CLASIFICACION DE PATRONES

M. Cabrera, J. Vidal
Dept. TSC
ETSETB
UPC
Febrero-Mayo 2007

Objetivo:

Aplicar algoritmos de clasificación a diversos problemas de clasificación que aparecen en diferentes campos o disciplinas de trabajo.

Misma Teoría Matemática: Múltiples Aplicaciones

TEMA 1: INTRODUCCIÓN

PROBLEMA BÁSICO:

Base de Datos:

- Vectores de Datos o de Características
- Pertenecientes a Diferentes Tipos o Clases

Objetivo:

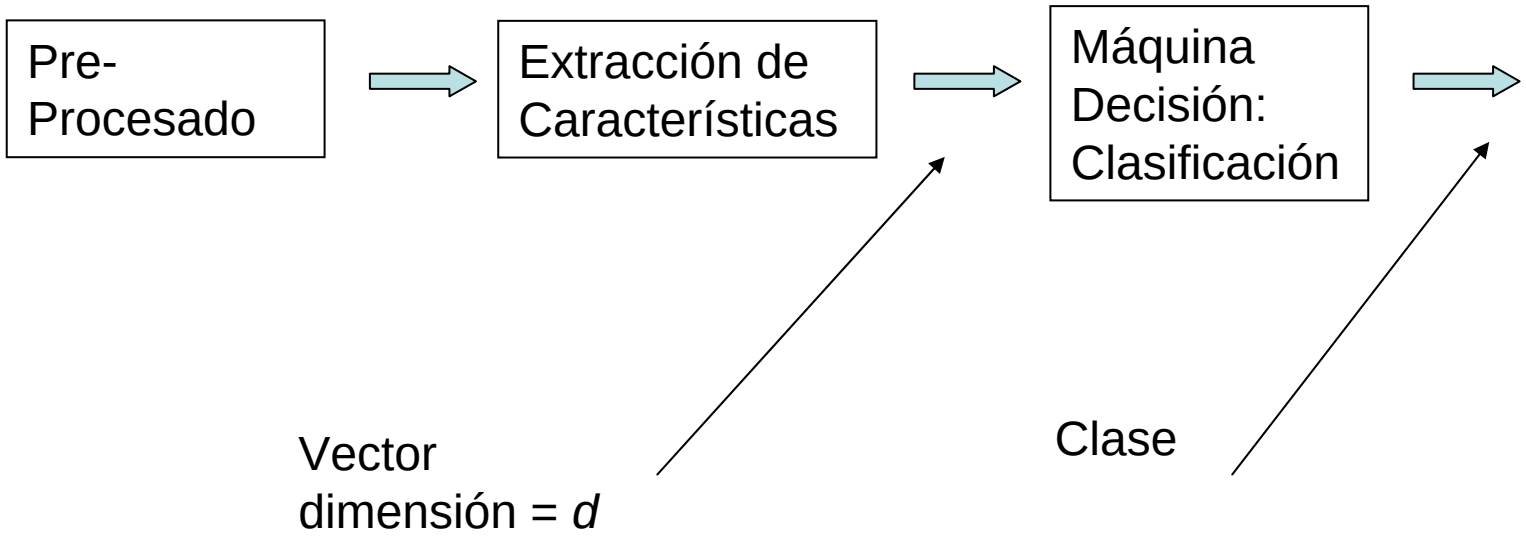
- Diseñar un Algoritmo para Clasificar un Nuevo Vector

Condicionantes:

- Se conoce a priori la estadística de los Vectores?
- Se conoce a priori la clase a la que pertenece cada uno de los vectores de la base de datos?

Etapas

CLASIFICACIÓN:



Temas

DIFERENTES GRADOS DE RESOLUCIÓN DEL PROBLEMA

1. **INTRODUCCIÓN**
2. **MODELOS BASADOS en f.d.p.**
3. **Selección de Características PCA-ICA**
4. **TECNICAS NO basadas en f.d.p. APRENDIZAJE SUPERVISADO**
5. **APRENDIZAJE NO SUPERVISADO**
6. **APRENDIZAJE INDEPENDIENTE DEL ALGORITMO**

- APLICACIONES:
  - Comunicaciones: Detección de Símbolos.
  - Reconocimiento de Voz
  - Clasificación de Imágenes.
  - Identificación biométrica.
  - Análisis de datos médicos: Detección de Enfermedades
  - OCR (Reconocimiento Óptico de Caracteres)
  - Identificación de ADN
  - SPAM: Reconocimiento de correo electrónico basura.
  - Etc...

- BASES DE DATOS DISPONIBLES EN LAB:
  - Comunicaciones: Detección de Símbolos.
  - Reconocimiento de Fonemas.
  - Análisis de datos médicos: Detección de Enfermedades
  - OCR (Reconocimiento Óptico de Caracteres)
  - Identificación de ADN
  - SPAM: Reconocimiento de correo electrónico basura.
  - Titanic
  - Base de datos: Voces (Señales de audio).
  - Base de datos: Brain: Imágenes de cortes cerebrales
  - Nuevas????????

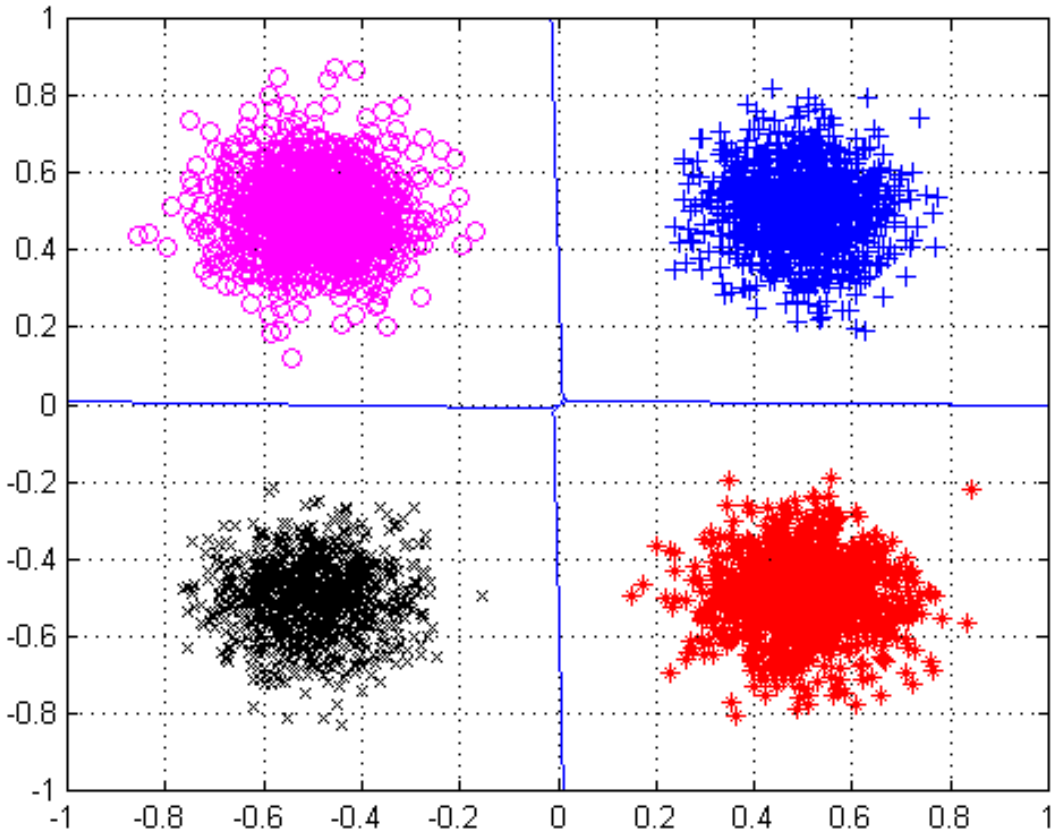
Ejemplo 1: Símbolo 1 ó Símbolo 2

Preprocesado:

Down-Conversión o Filtros adaptados

Extracción Características
Muestreo

Clasificación:
Detección MAP



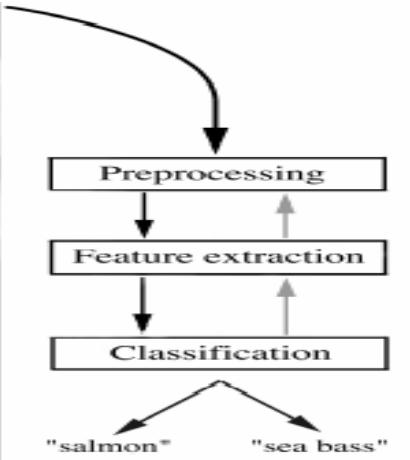
Ejemplo 2: Salmón o Lubina

Ejemplo:
Clasificar salmones o lubinas a partir de datos ópticos:

-Luminosidad

-Longitud

-Ancho



Ejemplo 3:
Diagnóstico de enfermedades cardiacas

Ejemplo:

BASE DE DATOS SHEART: utilizada para predicción del riesgo de contraer enfermedad cardiaca

VECTOR DE CARACTERÍSTICAS: (análisis de sangre, tabaquismo, antecedentes familiares, obesidad, consumo de alcohol, edad)

- Identificar las características más significativas para determinar la enfermedad.
- Predecir a partir de cada vector la probabilidad de sufrir un ataque de miocardio

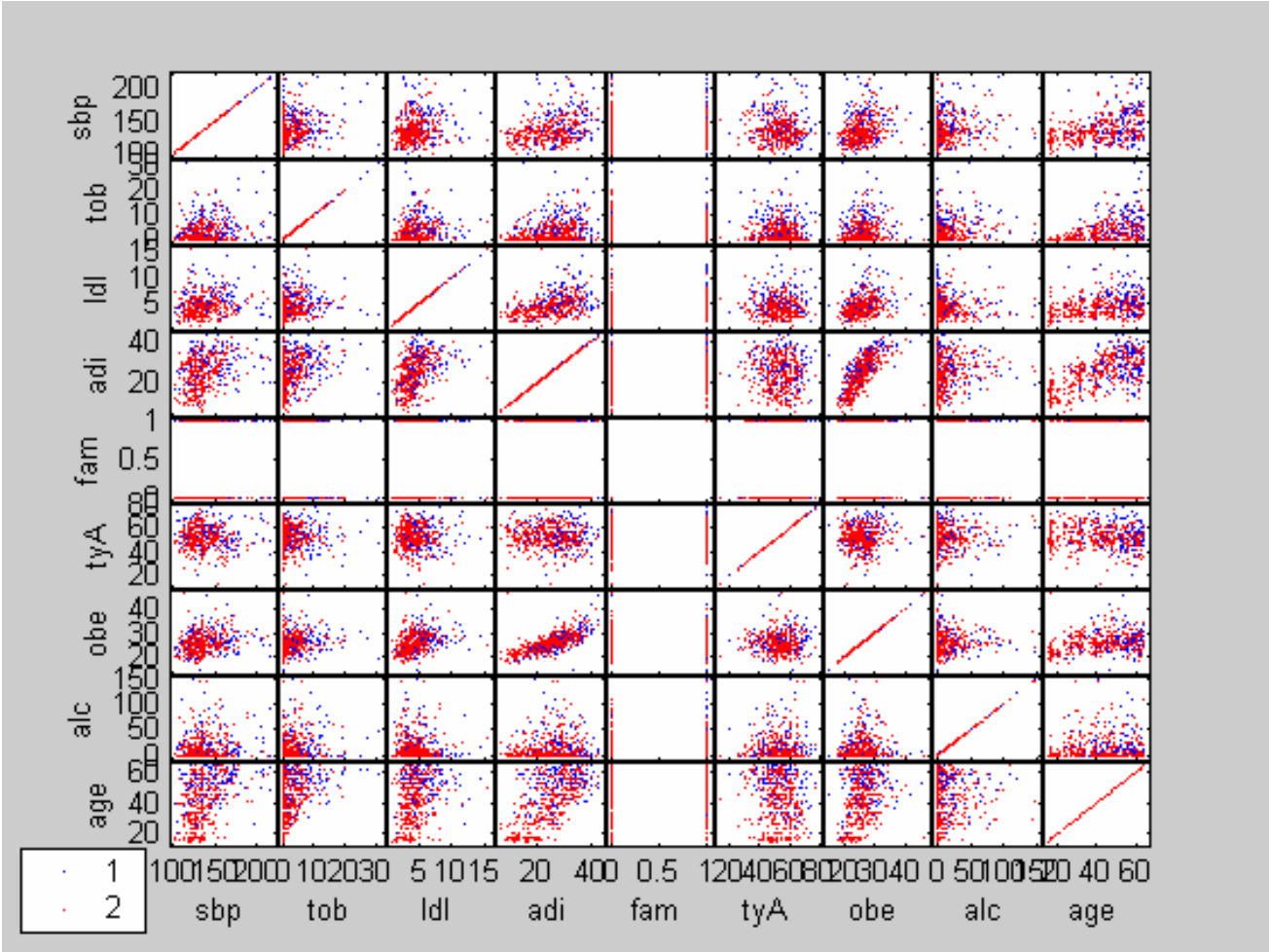
Ejemplo 3:Base de Datos Sheart

Sample of males in a heart-disease high-risk region of the Western Cape, South Africa.

sbp	systolic blood pressure
tobacco	cumulative tobacco (kg)
ldl	low densiity lipoprotein cholesterol
adiposity	
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior
obesity	
alcohol	current alcohol consumption
age	age at onset
chd	response, coronary heart disease

sbp	tob	ldl	Adip.	fa	A	Ob.	alc	age
160	12	5.73	23.11	1	49	25.3	97.2	52

Ejemplo 3:Base de Datos Sheart



Ejemplo 4:
Base de PHONEME

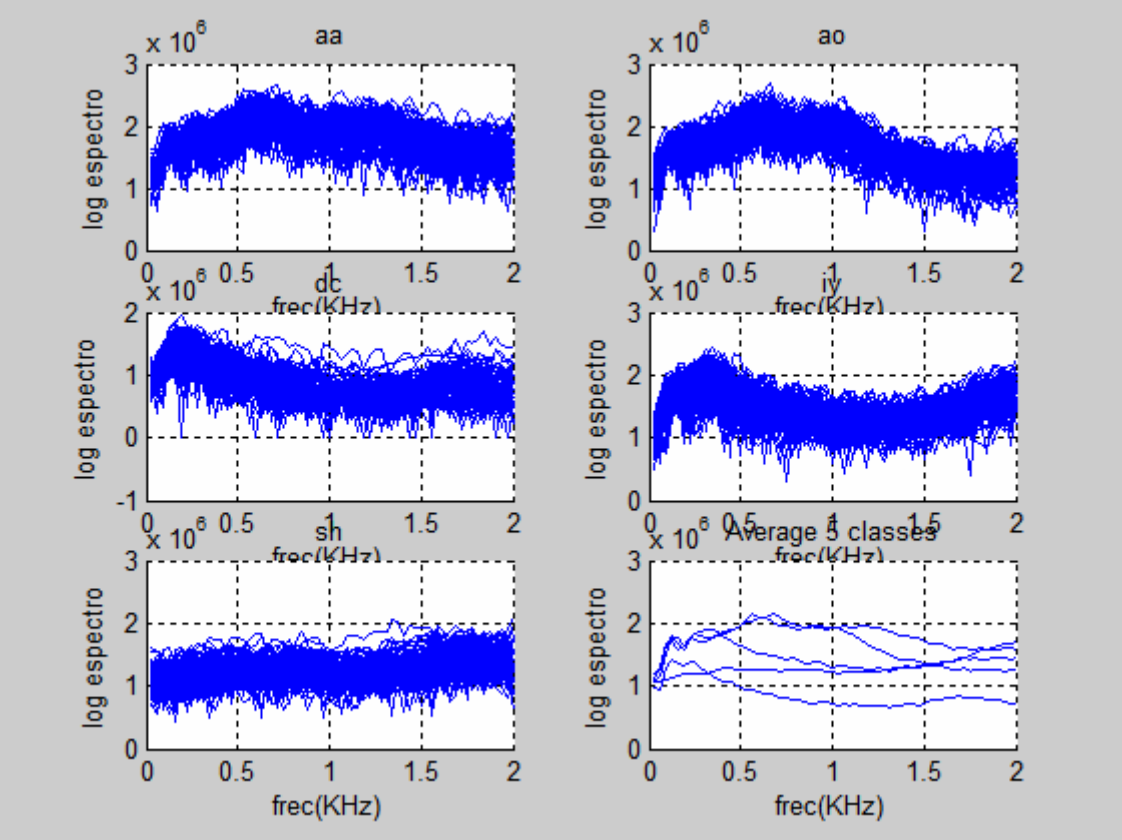
Samplig Frequency = 8KHz

LOG

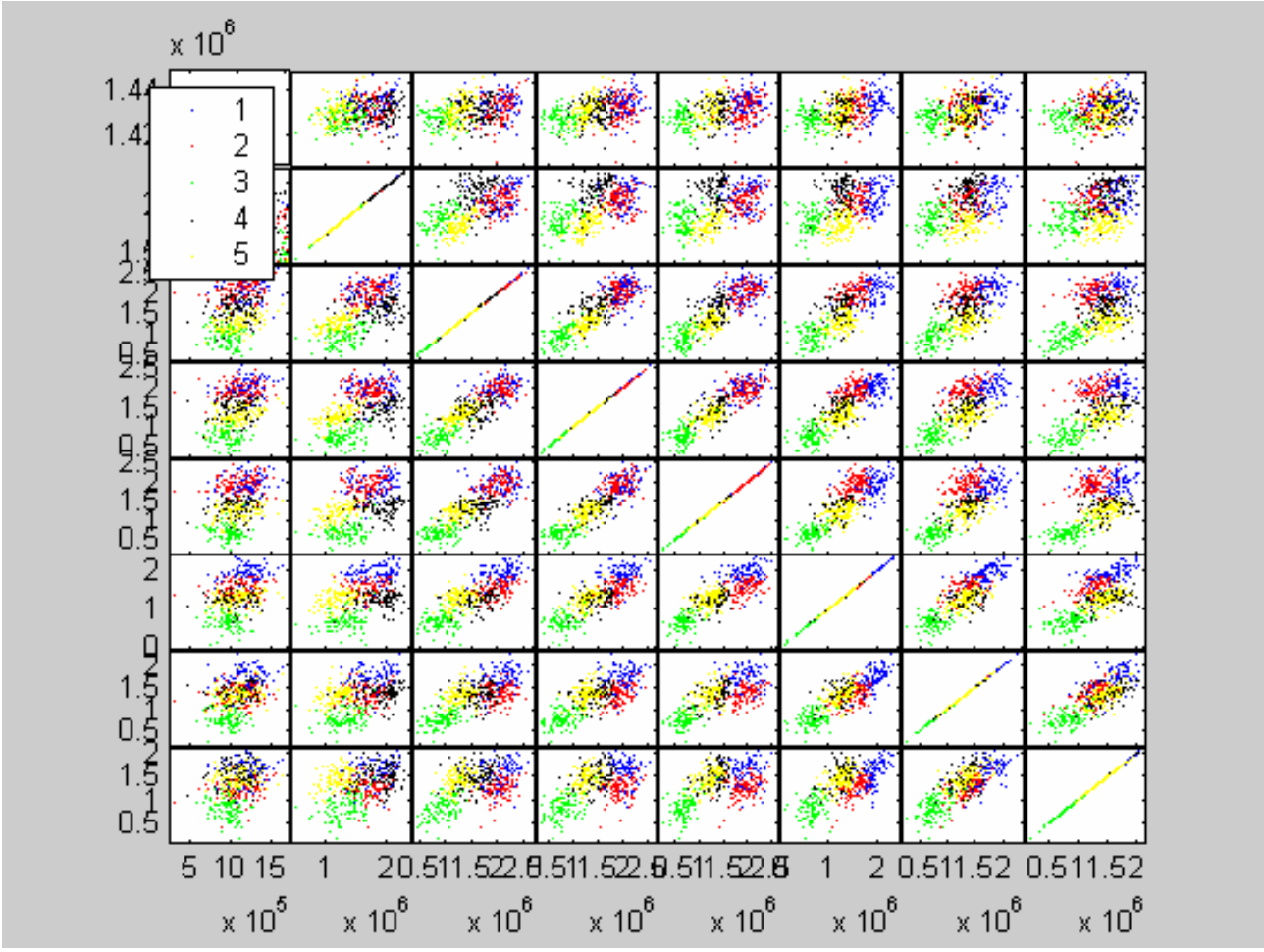
(ESPECTRUM)

aa ao dcl iy sh
695 1022 757
1163 872

Figure: Feature Vector Dimension = 64



Ejemplo 4:
Base de PHONEME, Scatter Plot



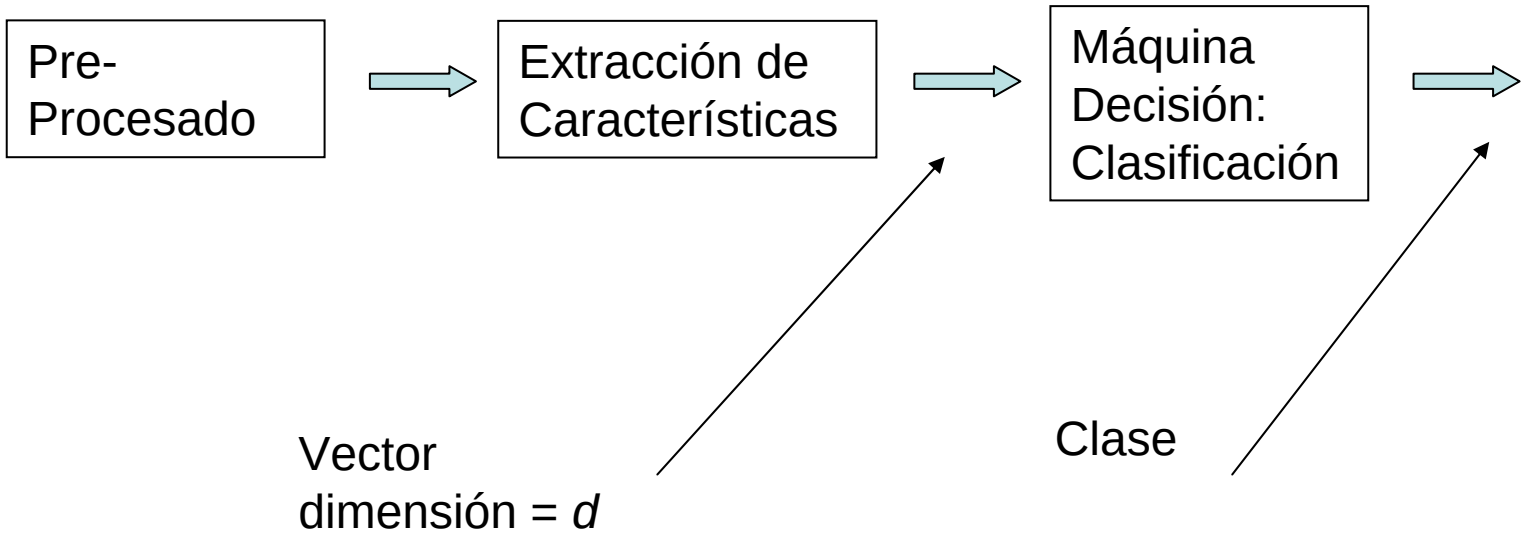
Ejemplo 5: Identificación Biométrica:

Tabla 8.1: Comparación de métodos biométricos.

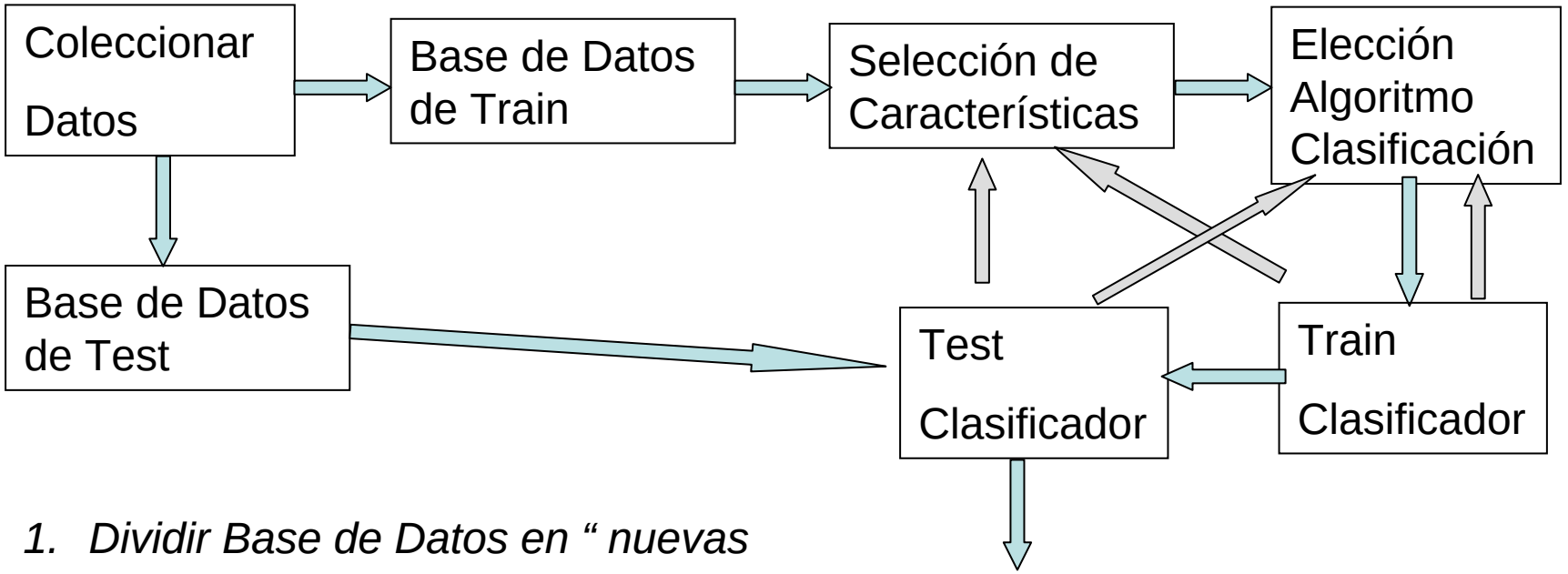
	Ojo - Iris	Ojo - Retina	Huellas dactilares	Geometria de la mano	Escritura - Firma	Voz
Fiabilidad	Muy alta	Muy alta	Alta	Alta	Alta	Alta
Facilidad de uso	Media	Baja	Alta	Alta	Alta	Alta
Prevención de ataques	Muy Alta	Muy alta	Alta	Alta	Media	Media
Aceptación	Media	Media	Media	Alta	Muy alta	Alta
Estabilidad	Alta	Alta	Alta	Media	Media	Media
Identificación y autenticación	Ambas	Ambas	Ambas	Autenticación	Ambas	Autenticación
Estándars	-	-	ANSI/NIST, FBI	-	-	SVAPI
Interferencias	Gafas	Irritaciones	Suciedad, heridas, asperezas ...	Artritis, reumatismo ...	Firmas fáciles o cambiantes	Ruido, resfriados ...
Utilización	Instalaciones nucleares, servicios médicos, centros penitenciarios	Instalaciones nucleares, servicios médicos, centros penitenciarios	Policia, industrial	General	Industrial	Accesos remotos en bancos o bases de datos

Etapas en el algoritmo de clasificación:

1. Pre - Procesado
2. Extracción de un vector de características (Feature Extractor)
3. Clasificación



Etapas en el Diseño del Sistema de Reconocimiento



1. *Dividir Base de Datos en " nuevas Bases de datos: Train, Test*
2. *Selección de Características*
3. *Selección Algoritmo de Clasificación*
4. *Entreno del Algoritmo*
5. *Test del Algoritmo*

Selección de Características:

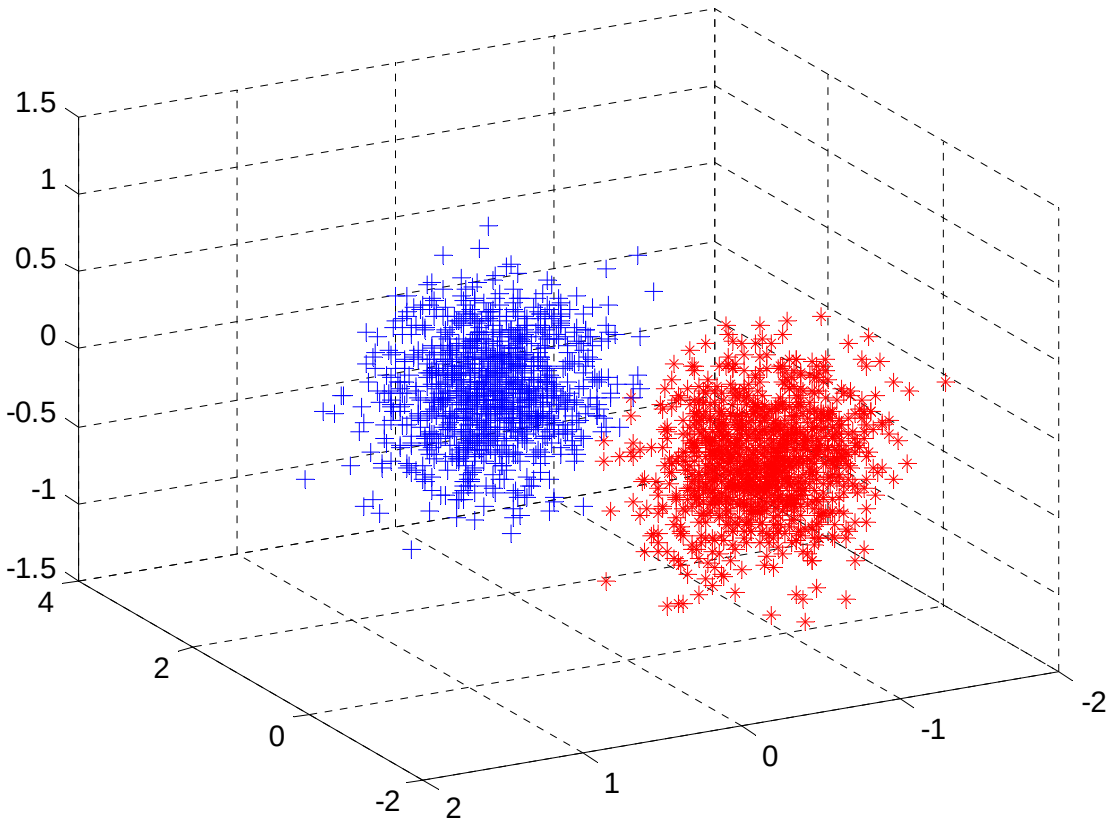
Ejemplo:

$V1=(0.1,0.5,1)^T$

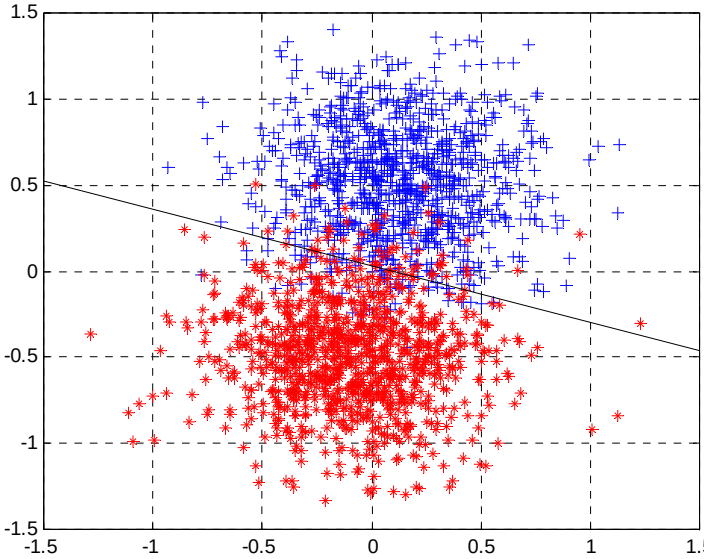
$V2=-(0.1,0.5,1)^T$

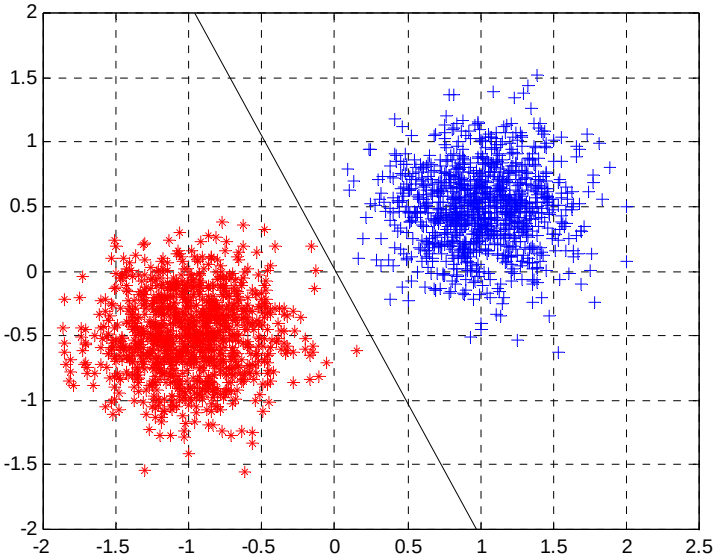
Varianza ruido =0.1

Se pueden clasificar datos con menos de 3 características??



Selección de 1 ó 2 características:

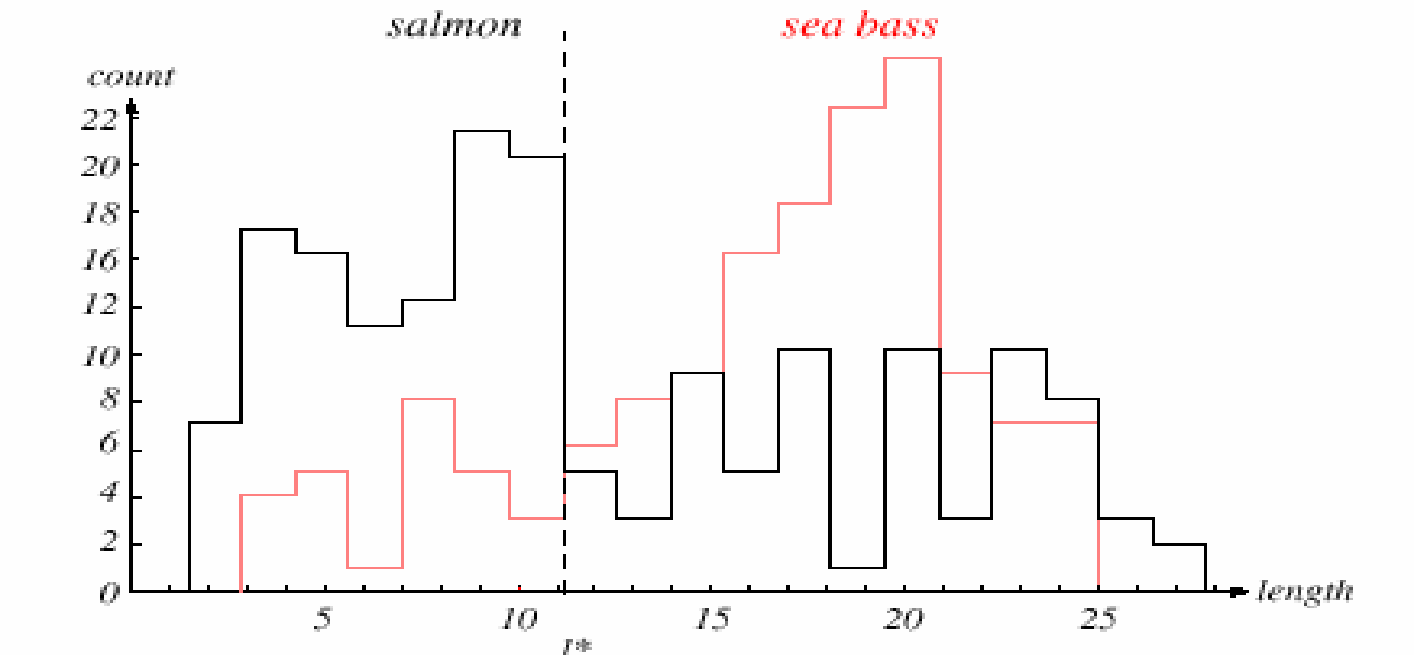




Selección de características:

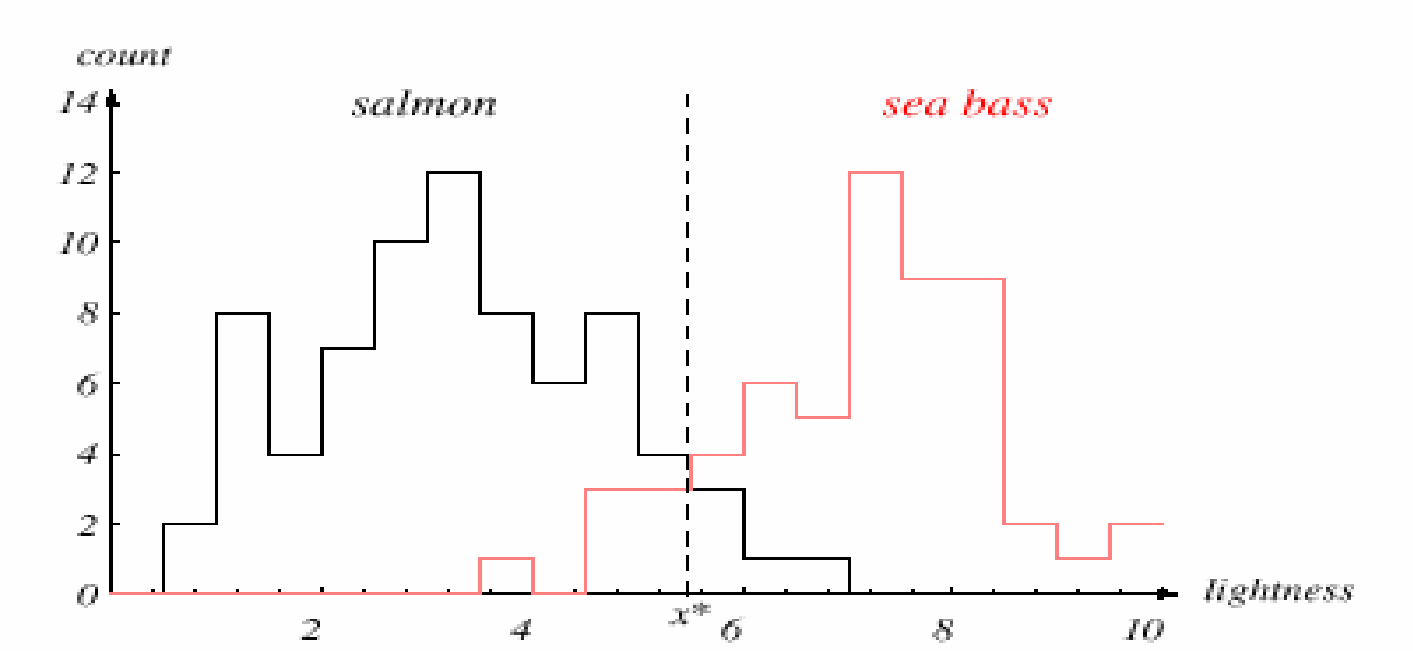
Longitud del pescado: Gran error de Clasificación

HISTOGRAMA:



Selección de características:

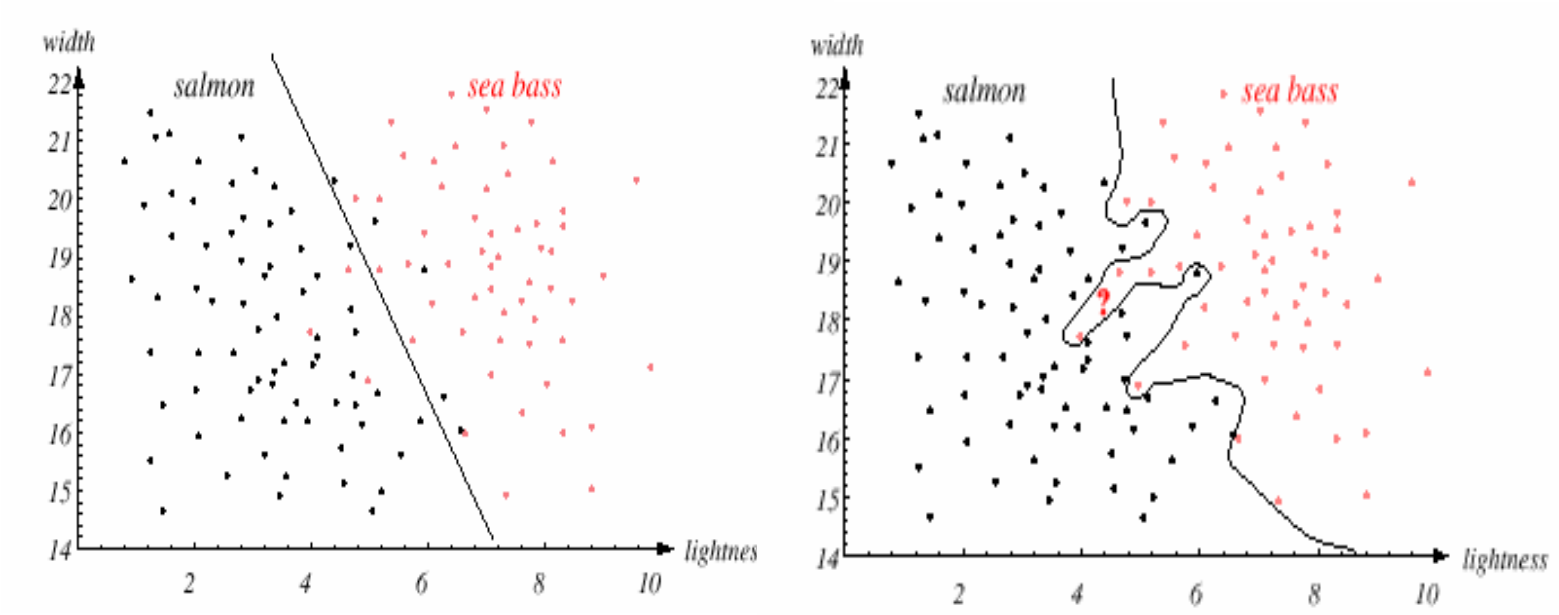
Luminosidad del pescado: Disminuye el error de Clasificación



Selección de características:

Luminosidad y ancho del pescado:

Disminuye el error de Clasificación



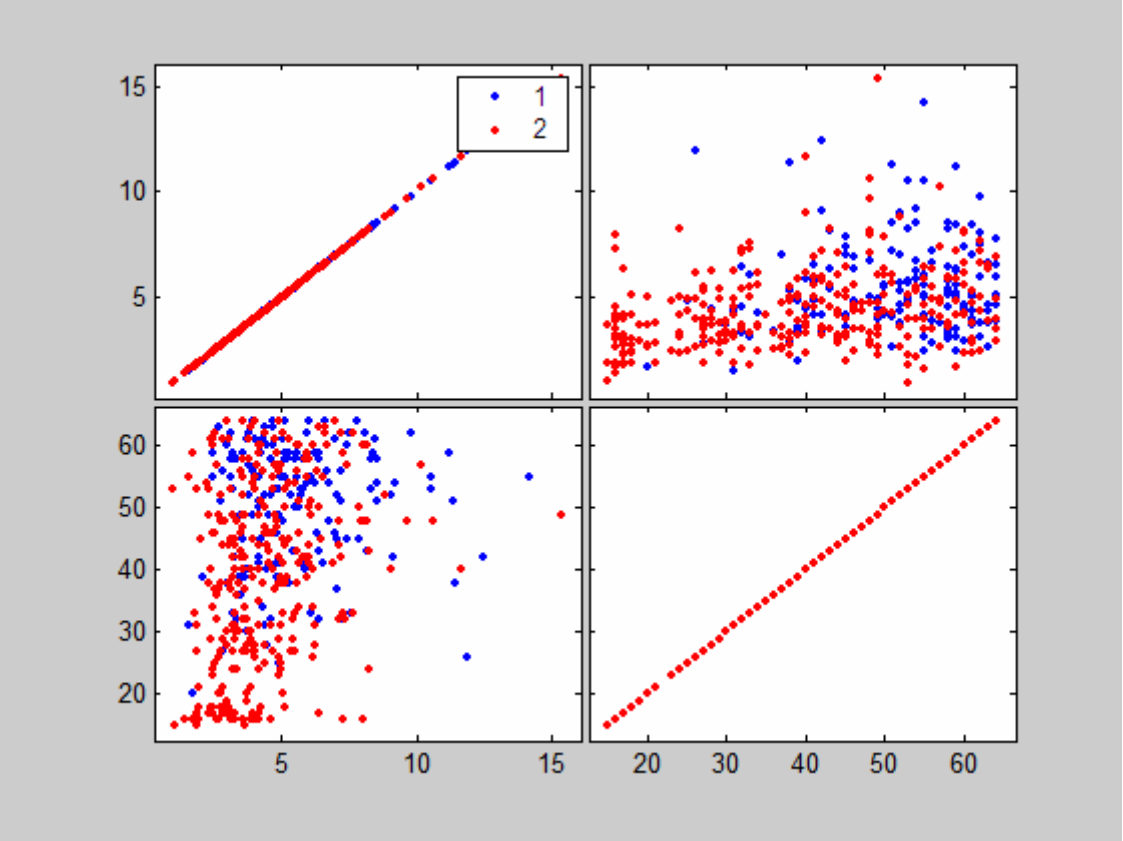
Selección de características:

Sheart

Cholesterol

and

AGE

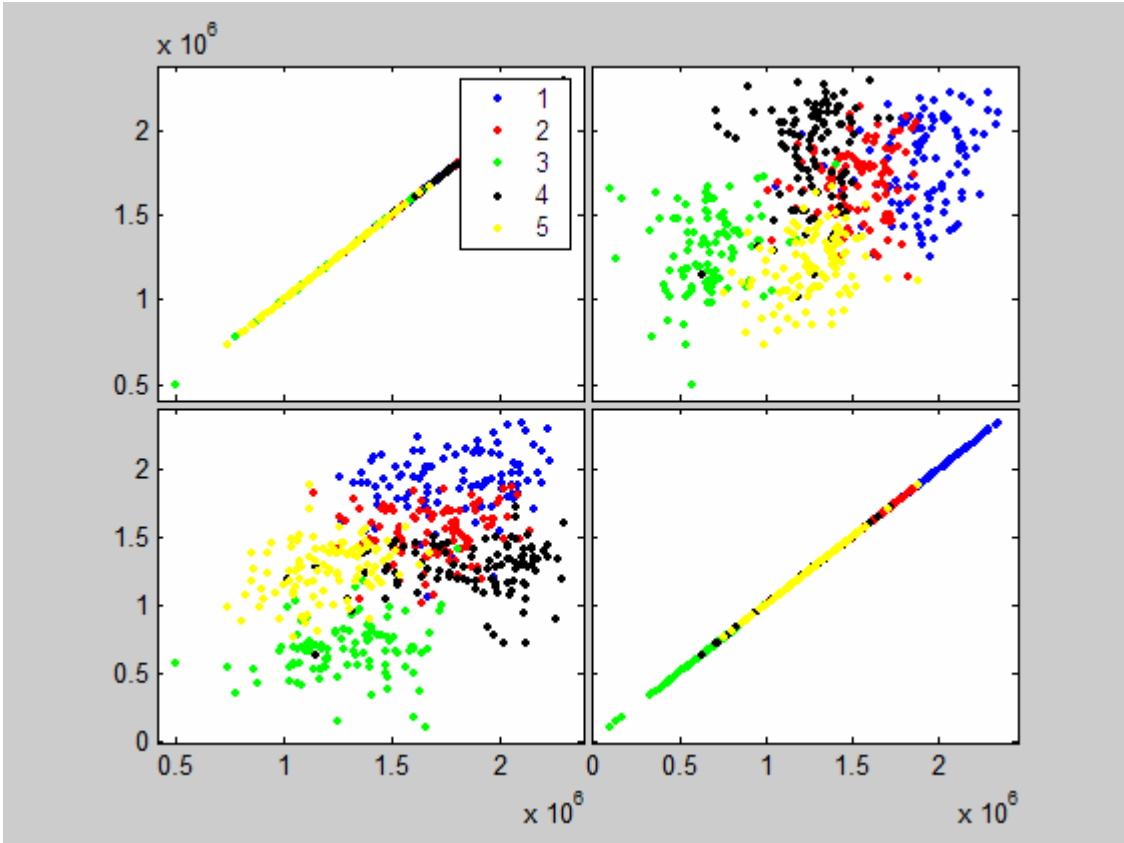


Selección de características:

Phonemas

5
Classes:

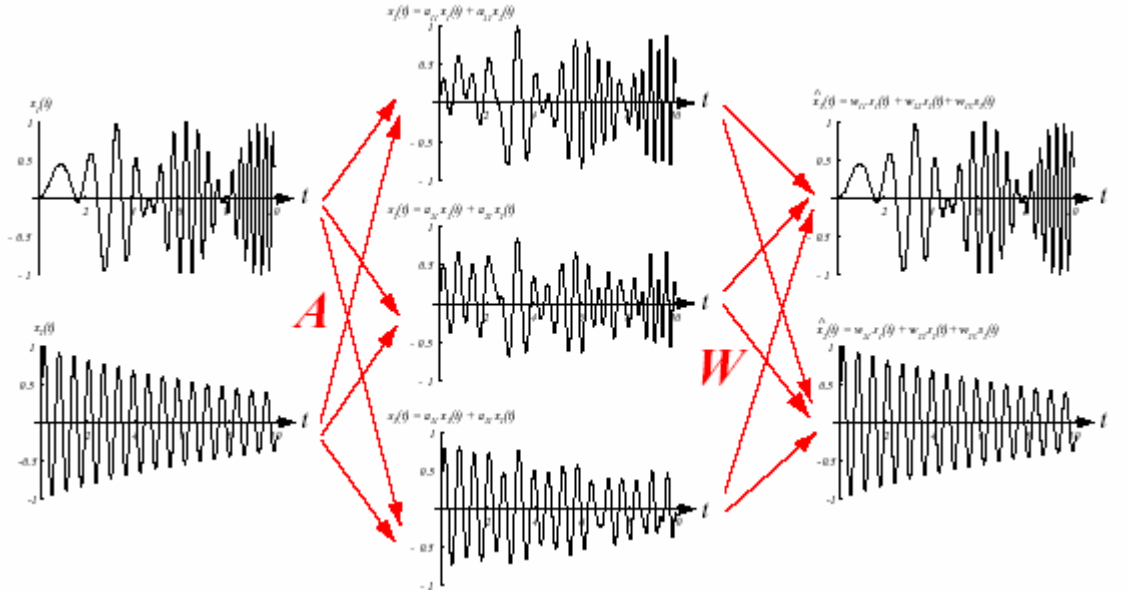
aa ao
dcl iy
sh



COMPONENT ANALYSIS

ICA Independent Component Analysis

- Application: Blind Source Separation (cocktail party problem).



Colección de Datos: Base de Datos

- La ó las Base de Datos disponibles para el diseño del clasificador se dividirán en:
  - Base de Datos (Vectores Muestra) de **Entrenamiento**
  - Base de Datos de **Test**

Selección de Características

- En ocasiones, el diseño de esta etapa puede resultar de mayor repercusión en el error de clasificación que el propio clasificador.
- Un **<u>número moderado de características</u>** influye en
  - Regiones de decisión más simples
  - Clasificador más fácil de entrenar
  - Así mismo se requiere que las características sean robustas a ruido, otros tipos de error, etc.

Selección de Características

- **Capacidad de Discriminación:** Las características deben constituir agrupaciones de valores medidos que se mantienen muy semejantes entre objetos de la misma categoría y diferentes entre objetos de distintas categorías.
  - Baja variabilidad Inter-clases
  - Alta variabilidad Intra-clases
- Interesan características **invariantes a transformaciones** irrelevantes del vector de entrada, como por ejemplo.
  - Escalado
  - Rotación
  - Traslación

Clasificador:

- El clasificador **asigna un objeto o categoría** a partir del vector de características medido.
- El grado de dificultad del diseño del algoritmo de clasificación propiamente dicho depende del **grado de variabilidad**, que interesa:
  - Baja variabilidad Inter-clases
  - Alta variabilidad Intra-clases
- El **ruido** que afecta a esta variabilidad se clasifica a groso modo en:
  - Ruido intrínseco a las clases, difícilmente modelables.
  - Ruido ajeno: Medida.

Clasificador:

- **Entrenamiento del Clasificador.** Se deben **ajustar una serie de parámetros** a la aplicación determinadaa partir de un **entrenamiento** mediante vectores de características de los cuales se conoce a priori la clase o categoría a la que pertenecen.
- **La evaluación del clasificador** debe hacerse en función del criterio de diseño elegido: Mínimo error de clasificación o Minimizar una función de coste o riesgo.
- **Eficiencia Computacional** del Clasificador
- **Aprendizaje Supervisado vs Aprendizaje NO Supervisado.** (No se siempre se dispone a priori de datos clasificados por clases o categorías).

Teoría asociada

- En general los clasificadores se basan en las **propiedades estadísticas de los vectores muestra**. Ya sea explícitamente o implícitamente las f.d.p. de los datos desempeñan un papel fundamental en los diferentes algoritmos de clasificación.

- Cuando los datos no son numéricos y en función de la aplicación las f.d.p. no constituyen la herramienta más adecuada. Así por ejemplo en reconocimiento de **modelos sintácticos**, que siguen determinadas normas lógicas, se requiere conocer reglas gramaticales que describan cada una de las decisiones a determinar.

- Idealmente como conocimientos previos en el desarrollo de un clasificador se requiere **(compromiso):**
  - Conocimiento Previo del Problema
  - Gran cantidad de datos de entrenamiento.

- El desarrollo a lo largo de esta asignatura se va a hacer trabajando con **vectores reales**, ya que en la mayoría de bases de datos, las muestras son reales. Tiene sentido trabajar con **muestras complejas** especialmente en las aplicaciones de comunicaciones, donde las señales I&Q se representan mediante números complejos. En general, al aplicar los diferentes algoritmos con muestras complejas se han de generalizar las diferentes estrategias de clasificación

Conclusiones

- En los ejemplos anteriores han aparecido implícitamente diferentes conceptos:
  - **Teoría de la Decisión**
  - **Función de coste** de la decisión (error de clasificación)
  - **Zonas de decisión** que dividen el espacio de características

Temas

1. INTRODUCCIÓN (2h)

2. MODELOS BASADOS en f.d.p.
- MAP, Caso Ideal en el que se conoce de forma ideal la f.d.p. de las categorías subyacentes. (4h)
- ML, Se conocen la forma de las diferentes f.d.p. salvo el valor particular de determinados parámetros (4h)

Duda, Temas 2,3

3. SELECCIÓN DE CARACTERÍSTICAS BASADA EN EL ANÁLISIS DE COMPONENTES.
- PCA, ICA (Independent Component Analysis) (4h)

4. TÉCNICAS NO basadas en f.d.p. APRENDIZAJE SUPERVISADO

No se tiene ningún conocimiento a priori sobre las f.d.p.
- K-NEAREST (2h)
- Funciones discriminantes Lineales (2h)
- Redes Neuronales (4h)
- Métodos de Árbol (Reglas lógicas) (2h)

Duda, Temas 4,5,6,7,8

5. APRENDIZAJE NO SUPERVISADO (2h)

No se tiene información previa,
- K Means, Clustering

Duda Tema 10 + Ref

6. APRENDIZAJE INDEPENDIENTE DEL ALGORITMO/VALIDACIÓN DE CLASIFICADORES
- Elección Forward-Backward de las características, Cross-Validation, Selección del mejor algoritmo de clasificación (2h)

Duda, Temas 9 + Ref

EVALUACIÓN

1. Asistencia a clases y desarrollo de las **prácticas** (Test de algoritmos, evaluación de resultados, programación en Matlab, nivel sencillo) –> 30%
2. Entrega y Realización de Ejercicios Propuestos en clase –> 15%
3. Presentación de **trabajo**: Prueba de los diferentes algoritmos de clasificación con una base de datos. –>15%
4. **Examen final**: Ejercicios propuestos en las clases teóricas. –> 40%

(15-JUNIO-2007, Viernes 8h)

Referencias, Facilidades

- **[Duda,2001]** R. O. Duda, P. E. Hart, D. G. Stork. “Pattern Classification”, Ed. Wiley Interscience, 2002.
- **[Stork,2004]** Computer Manual in MATLAB to Accompany Pattern Classification, 2nd Edition. David G. Stork, Elad Yom-Tov. Ed. Wiley Interscience, 2004
- **[T. Hastie, R. Tibshirani, J. H. Friedman 2001]** The Elements of Statistical Learning Springer Verlag, 2001
- **[Hyvarinen, 2001]** Independent Component Analysis, Aapo Hyvarinen, Juha Karhunen, Erkki Oja. Ed. Wiley Interscience, 2001.
- **[*Heijden, 2004*]** Classification, Parameter Estimation and State Estimation - An Engineering Approach Using MATLAB. Author: van der Heijden (John Wiley) [**Kuncheva, 2004]** Combining Pattern Classifiers: Methods and Algorithms, Ludmila I. Kuncheva ,July 2004, Ed Wiley
- **[Bishop,2006]** “Pattern Recognition and Machine Learning”, Christopher M. Bishop, Springer (2006).

Referencias, Facilidades

- http://gps-tc.upc.es/comm/marga/Wclass/

(Transparencias de las clases, Enunciados de Prácticas)

Software en Matlab

- **PRTools**, a Matlab Toolbox for Pattern Recognition Pattern Recognition Group
  - Department of Imaging Science and Technology, Faculty of Applied Sciences, Delft University of Technology, Lorentzweg,The NetherlandsPRTools, a Matlab Toolbox for Pattern Recognition Pattern Recognition Group
  - http://www.prtools.org/

- **The FastICA package for MATLAB:**
  - **http://www.cis.hut.fi/projects/ica/fastica/**

- **MATLAB GUI Tool:** [Stork,2004] Computer Manual in MATLAB to Accompany Pattern Classification, 2nd Edition

- **Bases de datos**: Datasets for "The Elements of Statistical Learning“: Department of Statistics at Stanford University
  - http://www-stat.stanford.edu/~tibs/ElemStatLearn/