Centro de
INNOVACIÓN y DESARROLLO
Tecnológico en Cómputo

Reconocimiento de Patrones

Lección 2: Introducción (parte 2)

Introducción

Tipos de Aprendizaje

Aprendizaje supervisado

En el aprendizaje supervisado se proporciona un conjunto de datos, usado para entrenar al sistema RP, el cual consta de patrones acompañados de sus clases (valor objetivo).

Es un enfoque dirigido por el concepto.

Si el objetivo del sistema RP es asignar una categoría entonces estamos ante una tarea de clasificación. Si, por otro lado, la salida del sistema consiste un 1 o más valores continuos entonces estamos ante una tarea de regresión.

Aprendizaje no Supervisado

- En aprendizaje no supervisado o *clustering,* no existe un experto. El sistema RP forma agrupaciones naturales basándose en los patrones de entrenada.

- Es un enfoque dirigido por los datos, cuyo objetivo es descubrir grupos de datos similares.

- Dado un determinado conjunto de datos o una función de costo, diferentes sistemas de *clustering* pueden conducirnos a diferentes agrupaciones de los datos.

- Generalmente el usuario debe suponer el número de *cluster* de antemano. ¿Cómo evitar una representación inadecuada de los datos?

Aprendizaje por Reforzamiento

- En aprendizaje por reforzamiento el objetivo es encontrar acciones adecuadas en una situación dada de forma que se maximice una recompensa.

- En este enfoque no se da un valor objetivo, sino que se debe encontrar la solución mediante un proceso de prueba y error.

- Alcanzar un balance entre explotar las acciones conocidas que producen una recompensa y explorar nuevas acciones.

- Ejemplo: un robot debe decir entre entrar a un cuarto a recolectar más basura o buscar un estación para recargar su batería.

Notación Básica

Símbolo	Significado	Consideraciones
$\boldsymbol{A}$	Conjunto genérico de donde se toman valores para las entradas de los patrones (vectores)	Ejemplos: $A = \{0,1\}$
$\boldsymbol{X}$	Banco de datos (conjunto de patrones)	
$\boldsymbol{N}$	Cardinalidad del conjunto de patrones	$N = \|X\|,\ N \in \mathbb{Z}^+$
$\boldsymbol{x^i}$	i-ésimo patrón del conjunto de patrones	$x^i \in X,\ i = \{1,2,\dots,N\}$
$\boldsymbol{n}$	Dimensión de los patrones, es decir número de rasgos o atributos	$n \in \mathbb{Z}^+$
$\boldsymbol{x_j^i}$	La j-ésima componente del patrón i	$i = \{1,2,\dots,N\}$ $j = \{1,2,\dots,n\}$
$\boldsymbol{c}$	Número de clases	$c \in \mathbb{Z}^+$
$\boldsymbol{\omega_k}$	k-ésima clase	$k = \{1,2,\dots,c\}$

Notación Básica

Símbolo	Significado	Consideraciones
$\boldsymbol{E}$	Conjunto de entrenamiento.	$E \cup P = X$ $E \cap P = \emptyset$
$\boldsymbol{P}$	Conjunto de prueba.	
$\boldsymbol{N_E}$	Cardinalidad del conjunto de entrenamiento	
$\boldsymbol{N_P}$	Cardinalidad del conjunto de prueba	
$\boldsymbol{N_E(\omega_k)}$	Número de patrones de la clase k en el conjunto de entrenamiento	
$\boldsymbol{N_P(\omega_k)}$	Número de patrones de la clase k en el conjunto de prueba	
$\tilde{\boldsymbol{x}}$	Patrón cuya clase se desconoce.	No es necesario agregar un índice, salvo que se requiera un ordenamiento.

Tipos de atributos

Dependiendo del tipo de valor que un atributo toma, es posible clasificarlo en 2 grandes grupos [7]:

- Numéricos
- Categóricos

Tipos de atributos (numéricos)

Un atributo numérico es aquel que toma valores que se encuentran en el dominio de los números enteros o reales. Ejemplos: edad (ℕ), largo del pétalo (ℝ+), temperatura (ℝ), por mencionar algunas [7].

Discretos: toman valores finitos o contables.

Continuos: aquellos que toman valores reales.

Binarios: tipo especial de atributo discreto que solamente toma valores de 0 y 1.

Tipos de atributos (categóricos)

Un atributo categórico es aquel cuyos valores son tomados de un conjunto de símbolos o cadenas. Ejemplos: sexo (F,M), estado civil (Soltero, Casado, Viudo) [7].

Nominal: son datos que no tienen un orden, por tanto, solo operaciones de comparación (igualdad) tienen sentido. Un ejemplo de esto es el estado civil.

Ordinal: son datos que si representan un orden, por tanto, aquí si es lógico el uso de comparaciones de igualdad o desigualdad (mayor o menor). Un ejemplo de esto seria nivel de educación (primaria, secundaria, posgrado).

Tipos de atributos (representación)

Si se considera un banco de datos con ***n*** atributos numéricos cada punto puede representarse como una tupla:

$$x^i = (x_1^i, x_2^i, \dots, x_n^i) \in \mathbb{R}^n$$

O pueden representarse como un vector columna:

$$x^i = \begin{pmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_n^i \end{pmatrix} = (x_1^i, x_2^i, \dots, x_n^i)^T \in \mathbb{R}^n$$

Tipos de atributos (representación)

Un atributo también puede ser representado como una cadena. Por ejemplo, una oración en un texto o una secuencia de ADN.

Human insulin gene: beta allele 5'flank (ullrich)

Sequence ID: J00267
Length: 889

```
Query  1 ~ 60
GAGGATGCCTGGGGGGCCTGGACGGAGCTGGGCCAGTGCACAGCTTCCCACACCTGCCCA
||||  |||||||||||||  ||||||||||||||||||||||||||||||||||||||||
GAGGGTGCCTGGGGGGCCAGGACGGAGCTGGGCCAGTGCACAGCTTCCCACACCTGCCCA
Sbjct  97 ~ 156

Query  61 ~ 120
CCCCCGGAGTCCTGCCGCCACCCCCAGATCACACGAAAGATGAGGTCCTAGTGGCCTGCT
|||||  |||||||||||||||||||||||||||||  ||||||||||||  ||||||||||||
```

Clases

El concepto de una clase puede verse como una categoría que se le asigna a un determinado objecto. Es decir que dicho objeto pertenece a esa clase.

Los problemas de clasificación pueden ser:

- Bi-clases (Ejemplo: problemas médicos, paciente sano y paciente enfermo).
- Multi-clase (Ejemplo: detección de intrusiones en una red, diferentes tipos de ataque).

Clases

Existen 2 tipos de problemas multi-clases:

- Clases no exclusivas: un objeto puede tener múltiples clases/categorías asignadas (multi-etiqueta).

Etiquetas

- Paisaje
- Lago
- Montañas
- Arboles

Clases

Existen 2 tipos de problemas multi-clases:

- Clase exclusivas: un objeto solo puede tener asignada una clase/categoría.

Clase: imagen color

Clase: imagen blanco/negro

No puede ser ambas al mismo tiempo.

Valores perdidos

Este problema se presenta cuando el valor para al menos un rasgo de un patrón en el conjunto de datos no se encuentra presente. Generalmente se representan con:

- Valor fuera de rango: -1 en un campo numérico que solo es positivo.
- Caracteres especiales: -, ? o espacio en blanco son comunes para campos nominales.

Causas: mal funcionamiento de equipos de medición, cambio en el diseño durante la captura de datos, imposibilidad de colectar los datos, entre otras.

Valores perdidos ¿Cómo lidiar con ellos?

- Eliminar los patrones cuyos atributos contienen valores perdidos.
- Adaptar el algoritmo para realizar alguna acción cuando se encuentre con valores perdidos.
- Si el atributo que tiene valores perdidos es numérico se rellenan los valores con el valor promedio o la media del atributo.
- Si el atributo que tiene valores perdidos es categórico se rellenan los valores con la moda del atributo.

Patrones atípicos (*outliers*)

Los patrones atípicos son datos que presentan una diferencia significativa del resto de elementos en un conjunto de datos o en una clase en particular.

Causas:

- Malas mediciones al capturar los datos.
- Mal etiquetado del patrón al asignarle una clase.
- Características propias del concepto que se aprende.

Patrones atípicos ¿Cómo lidiar con ellos?

- Eliminar patrones atípicos que presenten valores evidentemente imposibles, un ejemplo sería edades negativas.
- Tratar de normalizar los datos.
- Calcular la desviación estándar y seleccionar datos que se alejen un determinado número de desviaciones estándar de la media del atributo que se analiza para ser eliminados del conjunto de datos.
- Usar algoritmos que se vean menos afectados por valores atípicos, por ejemplo los *random forest*.
- Métodos de proximidad como *clustering*, densidad o vecinos mas cercanos.

Patrones atípicos ¿Cómo lidiar con ellos?

Edición de Wilson

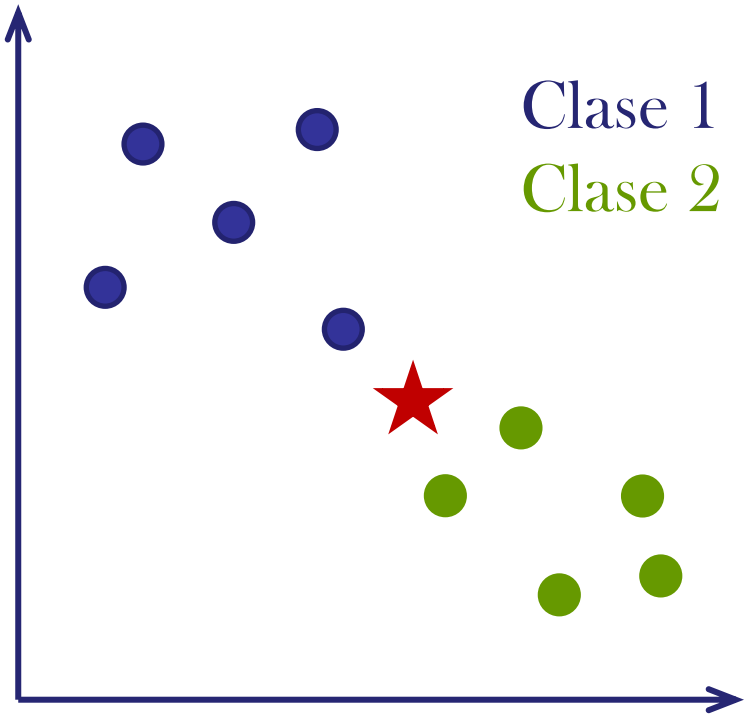
El objetivo general es identificar y eliminar patrones atípicos de un conjunto de datos.

Utiliza la regla del clasificador KNN (generalmente con k=3).

Los patrones cuyas clases no correspondan a la clase mayoritaria de sus *k* vecinos, serán eliminados del conjunto de datos.

Patrones atípicos ¿Cómo lidiar con ellos?

Edición de Wilson



Si la clase del dato que se analiza (★) es 1, el dato será removido del conjunto, en caso contrario se mantendrá en el conjunto de datos.

Patrones atípicos ¿Cómo lidiar con ellos?

Normalización

Ajustar los valores medidos en diferentes escalas con respecto a una escala común. Uno de los métodos mas sencillos es dividir cada atributo por el máximo valor que toma.

$$x^1 = (2,120) \longrightarrow x^{1'} = (0.11,0.11)$$
$$x^2 = (8,533) \longrightarrow x^{2'} = (0.44,0.48)$$
$$x^3 = (1,987) \longrightarrow x^{3'} = (0.06,0.88)$$
$$x^4 = (15,1121) \longrightarrow x^{4'} = (0.83,1)$$
$$x^5 = (18,1023) \longrightarrow x^{5'} = (1,0.91)$$

Al dividir cada rasgo por el valor mayor que toman, 18 y 1121 respectivamente, todos los valores estarán en un rango entre 0 y 1.

Patrones atípicos ¿Cómo lidiar con ellos?

Normalización

Existen otras formas de normalizar los datos, algunos ejemplos son:

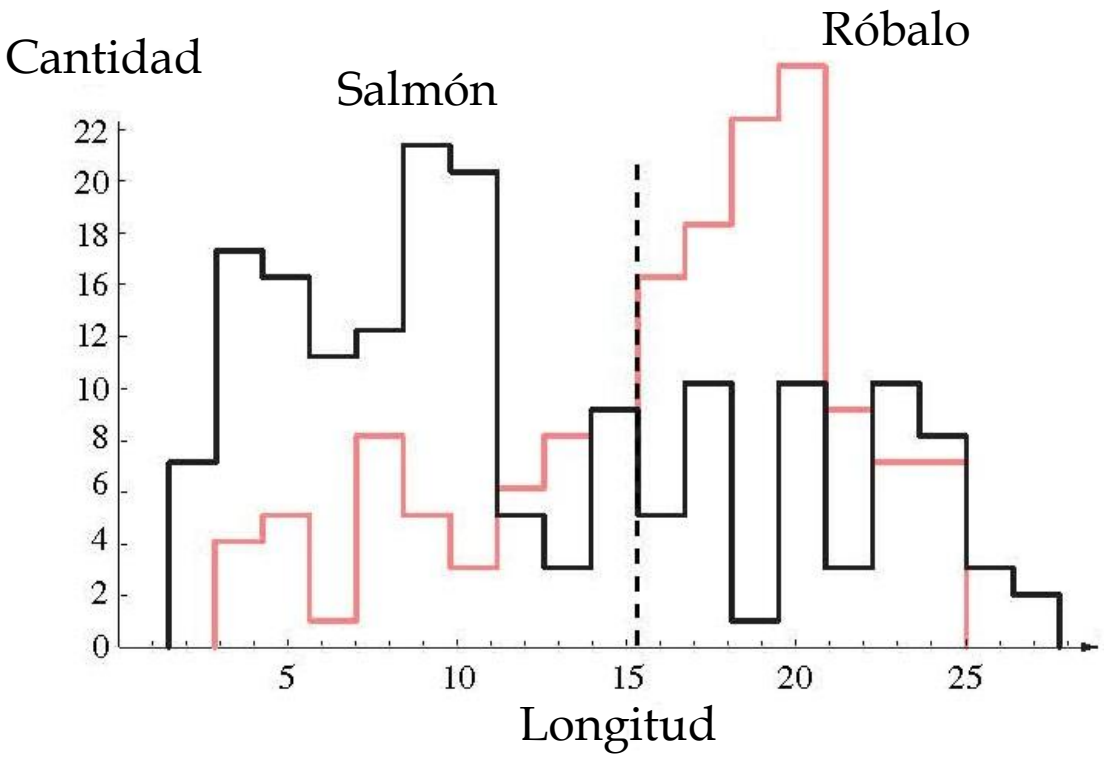
Puntuación estándar $$x' = \frac{x - \mu}{\sigma}$$

T de student $$x' = \frac{x - \bar{x}}{s}$$

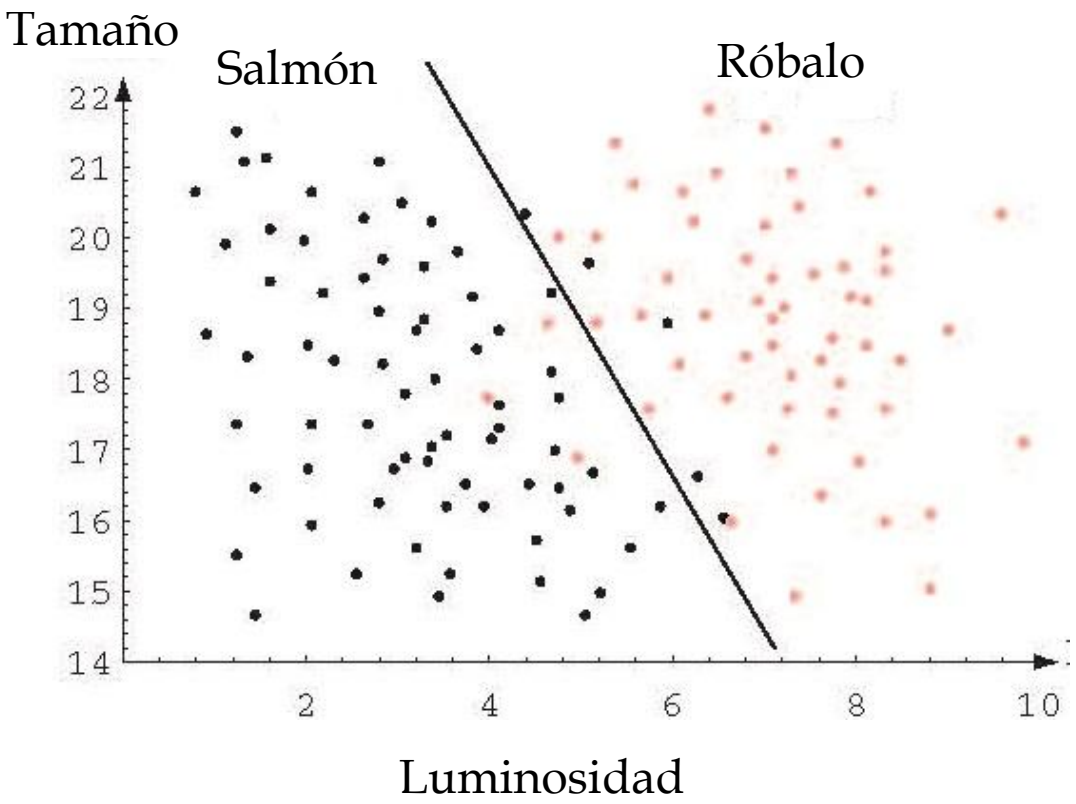
Normalización basada en la unidad $$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Solapamiento de clases

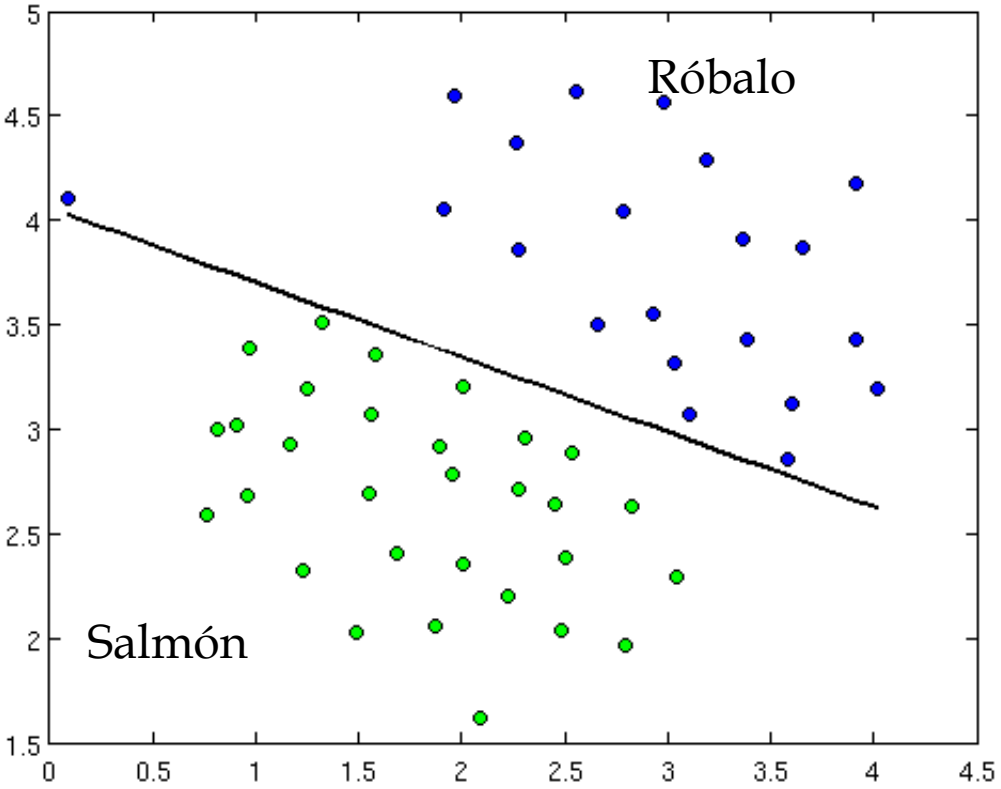
Este fenómeno se presenta cuando elementos que pertenecen a diferentes clases comparten características similares.



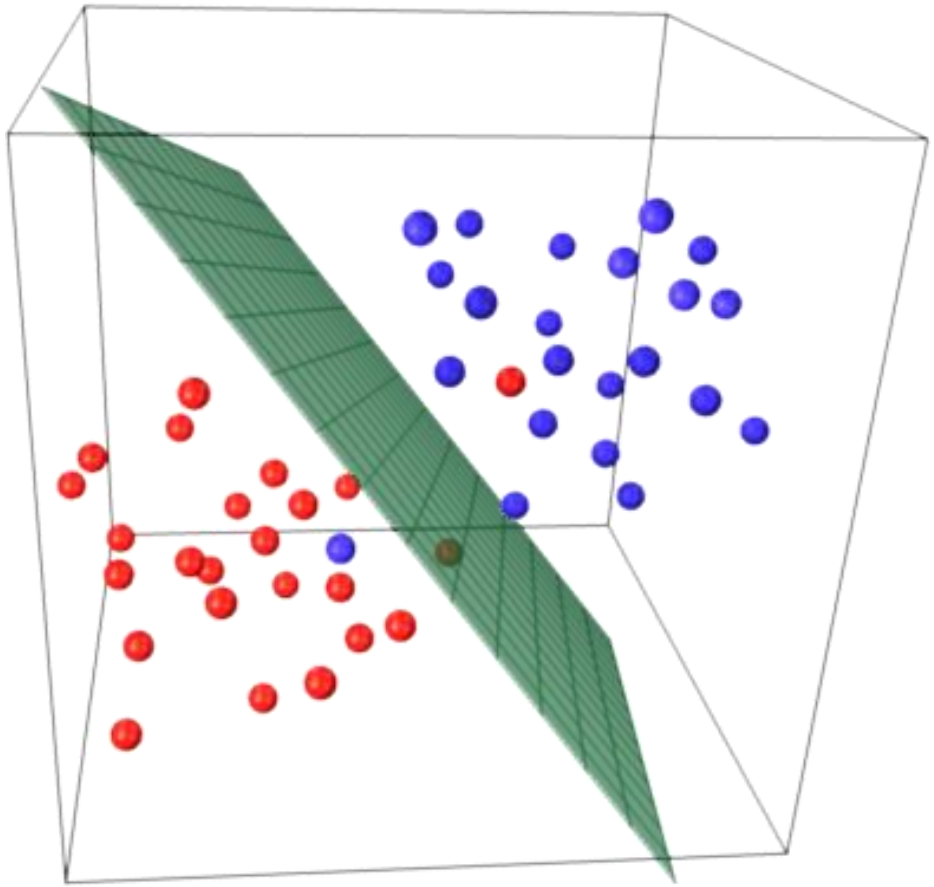
Separabilidad lineal



Separabilidad lineal



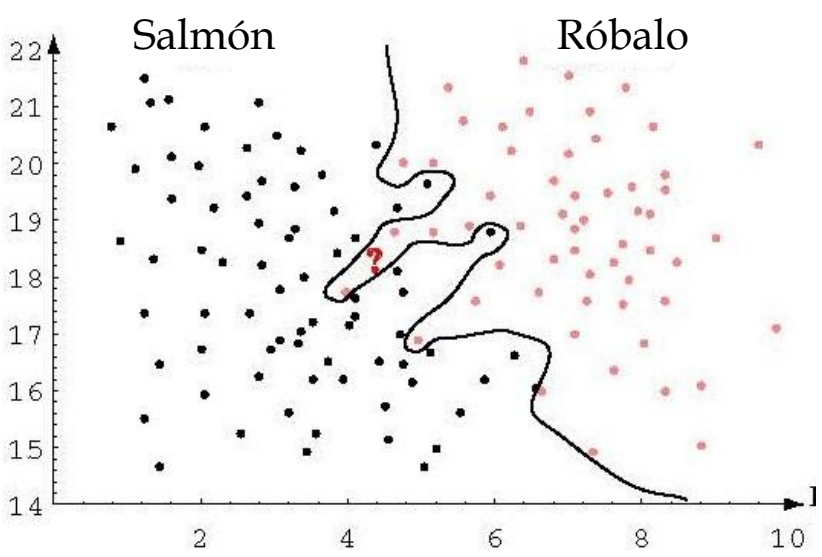
Separabilidad lineal



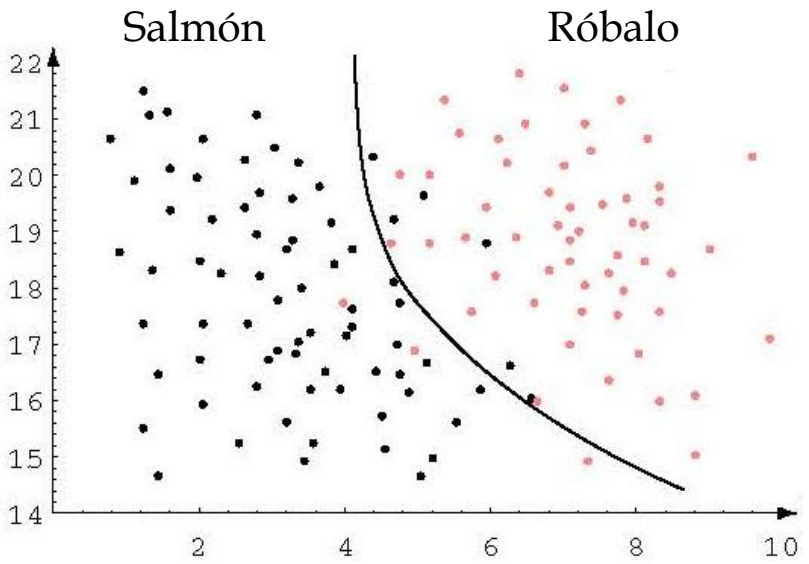
En caso de que de usar 3 rasgos, los patrones quedarían separados por un plano.

Al usar un espacio *n*-dimensional, se dice que los patrones deben quedar separados por un hiperplano de *n*-1 dimensiones.

Generalización



Overfitting



Generalización

Desbalance de Clases

Este fenómeno se presenta cuando el número total de instancias en una clase (clase mayoritaria) es significativamente mayor que el número de instancias de otra clase (clase minoritaria).

$$IR = \frac{\#\ instancias\ clase\ mayoritaria}{\#\ instancias\ en\ la\ clase\ minoritaria}$$

Un banco de datos se considera desbalanceado, si $IR > 1.5$; de lo contrario se dice que el banco de datos está balanceado.

Este es un problema que se presenta a menudo en situaciones comunes de la vida real.

Desbalance de Clases

Ejemplos típicos donde la clase minoritaria es la clase de interés en los problemas de clasificación, son evidentes en BD relacionados con problemas de salud y con problemas financieros.

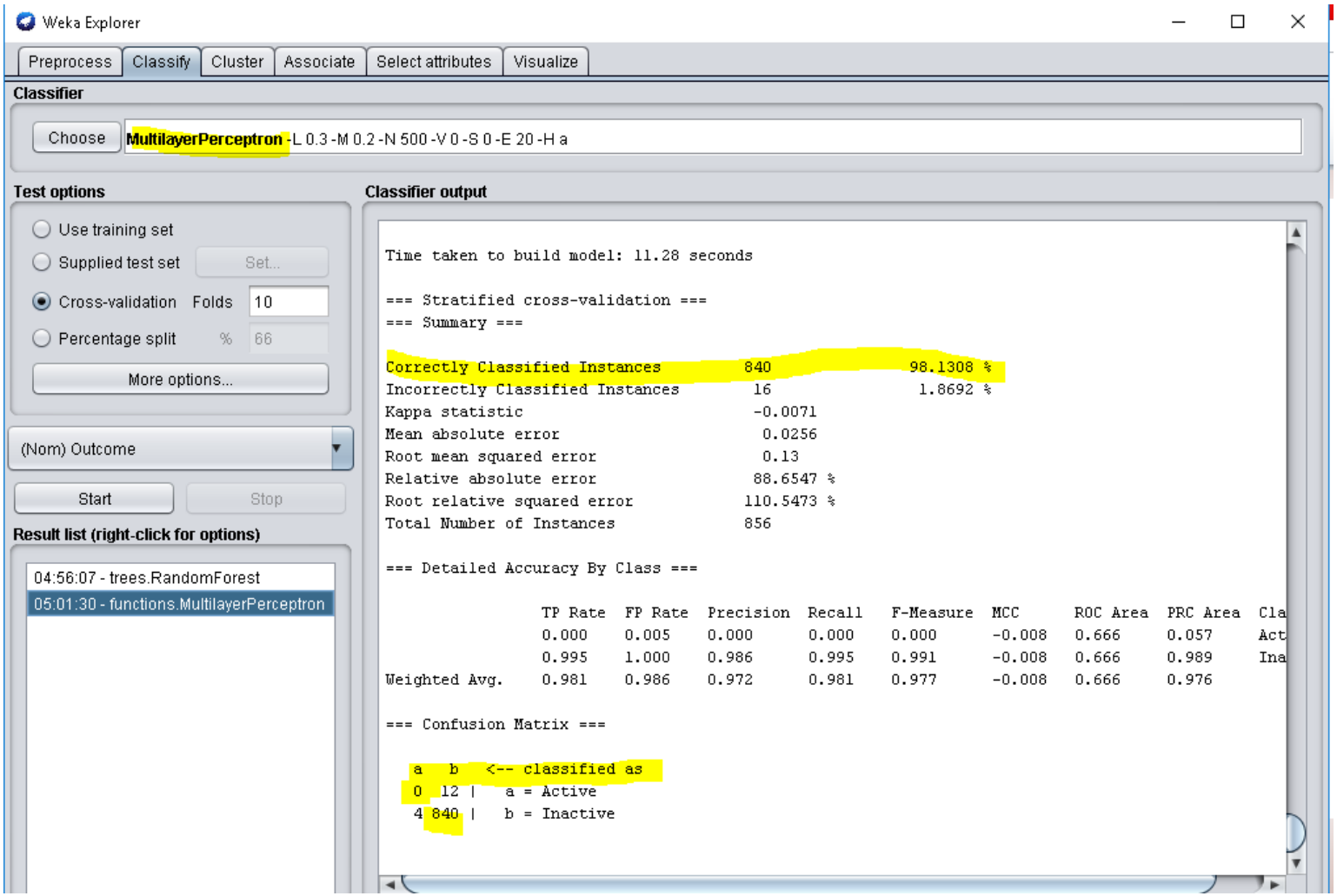
La clase de los enfermos es minoritaria respecto de la clase de los sanos.

La clase de fraudes es minoritaria respecto de la clase de no-fraudes.

Como medida de desempeño, *accuracy* no es útil en BD desbalanceados. ¿Pórque?

En ciertas aplicaciones (médicas), los falsos negativos son los peores casos posibles. Para penalizar los falsos negativos la métrica más adecuada es la sensibilidad (*recall*)

Desbalance de Clases (weka – unbalanced.arff)



Desbalance de Clases

Porque es importante atacar este problema [8]:

- Pobre desempeño de los clasificadores: regresión logística, maquinas de soporte vectorial y árboles de decisión.
- Proceso de aprendizaje guiado por métricas globales presenta un bias hacia la clase mayoritaria.
- Instancias de la clase minoritaria pueden ser ignorada.

Desbalance de Clases ¿Cómo lidiar con esto?

1. Trabajar el banco de datos tal cual esta: ajustar los algoritmos (funciones de costo).

2. Modificar el banco de datos: métodos de muestreo (*sampling*)

Undersampling: remover algunos patrones de la clase mayoritaria.
Ej.: *random under-sampling, condensed Nearest Neighbor (CNN), ADAptive SYNthetic Sampling (ADASYN).*

Oversampling: crear patrones sintéticos que se agregan a la clase minoritaria. Ej.: *random over-sampling, Synthetic Minority Over-sampling Technique (SMOTE), Bordeline-SMOTE.*

Híbrido: combinar *undersampling* y *oversampling*.

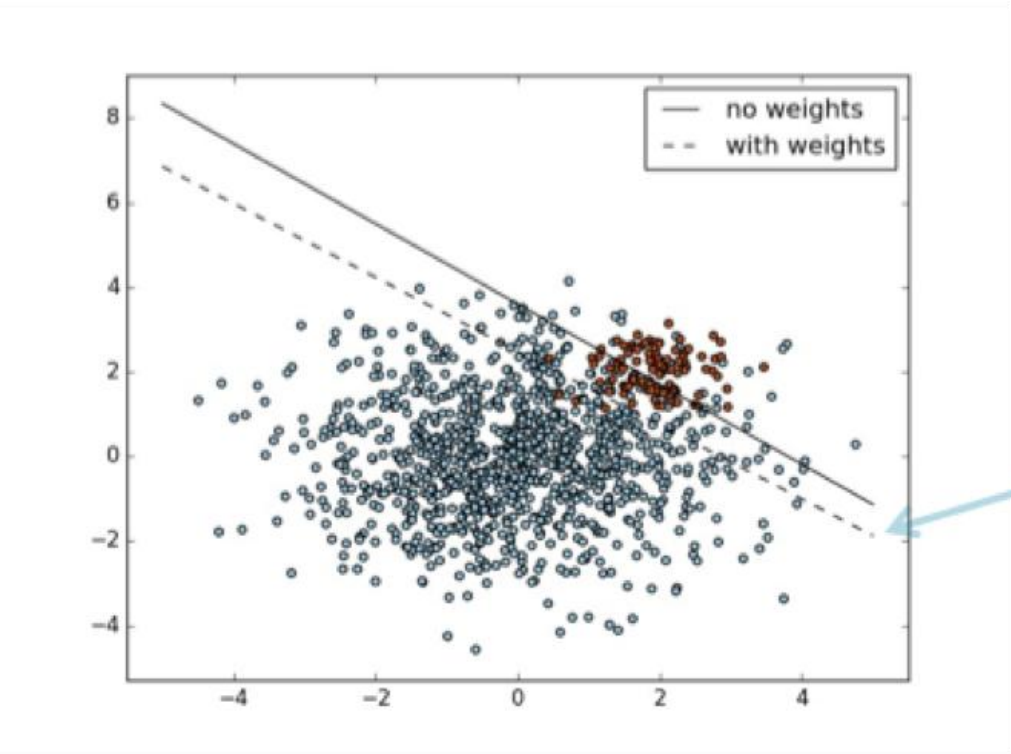
Desbalance de Clases ¿Cómo lidiar con esto?

3. Deshacerse de la clase minoritaria y tratar el problema con uno de detección de anomalías.

4. A nivel de algoritmo:

- Ajustar el peso de la clase.
- Ajustar el umbral de decisión.
- Modificar el algoritmo para que sea más sensitivo a clases raras.

Desbalance de Clases (Ajustar peso de clases)




```
import numpy as np
import pylab as pl
from sklearn import svm

# we create 40 separable points
rng = np.random.RandomState(0)
n_samples_1 = 1000
n_samples_2 = 100
X = np.r_[1.5 * rng.randn(n_samples_1, 2),
          0.5 * rng.randn(n_samples_2, 2) + [2, 2]]
y = [0] * (n_samples_1) + [1] * (n_samples_2)

# fit the model and get the separating hyperplane
clf = svm.SVC(kernel='linear', C=1.0)
clf.fit(X, y)

w = clf.coef_[0]
a = -w[0] / w[1]
xx = np.linspace(-5, 5)
yy = a * xx - clf.intercept_[0] / w[1]


# get the separating hyperplane using weighted classes
wclf = svm.SVC(kernel='linear', class_weight={1: 10})
wclf.fit(X, y)

ww = wclf.coef_[0]
wa = -ww[0] / ww[1]
wyy = wa * xx - wclf.intercept_[0] / ww[1]

# plot separating hyperplanes and samples
h0 = pl.plot(xx, yy, 'k-', label='no weights')
h1 = pl.plot(xx, wyy, 'k--', label='with weights')
pl.scatter(X[:, 0], X[:, 1], c=y, cmap=pl.cm.Paired)
pl.legend()

pl.axis('tight')
pl.show()
```

Tomado de: https://www.svds.com/learning-imbalanced-classes/

Alta dimensionalidad (Atributos)

En muchas ocasiones no encontramos con bancos de datos con gran cantidad de rasgo. En estos casos no necesariamente toda la información representada por los rasgos resulta relevante para la clasificación. Para atacar este problema se podría recurrir a:

- Asistencia de un experto en el tema.
- Método automáticos para la selección de características.

Selección de características

El objetivo de la selección de característica es elegir un subconjunto de atributos, del conjunto original, que maximice el porcentaje de clasificación correcta.

Sea F el conjunto original de atributos con cardinalidad $|F| = n$ y sea M una métrica de evaluación a ser optimizada entonces podríamos definir el proceso de selección de características como

$$|F'| = m < n. Encontrar\ F' \subset F\ tal\ que\ M(F')\ es\ maximizado$$

Encontrar un balance entre minimizar $|F'|$ y maximizar $M(F')$

Selección de características

Pero entonces ¿Cómo seleccionar un atributo?

- Uno generalmente no es descriptivo.
- Muchos pueden llegar a ser contraproducentes.
- Factores que afectan:
  - Relevancia
  - Co-dependencia
  - Variaciones en el tiempo
  - Incertidumbre y error
  - Valores perdidos
  - Dimensionalidad

Selección de características

¿Cómo decir si una característica es efectiva?

A través de una fase de entrenamiento. Entrenar un modelo con patrones típicos para descubrir:

- Efectividad
- Redundancia
- Agrupamientos
- Fronteras de decisión
- Etc.

Selección de características

En general los métodos de selección de características consisten un realizar una búsqueda a través de diferentes sub-conjuntos de atributos.

Métodos óptimos: búsqueda exhaustiva, *branch and bound Search*.

Métodos sub-óptimos: *sequential selection*, *stochastic search techniques* (algoritmos genéticos).

Referencias

[1] **Leondes, C.T. (2018)**. *Image Processing and Pattern Recognition*. California: Academic Press.
[2] **Duda, R.O., Hart, P.E. & Stork, D.G. (2001)**. *Pattern Classification*. 2nd edition. Wiley-Interscience.
[3] **Marques de Sá, J:P. (2001).** *Pattern Recognition: Concepts, Methods and Applications*. Berlin: Springer-Verlag.
[4] **Kuncheva, L. (2014)**. *Combining Pattern Classifiers: Methods and Algorithms*. 2nd edition. USA: Wiley.
[5] **Witten, I.H., Frank, E. & Hall, M.A. (2011).** *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. USA: Elsevier.
[6] **Murty, N.M. & Devi, V.S. (2011).** *Pattern Recognition: An Algorithmic Approach*. Springer.
[7] **Zaki, M.J. & Meira, W. (2014).** *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.
[8] **Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017)**. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73, 220-239.

¡ Gracias !

Thanks !

Obrigado

Xie xie ni

Domo arigatou

Спасибо

Merci

Grazie