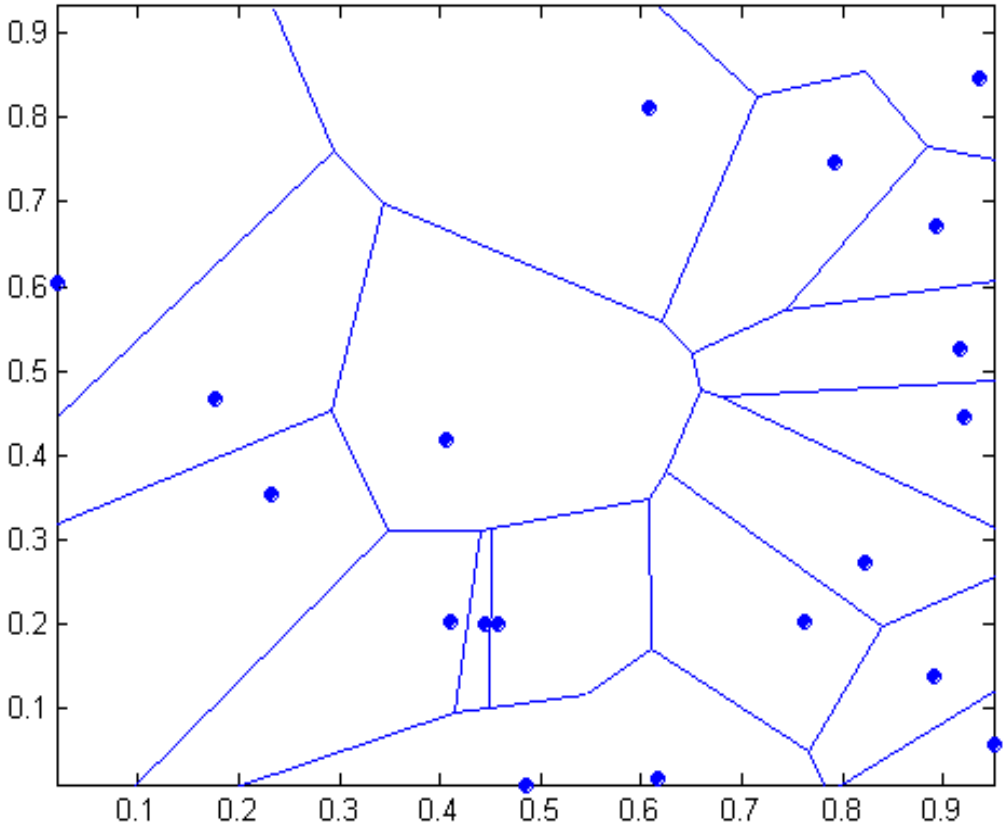
(a) 1-vecino más cercano **(b)** 2-vecinos más cercanos **(c)** 3-vecinos más cercanos

Los K-vecinos más cercanos de un registro ***X*** son los puntos cuya distancia es menor que la k-ésima menor distancia al punto ***X***

UNIVERSIDAD NACIONAL DE COLOMBIA

1 vecino más cercano

- Diagrama Voronoi



Clasificación de vecino más cercano

UNIVERSIDAD NACIONAL DE COLOMBIA

- Calcular la distancia entre dos puntos
  - Distancia Euclidiana

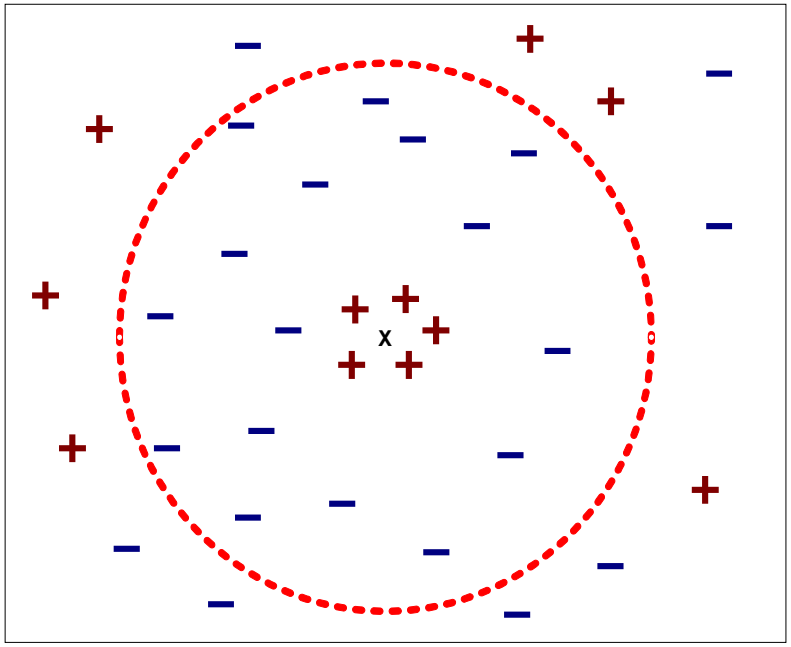
$$d(p,q)=\sqrt{\sum_i (p_i-q_i)^2}$$

- Determinar la clase de la lista de vecinos más cercanos
  - Tomar la clase como el voto de la mayoría entre los k-vecinos más cercanos
  - Dar pesos a los votos de acuerdo a la distancia
    - Factor de Peso, $w = 1/d^2$

Clasificación de vecino más cercano…

UNIVERSIDAD NACIONAL DE COLOMBIA

- Escoger el valor de k:
  - Si k es muy pequeño, es sensible al ruido
  - Si k es muy grande, la vecindad puede incluir puntos de otras clases



Clasificación de vecino más cercano…

UNIVERSIDAD NACIONAL DE COLOMBIA

- Problemas con la escala
  - Puede ser necesario cambiar la escala de algunos atributos para evitar que las distancias sean dominadas por uno de los atributos
  - Ejemplo:
    - la altura de una persona puede variar de 1,5 m a 1,8 m
    - el peso de una persona puede variar de 90 lb a 300 lb
    - los ingresos de una persona pueden variar de $10K a $1M

Clasificación de vecino más cercano…

UNIVERSIDAD NACIONAL DE COLOMBIA

- Problema con la distancia Euclideana
  - Alta dimensionalidad de los datos
    - Maldición de la dimensionalidad
  - Puede producir resultados no intuitivos

1	1	1	1	1	1	1	1	1	1	1	0			
0	1	1	1	1	1	1	1	1	1	1	1			

d = 1.4142

vs

1	0	0	0	0	0	0	0	0	0	0	0			
0	0	0	0	0	0	0	0	0	0	0	1			

d = 1.4142

- Solución: Normalizar los vectores a la longitud unitaria

Clasificación de vecino más cercano…

UNIVERSIDAD NACIONAL DE COLOMBIA

- Son clasificadores perezoso
  - No construyen modelos explícitos
  - Contrario a los clasificadores voraces como los árboles de decisión y sistemas basados en reglas
  - Clasificar registros desconocidos es relativamente costoso

UNIVERSIDAD NACIONAL DE COLOMBIA

Ejemplo: PEBLS

- PEBLS: Parallel Examplar-Based Learning System (Cost & Salzberg)
  - Funciona con características continuas o nominales
    - Para características nominales, se calcula la distancia entre dos valores nominales usando la métrica por diferencia de valor modificado (MVDM)
  - Se asigna a cada registro un factor de peso
  - Número de vecinos cercanos, k = 1

modified value d

Ejemplo: PEBLS

Tid	Reem-bolso	Estado Civil	Ingreso	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

Distancia entre atributos nominales:

d(Soltero,Casado)

= | 2/4 – 0/4 | + | 2/4 – 4/4 | = 1

d(Soltero,Divorciado)

= | 2/4 – 1/2 | + | 2/4 – 1/2 | = 0

d(Casado,Divorciado)

= | 0/4 – 1/2 | + | 4/4 – 1/2 | = 1

d(Reembolso=Sí,Reembolso=No)

= | 0/3 – 3/7 | + | 3/3 – 4/7 | = 6/7

Clase	Estado Civil		
	Soltero	Casado	Divorciado
Sí	2	0	1
No	2	4	1

Clase	Reembolso	
	Sí	No
Sí	0	3
No	3	4

$$d(V_1, V_2) = \sum_i \left| \frac{n_{1i}}{n_1} - \frac{n_{2i}}{n_2} \right|$$

Ejemplo: PEBLS

Tid	Reem-bolso	Estado Civil	Ingreso	Evade
X	Sí	Soltero	125K	**No**
Y	No	Casado	100K	**No**

Distancia entre el registro X y el registro Y:

$$\Delta(X,Y) = w_X w_Y \sum_{i=1}^{d} d(X_i, Y_i)^2$$

donde:

$$w_X = \frac{\text{Número de veces que X es usado para predecir}}{\text{Número de veces que X es predicha correctamente}}$$

$w_X \cong 1$ si X predice de forma exacta la mayoría de las veces

$w_X > 1$ si X no es confiable para hacer predicciones

UNIVERSIDAD NACIONAL DE COLOMBIA

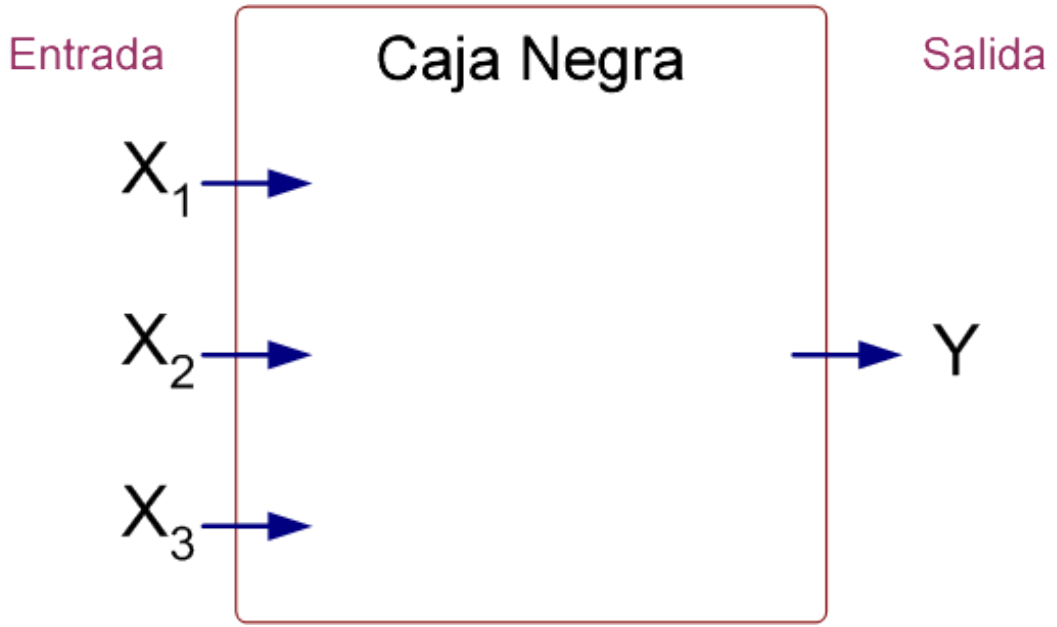
Agenda

- Clasificadores Basados en Ejemplos
- Redes Neuronales Artificiales
- Métodos de Ensamble
- Generalización y Sobreajuste

Redes Neuronales Artificiales (RNA)

UNIVERSIDAD NACIONAL DE COLOMBIA

X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0

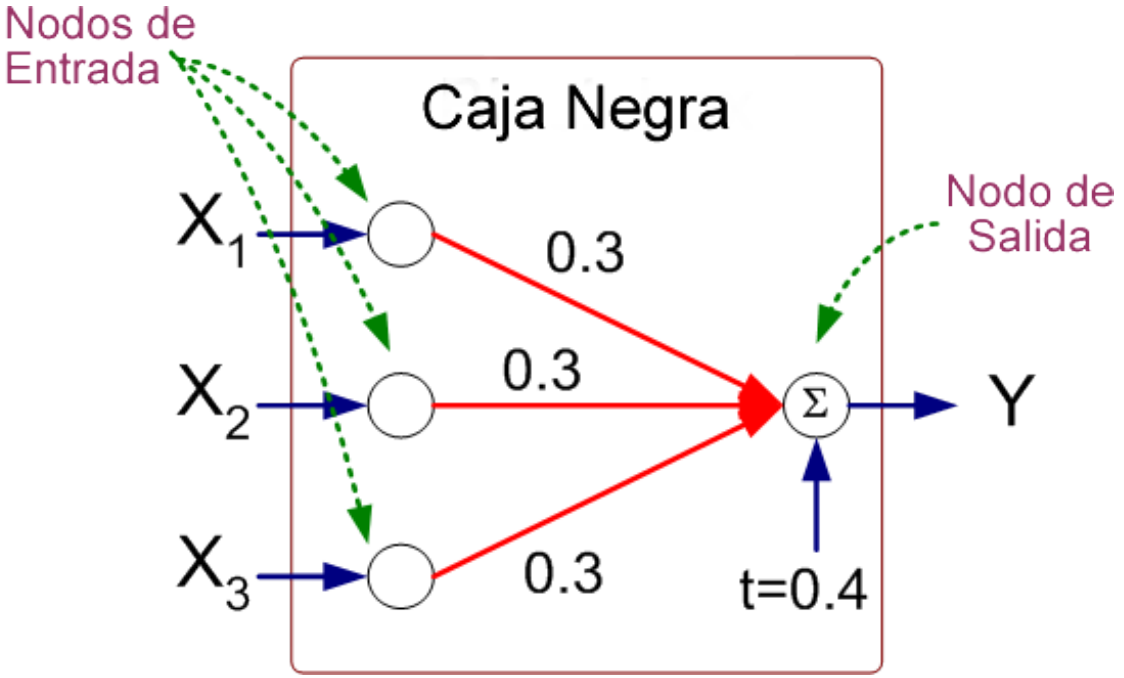


La Salida Y es 1 si por lo menos dos de las tres entradas es igual a 1.

Redes Neuronales Artificiales (RNA)

UNIVERSIDAD NACIONAL DE COLOMBIA

X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0



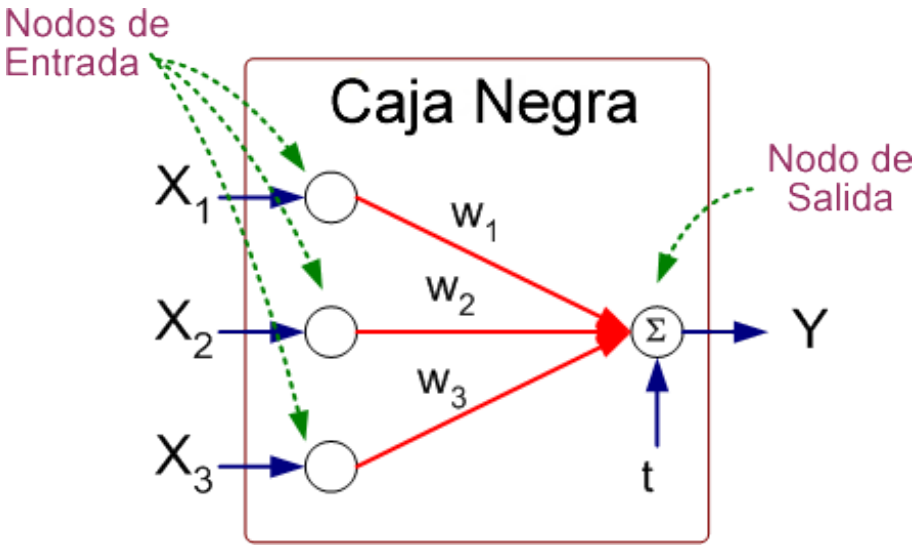
$$Y = I(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4 > 0)$$

$$\text{donde } I(z) = \begin{cases} 1 & \text{si } z \text{ es verdadero} \\ 0 & \text{en otro caso} \end{cases}$$

Redes Neuronales Artificiales (RNA)

UNIVERSIDAD NACIONAL DE COLOMBIA

- Modelo es un conjunto de nodos interconectados y enlaces ponderados
- El nodo de salida suma cada uno de sus valores de entrada de acuerdo a los pesos de sus enlaces
- Comparar el nodo de salida contra un umbral t

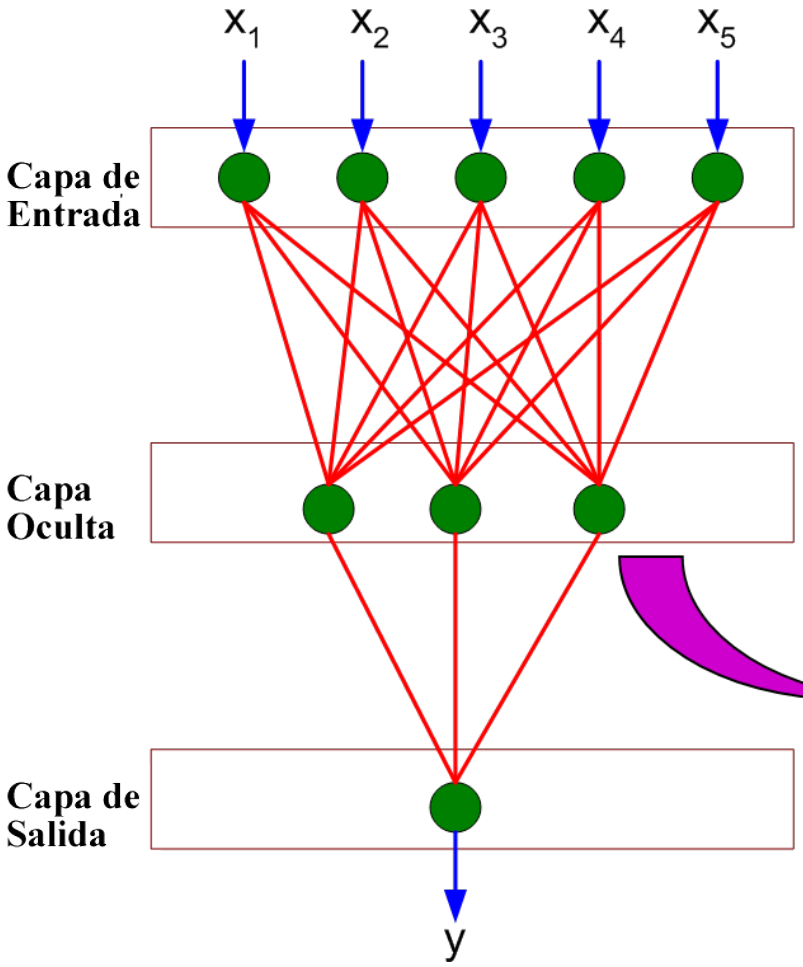


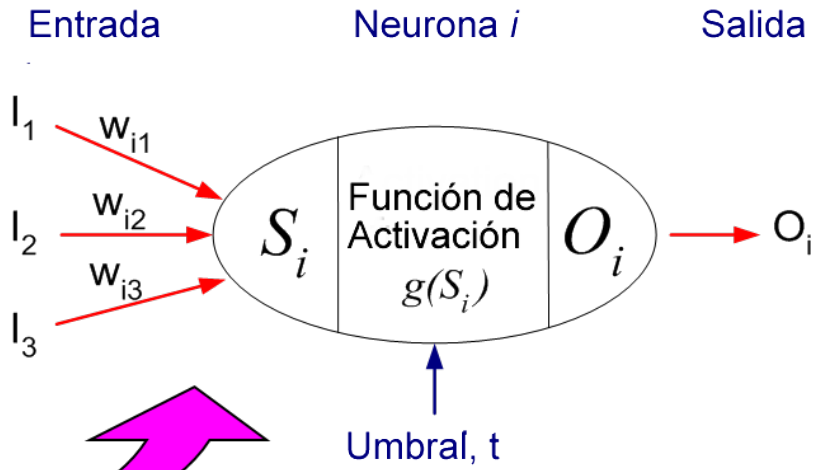
Modelo Perceptrón

$$Y = I\left(\sum_i w_i X_i - t\right)$$

$$Y = sign\left(\sum_i w_i X_i - t\right)$$

Estructura General de una RNA





Entrenar una RNA significa aprender los pesos de las neuronas

Algoritmo para aprender una RNA

UNIVERSIDAD NACIONAL DE COLOMBIA

- Inicializar los pesos (w_0, w_1, …, w_k)
- Ajustar los pesos de forma tal que la salida de la RNA sea consistente con las etiquetas de clase de los ejemplos de entrenamiento
  - Función objetivo: $E = \sum_i \left[Y_i - f(w_i, X_i) \right]^2$
  - Encontrar los pesos w_i's que minimicen la función objetivo
    - e.g. Algoritmo Backpropagation

UNIVERSIDAD NACIONAL DE COLOMBIA

Agenda

- Clasificadores Basados en Ejemplos
- Redes Neuronales Artificiales
- Métodos de Ensamble
- Generalización y Sobreajuste

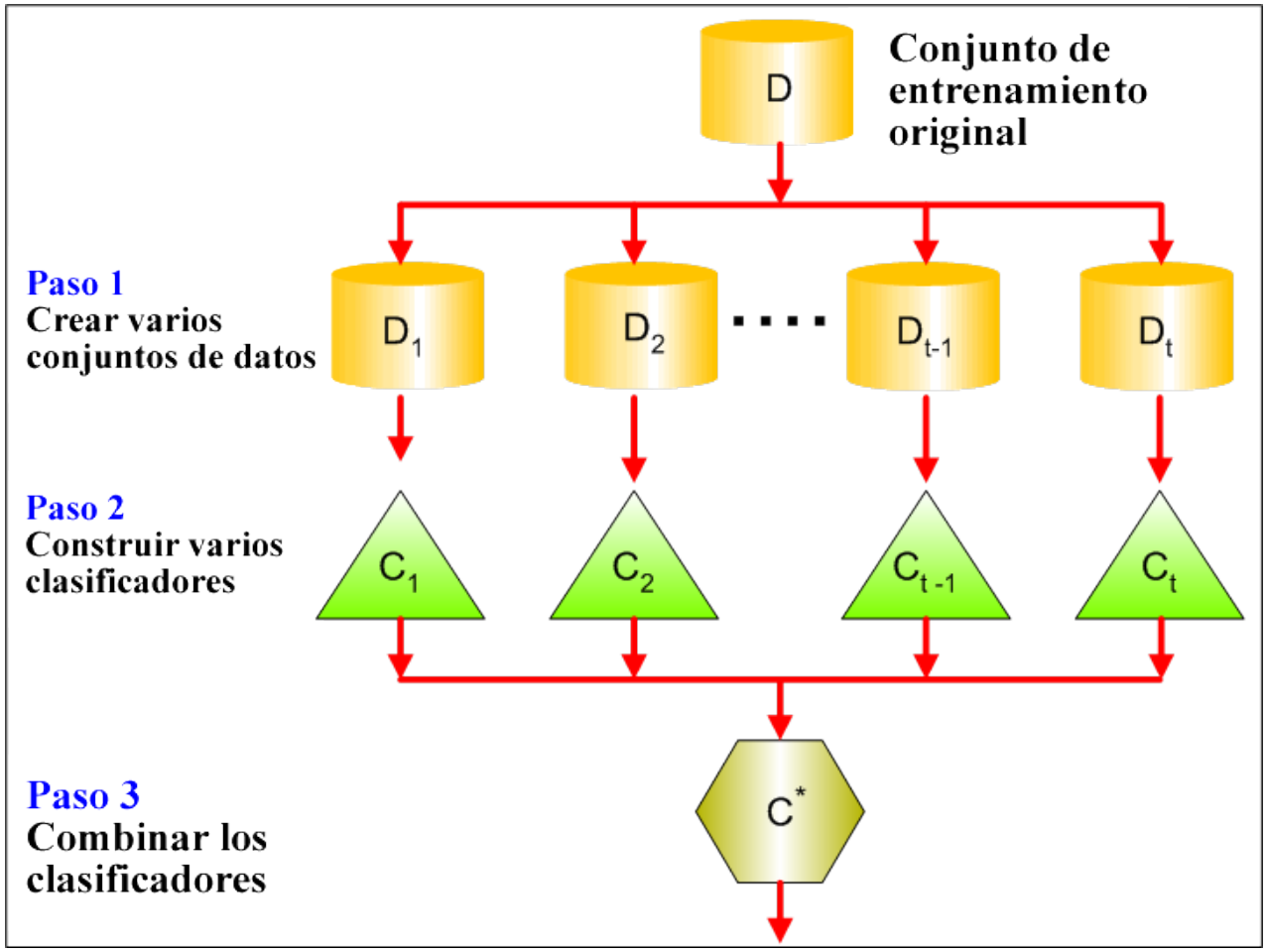
UNIVERSIDAD NACIONAL DE COLOMBIA

Métodos de Ensamble

- Construir un conjunto de clasificadores a partir de los datos de entrenamiento
- Predecir la clase de registros previamente desconocidos al agregar las predicciones hechas por varios clasificadores

UNIVERSIDAD NACIONAL DE COLOMBIA

Idea General



UNIVERSIDAD NACIONAL DE COLOMBIA

¿Por qué funcionan?

- Suponga que hay 25 clasificadores base
  - Cada clasificador tiene una tasa de error, $\varepsilon = 0.35$
  - Asuma que los clasificadores son independientes
  - La probabilidad de que el ensamble haga una predicción equivocada:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

Métodos de Ensamble: Ejemplos

UNIVERSIDAD NACIONAL DE COLOMBIA

- Cómo generar un ensamble de clasificadores
  - Bagging
  - Boosting

UNIVERSIDAD NACIONAL DE COLOMBIA

Agenda

- Clasificadores Basados en Ejemplos
- Redes Neuronales Artificiales
- Métodos de Ensamble
- Generalización y Sobreajuste

UNIVERSIDAD NACIONAL DE COLOMBIA

Generalización y sobre ajuste

- Error de clasificación
- Control del sobre-ajuste
- Complejidad del modelo

UNIVERSIDAD NACIONAL DE COLOMBIA

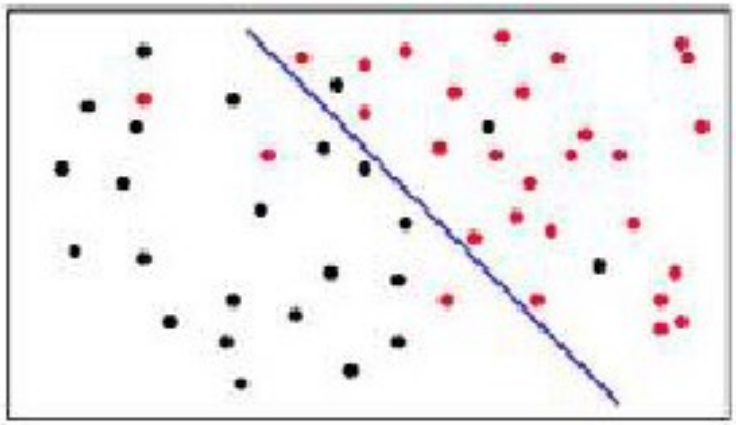
Error de Clasificación

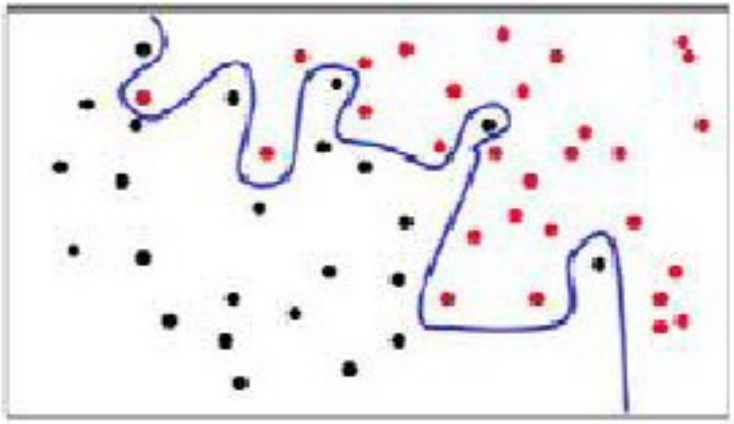
- Error de entrenamiento:
  - **e(modelo, datos)**
  - Número de ejemplos de entrenamiento clasificados incorrectamente
  - Conocido como error de re-substitución o error aparente
- Error de generalización:
  - **e'(modelo, datos)**
  - Error esperado del modelo en ejemplos no usados en el entrenamiento.
- Un buen modelo debe tener errores de entrenamiento y generalización bajos

UNIVERSIDAD NACIONAL DE COLOMBIA

Sobre-ajuste (Overfitting)

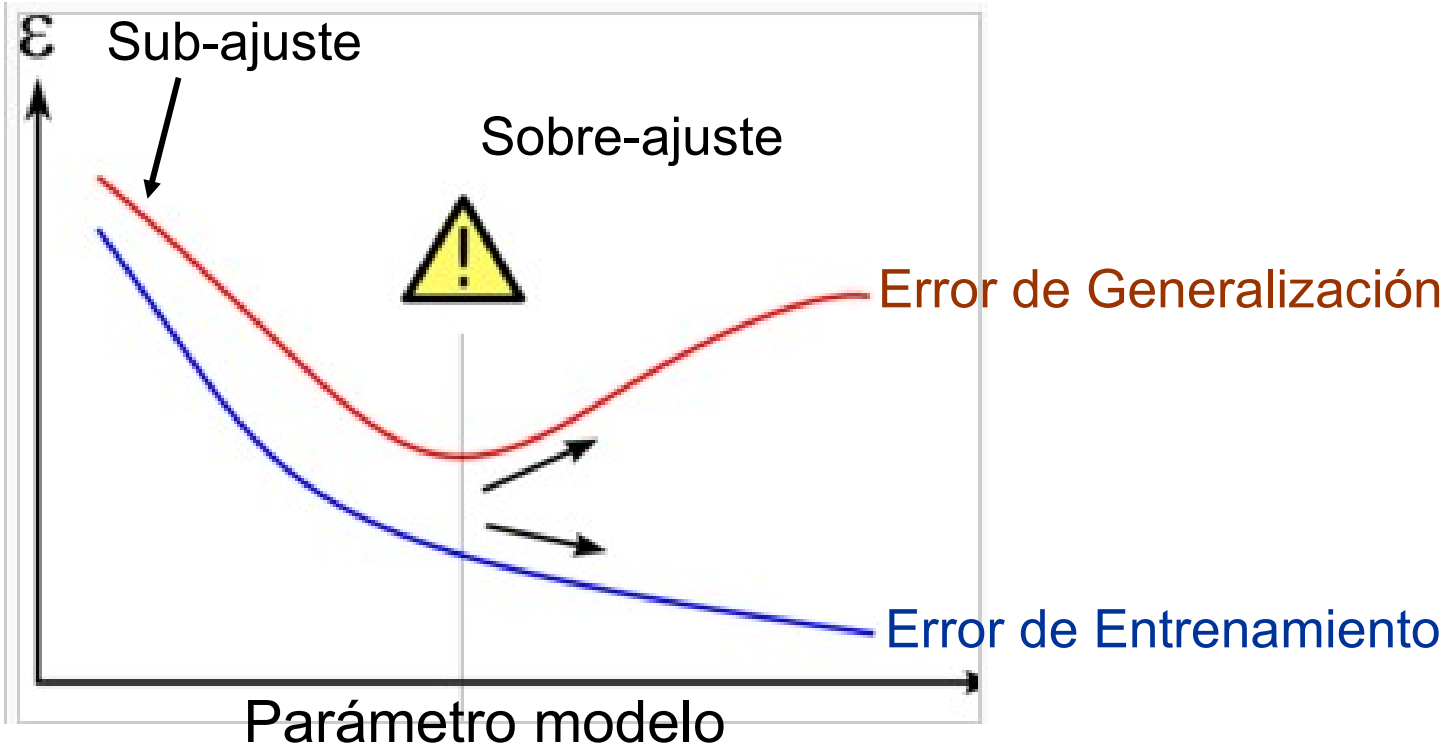
- Cuando el algoritmo de aprendizaje se ajusta tanto a los datos de entrada que pierde su capacidad de generalizar



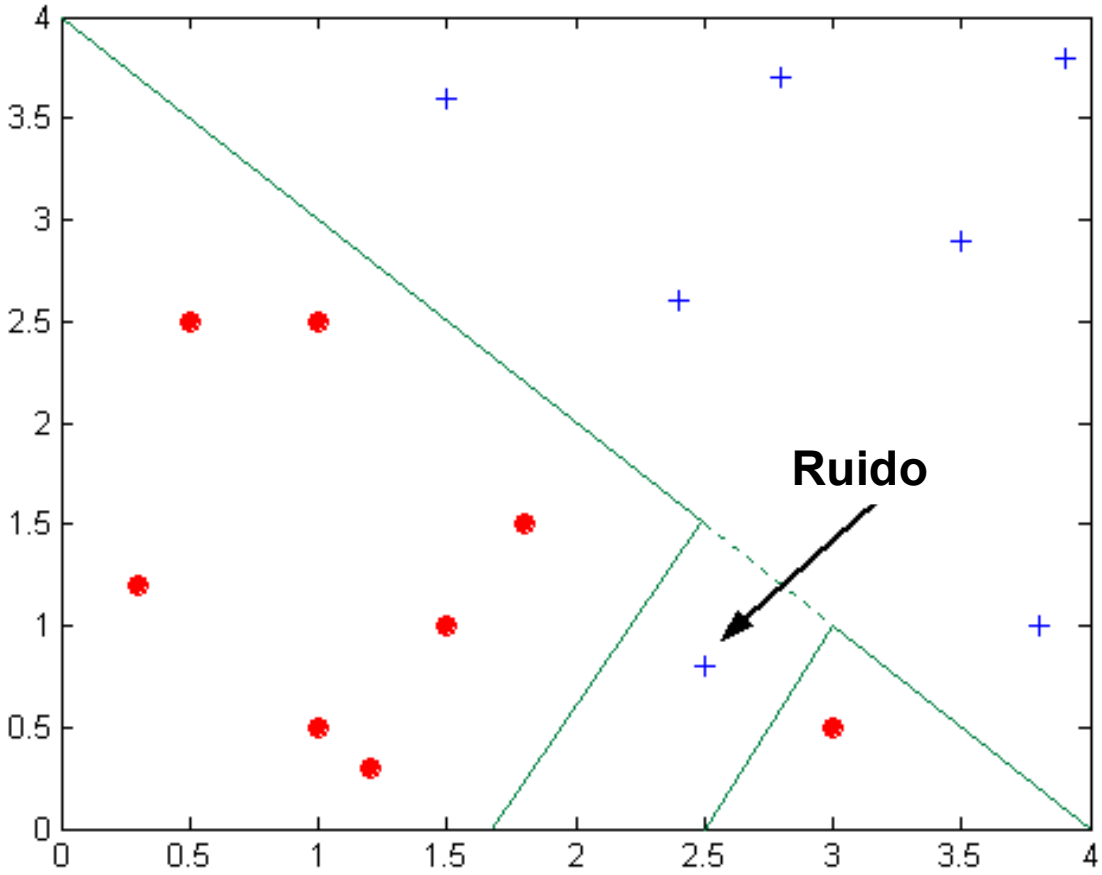


Sobre-ajuste

- **Sobre-ajuste:** Bajo error de entrenamiento pero error de generalización alto.
- **Sub-ajuste (underfitting):** Errores de entrenamiento y generalización altos

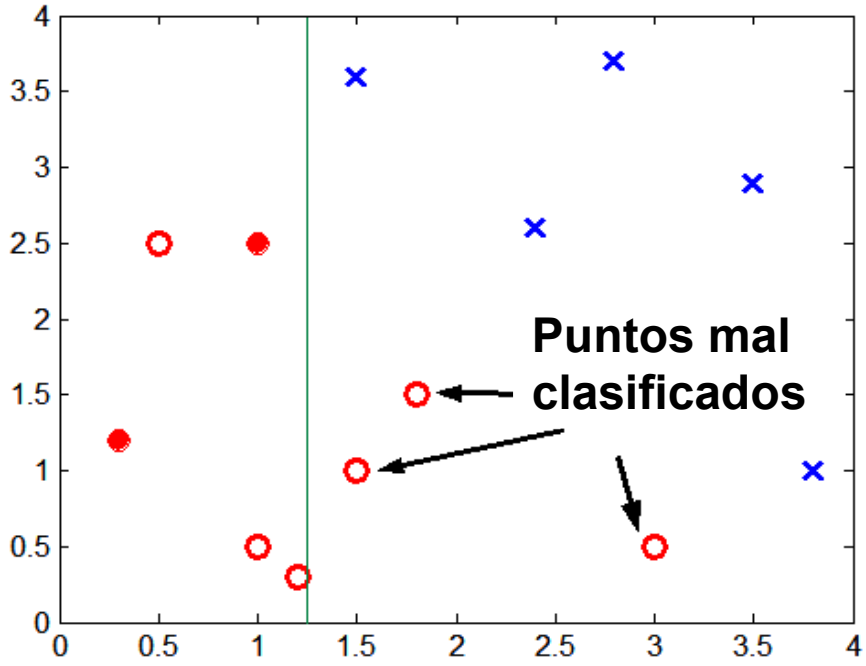


Causas sobre-ajuste: Presencia Ruido



La frontera de decisión es distorsionada por el ruido

Causas sobre-ajuste: 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

Causas sobre-ajuste: Procedimiento múltiple comparaciones

- Al comparar alternativas independientes y seleccionar la mejor
- Ejemplo:
  - Predecir cuando la demanda se incrementara o reducirá en los siguientes 10 días.
  - **Analista aleatorio:** Su probabilidad de acertar en 8 o mas días es

$(c(10,8) + c(10,9) + c(10,10)) / 2^{10} = 0.0547$

  - Seleccionar uno de entre 50 analistas aleatorios, la probabilidad de que uno de ellos acierte en 8 o mas días es:

$$1 - (1-0.0547)^{50} = 0.9399$$

Estimación del error de Generalización

UNIVERSIDAD NACIONAL DE COLOMBIA

- Estimación optimista: Usando re-substitución

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

- Incorporando la complejidad del modelo – Cuchilla de Occam

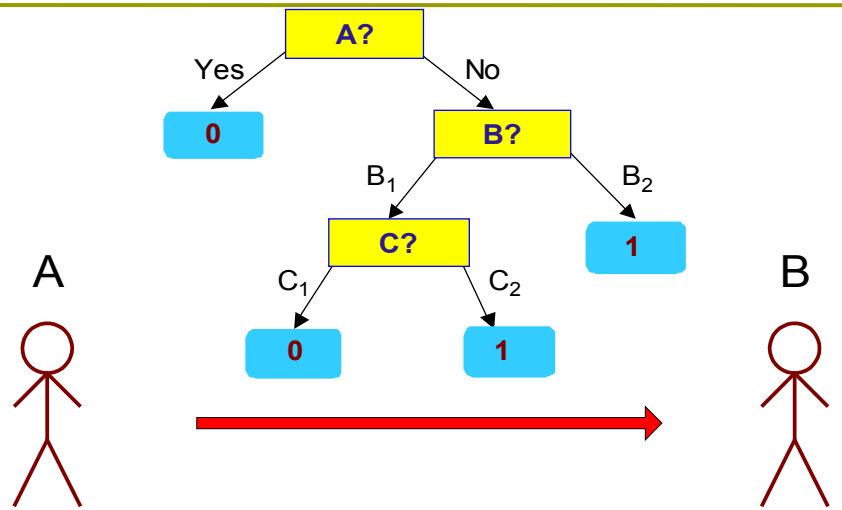
  e'(modelo, datos)=

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

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

Descripción de Mínima Longitud (MDL)

X	y
X_1	1
X_2	0
X_3	0
X_4	1
…	…
X_n	1



X	y
X_1	?
X_2	?
X_3	?
X_4	?
…	…
X_n	?

- **costo(modelo,datos) =**
  costo(datos|model) + costo(modelo)
  - Costo es medido en el numero de bits necesarios para su codificación.
  - El problema es el de buscar el modelo de menor costo.
- **costo(datos|modelo)** para codificar los errores de clasificación.
- **costo(modelo)** para codificar el modelo.

UNIVERSIDAD NACIONAL DE COLOMBIA

Métodos de Estimación

- **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

UNIVERSIDAD NACIONAL DE COLOMBIA

Complejidad del modelo

- 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

UNIVERSIDAD NACIONAL DE COLOMBIA

Bibliografía

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