Aprendizaje de Máquinas: Introducción

Carlos Hurtado L.

Depto de Ciencias de la Computación, Universidad de Chile

Nociones de Aprendizaje

- “Learning denotes changes in a system that ... enable a system to do the same task more efficiently the next time.” – Herbert Simon
- “Learning is constructing or modifying representations of what is being experienced.” –Ryszard Michalski

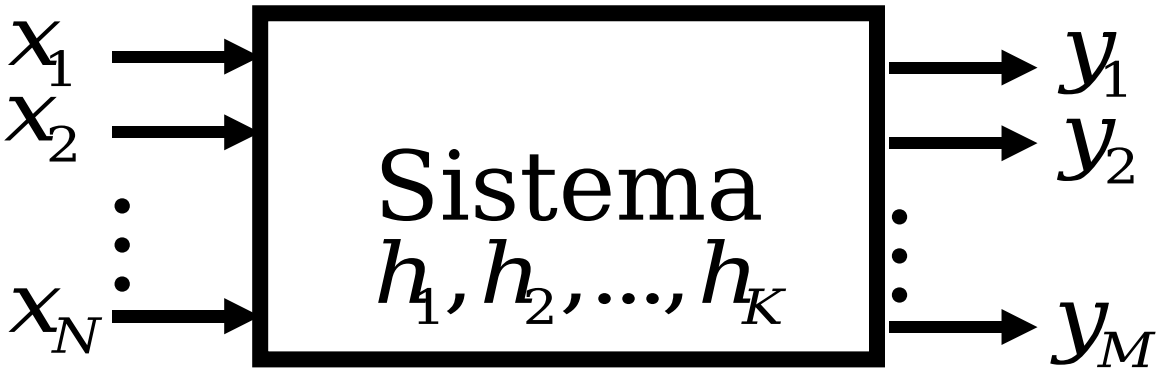
Aprendizaje de Máquina

- “Machine Learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system” – Greg Grudic
- “Samples”: observaciones de un sistema, datos, eventos.
- Las relaciones entre variables son capturadas por un modelo o patrón.
- El problema de aprendizaje que estudiaremos consiste en “inducir” un modelo de las observaciones.

Aprendizaje de Máquinas

- “Machine learning (inductive) creates computer programs by extracting rules and patterns out of massive data sets” - Wikipedia

Sistema Observado



Variables de entrada: $\mathbf{x} = (x_1, x_2, \ldots, x_N)$

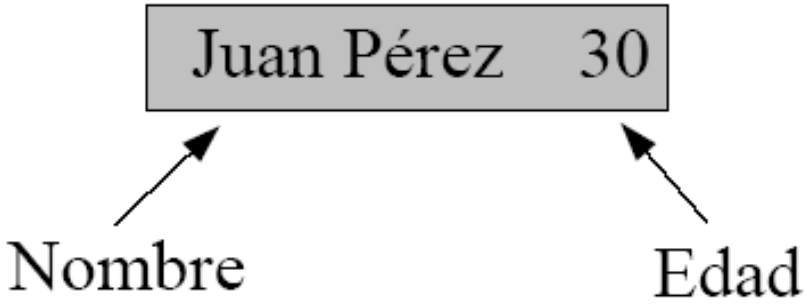
Variables ocultas: $\mathbf{h} = (h_1, h_2, \ldots, h_K)$

Variables de salida: $\mathbf{y} = (y_1, y_2, \ldots, y_K)$

Observaciones: “datos”, eventos del sistem

Datos

- Datos son instancias de un vector de variables.
- Terminología: dato, registro, evento, objeto, punto, etc.



Terminología: variable, atributo, etc.

Ejemplo (Weka): weather.nominal

Outlook	Temp.	Humidity	Windy
Sunny	Hot	High	FALSE
Sunny	Hot	High	TRUE
Overcast	Hot	High	FALSE
Rainy	Mild	High	FALSE
Rainy	Cool	Normal	FALSE
Rainy	Cool	Normal	TRUE
Overcast	Cool	Normal	TRUE
Sunny	Mild	High	FALSE
Sunny	Cool	Normal	FALSE
Rainy	Mild	Normal	FALSE
Sunny	Mild	Normal	TRUE
Overcast	Mild	High	TRUE
Overcast	Hot	Normal	FALSE
Rainy	Mild	High	TRUE

Tipos de Variables

- Numéricas
  - Valores medidos en una escala numérica
  - Reales, enteros, etc.
- Categóricas
  - Ordinales: valores poseen un orden
  - Nominales: valores son sólo nombres

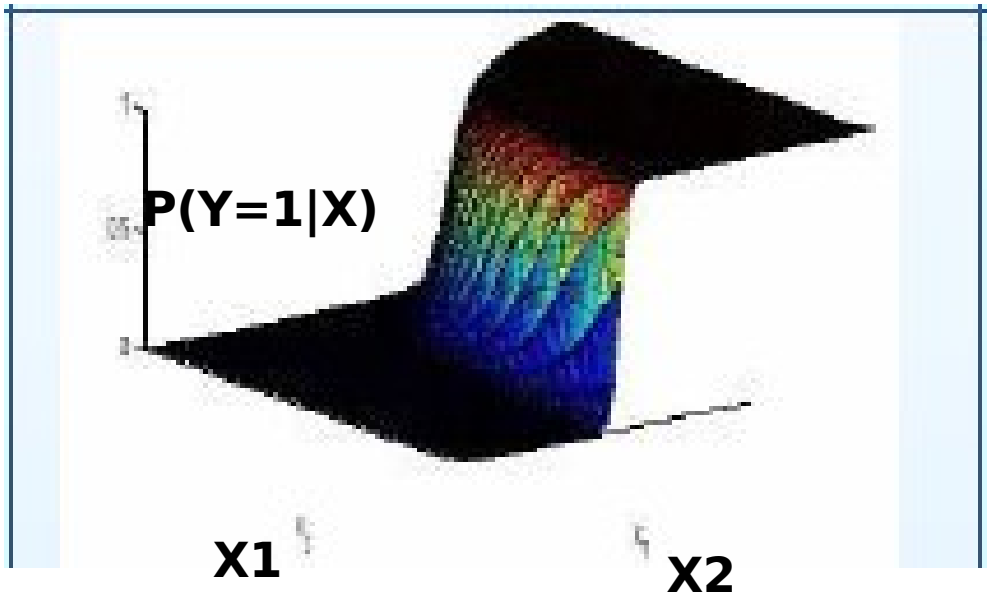
Aprendizaje supervisado

- Inducir una función f(X)=Y a partir de datos X,Y.
- La función debe “ajustarse” a los datos de manera de minimizar alguna noción de error.
- Variantes:
  - **Clasificación**: f es una funcion discreta.
  - **Regresión**: f es una función numérica arbitraria.

Sistema observado

- Los datos son observaciones de un sistema de variables X e Y
- La variable Y no es necesariamente función de X
- X,Y ~ P(X,Y)
- X,Y ~ P(X), P(Y|X)

Ejemplo: Y es variable binaria



Aprendizaje no Supervisado

- Inducir f(X)=Y a partir de datos X.
- Variantes
  - **Segmentación**: f modela una segmentación de los datos
  - I**nferencia Estadística**: f modela una distribución de probabilidades
  - **Reglas de Asociación**: f modela asociaciones o correlaciones.

Clasificación

- Cuando f(X) es discreta se denomina un modelo de clasificación.
- El problema se puede explicar simplemente:
  - Dado un conjunto de datos, donde cada dato pertenece a una clase.
  - Construir un modelo que permita predecir la clase de un nuevo dato.
- En este curso nos centraremos en clasificación.

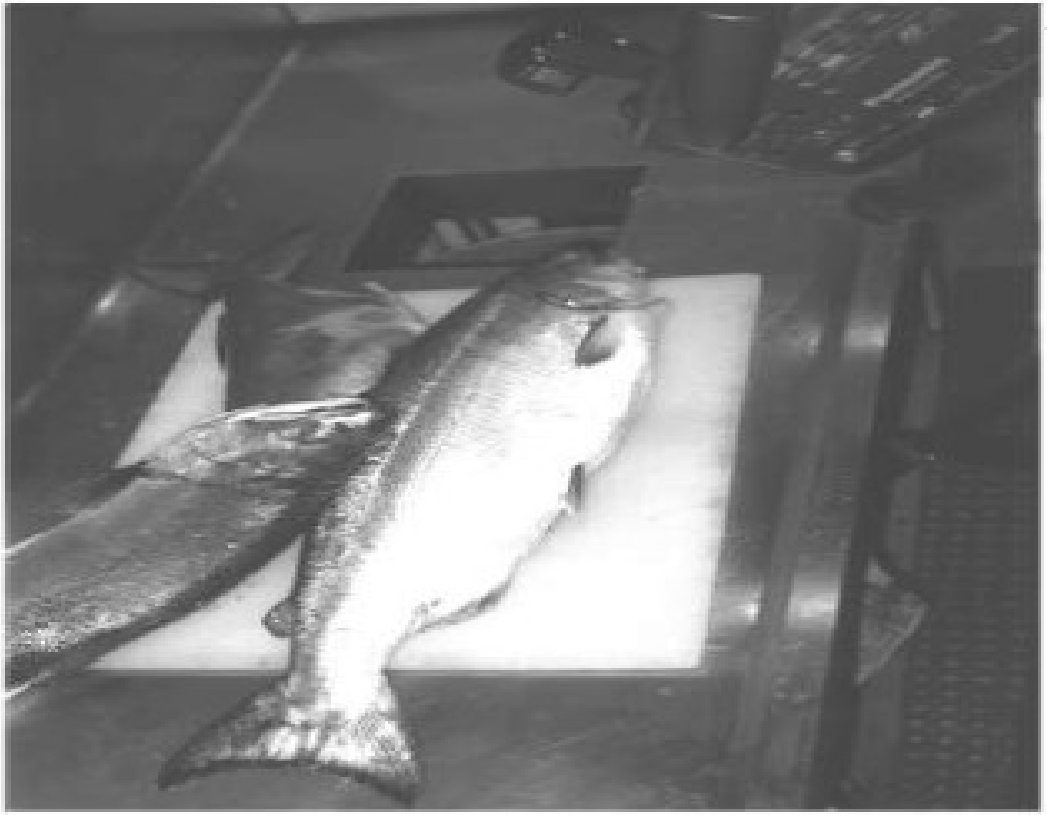
Etapas en un proceso de Clasificación

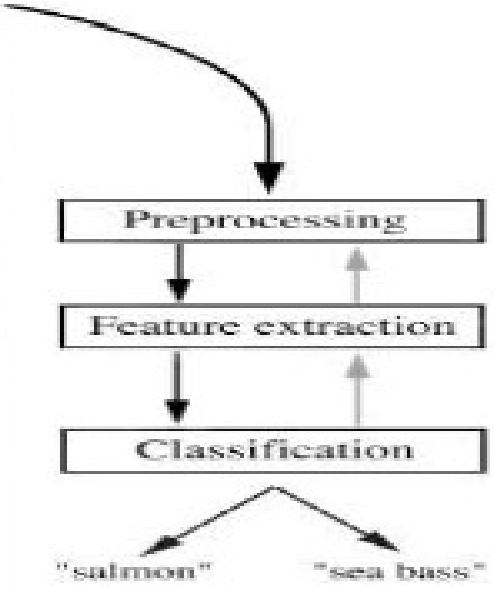
- Aprendizaje
  - Construcción del modelo
  - Se usan datos de entrenamiento
- Prueba
  - Evaluación del modelo
  - Se usan datos de prueba
- Uso
  - Aplicación del modelo para predecir
  - Datos de uso

Clasificación: Tipos de Modelos

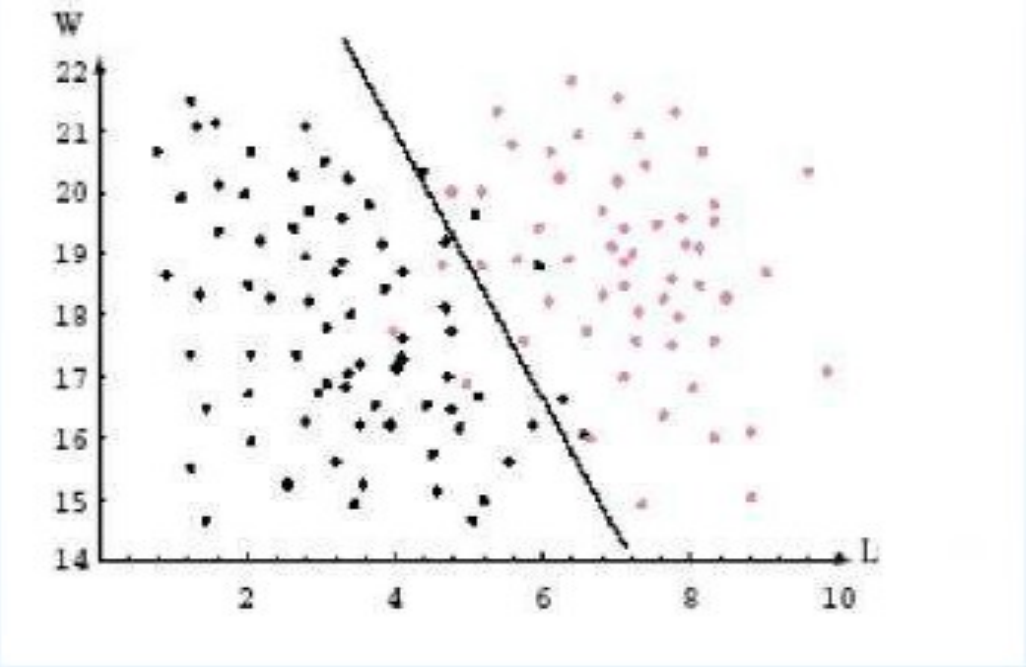
- Enfoque Discriminante
  - Arboles de Decisión
  - Reglas de Decisión
  - Discriminantes lineales
- Enfoque Generativo
  - Clasificador Bayesiano Naive
  - Redes Bayesianas
- Enfoque de Regresión
  - Redes Neuronales
  - Regresión Logística

Clasificación: Ejemplo 1





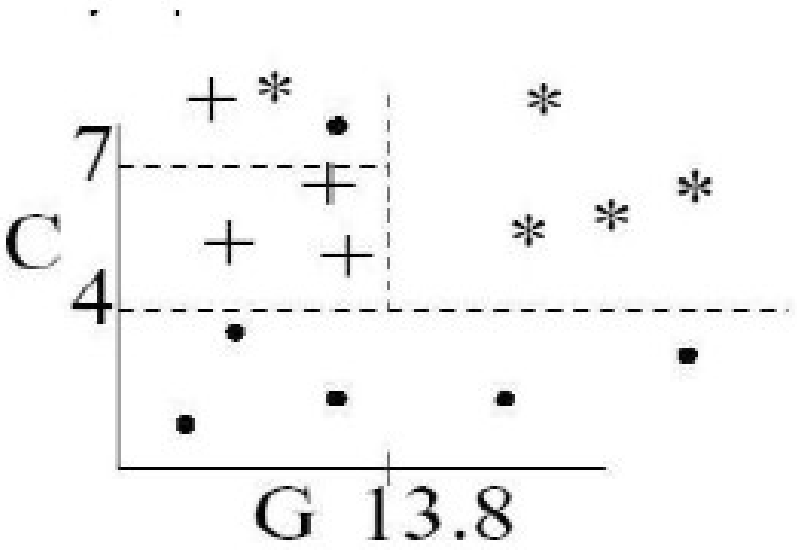
Clasificación: Ejemplo 1 (cont.)

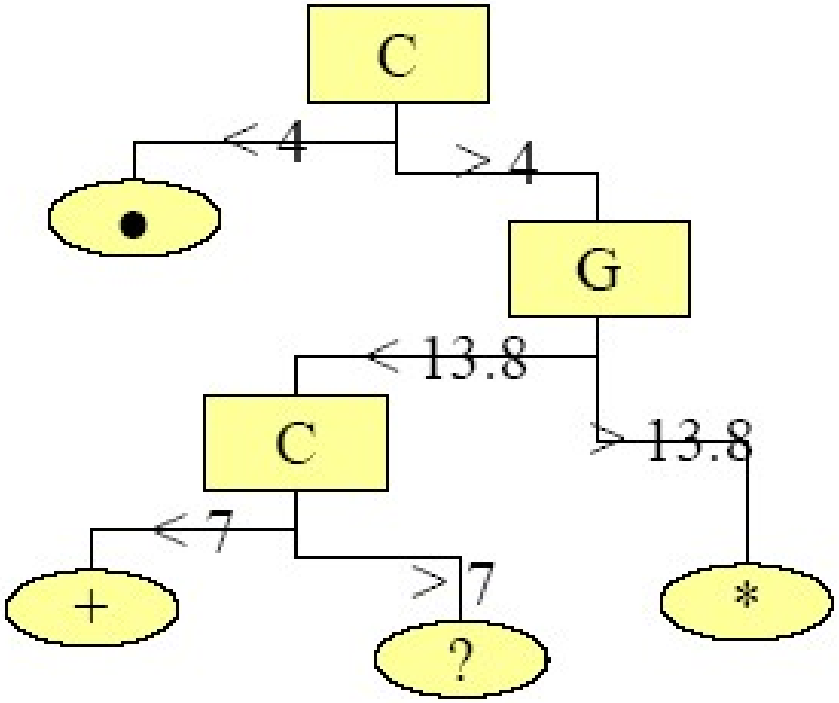


En este caso el modelo es la función binaria:

$$f(X) = \begin{cases} 1 & \text{si } W + 4{,}5\ L - 39 < 0 \\ 0 & \text{si no.} \end{cases}$$

Clasificación: Ejemplo 2





En este caso el modelo es un árbol de decisión

Aplicaciones de Clasificación (I)

- Detección de Fraude:
  - Objetivo: predecir uso fraudulento en tarjetas de crédito
  - Método:
    - Usar transacciones de compras en un determinado período e info. en cuentas.
    - Definir atributos de entrada como: cuándo se compra, frecuencia de compra, frecuencia de pagos, etc.
    - Inducir un modelo para predecir clase de nuevas transacciones de compra.

Aplicaciones de Clasificación (II)

- Predicción de Deserciones (“churn”):
  - Objetivo: en un determinado mes, predecir qué clientes abandonarán un determinado servicio (ej., teléfono celular)
  - Método:
    - Definir atributos de entrada a partir de transacciones del cliente (ej., llamados telefónicos, frecuencia, último llamado, interrupciones, etc.) e info. descriptiva de usuarios.
    - Usar info. de meses anteriores y definir atributo de clase como abandono o no abandono del servicio.
    - Inducir un clasificador.

Aplicaciones de Clasificación (III)

- Clasificación de Documentos:
  - Objetivo: Predecir tópico de un nuevo documento
  - Método:
    - Definir atributos de entrada a partir del contenido del documento (e.g., vector de términos)
    - Usar documentos pre-clasificados como datos de entrenamiento.
    - Inducir un clasificador.

Aplicación III: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.
- **Sentimientos:** connotacion positiva, negativa, neutral, emotiva, ofensiva, etc.
- **Spam:** entre 3000 - 7000 splogs (blogs falsos o spam) se crean diariamente (fuente: Technorati).
- **Contenido no apto:** pornografía, violencia, etc.
- **Comunitario:** contenido de interés para una comunidad de usuarios.
- **Personalizado:** contenido de interés para un usuario único.
- **Categorías del lenguaje** opiniones vs. hechos.

Referencias para el Curso

- Machine Learning. Tom M. Mitchell. 1997. McGraw-Hill Companies, Inc.
- Data Mining. Practical Machine Learning Tools and Techniques. Ian Witten, Eibe Frank. Elsevier, 2005.
- Weka: colección de algoritmos de aprendizaje de maquina (GNU/GPL).
  - http://www.cs.waikato.ac.nz/ml/weka/