4a. Escuela de Verano en Biomatemáticas UAEH

31 de Mayo – 4 de Junio
2010

Análisis y clasificación de datos

Jorge Viveros

CIMA-MF1-01 · jviveros@uaeh.edu.mx · CIMA · UAEH · MX

Mensaje importante

La presente es una panorámica sobre algunos temas muy específicos acerca del análisis y clasificación de datos, la cual se presentó los dias 31 de Mayo y 1 de Junio del 2010 en el Centro de Investigación en Matemáticas de la Universidad Autónoma del Estado de Hidalgo.

Uno de los propósitos de esta serie de pláticas fué el de presentar a los estudiantes de semestres avanzados en la licenciatura en Matemáticas, problemas para una tesis de licenciatura. Aquellas personas interesadas están invitadas a escribirme un correo electrónico para obtener mas información acerca de esta tema. La exposición está entonces orientada hacia estudiantes de licenciatura.

Las imágenes que aparecen en esta presentación fueron, en su mayoría, obtenidas de otras fuentes sin previa autorización, créditos han sido otorgados explícitamente, cualquier omisión es accidental. A aquellas personas que deseen utilizar algunas de las imágenes incluidas, se les pide atentamente incluyan la fuente como aparece reportada aquí.

En algunos casos se ha incluido trabajo no publicado del autor cuando este era asistente de investigación en el Wallace H. Coulter Department of Biomedical Engineering en Georgia Tech (2005-2006).

Contenido (1/2)

1. Introducción: ejemplos de problemas de clasificación de datos.

   1.1 Clasificación de textos,
   1.2 Análisis y reconocimiento de imágenes,
   1.3 El problema del correo postal,
   1.4 Clasificación de proteínas en grupos de homologías,
   1.5 Expresión genética (clasificación de datos tipo microarreglo.)

2. Técnicas del Análisis de datos

   2.1 Análisis lineal de componentes principales (L-PCA)

3. Métodos de clasificación de datos

   3.1 El problema abstracto de clasificación binaria
   3.2 Clasificadores lineales (y el problema de la dimensionalidad alta)
   3.3 El truco de la nucleación del espacio
   3.4 La idea del análisis de componentes principales en el caso no lineal (k-PCA)

Contenido (2/2)

1. Introducción

A continuación se mencionan 4 problemas de interes práctico y científico los cuales involucran datos de un tipo muy especial en el cual estamos interesados, estos problemas son:

1. análisis de textos
2. el problema del correo postal
3. clasificación de proteinas en grupos de homología
4. clasificación de datos tipo microarreglo

Hemos inlcuido referencias selectas para ampliar la información. Estas referencias no son, necesariamente, exhaustivas ni tampoco las mas sobresalientes, sinembargo le proporcionarán al lector interesado un buen punto de partida para adentrarse o ampliar su conocimiento sobre el tema.

1. Introducción

1.1: Clasificación automatizada de textos

Problema: asignación automática de textos dentro de clases predeterminadas con base en su contenido.

Aplicaciones:

* Clasificación de **noticias electrónicas** (Reuters, Associated Press, Yahoo!, etc) para su publicación en-línea y su posterior archivamiento.

* Categorización del **ingreso continuo de textos** a centros de información de acuerdo con bancos de datos ya existentes, o bien **creación de nuevos bancos de información** (Biblioteca del Congreso (EEUU), etc.)

* Clasificación de **artículos técnicos** para su pronta localización dentro de grandes bancos de información (ArXiv y revistas de divulgación e/o investigación con páginas electrónicas que permitan su escrutinio desde un ordenador).

* Clasificación de **textos electrónicos** e incluso **videos** en páginas electrónicas con "acceso abierto" (i.e., bajo la adminstración de los propios usuarios: Wikipedia, Scholarpedia, Gigapedia, Youtube, etc).

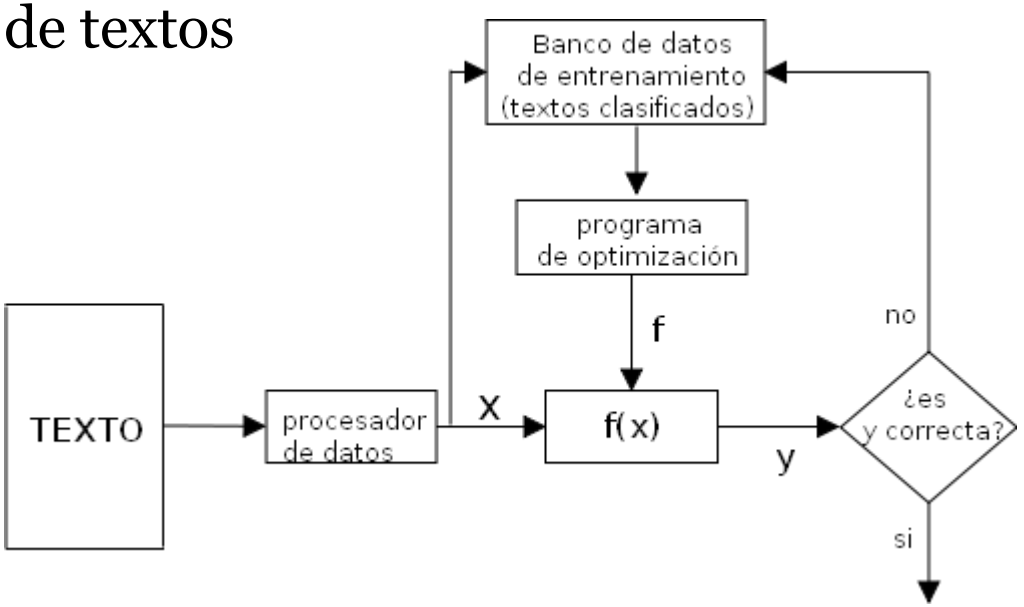
... muchas aplicaciones más.

1.1: Clasificación automatizada de textos

Estrategia:

- asignar a cada texto un vector $X \in \mathbb{R}^n$

- substituir X en el argumento de una función clasificadora $f : \mathbb{R}^n \to Y$ previamente calculada.



- con base en el resultado de la evaluación y=*f(X)*, X se asigna a una o varias categorías en las cuales se dividide $\mathbb{R}^n$

- validar o refutar el resultado de la clasificación. Si la clasificación fué adecuada nuestro grado de confianza en *f* aumenta. Si la clasificación no es adecuada se determinan las causas (el problema es el texto, la función clasificadora o ambos).

- Si el problema es el texto, este se clasifica manualmente o se elimina. Si el problema es la función clasificadora, esta se recalcula con base en la información subministrada.

La construcción de la función clasificadora es el problema central de la teoría del aprendizaje (“learning theory”)

1.1: Clasificación automatizada de textos

*Asignación de vectores a **textos**:* cada texto es asignado un vector $X \in \mathbb{R}^n$ ($n \in \mathbb{N}$ predeterminado).

Las componentes de X son **pesos** asignados a (n) palabras características dentro del texto. Los pesos se calculan, básicamente, con base en la **frecuencia** con la que las palabras aparecen en el texto, n es típicamente del orden de varios cientos, e.g. n~300.

*Asignación de vectores a **categorías**:* cada categoría es asignada un vector $w \in \mathbb{R}^n$.

1. formar banco con palabras extraidas de los textos en la categoria en cuestión.
2. eliminar palabras exclusivas de textos individuales.
3. (**"feature selection"**) escoger un subconjunto de palabras observando su frecuencia dentro de la colección y la información que contienen y que las distinguen como perteneciente a la categoria en cuestión.

La función clasificadora cumple su objetivo, básicamente, observando el signo del producto interno (w, X).

1.1: Clasificación automatizada de textos

Referencias y fuentes de información:

Dumais et al (1998)

Banco de datos = **colección Reuters-21578**: 12,902 textos o historias previamente clasificados en 118 grupos. Se seleccionaron 9,603 historias (conjunto de práctica o "training set") para construir una función clasificadora por categoría. La acertividad de las funciones clasificadoras se examina en las restantes 3,299 historias.

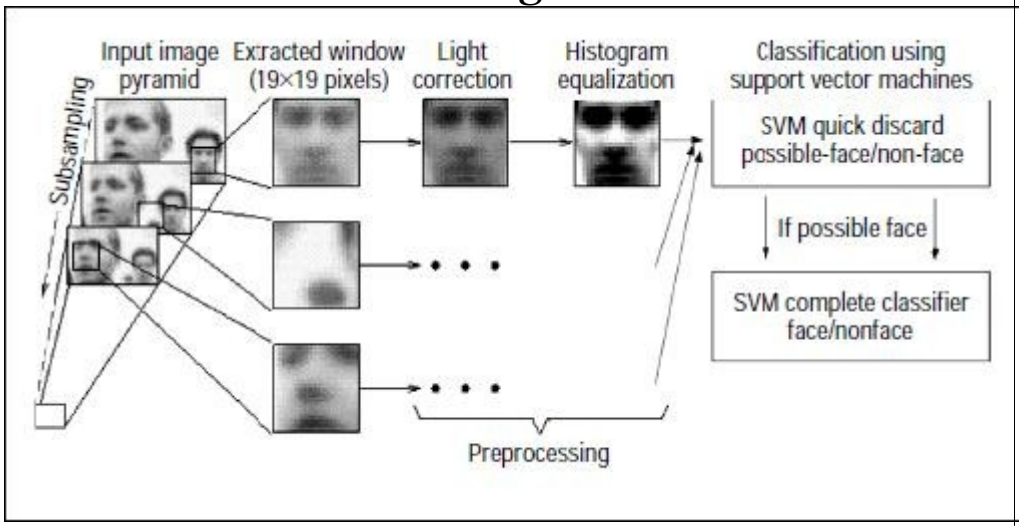
Cada texto de prueba es asociado con un vector cuyas componentes representan la frecuencia con que ciertas palabras "clave" aparecen en el texto. La selección de tales palabras se hace de manera que estas maximicen la información mutua (**"mutual information"**) que tiene cada una con las palabras características de cada categoría (no es un problema trivial).

Colección de textos Reuters-21578

1.2: Análisis y reconocimiento de imágenes

Osuna, Freund, Girosi (1998)

Problema: identificación automatizada de rostros en fotografias



Fuente: Osuna, Freund, Girosi 1998

Procesamiento de imágenes:

Cada imagen es escaneada exhaustivamente en distintas escalas.
De cada imagen escaneada se cortan cuadros solapados de 19x19 pixeles
Cada cuadro es sometido al siguiente proceso:

"Masking:" remoción de pixeles cercanos a la orilla de la cuadro (reducción de la dimensión del espacio ambiente de 19x19=361 a 283).

Corrección del gradiente de iluminación (reducción de luz en zonas oscuras)

Aplicación de un histograma de ecualización (compensa diferencias en brillo e iluminación entre imágenes).

Traducción de cada cuadro en un vector *X* propio para la clasificación (e.g. método De Wijewickrema et al –ver más adelante)

1.3: El problema del correo postal

Problema: identificación, reconocimiento y lectura automática de códigos postales

Fuente: Pfitze, Behnke, Rojas (2000)

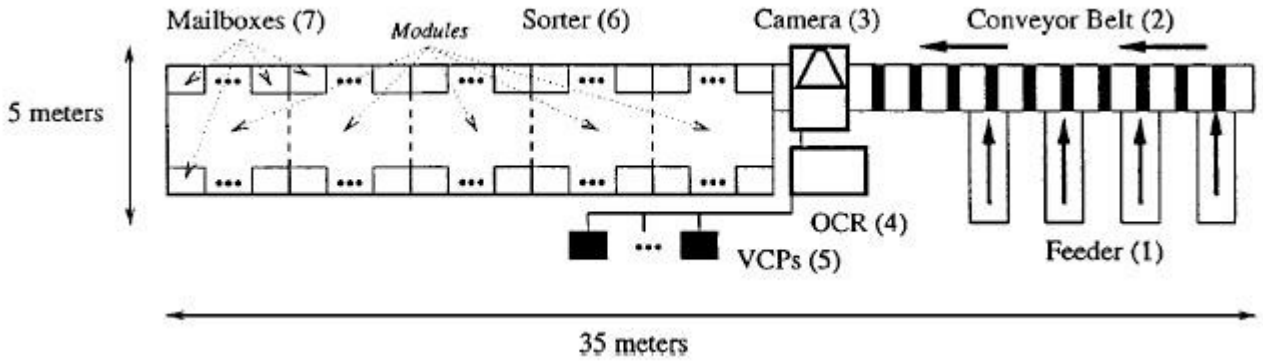
Algunos de los retos principales:

• No existe una sobre universal de tamaño estándar con una zona reservada exclusivamente para anotar direcciones de destinatario y remitente.

• Propaganda en el reverso de tarjetas postales que puede obstruir lectura de direcciones.

• El uso de distintos instrumentos de escritura puede hacer de la lectura un ejercicio aún más difícil.

1.3: El problema del correo postal



Fuente: Pfitze, Behnke, Rojas (2000)

Großbrief-Sortierenlage Siemens

Oficina del Servicio Postal,
(en algún lugar de) Alemania.

150 máquinas en
80 centros del servicio postal alemán

Capacidad para procesar 20K sobres/hr (6 sobres/seg)

85% de certeza

1.3: El problema del correo postal

Una imagen en blanco y negro es capturada por la cámara de video. El problema ahora consiste en encontrar y leer las direcciones en la imagen obtenida.

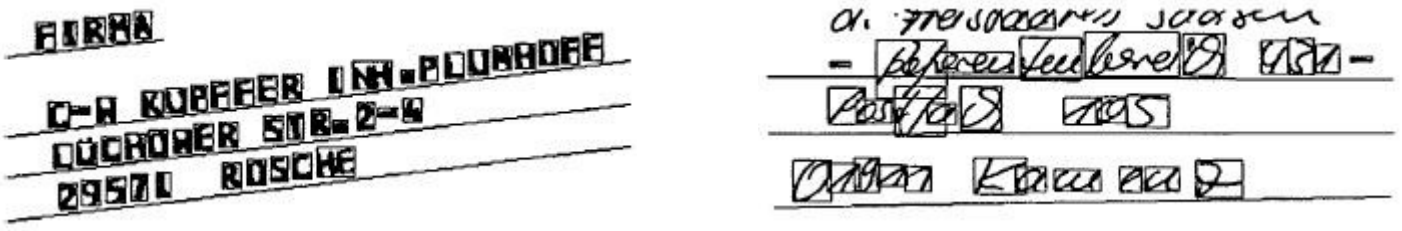
Detección de areas de interés: los pixeles de las imágenes son agrupados en "superpixeles," los cuales son clasificados de acuerdo con características tales como: contenido de rasgos de 1er plano (caracteres escritos, impresos, etc), contenido de rasgos de 2o. plano (e.g., imágenes), etc.

Categorización de las areas de interés: asignación de grado de importancia con base en parámetros tales como su ubicación geométrica dentro del sobre, etc. De acuerdo a su rango, las areas de interés pasan a una segunda etapa de análisis cuya finalidad es localizar la dirección.

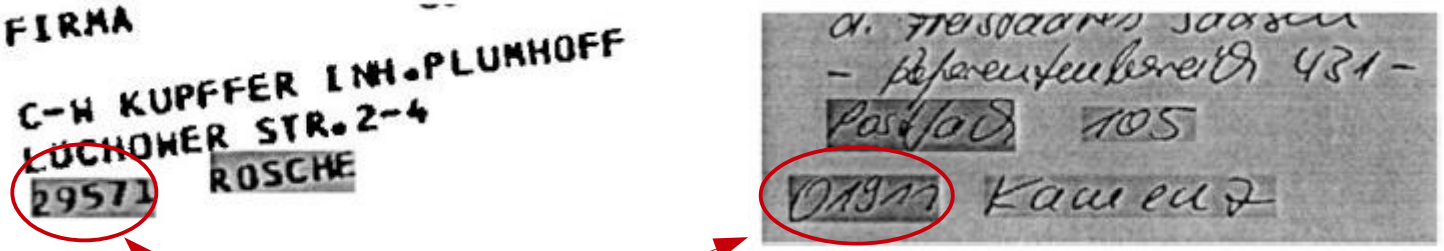
Ubicación del código postal: extracción de las imágenes de 2o. plano del area que contiene la dirección, conversión de la imagen restante a blanco y negro.

Detección de las componentes conexas de la imagen: agrupación de *renglones* de pixeles (marcando inicio y fin),

1.3: El problema del correo postal

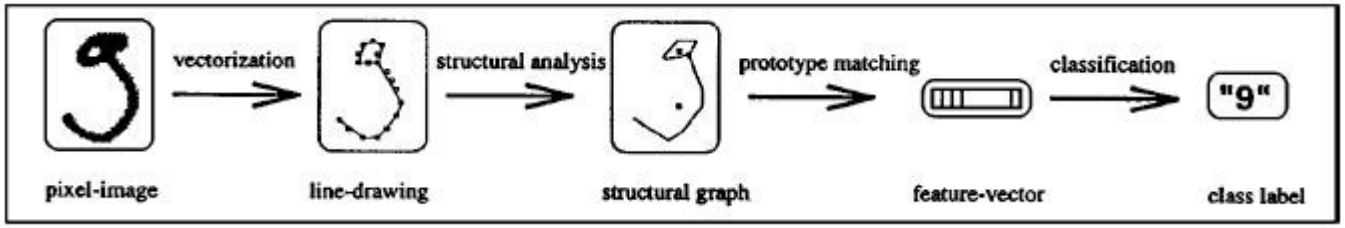


Fuente: Pfitze, Behnke, Rojas (2000)



Segmentación de las areas que contienen el código postal: algoritmos de redes neuronales.

Aplicación de algoritmos de reconocimiento e interpretación de patrones.

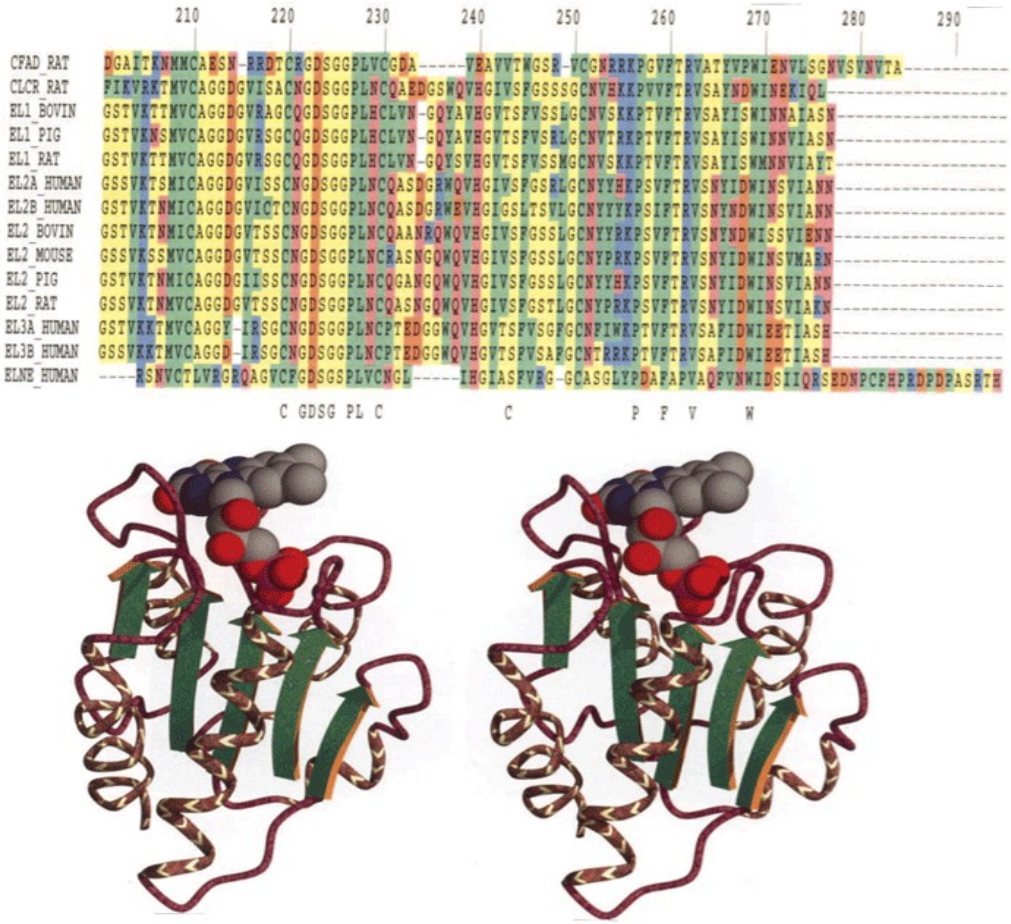


Fuente: Pfitze, Behnke, Rojas (2000)

1.4: Clasificación de proteinas en grupos de homologías

Problema: inferir modelo tridimensional de una proteína, dada la similaridad de su secuencia genética con la secuencia genética de otras proteinas cuyas estructuras son conocidas.

Este es un problema del dominio de la biología estructural computacional. Support Vector Machines (ver más adelante) son una herramienta (entre otras) utilizada con la finalidad de determinar grupos de homología entre proteinas.



Fuente: Bader et al, Ctwatch Quarterly, Nov. 2006 B

1.4: Clasificación de proteinas en grupos de homologías

Algunas referencias y sitios de interés:

Molecular Modeling of Proteins & Nucleic Acids. Dept. of Biochemistry, Virginia Tech.

1. Saigo, H.; Vert, J-P; Ueda, N.; Akutsu, T. Protein homology detection using string alignment kernels. *Bioinformatics* **20,** no. 11 (2004) pp. 1682-1689.
(Datos y programas suplementarios)

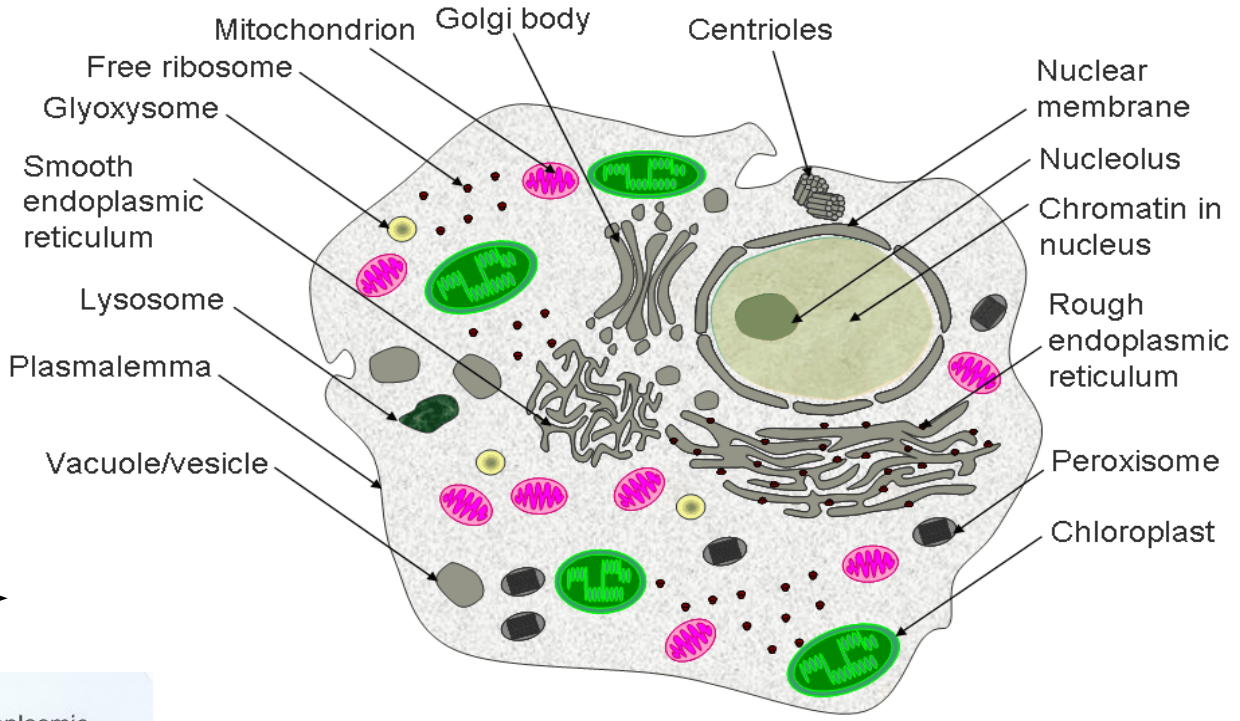
2. Sonego, P; et al. A protein classification benchmark collection for machine learning. *Nucleic Acids Research* ***35*** (2007)
The Protein Classification Benchmark Collection (ICGEB/EMBnet)

3. Rangwala, H.; Karypis, G. Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics* **21,** no. 23 (2005) pp. 4239-4247

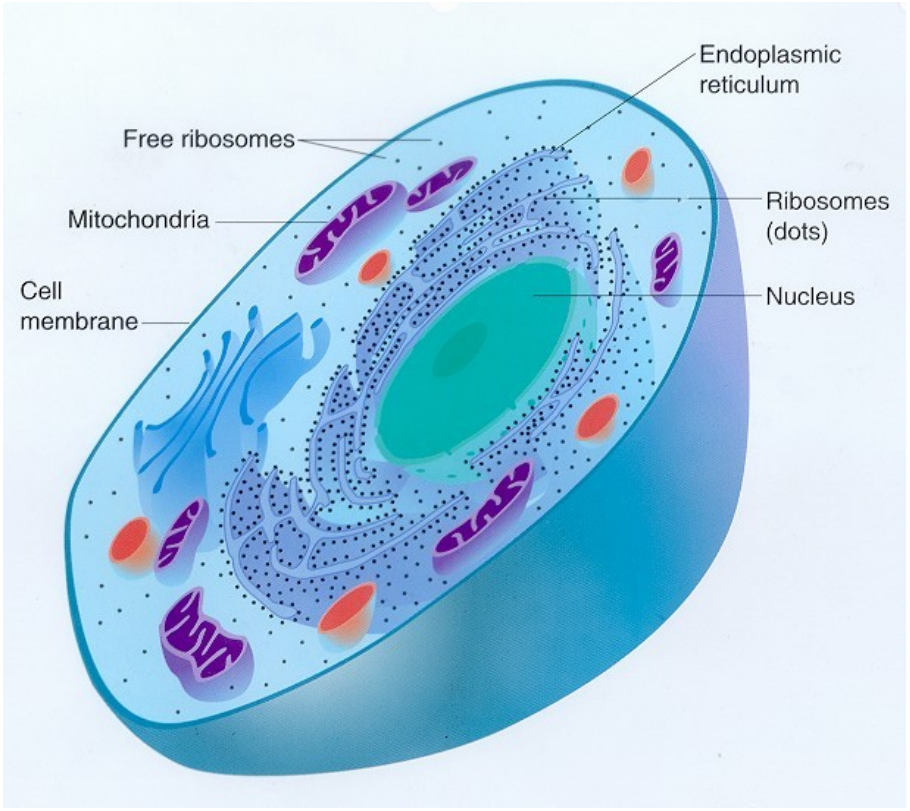
4. Lingner, T.; Meinicke, P. Remote homology detection based on oligomer distances. *Bioinformatics* **22,** no. 18 (2006) pp. 2224-2231

1.5: Expresión genética

1.5.1 Preliminares



~100µm



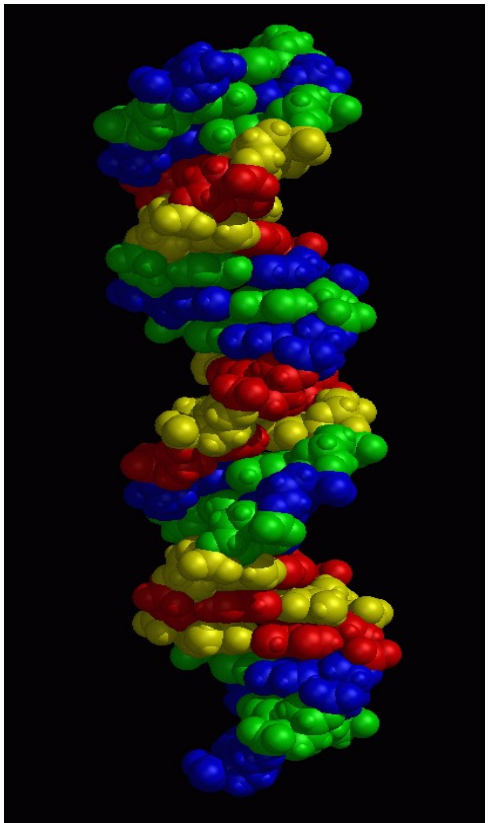
Típicas células Eucariotas ("Eukaryotic"), animales, plantas y hongos ("fungi"). Interior compartamentalizado.

Fuentes:
Fluwiki Glossary
http://www.fluwiki.info/pmwiki.php?n=Science.Glossary
Williams Class
http://www.williamsclass.com/SeventhScienceWork/CellTheoryParts.htm

1.5.1 Preliminares: glosario de términos y conceptos básicos

Moléculas = cadenas más pequeñas de átomos unidos mediante enlaces covalentes.
También llamadas *polímeros* = cadenas covalentes de *monómeros*.

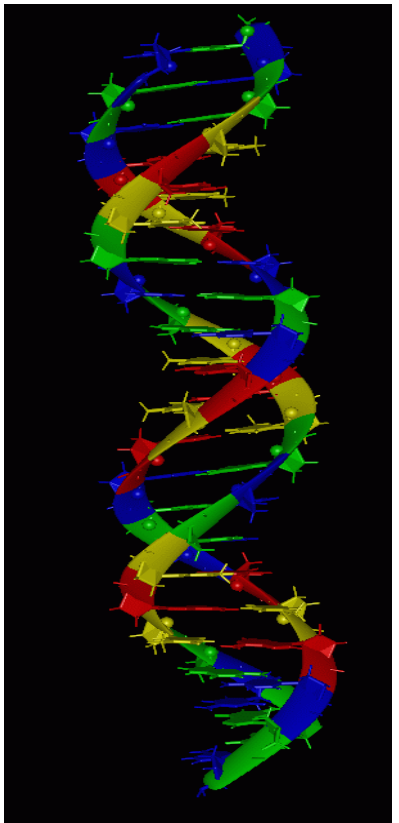
ADN = (Ácido desoxiribonucléico/DNA: deoxyribonucleic acid). "El genoma de la célula,"
Contiene la mayor parte de la información hereditaria de la célula.
Cadena molecular lineal compuesta por 4 tipos de nucleótidos: **A**denina (Adenine), **C**itosina (Cytosine), **G**uanina (Guanine), **T**iamina (Thymine). Tiene orientación y puede ser descrita mediante un texto sobre un alfabeto de cuatro letras.



Modelos de visualización del ADN

Modelo "spacefill": cada átomo está representado por una esfera

Modelo covalente: cada peldaño es un enlace covalente entre átomos pesados.



fuente:
http://www.imb-jena.de/image_library/DNA/DNA_models/B-DNA/bdna.html

1.5.1 Preliminares: glosario de términos y conceptos básicos

Uniones entre nucleótidos (complementarios)

AT	**GC**
2 enlaces de hidrógeno	3 enlaces de hidrógeno

cADN (cDNA) ("molécula complementaria del ADN") = secuencia de bases complementarias a las bases asociadas con una secuencia de ADN molecular, en orden inverso.

GGGGaaaaaaaatttatatat → atatataaatttttttCCCC

(secuencia de ADN) (secuencia complementaria invertida o cADN)

GGGGaaaaaaaa t t t a t a t a t
CCCC t t t t t t t t aaa t a t a t a

} Estructura de la doble hélice

Hibridización = unión de dos moléculas complementarias de ADN para formar la doble hélice.

El genoma humano contiene más de 3 billones de nucleótidos y se encuentra separado en 23 moléculas de ADN, cada una de las cuales mide 5.0 cm aprox (5000 veces el diámetro de la célula).

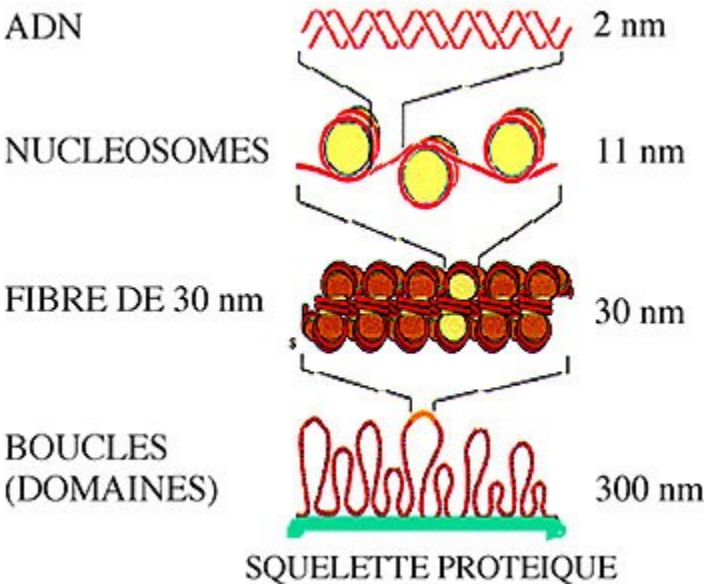
http://www.bioinformatics.org/sms/rev_comp.html

1.5.1 Preliminares: glosario de términos y conceptos básicos

ADN se enrolla aprox. 1.5 veces sobre proteinas llamadas histones (adheridas al ADN) para formar estructuras llamadas nucleosomas (“nucleosomes”).

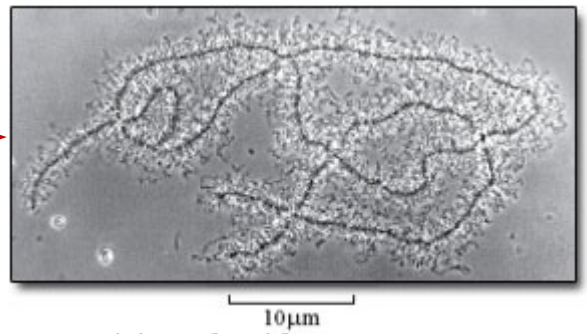
Los nucleosomas se apilan unos sobre otros para formar una estructura fibrosa llamada cromatina (“chromatin”).

A su vez, a lo largo de su estructura, las cromatinas forman circuitos (rollos) formando a los cromosomas, los cuales se encuentran anclados en regiones específicas.

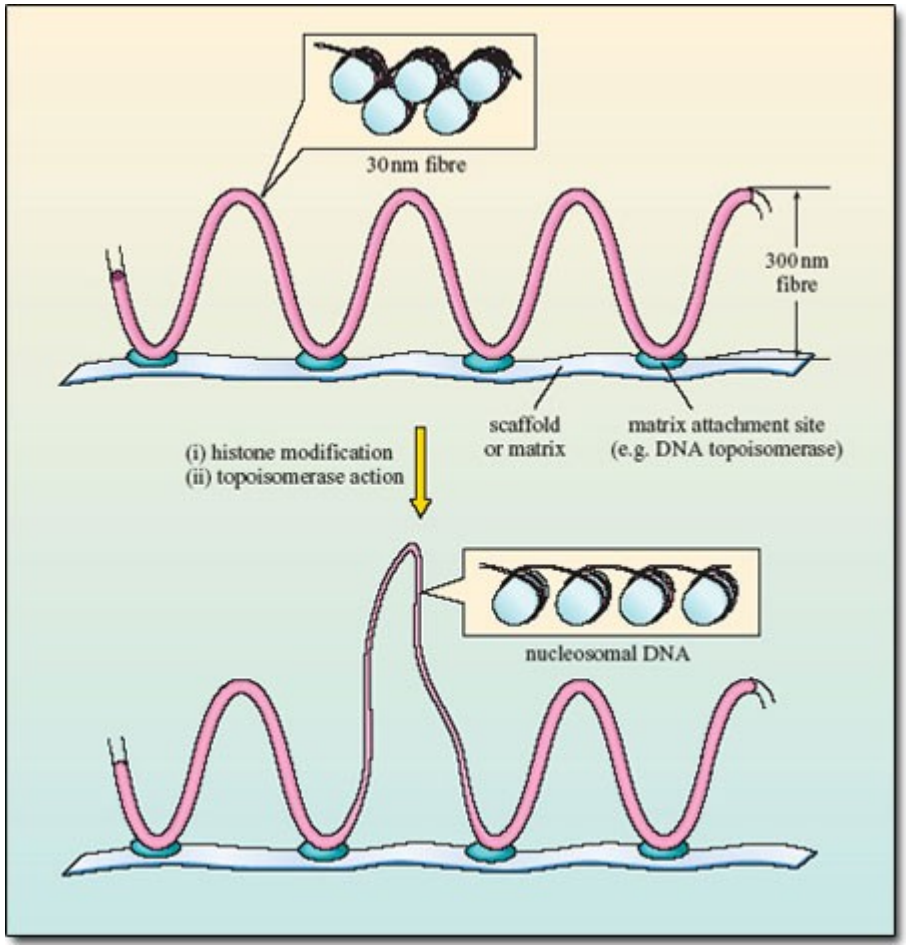


Fuente: http://www.igr.fr/index.php?p_id=737

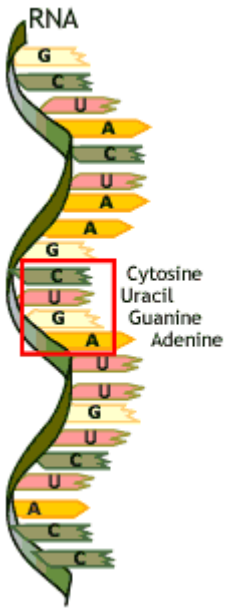
Lampbrush chromosomes from amphibian oocyte



Fuente: http://openlearn.open.ac.uk/mod/resource/view.php?id=172631



1.5.1 Preliminares: glosario de términos y conceptos básicos



fuente

ARN (RNA) (Ácido ribonucleico) = cadena de nucleótidos que usa al nucleótido **U**racil en lugar de la **T**iamina (ADN) . Usualmente consisten en una sola cadena o hélice, lo cual les proporciona versatilidad para formar estructuras con funciones complejas (ribosomas). De acuerdo con las funciones que desempeñan, hay 3 clases principales de ARNs:

Mensajero (mARN) – transfiere información sobre (parte de) la secuencia de aminoácidos en el ADN para la síntesis de proteinas.

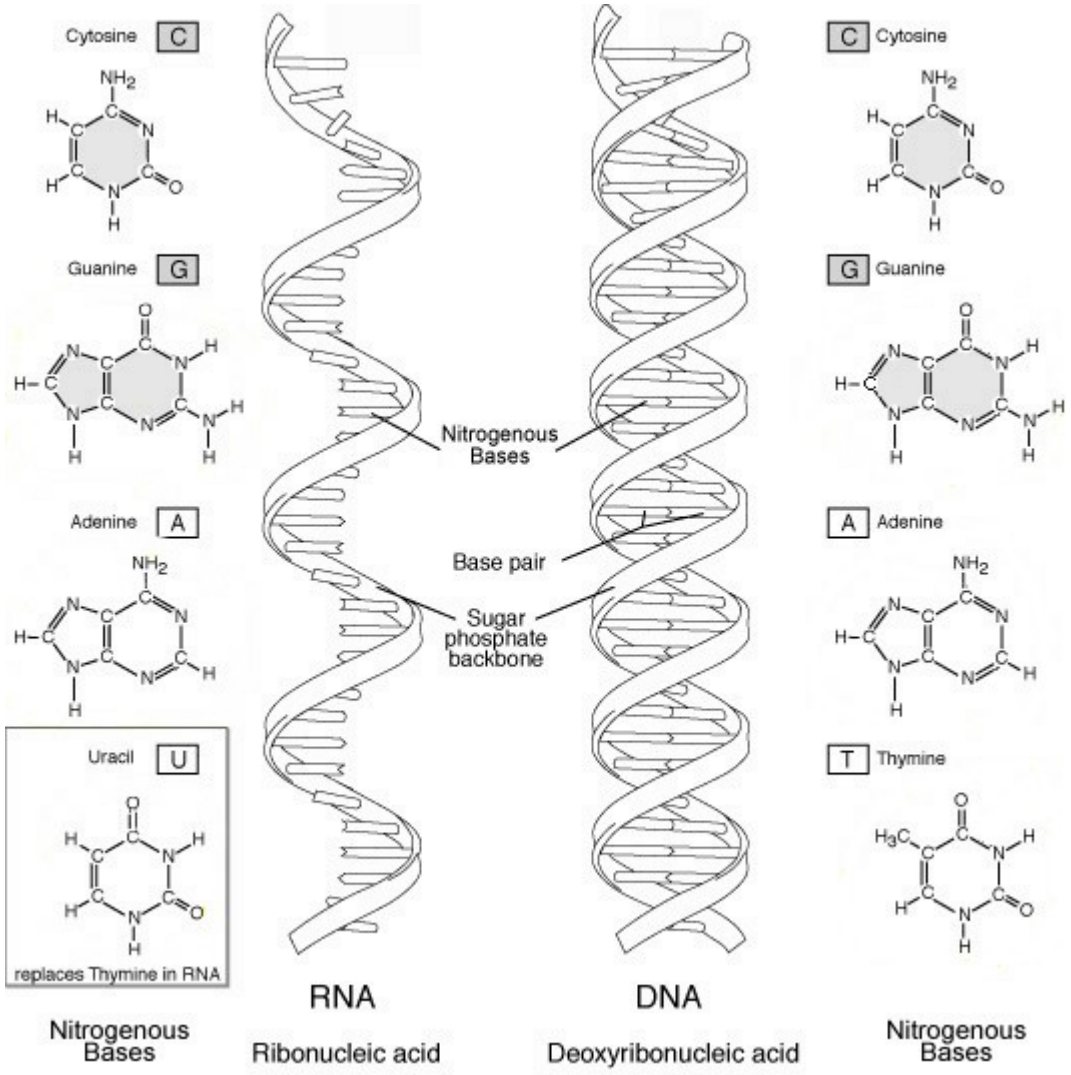
Ribosomal (rARN) – junto con proteinas, constituye al ribosoma.

Transferencia (tARN) – transfiere aminoácidos al ribosoma para la síntesis de proteinas (= polímeros compuestos por aminoácidos).

http://nobelprize.org/educational_games/medicine/dna/index.html

1.5.1 Preliminares: glosario de términos y conceptos básicos

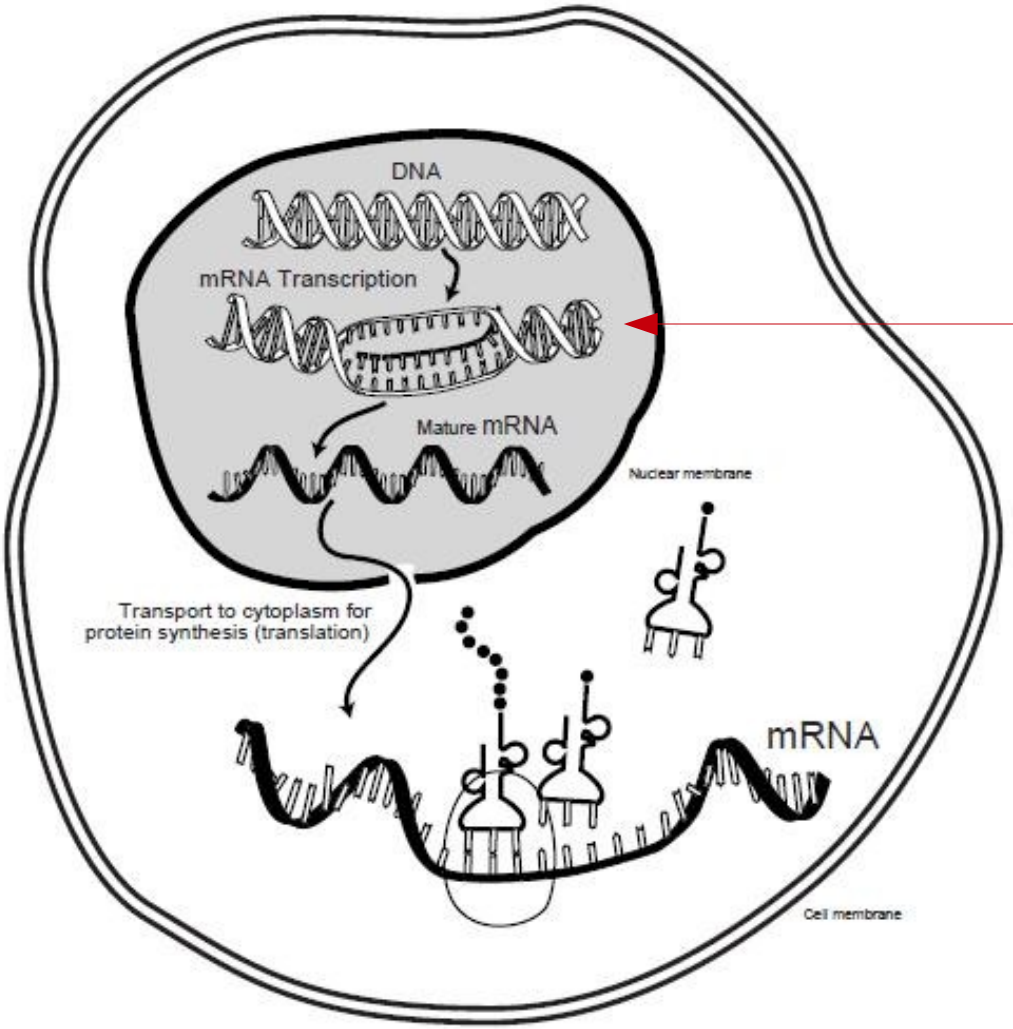
ARN

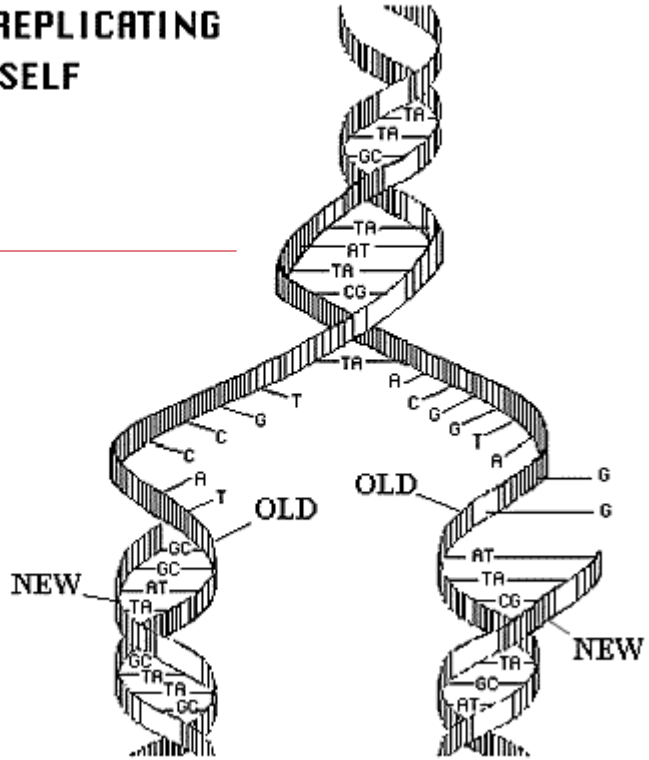


Fuente: http://www.accessexcellence.org/RC/VL/GG/

1.5.1 Preliminares: glosario de términos y conceptos básicos

mARN

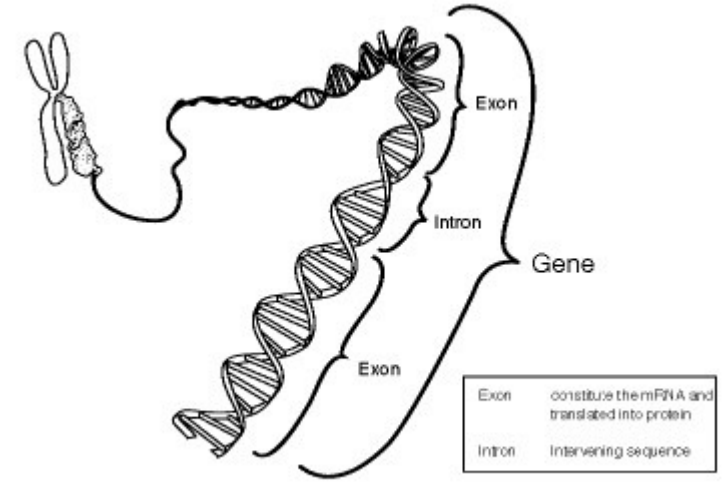




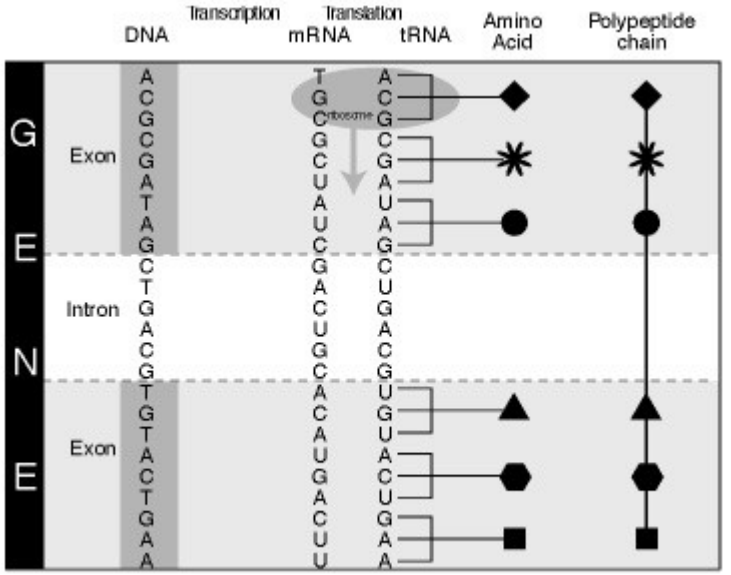
Fuente: http://www.accessexcellence.org/RC/VL/GG/

1.5.1 Preliminares: glosario de términos y conceptos básicos

Gen = subcadena del genoma responsable de la producción de uno o dos tipos de ARNs, su función principal es la de codificar instrucciones para la producción de proteínas.



Expresión genética (gene expression) = las primeras dos etapas del proceso de síntesis de RNA complementario a una cadena de ADN.



Fuente: http://www.accessexcellence.org/RC/VL/GG/

1.5.1 Preliminares: glosario de términos y conceptos básicos

El proceso de expresión de genes y proteinas (síntesis y procesamiento de proteínas)

Expresión de genes

1. Transcripción: producción de pre-mARN en el núcleo de la célula.

2. Juntura ("Splicing"): partes del pre-mARN son removidas ("introns"). Las partes restantes ("exons") son reconectadas para formar el mARN maduro el cual viaja fuera del núcleo a través de la membrana doble de este último.

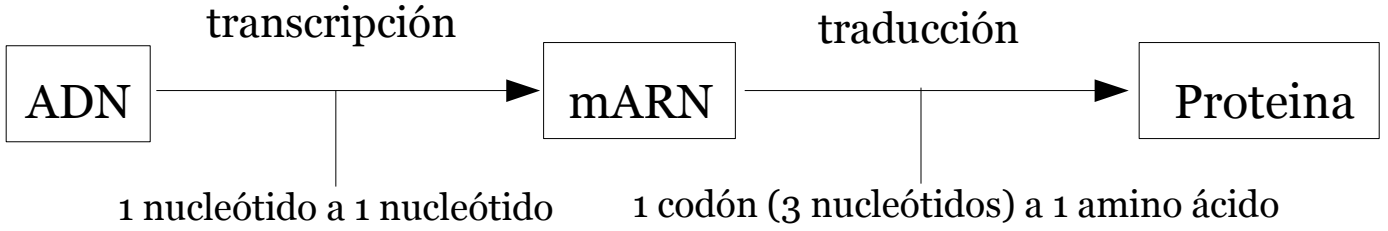
3. Traducción ("Translation"): ribosomas aguardan la llegada de los mARN para sintetizar proteinas.

4. Modificación(*).

5. Transporte ("Translocations")(*)

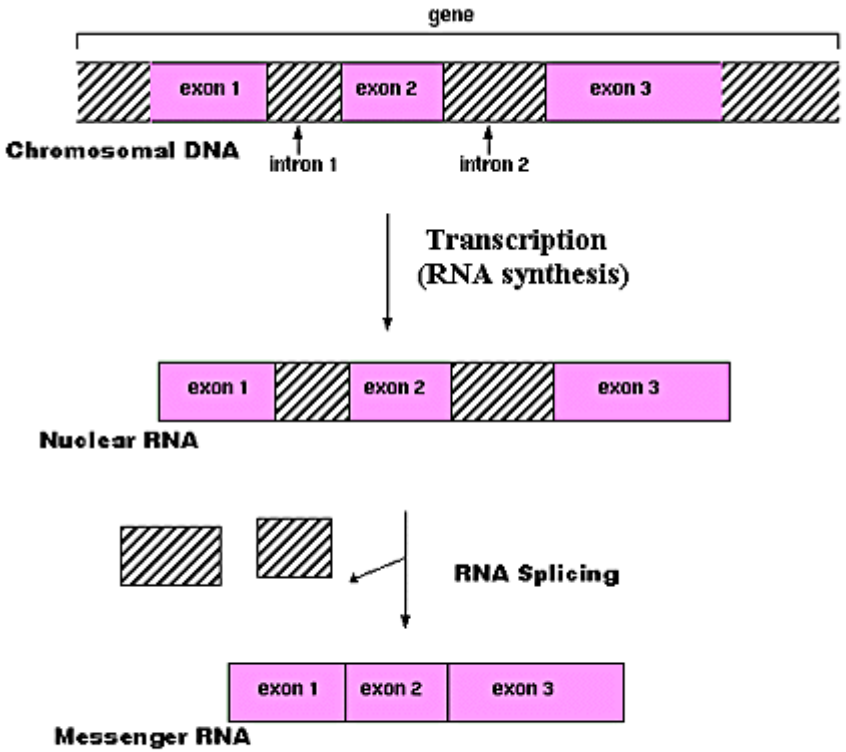
6. Degradación: ruptura (digestión) de proteinas en sus amino ácidos.

Expresión de proteinas

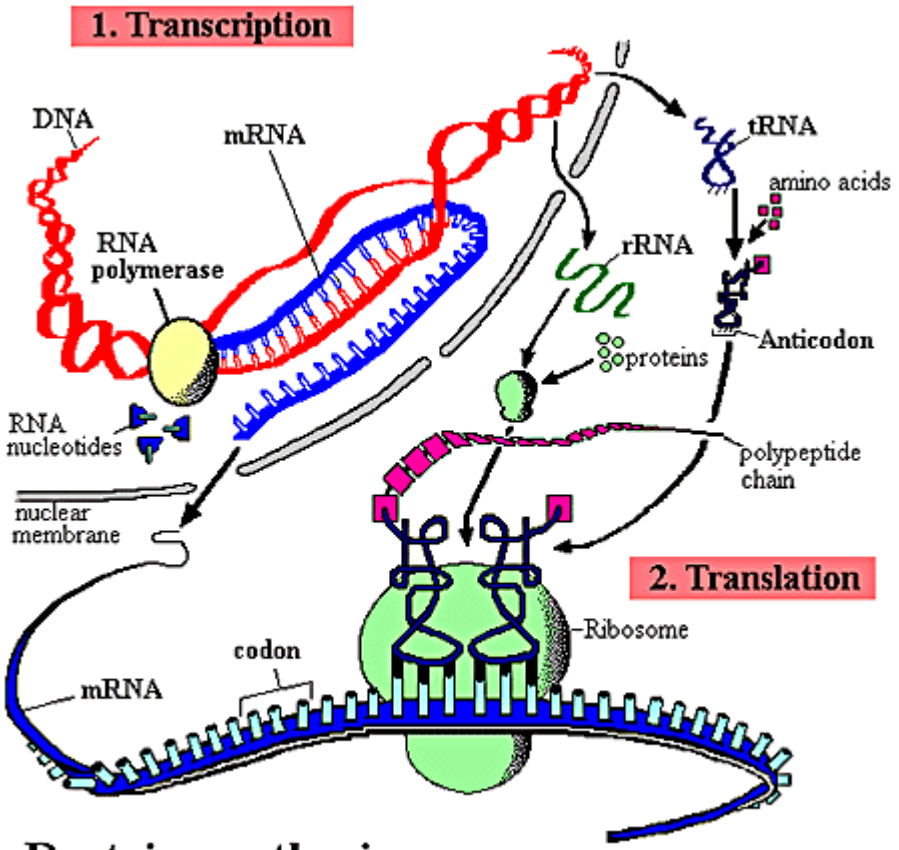


1.5.1 Preliminares: glosario de términos y conceptos básicos

Expresión de genes y proteinas



RNA synthesis and processing



Protein synthesis

Fuente: http://www.accessexcellence.org/RC/VL/GG/protein_synthesis.php

1.5.1 Preliminares: glosario de términos y conceptos básicos

Datos de expresión genética ("gene expression data") = arreglos rectangulares de valores numéricos (matrices!) que representan los niveles de expresión para cada gen en un cierto conjunto a lo largo de una colección de muestras celulares.

Metodologías para la medición de valores de expresión genéticos:

1. **Microarreglos** ⬅

2. Otras tecnologías: (utilizan los siguientes pasos...)

   Secuenciamiento ("sequencing"): escoger aleatoriamente moléculas de mARN, revertir su transcripción para generar cDNA, amplificar y determinar sus secuencias.

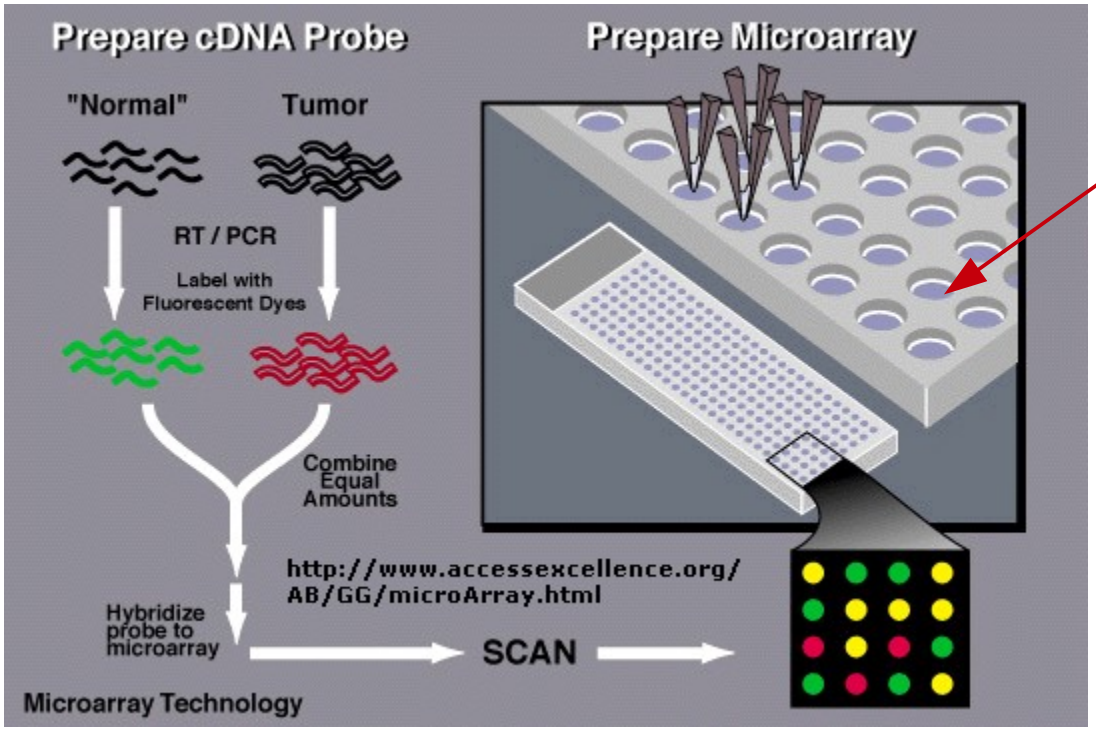
   Agrupación ("clustering"): en secuencias que corresponden a un mismo gen.

   Conteo: el tamaño de los grupos determina el nivel de expresión de sus correspondientes genes.

1.5.1 Preliminares: glosario de términos y conceptos básicos

Microarreglo = superficie plana que contiene un arreglo rectangular de “lugares” (“spots”) (aprox. 2000) cada uno de los cuales está asociado con un gen predeterminado.

Los microarreglos utilizan la técnica de hibridación para determinar, simultaneamente, los niveles de expresión de distintos genes en varias muestras celulares. Los genes cuyos niveles de expresión se quieren definir deben de ser conocidos de antemano.



En cada lugar hay un fragmento de cADN (oligonucleótido) complementario a una subsecuencia específica de un solo gen.

Renglones = genes

Columnas = muestras (pacientes o experimentos)

Fuente: http://employees.csbsju.edu/hjakubowski/classes/ch331/bind/oldrugdevel.html

1.5.1 Preliminares: glosario de términos y conceptos básicos

El proceso (tecnológico) de hibridación:

1. Transcripción inversa de mARN de muestra celular a cADN.

2. cADN es etiquetado (fluorescente o radioactivo).

3. Aplicación de las muestras transcritas y etiquetadas sobre el chip microarreglo (toma lugar hibridación de las muestras con las cadenas complementarias (cADN) en cada lugar del chip). *# hibridaciones proporcional a concentración de* mARN *celular*.

4. Lavado del arreglo.

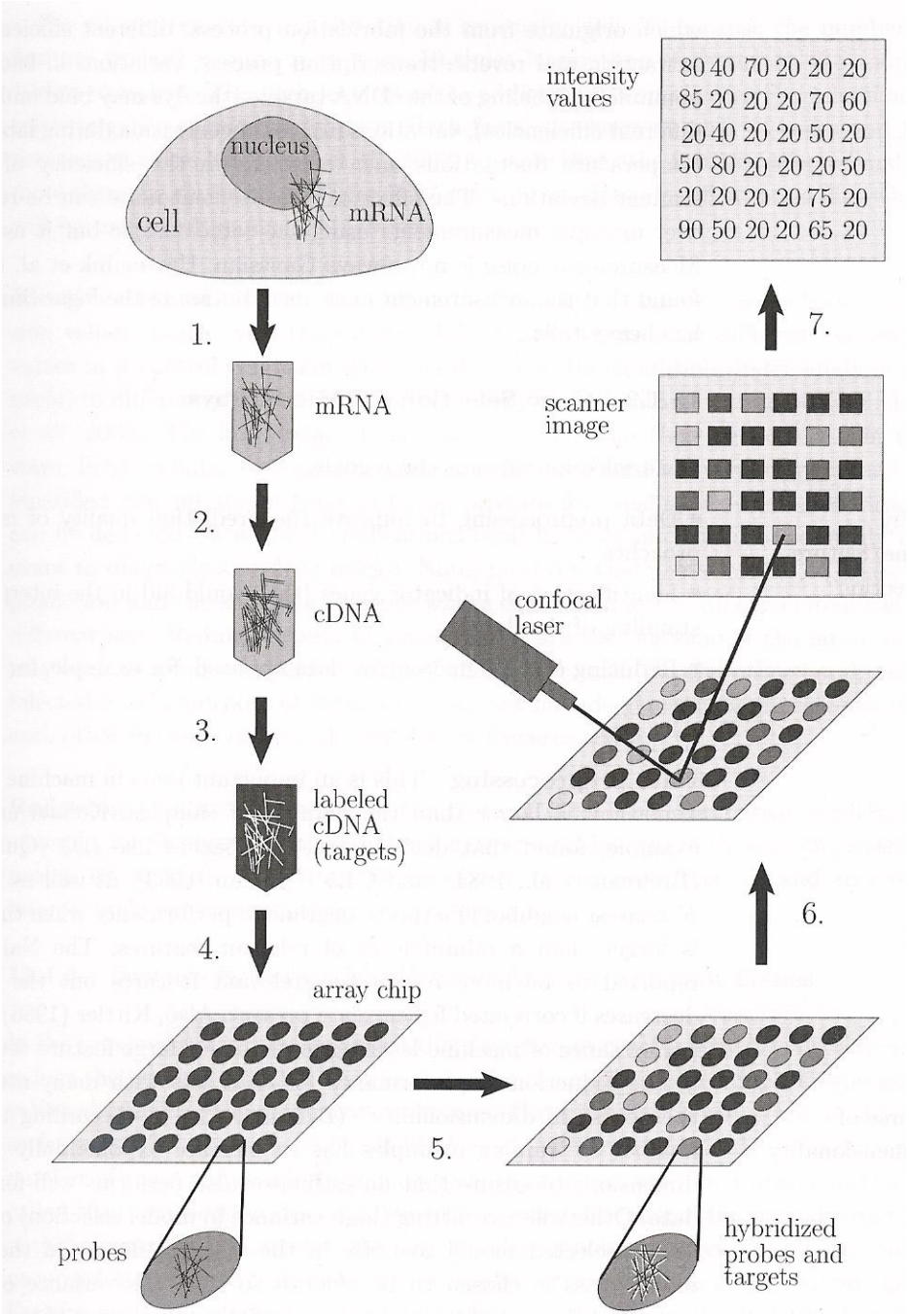
5. Medición de la intensidad de la señal emitida por la etiqueta para determinar la concentración de mARN en la muestra.

Chip microarreglo ("microarray chip" o "DNA chip") con capacidad de medir decenas de miles de expresiones genéticas simultáneamente.

Precio (aprox): US$1,000 (en el 2006!).

Fuente: http://www.carleton.ca/catalyst/2006s/hms7.html

1.5.1 Preliminares: glosario de términos y conceptos básicos



La técnica del microarreglo (otra vez)

(Southern, 1988; Lysov et al, 1988;
Drmanac et al 1989; Bains & Smith, 1988)

1. Extracción de mARN.

2. Transcripción inversa (mARN→cADN)

3. Etiquetado

4. Aplicación de muestras al chip

5. Hibridación

6. Medición de la intensidad de la señal emitida por la etiqueta

7. Procesamiento de mediciones (corrección por intensidad de fondo, etc)

Fuente: Schölkopf, Tsuda, Vert (eds);
Kernel methods in computational biology (2004)

1.5.1 Preliminares: glosario de términos y conceptos básicos

Hybstation por Digilab

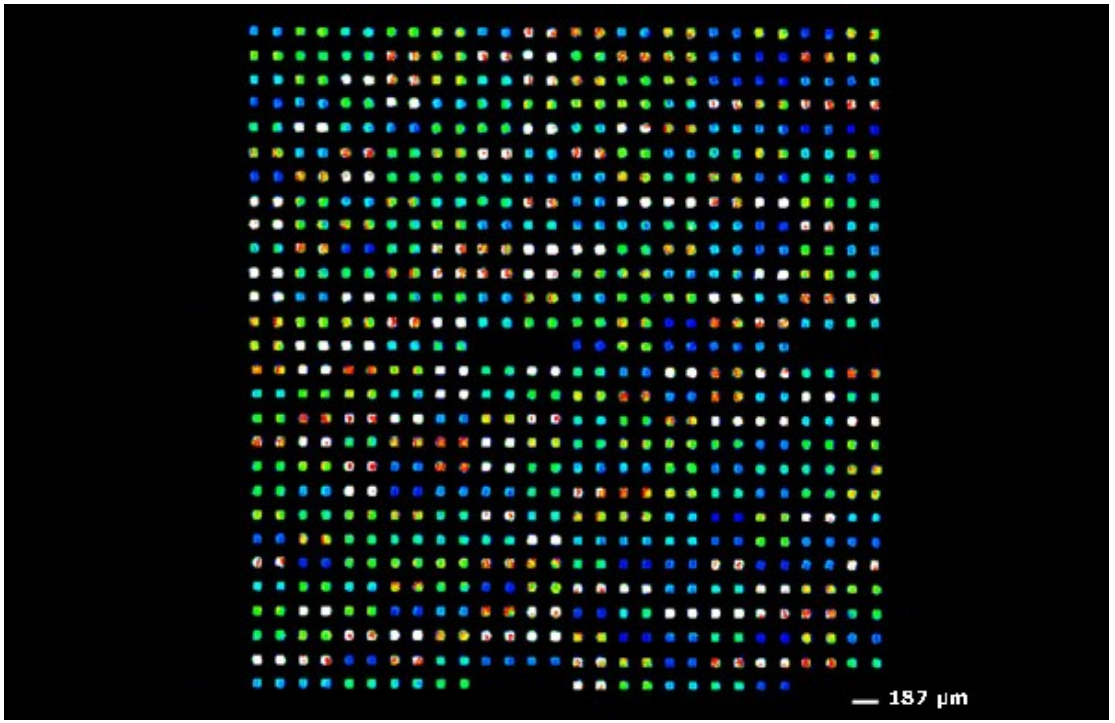
Ancho: 61 cm. Profundidad: 53.4 cm. Altura: 58.5 cm

Fuente: http://www.digilabglobal.com/UploadedFiles/Files/11DHybstationBrochure.pdf

1.5.1 Preliminares: glosario de términos y conceptos básicos

Se estima que el genoma completo del ser humano contiene entre 20K a 25K genes[1]

Estimación del tamaño de la matriz de datos.



10K - 22K

 # pacientes, apenas del orden de cientos

[1] Human Genome Project

Fuente: http://www.arrayit.com/Products/DNA_Microarrays/DNA_Microarrays_DCH/dna_microarrays_dch.html

1.5.1 Preliminares: glosario de términos y conceptos básicos

¿Cuáles son los objetivos principales al resolver problemas en esta area?

1. Determinación del diagnóstico y prognosis de una enfermedad así como tratamiento.

2. Identificación de los genes causantes de una enfermedad y estimación de la probabilidad de que un individuo la desarrolle.

3. Prevención (cura) de enfermedades (congénitas, neurológicas, cáncer, etc ...)

Los objetivos requieren del análisis y clasificación de datos (trabajo del matemático, ingenieros en computación, biomedicina, bioinformática, etc.)

Algunos problemas difíciles de tratar a la vez que interesantes e importantes:

a. contaminación de datos: *biológica* – muestras no se encuentran en el mismo "estado de expresión;" *tecnológica* – propiedades de tolerancia del chip (fabricación); *técnica* – deficiencias del método de transcripción, etc

b. dificultades computacionales – las matrices de datos son muy grandes!!, (lo cual afecta el desempeño de los algoritmos de clasificación).

c. selección de genes relevantes ("feature selection" o "gene selection").

2. Análisis de Componentes Principales (PCA)

2.1: Caso Lineal

Generalidades:

a. Concierne al álgebra lineal.

b. Determina las direcciones en el espacio ambiente ("input space") de variación máxima de los datos.

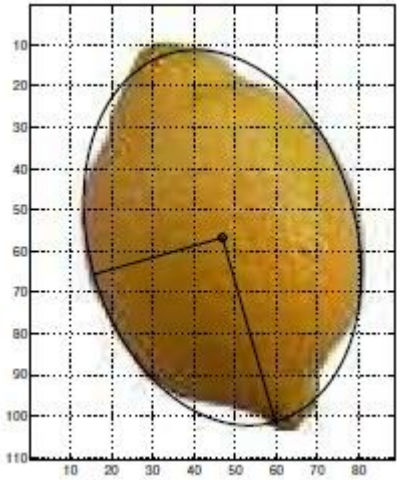
c. La 1a componente principal es la dirección a lo largo de la cual la varianza de los datos es máxima.

d. La 2da componente principal es una dirección perpendicular a la primera a lo largo de la cual la varianza de los datos es máxima (dirección de varianza máxima en el complemento ortogonal de la primera componente).

e. Hay tantas componentes principales como dimensiones tenga el espacio ambiente.

f. Limitaciones geométricas y computacionales.

2.1: PCA: caso lineal



Fuente: Wijewickrema & Paplinski (2004)

Wijewickrema & Paplinski (2004)

Problema: ajustar una elipse a imágenes bidimensionales de objetos aproximadamente convexos, de forma tal que la elipse aproxime "lo más fielmente posible" el contorno del objeto en la imagen.

Algoritmo:

1. Adquisición y procesamiento de los datos (imágenes).

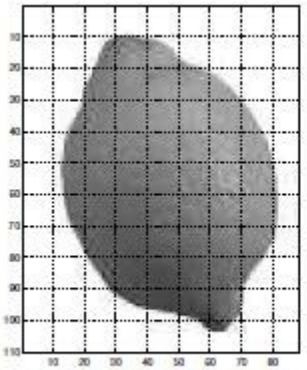
2. Cálculo del promedio de los datos (centro de la figura = centro de la elipse aproximante)

3. Cálculo de la matriz de covarianza (matriz cuadrada de tamaño dos)

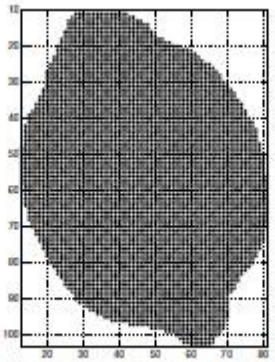
4. Cálculo de los eigenvectores y eigenvalores de la matriz de covarianza (componentes principales).

2.1: PCA: el caso lineal

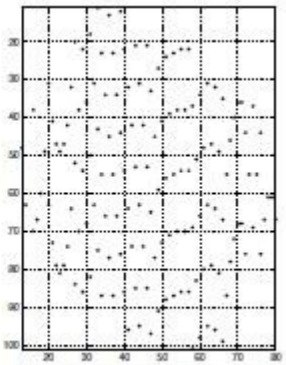
Fuente: Wijewickrema & Paplinski (2004)



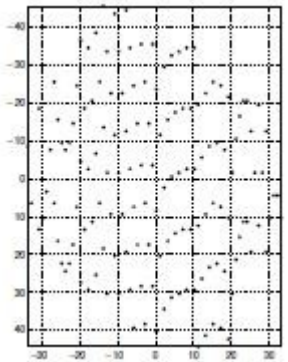
(a) Grayscale image



(b) Segmented image



(c) Decimated image



(d) Distribution after mean removal

Adquisición y procesamiento de imágenes:

1. extraer un conjunto de puntos representativos: convertir imagen a escala de grises. (iluminación uniforme y fondo claro c.r. al color del objeto)

2. Discriminación básica por umbral del brillo ("global brightness thresholding").

3. Segmentación de imagen.

4. Decimación ("decimation"): de cada segmento de imagen seleccionar un subconjunto de puntos (aleatoriamente o con algún otro método). En la figura se seleccionó uno de cada p pixeles, en donde p es coprimo con el número de renglones (o columnas).

5. Remover el promedio de los datos (traslación o centrado de la imagen)

2.1: PCA: el caso lineal

Programas (Matlab) disponibles en Wijewickrema & Paplinski (2004)

Matriz de datos $X = [X_1 \dots X_m] \in \mathbb{R}^{n \times m}$ *(n=2)* $\sum_{i=1}^{m} X_i = 0$ (datos centrados)

Matriz de covarianza: (medida de la correlación entre los componentes de los datos)

$$C = \frac{XX'}{m-1} \in \mathbb{R}^{n \times n}$$

Ejercicio:
(a) *C* es una matriz cuyas componentes son las sumas de todos los productos entre las componentes de los datos.
(b) *C* es simétrica y positiva definida.
(c) Para *n=2* verificar que:

$$C = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \qquad \sigma_{ii} = \frac{1}{m-1} \sum_{k=1}^{m} (X_k(i))^2 \qquad \sigma_{12} = \sigma_{21} = \frac{1}{m-1} \sum_{k=1}^{m} X_k(1) X_k(2)$$

i=1,2

2.1: PCA: el caso lineal

Ejercicio (cont.):
(d) verificar las sig. formulas explícitas para los eigenvalores de *C:*

$$\lambda_{1,2} = \frac{1}{2}\left(\sigma_{11}\,\sigma_{22} \pm 1\sqrt{(\sigma_{11}-\sigma_{22})^2 + 4\,\sigma_{12}^2}\right)$$

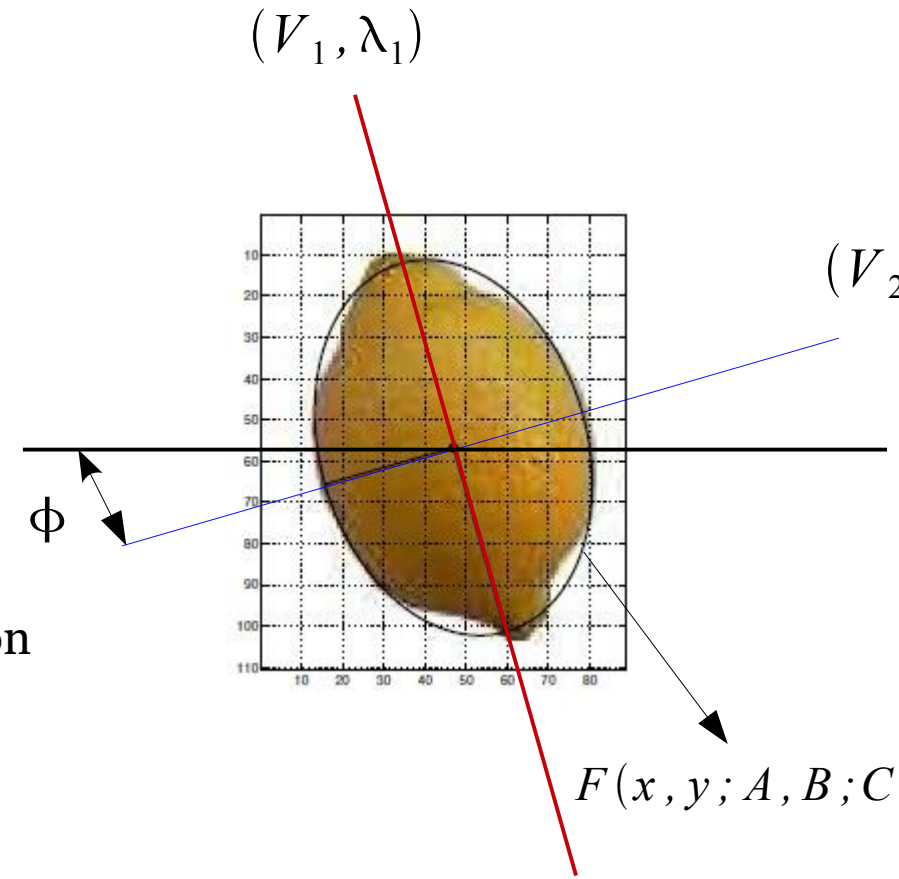
(e) verificar las sig. formulas explícitas para los eigenvectores de *C*:

$$V_1 = \begin{bmatrix} x \\ y \end{bmatrix}, \quad V_2 = \begin{bmatrix} y \\ -x \end{bmatrix} \qquad \text{en donde} \quad x = \frac{\sigma_{12}}{\sqrt{(\lambda_1 - \sigma_{11})^2 + \sigma_{12}^2}}, \quad y = \frac{\lambda_1 - \sigma_{11}}{\sqrt{(\lambda_1 - \sigma_{11})^2 + \sigma_{12}^2}}$$

Obs: (el problema con la dimensionalidad alta) comparado con la cantidad de procesamiento de las imágenes, el cálculo de los eigenvals. y eigenvecs. de *C* requiere de un mínimo de cálculos (caso 2-diml). En dims. mayores que dos (datos tipo microarreglo!) este mismo cálculo suele ser "caro" en términos computacionales, más aún, está sujeto a errores numéricos, etc.

2.1: PCA: el caso lineal

Ajuste de la elipse:



$\lambda_1 \geqslant \lambda_2 \geqslant 0$

$$\tan(\phi) = \frac{x}{y}$$

ánglulo de inclinación

$$F(x, y; A, B; C; D) = Ax^2 + 2\mathrm{Bxy} + Cy^2 + D = 0$$

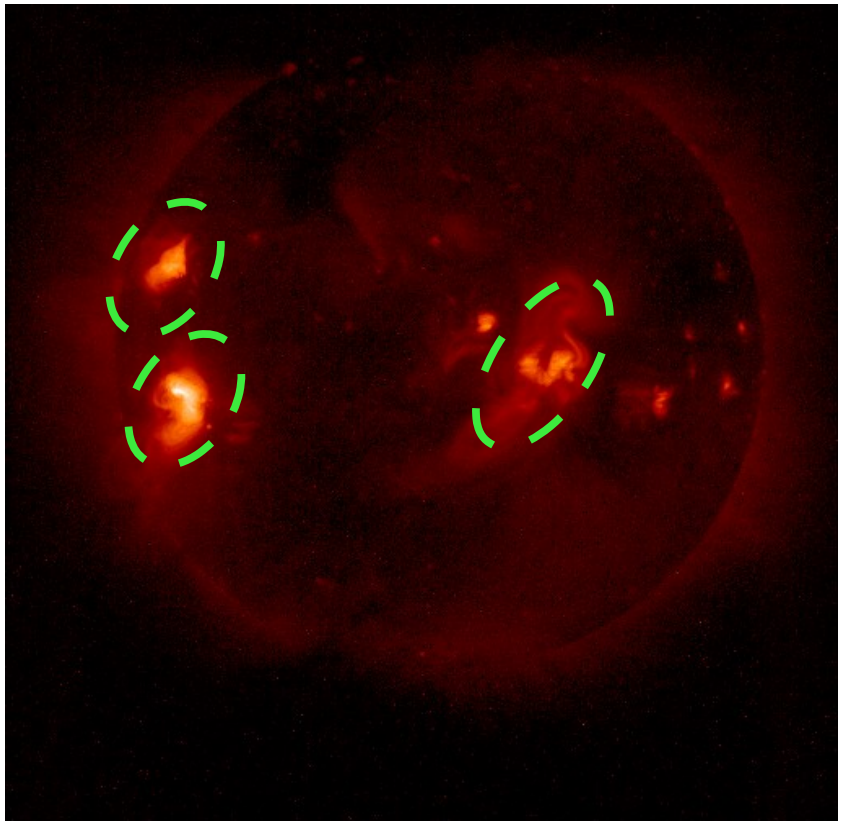
o paramétricamente:

$$x(t) + iy(t) = 2\left(\sqrt{\lambda_1}\cos(t) + i\sqrt{\lambda_2}\sin(t)\right)e^{i\Phi}$$

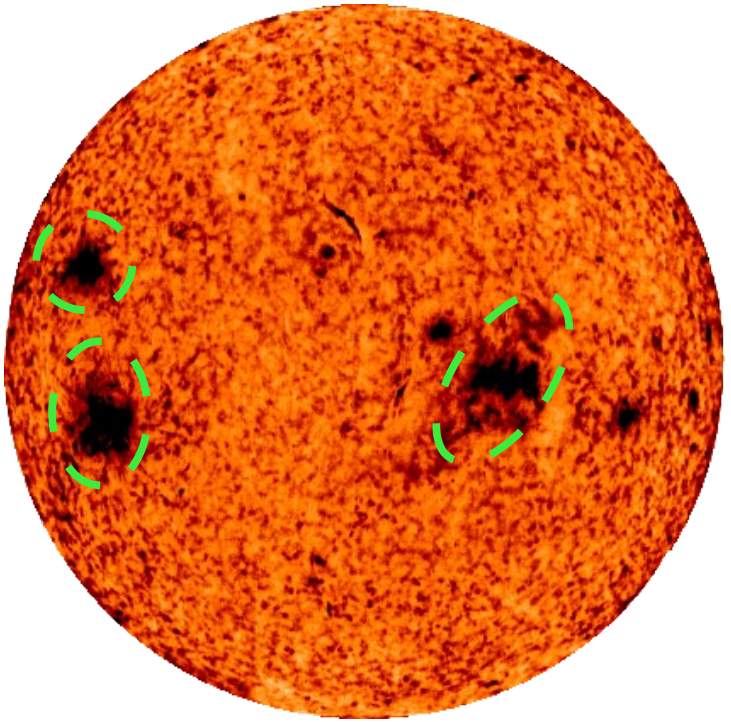
Fuente: Wijewickrema & Paplinski (2004)

2.1: PCA: el caso lineal

Aplicaciones potenciales: *análisis automatizado de imágenes*
Problema: se tiene una gran cantidad de imágenes, c/u es segmentada en subimágenes con el propósito de encontrar y categorizar sus características (sin intervención humana)



Fotografía de rayos X de la corona solar fuente: vathena.arc.nasa.gov

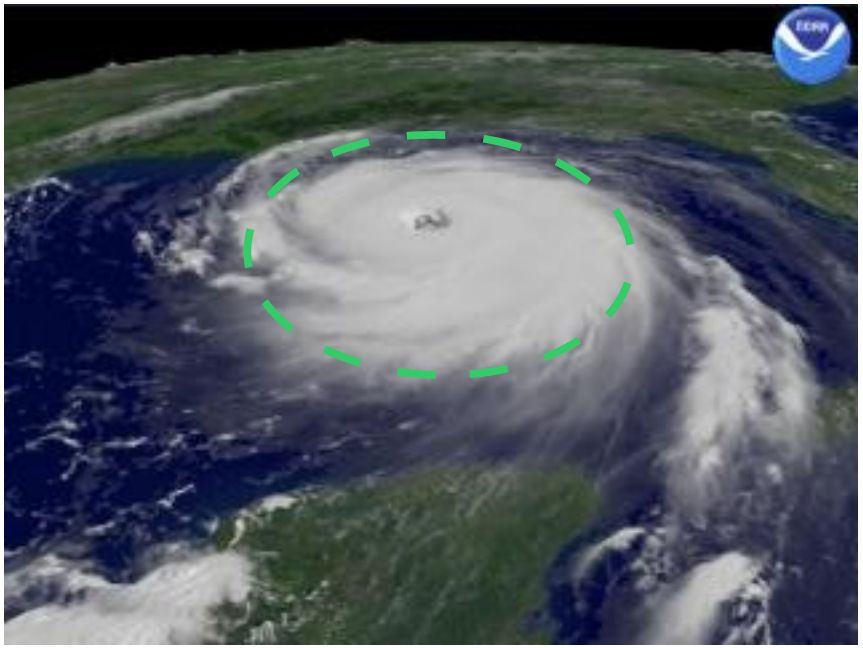


Fotografía infraroja de la corona solar
Fuente: vethena.arc.nasa.go

Observación del Sol durante un periodo de tiempo *T*. **Objetivo:** determinar zonas altamente activas y agujeros en la corona solar así como la formación de estos.

2.1: PCA: el caso lineal

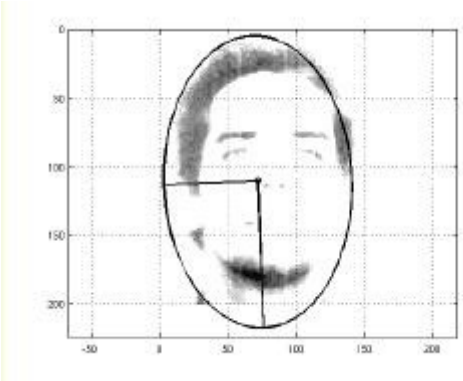
Detección y monitoreo de fenómenos meteorológicos.



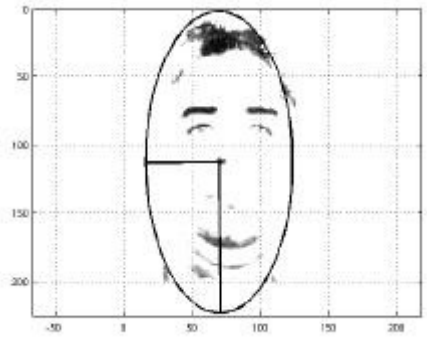
Huracán Katrina (2005) latitud: 23:13:59N, longitud: 88:08:03W
Fuente: nnvl.noaa.gov

2.1: PCA: el caso lineal

Localización de rostros en imágenes y cálculo de sus parámetros dimensionales



Glen Stewart Godwin
10 más buscados por le FBI (2006)
(asesinato y tráfico de drogas)
fbi.gov



(En el ajuste de la elipse se utilizó el código Matlab en Wijewickrema y Paplinsky (2004).)

2.1: PCA: el caso lineal

Definición de la medida del error del ajuste:

B = conjunto de puntos en el perímetro de la figura decimada.
E = error del ajuste:

$$E := \frac{1}{|B|} \sum_{X \in B} F^2(X\,;A\,,B\,,C\,,D)$$

Nota: E depende del parámetro d empleado cuando se decimó la figura. Este parámetro debe escogerse de manera óptima tal que mínimice E y mantenga el tiempo de procesamiento de las imágenes dentro de un rango razonable.

En las simulaciones de Wijewickrema & Paplinksi,

$$d = \left\lfloor \frac{R \times (3 - \sqrt{5})}{2} \right\rfloor$$

2.1: PCA: el caso lineal

Comentarios al trabajo de Wijewickrema & Paplinski

1. En el ajuste de la elipse solamente son de relevancia los puntos sobre el borde de la figura. El costo de la producción de una figura decimada podría reducirse substancialemente si solamente se trabaja con puntos en el borde de la figura.
La implementación del método de W&P con algoritmos de detección de bordes de imágenes no se ha hecho.

2. El error de ajuste de W&P no está confinado a un intervalo, por lo tanto no proporciona una medida acerca de la "bondad del ajuste" ("goodness of fit"). Se propone entonces la siguiente modificación:

$$\text{Error} = \frac{\textit{area entre los contornos}}{\textit{area total del objeto}}$$

La dependencia de esta nueva función de error con respecto al parámetro de decimación *d* no se conoce.

2.1: PCA: el caso lineal

3. La extensión de este método de ajuste de perímetros a cuerpos no convexos no se ha estudiado sistemáticamente (resultados parciales solamente: A.Tantia, J. Viveros, 2006)

4. Similaremente: la explotación de este método en la detección semi-automatizada (supervisada) de patrones en imágenes (problema de clasificación) tampoco ha sido estudiada a fondo (A. Tantia, J. Viveros, 2006).

5. Una gran cantidad de imágenes (e.g., Hubble, satélites alrededor de la tierra, imágenes microscópicas o médicas, etc.) muestran patrones **no lineales.** La detección e identificación de dichos patrones es un campo muy fértil (y con demucha demanda) en Ingeniería Biomédica y en el campo del procesamiento digital de señales. Para este problema el empleo de la técnica de PCA no-lineal ("kernel-PCA" --ver más adelante) dentro del campo del análisis armónico ha resultado ser una herramienta con muchas aplicaciones y ramificaciones interesantes.

3. Métodos de clasificación de datos

3.1 El problema abstracto de clasificación binaria

(X,y) par de variables aleatorias, *X* toma valores en $\mathbb{R}^n$ y *y* toma valores en $\mathbb{R}$; es decir, son funciones medibles en un espacio de probabilidad (espacio de medida finita e igual a uno).

Problema: dado que *X* puede ser observada (i.e., sus valores están disponibles), se desea predecir el valor asumido por *y* dado que el valor asumido por *X* es conocido.

En este documento nos enfocaremos solamente en el caso de clasificación binaria, es decir, $y \in \{-1,+1\}$

La predicción se los valores de *y* se hace por medio de una función clasificadora,

$$g : \mathbb{R}^d \to \{-1,+1\}$$

$$X \to y$$

3.1 El problema abstracto de clasificación binaria

La función clasificadora se escoge dentro de un conjunto de funciones (generalmente un espacio de funciones) con ciertas propiedades P_0,

$$g \in G = \{ g : \mathbb{R}^n \to \{-1, +1\} \mid P_0 \}$$

tal que minimice el error de generalización; mas precisamente, queremos encontrar $g^* \in G$ tal que

$$g^* = \underset{g \in G}{arg\ min}\ P\{ g(X) \neq y \}$$

nn donde P representa la probabilidad del evento en el que se ha cometido un error de clasificación. Para calcular esta probabilidad hace falta conocer la función de distribución conjunta de X y y, esto es imposible en la práctica, entonces la expresión anterior tiene utilidad teórica únicamente.

Si pudiésemos trabajar con todas las funciones del espacio, i.e., si $G = \{ g : \mathbb{R}^n \to \{-1, +1\} \}$ entonces

$$g^* = g_B = \begin{cases} +1 & E\{Y \mid X = x\} \geqslant 0 \\ -1 & E\{Y \mid X = x\} < 0 \end{cases}$$

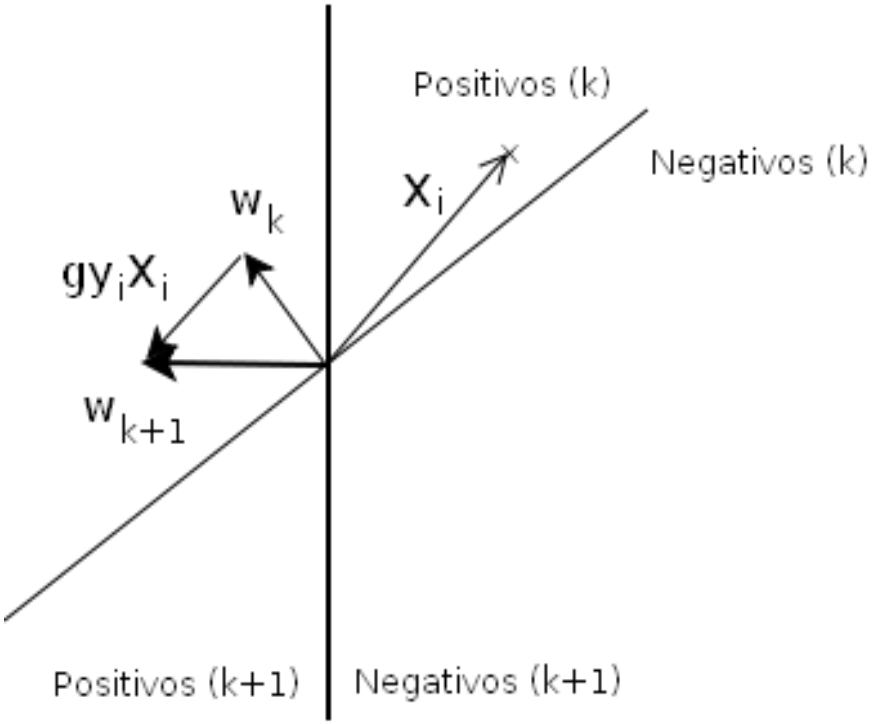
Clasificador de Bayes
(E denota la esperanza del evento dado)

3.2: Clasificadores lineales (y el problema de la dimensionalidad alta)

3.2.1 **El Perceptron de Rosenblatt** (clasificación binaria)

Frank Rosenblatt (1956): (1) en-línea

(2) guidado por errores de clasificación: por cada dato mal clasificado, los parámetros del plano discriminante son actualizados.



$$S=\{(X_i, y_i): X_i \in \mathbb{R}^n,\ y_i \in \{-1,+1\}, i=1,\ldots,m\}$$

(X_i, y_i) tal que $y_i=-1$ (negativo),

Hiperplano (w_k, b_k) asigna etiqueta

$\hat{y}_i=+1$ (error).

El plano discrimiante es actualizado:

(w_{k+1}, b_{k+1})

En la figura, el plano (w_k, b_k) cometió un error al clasificar a X_i como positivo.

3.2: Clasificadores lineales (y el problema de la dimensionalidad alta)

3.2.1 El Perceptron de Rosenblatt

¿Por qué nos interesa?

- Históricamente fué uno de los primeros métodos empleados para la clasificación de datos.

- Tiene la mayoría de las características de los métodos de clasificación modernos:

1. Posee una formulación dual
2. Es posible obtener estimaciones del máximo error cometido durante la clasificación
3. Puede utilizarse para inferir datos (genes) importantes para la clasificación ("feature selection").

3.2: Clasificadores lineales (y el problema de la dimensionalidad alta)

Def.: Sea $S=\{(X_i, y_i): X_i \in \mathbb{R}^n,\ y_i \in \{-1,+1\}, i=1,\ldots,m\}$ un conjunto muestra

S es linealmente separable si existe un hiperplano (w, b), $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ tal que

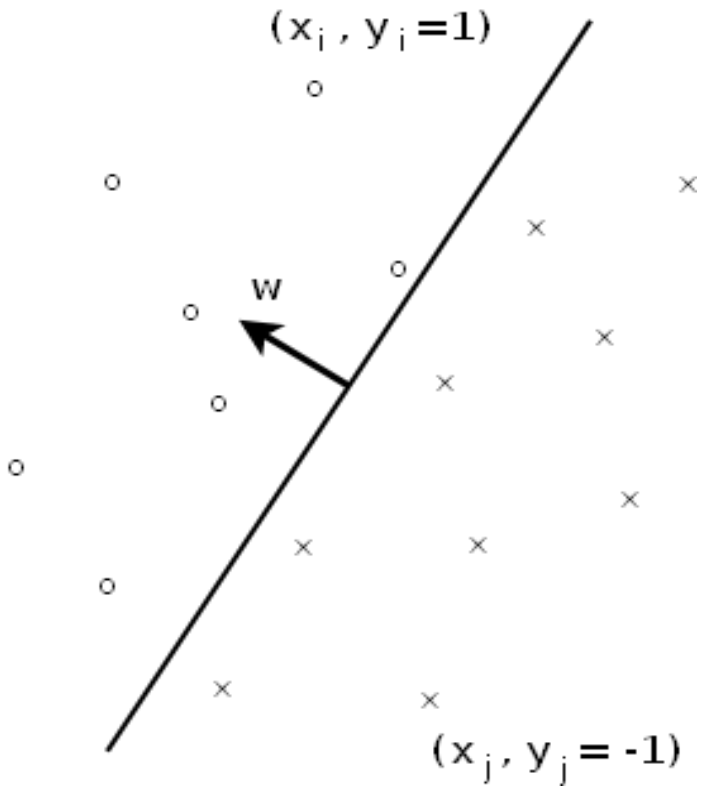
$$y_i((w, X_i)+b) \geq 0 \ \ \forall\ i$$

Def: para cualesquier $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ defínase

$g_{w,b}: S \to \mathbb{R}$ tal que

$$\hat{y}_i = g_{w,b}(X_i) = \begin{cases} -1 & \text{si } (w, X_i)+b < 0 \\ +1 & \text{si } (w, X_i)+b \geq 0 \end{cases}$$

Función clasificadora asociada al plano (w, b).



3.2.1 El perceptron de Rosenblatt

Algoritmo: (datos separables linealmente) $\eta \in \mathbb{R}^+$ "razón de aprendizaje"

$w_0 \leftarrow 0 \,;\, b_0 \leftarrow 0 \,;\, k \leftarrow 0$

$R \leftarrow max\{\|X_i\| : 1 \le i \le m\}$

repetir

for $i = 1$ to m

if $y_i((w_k, X_i) + b_k) \le 0$

then $w_{k+1} \leftarrow w_k + \eta\, y_i X_i$

$b_{k+1} \leftarrow b_k + \eta\, y_i R^2$

$k \leftarrow k + 1$

end if

end for

hasta que el número de errores dentro de la rutina "for" sea igual a cero
regresar *k=número de errores* y (w_k, b_k)

3.2.1 El perceptron de Rosenblatt

Obs.: el último plano arrojado por el algoritmo (w_k, b_k) es el primer plano que clasifica al conjunto de práctica *S* correctamente ("pérdida de generalidad")

Justificación del algoritmo (método del gradiente o descenso más rápido)

$$R_{w,b}(X) = y((w, X)+b)$$

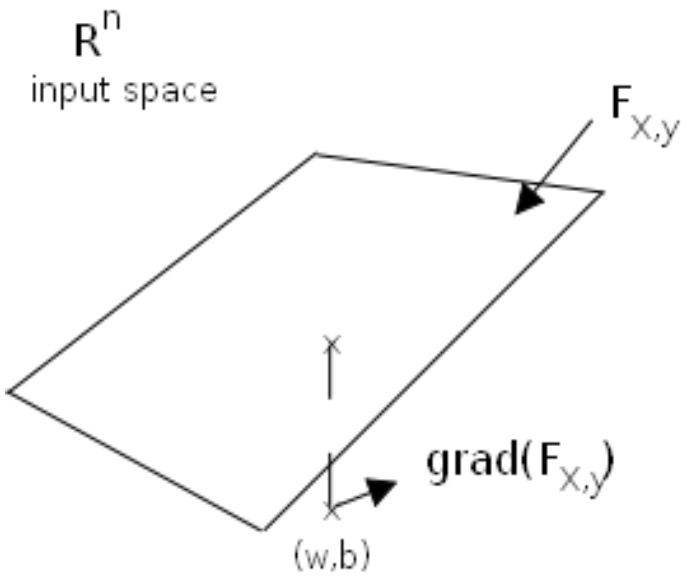
$$X \in \mathbb{R}^n, \quad y \in \{-1, +1\}$$

Si $R_{w,b}(X) \leqslant 0$ entonces el dato X ha sido mal clasificado.

$$F_{X,y}(w, b) := y((w, X)+b)$$

$$\nabla_{w,b} F = \begin{bmatrix} \frac{\partial F}{\partial w} \\ \frac{\partial F}{\partial b} \end{bmatrix} = \begin{bmatrix} yX \\ y \end{bmatrix} \in \mathbb{R}^{n+1}$$

Dirección de máximo crecimiento



$$\begin{pmatrix} w_{k+1} \\ b_{k+1} \end{pmatrix} = \begin{pmatrix} w_k \\ b_k \end{pmatrix} + \alpha \begin{pmatrix} yX \\ y \end{pmatrix}$$

$\alpha =$ "razón de aprendizaje"

Para clasificar X correctamente necesitamos movernos en la dirección del gradiente (ascenso más rápido).

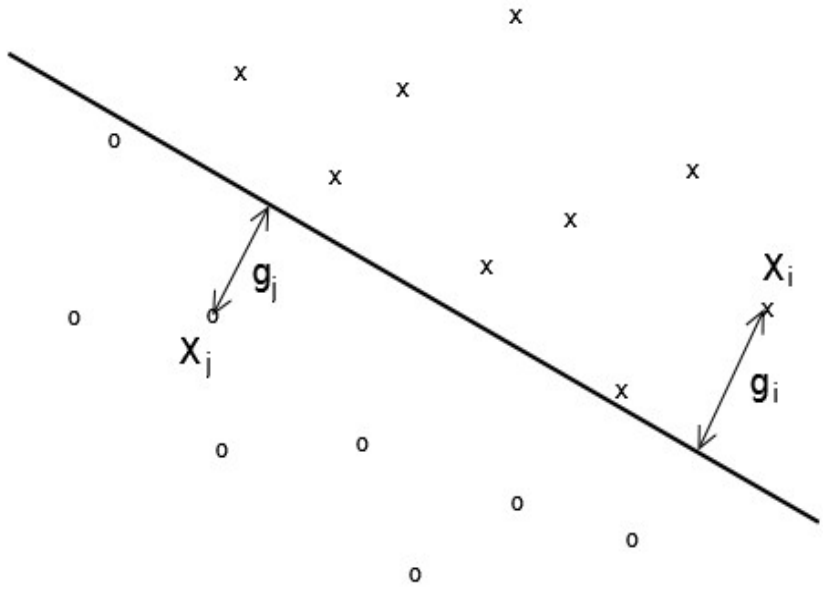
3.2: Clasificadores lineales (y el problema de la dimensionalidad alta)

Def: dado (w, b) se define el margen de la muestra X_i como:

$$\gamma_i := y_i((w, X_i)+b)$$

Def: margen de (w, b) con respecto al conjunto muestra $S=\{X_i : i \in I\}$ se define como

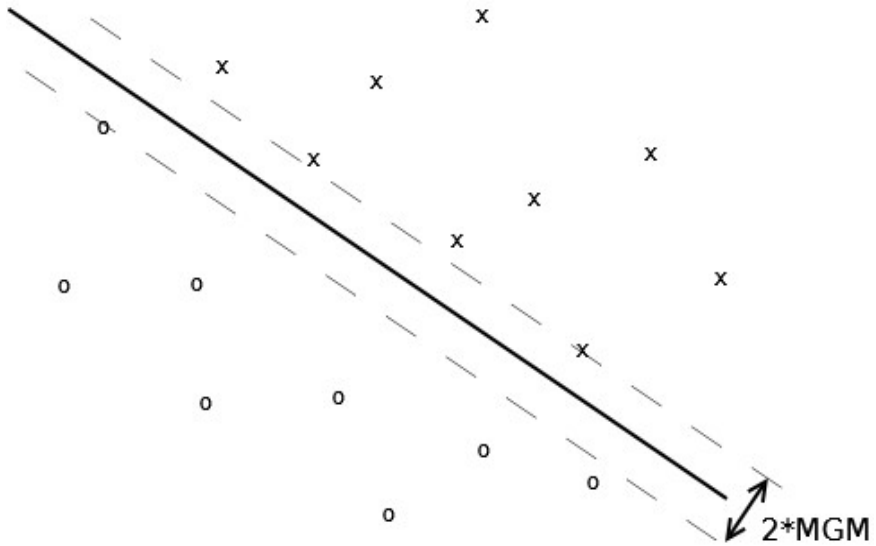
$$\gamma := min_i \gamma_i$$



Si margen muestral = distancia al hiperplano, entonces se le llama margen geométrico.

Def: el margen de un conjunto muestral es el *margen geométrico máximo* (MGM) sobre todos los hiperplanos discriminantes.

Def: plano discriminante de margen máximo es aquel cuyo margen es el MGM.



3.2.1 El perceptron de Rosenblatt

Teorema: (Novikoff / Block 1962)

Sea $S=\{(X_i, y_i): 1\leq i\leq m\}$ conjunto de prueba no trivial y $R=max\{\|X_i\|: 1\leq i\leq m\}$

Supongamos que existe w^* tal que $\|w^*\|=1$ y

$$y_i((w^*, X_i)+b^*)\geqslant \gamma$$

$\forall\ 1\leqslant i\leqslant m$. Entonces el número de errores que comete el perceptron al correr *una sola vez* ("una época") sobre *S* está acotado superiormente por

$$\overline{E}:=\left(\frac{2R}{\gamma}\right)^2$$

Corolario: si *S* es linealmente separable, entonces el perceptron de Rosenblatt converge en un número finito de pasos a un plano que discrimina todos los puntos de *S* correctamente.

Obs: $\overline{E}$ es independiente de η

3.2.1 El perceptron de Rosenblatt

Prueba: cambio de coords, $\hat{X}_i := (X_i', R)'$ $\quad \hat{w} := (w', b/R)'$

$$f_{w,b}(X_i) := w' X_i + b = \hat{w}' \hat{X}_i =: f_{\hat{w}}(\hat{X}_i)$$

$\hat{w}_0 = 0$

$\hat{w}_{k-1}$ vector de costos antes del k-ésimo error

corregir $\hat{w}_{k-1} \to \hat{w}_k$ cuando ocurra un error: $y_{i+1}(\hat{w}_{k-1}' \hat{X}_{i+1}) \leq 0$

$$\hat{w}_k := (w_k', b_k/R)' = (w_{k-1}', b_{k-1}/R)' + \eta\, y_i (X_i', R)' = \hat{w}_{k-1} + \eta\, y_i \hat{X}_i$$

$$(\hat{w}_k, \hat{w^*}) = (\hat{w}_{k-1}, \hat{w^*}) + \eta\, y_i (\hat{X}_i, \hat{w^*}) \geqslant (\hat{w}_{k-1}, \hat{w^*}) + \eta\, \gamma$$

$$y_i((w^*, X_i) + b^*) = y_i(\hat{w^*}, \hat{X}_i) \geqslant \gamma$$

Por lo tanto: $(\hat{w}_k, \hat{w^*}) \geqslant k\, \eta\, \gamma$

$$\|\hat{w}_k\|^2 = \|\hat{w}_{k-1}\|^2 + 2\eta\, y_i(\hat{w}_{k-1}, \hat{X}_i) + \eta^2 \|\hat{X}_i\|^2 \leqslant \|\hat{w}_{k-1}\|^2 + \eta^2 \|\hat{X}_i\|^2$$

$$\leqslant \|\hat{w}_{k-1}\|^2 + \eta^2 (\|X_i\|^2 + R^2)$$

$$\leqslant \|\hat{w}_{k-1}\|^2 + 2\eta^2 R^2$$

3.2.1 El perceptron de Rosenblatt

Prueba: (cont.)

Por lo tanto: $\|\hat{w}_k\|^2 \leqslant 2k\,\eta^2 R^2$

Juntando resutlados: $\sqrt{2k}\,\eta R\|\hat{w}^*\| \geqslant \|\hat{w}_k\|\,\|\hat{w}^*\| \geqslant (\hat{w}_k, \hat{w}^*) \geqslant k\,\eta\,\gamma$

manipulando,

$$\sqrt{2}\frac{R}{\gamma}\sqrt{\|w^*\|^2+(b^*/R)^2} \geqslant \sqrt{k}$$

$\|w^*\|^2 = 1$, $b^* \leqslant R$ **(ejercicio)**

Entonces, $\left(\frac{2R}{\gamma}\right)^2 \geqslant k$

Si *S* no es linealmente separable el alg. de Rosenblatt no converge.
¿En tal caso, cuantos errores puede cometer el algoritmo en una primera ejecución sobre *S*?

3.2.1 El perceptron de Rosenblatt

Def.: sean $\gamma > 0$ (margen objetivo) y un hiperplano (w, b), dados.

Entonces, para cualquier dato de prueba (X_i, y_i) se define su margen de holgura ξ_i de la sig. manera:

$$\xi_i := max\{0, \gamma - y_i((w, X_i) + b)\}$$

Notas.: ξ_i es una medida de qué tan cerca está X_i de tener margen γ c.r. a (w, b)

Si $\xi_i > \gamma$ entonces el algoritmo cometió un error clasificando X_i

Si $\gamma > \xi_i > 0$ entonces (w, b) es un hiperplano discriminante cuyo margen es inferior a γ

Def: $D := \sqrt{\sum_{i=1}^{m} \xi_i^2}$ (medida de la desviación de S de tener margen γ).

3.2.1 El perceptron de Rosenblatt

Teorema: (Freund & Schapire, 1999)

Sea $S=\{(X_i, y_i): i=1, \ldots, m\}$ un conjunto de prueba t. q. $\|X_i\| \leqslant R \; \forall \; i$.

Sean w un vector unitario, $\gamma > 0$ (margen objetivo), dados, y las cantidades antes defs:

$\xi_i := max\{0, \gamma - (y_i(w, X_i)+b)\}$ (*i*-ésimo margen de holgura) y $D^2 := \sum_{i=1}^{m} \xi_i^2$

Entonces el número de errores del algoritmo de Rosenblatt al correr *una sola* vez ("una época") sobre *S* está acotado superiormente por

$$\left(\frac{R+D}{\gamma}\right)^2$$

Idea principal de la prueba: embeber el conjunto de prueba (el cual no es linealmente separable) en un espacio de dimensión mayor en el cual separación lineal es posible y el algoritmo de Rosenblatt produce planos discriminantes que tienen la misma "funcionalidad" en el espacio original durante la primera época.

3.2.1 El perceptron de Rosenblatt

Sean como antes: $X_i \to \tilde{X}_i := [X_i' \ \Delta e_i(p)]' \in \mathbb{R}^{n+m}$, $e_i(p) \in \mathbb{R}^m$

$$w \to \tilde{w} := \left[\frac{w'}{Z} \quad \frac{y_i \xi_i}{Z\Delta} 1'(p)\right]' \qquad Z = \sqrt{1 + \left(\frac{D}{\Delta}\right)^2}$$

unitario

Note que: $y_i((\tilde{w}, \tilde{X}_i) + \frac{b}{Z}) = y_i((w, X_i) + y_i \xi_i + b)\frac{1}{Z}$

$\tilde{b}$

$$= (y_i((w_i, X_i) + b) + \xi_i)\frac{1}{Z}$$

$$\geqslant (y_i((w, X_i) + b) + \gamma - y_i((w, X_i) + b))\frac{1}{Z}$$

$$\geqslant \frac{\gamma}{Z} =: \tilde{\gamma} > 0$$

3.2.1 El perceptron de Rosenblatt

$(\tilde{w}, \tilde{b})$ separa los datos $\tilde{S}=\{(\tilde{X}_i, y_i)\}$ con margen $\tilde{\gamma}$, entonces, del teorema anterior, el número de errores cometidos por el alg. de Rosenblatt sobre $\tilde{S}$ (en una época) está acotado superiormete por

$$\left(\frac{2\tilde{R}}{\tilde{\gamma}}\right)^2 = \frac{4(R^2+\Delta^2)(1+\left(\frac{D}{\Delta}\right)^2)}{\gamma^2} = F(\Delta^2)$$

Ejercicio: terminar la prueba siguiendo los siguientes pasos:

(a) demuestre que el mínimo de *F* se alcanza cuando $\Delta^2 = RD$

(b) sea $w_{|i}$ el vector discriminante al momento de clasificar X_i, obtenido utilizando el perceptron sobre *S*. Sea $\tilde{w}_{|i}$ el vector discriminante al momento de clasificar $\tilde{X}_i$, obtenido utilizando el perceptron sobre $\tilde{S}$. Entonces, las primeras *n* coordenadas de $\tilde{w}_{|i}$ son iguales a las de $w_{|i}$; más aún, la *(n+i)*-ésima coordenada de $\tilde{w}_{|i}$ es igual a cero.

(c) la predicción de $\tilde{w}_{|i}$ es la misma que la de $w_{|i}$ ($\hat{y}_i = \hat{\tilde{y}}_i$).

3.2.1 El perceptron de Rosenblatt

Nota: *el problema de encontrar el plano discriminante que cometa el menor número de errores sobre S cuando este conjunto no es linealmente separable es* ***NP-completo.***

NP= "non-deterministic polinomial time."
Este tipo de problemas tiene dos caracterícas:

(i) cualquier solución del problema puede ser verificada "rápidamente" en tiempo polinomial.
(ii) si A y B son problemas NP-completos, existe una transformación de uno en el otro la cual puede realizarse en tiempo polinomial.

A la fecha, no se conoce ningún algoritmo eficiente para resolver los problemas NP-c. Típicamente, el tiempo para resolver un problema NP-c con cualquier algoritmo conocido, aumenta "rápidamente" más allá de toda utilidad práctica a medida que el tamaño del problema crece (e.g., a medida que la dimensión del espacio ambiente incrementa).

Ejemplos de problemas NP-c:

Problema del vendedor viajero ("traveling salesman problem")
Colorear los vértices de una gráfica con un número *k* dado de colores ("Graph coloring")
... muchos mas

3.2.1 El perceptron de Rosenblatt

Algunas preguntas y notas interesantes:

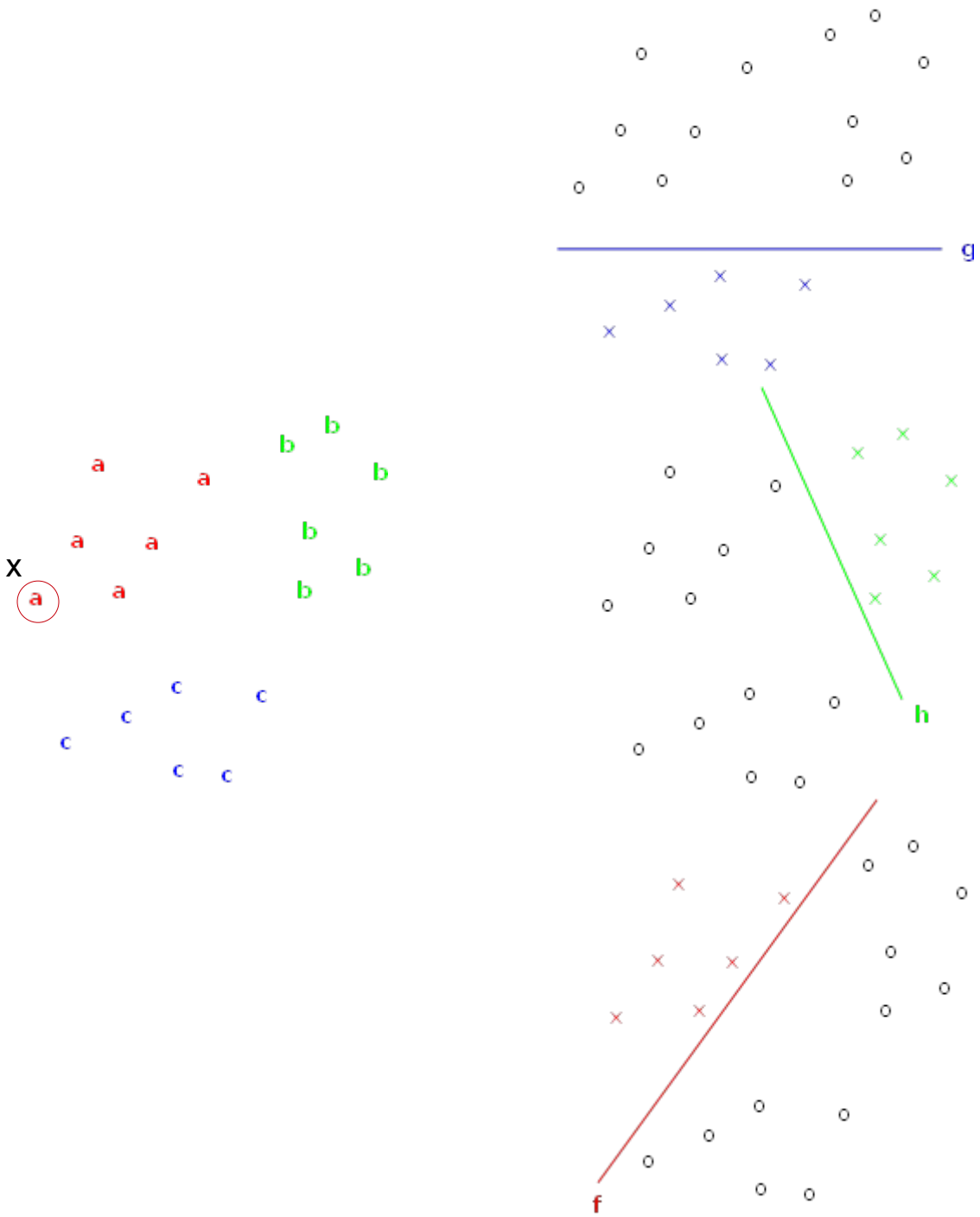
Caso linealmente separable:
(a) el número de errores cometidos en cada época disminuye hasta alcanzar el valor cero. **¿Cómo puede describirse cuantitativamente esta disminución en el número de errores?** $¿\, \bar{E}(k) - \bar{E}(k+1) = F(\gamma, R)\,?$

(b) ¿Cuál es una cota superior para el número de épocas necesarias para alcanzar convergencia?

Caso no separable linealmente:
(a) "El algoritmo del bolsillo" ("pocket algorithm"): en este algoritmo, se almacenan todos los planos discriminantes durante un número de épocas T fijo, al final de las cuales se escoge el plano discriminante que haya sobrevivido el mayor número de actualizaciones consecutivas.

(b) ¿Cuál es la frecuencia con la que las actualizaciones toman lugar conforme $T \to \infty$ **?**

Separación lineal multiclase: método de votación



	a	b	c
0<f(X)	1	0	0
0>g(X)	1	0	1
0>h(X)	1	1	0
	3	0	1

En caso de empate seleccionar de manera aleatoria una clase entre las que tienen el mayor número de votos.

Separación lineal multiclase: votación

Ejercicio:

En el caso en que los datos no son linealmente separables sino “casi linealmente separables,” ¿cómo podría definir un método de clasificación que tuviese el menor número de errores de clasificación? ¿Existe un algoritmo análogo al algoritmo del bolsillo?

3.2.1 **El perceptron "pesado" de Freund y Schapire** (1999)

Entrenamiento ("training"):

Inputs: conjunto de práctica $S=\{(X_i, y_i): y_i \in \{-1,1\}, i=1,\dots,m\}$
número de épocas T

Output: perceptrons pesados $\{(w_i, b_i; c_i): i=1,\dots,k\}$

Inicializar $k=0,\ w_1=0,\ c_1=0$

Repetir T veces:
for i = *1, … , m:*
calcular predicción: $\hat{y}_i = sgn((w_k, X_i)+b_k)$
si $\hat{y}_i = y_i$ entonces $c_k = c_k + 1$
de lo contrario $v_{k+1} = v_k + y_i X_i$
$c_{k+1} = 1$
$k = k+1$

El peso es mayor
Entre mas acertado
El plano discriminante

Predicción:

Dados: X dato sin clasificar
lista de perceptrons pesados $\{(w_i, b_i; c_i): i=1,\dots,k\}$

Calcular etiqueta $\hat{y}$ de la sig. forma: $s = \sum_{i=1}^{k} c_i\, sgn((w_i, X)+b_i)$ luego $\hat{y} = sgn(s)$

3.2.1 **El perceptron de Freund y Schapire** (1999)

Comentarios:

1. para datos linealmente separables: el perceptron de Freund y Schapire converge, conforme $T \to \infty$ al Perceptron de Rosenblatt (plano discriminante=último plano obtenido).

2. Intuitivamente: el algoritmo debería de ser menos sensible a datos contaminados ("outliers" y datos con largas desviaciones). **(Falta estudio formal.)**

3. Problema abierto: el perceptron de Freund y Schapire mejora su número de aciertos ("mejor desempeño") después de *T=1*. **(No se tiene explicación teórica para esto.)**

4. Cálculo de cotas superiores al error esperado (ver más adelante) del perceptron de Freund y Schapire son comparables (aunque superiores) a las de los clasificadores de rango máximo de Vapnik ("support vector machines" o SVMs --ver más adelante). **(Falta estudio somparativo sistemático exhaustivo)**

3.2.1 **El perceptron de Freund y Schapire** (1999)

5. El perceptron de Freund y Schapire tiene un mejor desempeño que el de Rosenblatt, y aunque no es tan acertado como SVMs y requiere de más memoria virtual, es sencillo de programar y usualmente requiere de menos tiempo que SVMs para clasificar.

6. Método es extendible al problema de separación no lineal (métodos de nucleación o “kernel methods” --ver más adelate).

7. El problema de la generalización: un clasificador que tenga un buen desempeño en un conjunto de práctica *S* no necesariamente tendrá el mismo desempeño en otros conjuntos con las mismas características que *S*, de hecho, como regla general, un clasificador que tenga un 100% de aciertos en *S* tendrá un porcentaje muy bajo de aciertos en otro conjunto similar (este problema se conoce como “overfitting”).

3.2.1 **El perceptron de Freund y Schapire** (1999)

Estimaciones teóricas del error de clasificación (ejemplos de resultados deseables)

Teorema. (Freund, Schapire, 1999):
Asuma que los datos son variables aleatorias independientes e idénticamente distribuidas (iid). Sea $S=\{(X_i, y_i): i\in I=\{1,\dots,m\}\}$ una sucesión de datos de entrenamiento y (X_{m+1}, y_{m+1}) un dato de prueba. Sea $R=max\{\|X_i\|: i\in I\}$. Dados u vector unitario y $\gamma>0$ defínase

$$D_{u,\gamma}=\sqrt{\sum_{i=1}^{m+1}(max\{0,\gamma-y_i(u, X_i)\})^2}$$

Entonces la probabilidad (sobre todas las m+1 variables) de que el perceptron de Freund y Schapire se equivoque al clasificar X_{m+1} durante la primera época es menor o igual a

$$\frac{2}{m+1}E\left[inf\left\{\left(\frac{R+D_{u,\gamma}}{\gamma}\right)^2 : \|u\|=1\,;\gamma>0\right\}\right]$$

(Este resultado puede mejorarse: el cálculo de la esperanza puede restringirse solamente a aquellas variables que fueron clasificadas incorrectamente.)

3.2.1 El perceptron de Rosenblatt

Observaciones.:

1. El problema de encontrar el hiperplano discriminante utilizando el algoritmo de Rosenblatt no está bien planteado ("ill-posed"), la solución podría no ser única.

2. Dado que el algoritmo comienza con $w_0 = 0$, los vectores discriminantes son únicamente combinación lineal de los vectores que fueron clasificados incorrectamente, es decir, el vector que construye el algoritmo de Rosenblatt puede escribirse como sigue:

$$w = \sum_{i=1}^{m} \alpha_i y_i X_i$$

En donde α_i es proporcional al # de veces que X_i fué clasificado incorrectamente.

En general, $w_{T=1} = \sum_{i \in I} \alpha_i y_i X_i$ (vector obtenido al término de la primera época),

en donde $I \subseteq \{1, \ldots, m\}$ es el conjunto de índices correspondientes a los datos mal clasificados. Esto da pauta para una formulación alternativa o "dual" del algoritmo.

3.2.1 El perceptron de Rosenblatt (forma dual)

Formulación dual (no es realmente una formulación dual ya que, en principio, distintos vectores w podrían tener el mismo vector de pesos $\alpha=(\alpha_1, \ldots, \alpha_m)$)

Función de decisión dual: $h(X) := sgn((w, X)+b) = sgn(\sum_{i=1}^{m} \alpha_i y_i (X_i, X)+b)$

$S=\{(X_i, y_i): i\in\{1, \ldots, m\}\}$

$\alpha \leftarrow 0\,,\ b \leftarrow 0$

$R := max_i \|X_i\|$

Repetir *for i=1 to m*

if $y_i\left(\sum_{i=1}^{m} \alpha_i y_i (X_j, X_i)+b\right)\leqslant 0$

then $\alpha_i \leftarrow \alpha_i + 1$

$b \leftarrow b + y_i R^2$

end if

end for

hasta que # de errores dentro del loop del *for* sea cero

Devolver $(\alpha, b) \Rightarrow h(X) = sgn((w, X)+b) = sgn(\sum_{i=1}^{m} \alpha_i y_i (X_i, X)+b)$

Forma "dual" del algoritmo de Rosenblatt

3.2.1 El perceptron de Rosenblatt (forma dual)

Propiedades interesantes de la formulación dual:

1. Datos pueden categorizarse de acuerdo con el # de errores cometidos durante su clasificación.
2. # actualizaciones = # de errores, por lo tanto:

$$\|\alpha\| \leqslant \left(\frac{2R}{\gamma}\right)^2$$

medida de la "complejidad" de la función clasificadora

3. En el cálculo de α y b los vectores de entrenamiento solamente influyen mediante evaluaciones de productos internos (X_i, X_j), lo cual motiva la siguiente definición:

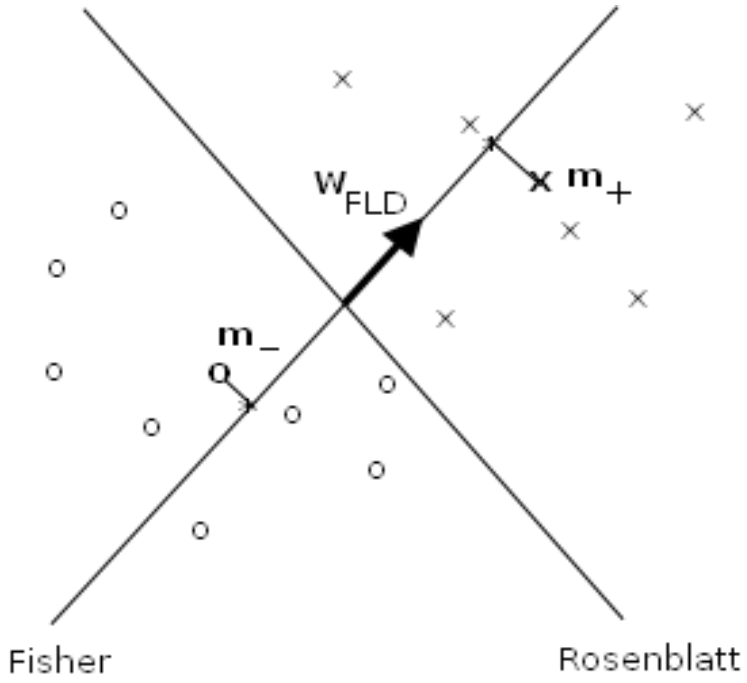
$$G \in \mathbb{R}^{m \times m}, \quad (G)_{ij} := (X_i, X_j)$$ Matriz de Gram

(simétrica, positiva definida)

3.2.2 Discriminante lineal de Fisher (Fisher's linear discriminant --FLD)

Aquellos clasificadores que resuelven problemas de optimización convexa son soluciones únicas de los programas que satisfacen. Convexidad es una propiedad muy deseable.

FLD responde responde al problema de encontrar una linea (dirección) en el espacio ambiente (input space) tal que las proyecciones sobre dicha linea de los centros de masa de las clases de datos tienen separación máxima.



$$h_{FLD}(X) = \begin{cases} +1 \;\; si \;\; (w_{FLD}, X) + b \geqslant \theta \\ -1 \;\; si \;\; (w_{FLD}, X) + b < \theta \end{cases}$$

θ = "umbral" (parámetro) su selección depende de los datos.

w_{FLD} corresponde a una dirección a lo largo de la cual es más "sencillo" separar los datos.

3.2.2 Discriminante lineal de Fisher (Fisher's linear discriminant --FLD)

FLD satisface un problema de optimización.

Sea

$$J(w) := \frac{(w'(m_+ - m_-))^2}{\sum_{X \in S_+} (w'(X - m_+))^2 + \sum_{X \in S_-} (w'(X - m_-))^2} = \frac{(\tilde{m}_+ - \tilde{m}_-)^2}{\tilde{\sigma}_+^2 + \tilde{\sigma}_-^2}$$

Veremos a continuación que $w_{FLD} = arg\ max\ J(w)$

Primero obsérvese que

$(w'(m_+ - m_-))^2 = w'(m_+ - m_-)(m_+ - m_-)'w =: S_B$ matriz de dispersión entre clases

$$S_+ := \sum_{X \in S_+} (X - m_+)(X - m_+)', \quad S_- := \sum_{X \in S_-} (X - m_-)(X - m_-)'$$

$S_W := S_+ + S_-$ matriz de dispersión interna

De forma que $J(w) = \frac{w' S_B w}{w^t S_w w}$

3.2.2 Discriminante lineal de Fisher (Fisher's linear discriminant --FLD)

Resolviendo $\nabla J(w)=0$ se encuentra que:

$S_B w = \lambda S_w w$ *problema generalizado de valores y vectores propios* (**ejercicio: verificar**)

Si S_w es invertible: $S_w^{-1} S_B w = \lambda w$ problema clásico de valores y vectores propios.

De hecho, $w = S_w^{-1}(m_+ - m_-)$ (**ejercicio: verificar**)

Generalización al caso multiclase: cf. Duda, Hart & Stork (2001)

3.2.3 Métodos de regresión

Datos: (X, y), $y \in \mathbb{R}$ las etiquetas no pertencen a un conjunto discreto necesariamente.

Problema: encontrar la función lineal $f_{w,b}(X) = (w, X) + b$ que mejor interpole las etiquetas (ajuste de rectas o planos).

s. XVIII (Gauss, Legendre): $f_{w,b}$ es la función que minimiza la suma de los cuadrados de las distancias verticales entre los datos y su gráfica. (Ajuste por mínimos cuadrados o "least squares.")

Formulación: $S = \{(X_i, y_i) : i \in I, y_i \in \mathbb{R}\}$ Conjunto de práctica

$$\hat{y} := f_{w,b}(X) = (w, X) + b$$

$$L(w, b) = \sum_{i=1}^{m} (y_i - (w, X_i) - b)^2$$ función costo

$$(w^*, b^*) = arg\ min\ L(w, b)$$ valores óptimos

3.2.3 Métodos de regresión

Ejercicio:
(a) verificar que el cambio de variables $\bar{w}=[w'\ b]'$, $\bar{X}=[X'\ 1]'$

permite escribir $L(w,b)=\|Y-\Xi'\bar{w}\|^2=(Y-\Xi'\bar{w})'(Y-\Xi'\bar{w})$

(b) defina $\Xi:=[\bar{X}_1\cdots\bar{X}_m]$

verifique que $\dfrac{\partial L}{\partial \bar{w}}=-2\,\Xi\,Y+2\,\Xi\,\Xi'\bar{w}=0\ \Rightarrow\ \boxed{\Xi\,\Xi'\bar{w}=\Xi\,Y}$ Ecuaciones normales

Si $\Xi\,\Xi'$ es invertible, entonces $\bar{w}=(\Xi\,\Xi')^{-1}\Xi\,Y$

En caso de no ser invertible las ecuaciones normales pueden resolverse calculando la matriz pseudoinversa de $\Xi'\Xi$ o bien utilizando el método de regresión de Ridge (ver siguiente método).

3.2.3 Métodos de regresión

Método de regresión de Ridge: si $\Xi\,\Xi^{-1}$ es singular, substituir las ecs. normales por:

$$(\gamma\, I_m + \Xi\,\Xi')\,\bar{w} = \Xi\, Y$$

Ejercicio: demostrar que las ecs. anteriores son las ecs. normales del sig. problema:

$$\bar{w} = arg\ min\left\{L_\gamma(\bar{w}) := \gamma\,\|\bar{w}\|^2 + \sum_{i=1}^{m}(y_i - (\bar{w}, \bar{X}_i))^2\right\}$$

Notas:

1. γ controla el equilibrio entre norma mínima ("complejidad") y costo cuadrático mínimo.

2. manipulando las ecuaciones normales se obtiene:

$$\gamma\,\bar{w} = -\sum_{i=m}^{m}((\bar{w}, \bar{X}_i) - y_i)\,\bar{X}_i \Rightarrow \bar{w} = \sum_{i=1}^{m}\alpha_i\,\bar{X}_i \qquad \text{en donde} \qquad \alpha_i = \frac{-1}{\gamma}((\bar{w}, \bar{X}_i) - y_i)$$

3. **Ejercicio:** verifique que substituyendo la expresión para w en L_γ se obtiene:

$$L_\gamma = \gamma\,\alpha'\,G\,\alpha + \alpha'\,G^2\,\alpha - 2\mathrm{Y}'\,G\,\alpha + Y'\,Y$$

En donde $G = \Xi'\,\Xi \in \mathbb{R}^{m\times m}$ matriz de Gram (matriz de productos internos)

3.2.3 Métodos de regresión

Ejercicio: verificar que tomando $\frac{\partial L_\gamma}{\partial \alpha}=0$ conduce a las ecs. normales: $(\gamma I_m+G)\alpha=Y$

y que a su vez las ecs. normales dan como resultado la función dual discriminante:

$$f(X)=Y'(\gamma I_m+G)^{-1}\Xi' X$$

Obs.: la función clasificadora dual *f* está definida a partir de la matriz de Gram *G*, este también era el caso en el algoritmo de Rosenblatt.

Ejercicios:
1. ¿cómo se modifica el procedimiento anterior si $\gamma=0$? (Puede hacer lo mismo para el caso de ajuste por mínimos cuadrados?)
2. ¿Puede hacer algo similar para el caso de FLD? (Si su respuesta es afirmativa plantee una formulación dual y en caso negativo explique por qué no.)

3.2.3 Métodos de regresión

El método simplex (programa de optimización lineal –ver referencias)

Problema: dado (X, y), dato de prueba, encontrar una función lineal (clasificadora o discriminante) tal que prediga correctamente el valor de la etiqueta $y \in \mathbb{R}$

Función discriminante: $\hat{y} = f_{w,b}(X) = (w, X) + b$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ parámetros por determinar.

Idea: para cada X_i defínase error del ajuste como $|y_i - \hat{y}| = |y_i - (w, X_i) - b|$

Para cada *i* se quiere encontrar un margen $\varepsilon_i \geqslant 0$ tal que sea lo más pequeño posible y además,

$$|y_i - (w, X_i) - b| \leqslant \varepsilon_i \Leftrightarrow \begin{cases} y_i - (w, X_i) - b \geqslant -\varepsilon_i \\ \text{y} \\ -y_i + (w, X_i) + b \geqslant -\varepsilon_i \end{cases}$$

3.2.3 Métodos de regresión: el método simplex

Se tiene entonces el sig. programa lineal para $w, b, \varepsilon = [\varepsilon_1, \ldots, \varepsilon_m]$

$$\min_{\substack{w,b \\ \varepsilon_1,\ldots,\varepsilon_m}} \quad z = \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_m$$

tal que

$$\begin{bmatrix} \Xi' & 1 & I_m \\ -\Xi' & -1 & I_m \end{bmatrix} \begin{bmatrix} w \\ b \\ \varepsilon \end{bmatrix} \geqslant \begin{bmatrix} Y \\ -Y \end{bmatrix}$$

$$w \in \mathbb{R}^n, \quad b \in \mathbb{R}, \quad 0 \leqslant \varepsilon \in \mathbb{R}^n$$

Este programa lo aprendimos a resolver en programación lineal

Problemas abiertos/ejercicios:

1. Hace falta un **estudio sistemático** que compare el desempeño del método simplex, vs, por ejemplo, mínimos cuadrados (o cualqueir otro método de regresión lineal –SVM, SDF, DWD, FLD, etc) ya sea con datos de laboratorio o bien con datos sintéticos.

interés práctico

interés teórico

Signed **D**istance **F**unction (ver referencias)

Distance **W**eighted **D**iscrimination

3.2.3 Métodos de regresión

Problemas abiertos/ejercicios:

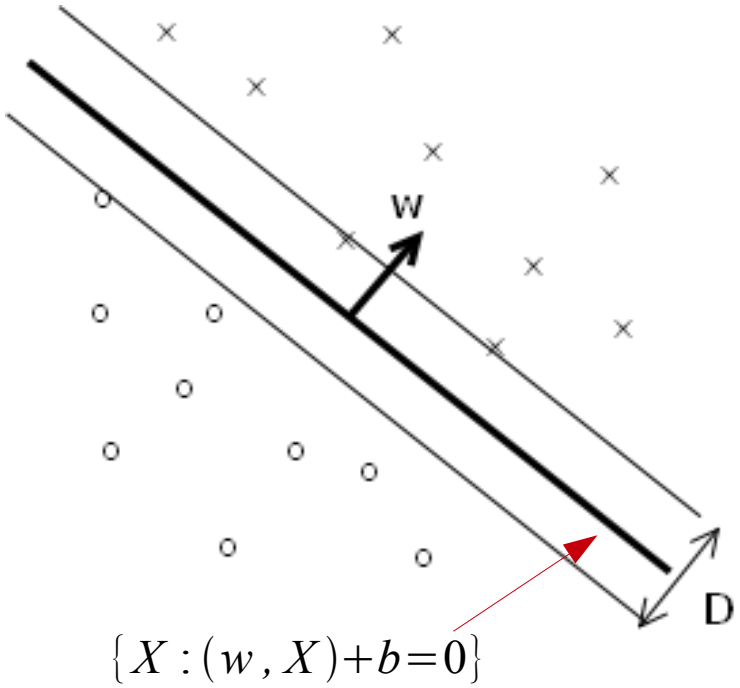
2. SVM (ver método siguiente), simplex y regresión de Ridge tienen formulaciones duales. ¿Cómo se comparan los desempeños de las **versiones duales** en problemas específicos de clasificación?

3. El método simplex parece ser mas afectado que el método de mínimos cuadrados por "outliers," ¿cuál es la razón de esto? (Este fué uno de los resultados de los ejercicios que hicieron los estudiantes de programación lineal del semestre pasado (UAEH).)

3.2.4 Support Vector Machines (SVMs –Clasificadores de margen máximo)

Estos son los clasificadores que hoy dia son mas usados debido a su alto grado de confiabilidad (pero no son los mejores en todo escenario).

$S=\{(X_i, y_i): i\in\{1,\dots m\}, y_i\in\{-1,+1\}\}$ Conjunto de práctica (clasif. binaria)



Caso separable: encontrar $(\tilde{w}, \tilde{b})$ t.q.

$$y_i((\tilde{w}, X_i)+\tilde{b}_i)\geq D>0 \ \forall\ i$$

$$\Rightarrow \quad y_i((w, X_i)+b)\geq 1 \ \forall\ i, \qquad b=\frac{\tilde{b}}{D}, \ w=\frac{\tilde{w}}{D}$$

$$(*)\ \begin{cases} \min\limits_{w,b} \ \|w\|^2 \\ \text{tal que} \ \ y_i((w, X_i)+b)\geq 1 \ \forall\ i \end{cases}$$

Por lo tanto, el problema de encontrar el plano de margen máximo que separa los datos es equivalente a encontrar el plano discriminante con vector de dirección w de norma mínima.

3.2.4 Support Vector Machines (SVMs –Clasificadores de margen máximo)

Obs.: a pesar de ser un clasificador lineal, el plano discriminante se encuentra resolviendo un programa no-lineal.

Caso no separable linealmente: definir variables de holgura $\xi_i \geqslant 0$ tales que

$$y_i((w, X_i)+b) \geqslant 1-\xi_i, \quad \forall\ i$$

Sea $\xi := [\xi_1, \ldots, \xi_m]'$ entonces **(*)** puede reemplazarse por el siguiente

programa cuadrático primario **(PQ²)**:

$$\min_{w, b, \xi} \quad z = \|w\|^2 + \gamma \|\xi\|^2$$

tal que

$$\left[D(Y)\Xi' \quad Y \quad I_m \right] \begin{bmatrix} w \\ b \\ \xi \end{bmatrix} \geqslant 1$$

$$w \in \mathbb{R}^n, \quad b \in \mathbb{R}, \quad \xi \geqslant 0$$

$\gamma \geqslant 0$ constante de regularización o penalización

$D(Y)$ = diag $(\xi_1, \ldots, \xi_m)$ (matriz diagonal)

Obs: la condición $\xi \geqslant 0$ puede reemplazarse por $\xi \in \mathbb{R}^n$

3.2.4 Support Vector Machines (SVMs –Clasificadores de margen máximo)

Nota: estimaciones teóricas del error cometido por SVM, sugieren que $\gamma \sim R^{-2}$ en donde R es el radio de una bola que contiene el soporte de la función de distribución conjunta de X y y, $B_R(0) \supset spt\{\rho_{Xy}\}$

Ejercicio: demostrar que si w^* resuelve **(PQ²)** entonces $w^* \in sp\{X_1, \ldots, X_m\} = Z$ (hint: si w^* tuviese una componente perpendicular al espacio generado por los vectores de entrenamiento, dicha componente se cancela al evaluar los productos internos que constituyen las restricciones del programa cuadrático.)

El hecho del ejercicio anterior permite una formulación dual de **(PQ²)**

Alternativamente, (PQ²) tiene una versión de norma 1:

$$\min_{w,b,\xi} \quad z = \|w\|^2 + \gamma \|\xi\|_1$$

tal que

$$\left[D(Y)\Xi' \quad Y \quad I_m \right] \begin{bmatrix} w \\ b \\ \xi \end{bmatrix} \geqslant 1$$

(PQ¹) $\qquad w \in \mathbb{R}^n, \; b \in \mathbb{R}, \; \xi \geqslant 0$

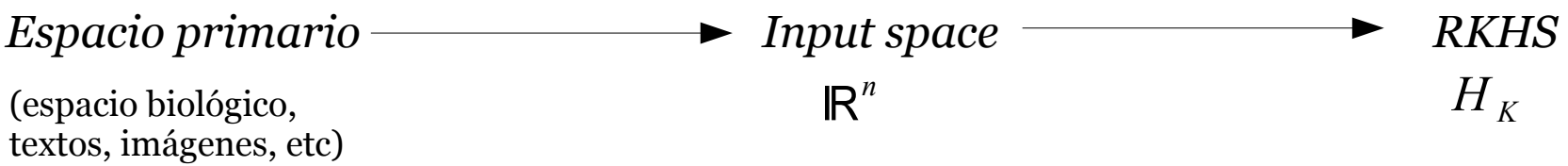
En donde $\|\xi\|_1 = \sum_{i=1}^{m} |\xi_i|$

Ejercicio: obtener la forma dual de (PQ²)

Las formas duales son deseables ya que, usualmente, tienden a estar mejor condicionadas que sus forma primarias.

3.3 El truco de la nucleación del espacio

Idea general: hacer una transformación del espacio ambiente ("input space") a un espacio de dimensión más alta, en general, un espacio de Hilbert ("RKHS=reproducing kernel Hilbert space"), que tiene la propiedad de que el cálculo de los productos internos o punto que consumian memoria y tiempo-máquina, puede hacerse rápidamente mediante evaluaciones de una función llamada kernel.



Def. kernel o núcleo de Mercer es una función $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ con las sig. props.:

1. continua

2. simétrica: $K(a,b) = K(b,a)$

3. positiva definida: dado un conjunto cualquiera de vectores $\{X_1, \ldots, X_m\}$

se cumple que si $C \in \mathbb{R}^m$, arbitrario y $K \in \mathbb{R}^{m \times m}$ tal que $(K)_{ij} = K(X_i, X_j)$, entonces

$$C' KC > 0 \ \forall \ C \neq 0$$

3.3 El truco de la nucleación del espacio

Ejemplos de núcleos usados con mucha frecuencia:

1. Gaussiano: $K(u,v) = e^{-\|u-v\|^2/\sigma^2}$, $\sigma \geqslant 0$ es un parámetro.

2. polinomial: $K(u,v) = (u,v)^r$, $r \in \mathbb{N}$

Nota: *en la práctica la elección de parámetros no es sencilla y su determinación tiene efectos importantes en el desempeño de la función clasificadora.*

Construccion del espacio de Hilbert H (RKHS):

Sea $\{\xi_i : i = 1, \ldots, s\}$ una colección de vectores en $\mathbb{R}^n$ los cuales son todos no cero y distintos entre sí.

Defínanse funciones $K_i : \mathbb{R}^n \to \mathbb{R}$

$$X \to K_i(X) := K(\xi, X), \quad i = 1, \ldots, s$$

y el espacio lineal generado por tales funciones: $\Psi := span\{K_i : i = 1, \ldots, s\}$

3.3 El truco de la nucleación del espacio

Definición de un producto interno en Ψ: $(\cdot,\cdot)_\Psi : \Psi \times \Psi \to \mathbb{R}$

1. $(K_i, K_j)_\Psi := K(\xi_i, \xi_j)$

2. sean funciones en Ψ, $f_1 = \alpha_1 K_1 + \ldots + \alpha_s K_s$ y $f_2 = \beta_1 K_1 + \ldots + \beta_s K_s$ entonces

$$(f_1, f_2) := \alpha^t K \beta \; ; \quad (K)_{ij} := K(\xi_i, \xi_j)$$

Ejercicio: verificar que $(\Psi, (\cdot,\cdot)_\Psi)$ es un espacio con producto interno, de hecho, un espacio normado con la norma definida de la siguiente manera:

$$\|f\|_\Psi^2 := (f, f)_\Psi$$

Def.: (RKHS) $H_K := cl_{\|\cdot\|_\Psi}(\Psi)$ (clausura de Ψ)

H_K es un espacio de funciones continuas.

3.3 El truco de la nucleación del espacio

Notas:

1. El cálculo de productos internos puede hacerse, básicamente, evaluando la función K.

2. Si $f \in H_K$ entonces $(f, K_i)_\Psi = f(\xi_i)$

3. En la práctica, H_K es un espacio de dimensión finita (y por lo tanto, isomorfo a $\mathbb{R}^d$ para algún $d>n$.

3.4 kernel-PCA

Análisis de componentes principales en el caso no lineal (idea general)

$$\mathbb{R}^n \xrightarrow{\Phi} H_K$$

input space RKHS

El mapeo Φ no necesita conocerse explícitamente, en general, es no lineal y complicado.

$$\{X_i : i = 1, \ldots, m\} \quad \Rightarrow \quad X_i \to \hat{\Phi}(X_i) := \Phi(X_i) - \frac{1}{m}\sum_{i=1}^{m} \Phi(X_i)$$

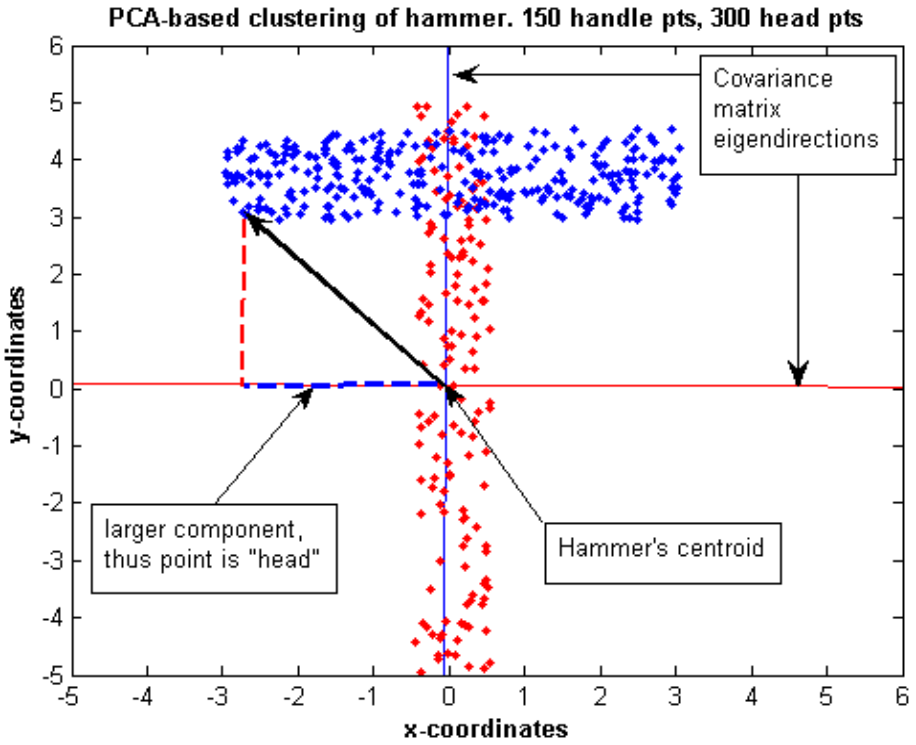
La matriz de covarianza en RKHS es un operador lineal: $\hat{C} : H_K \to H_K$

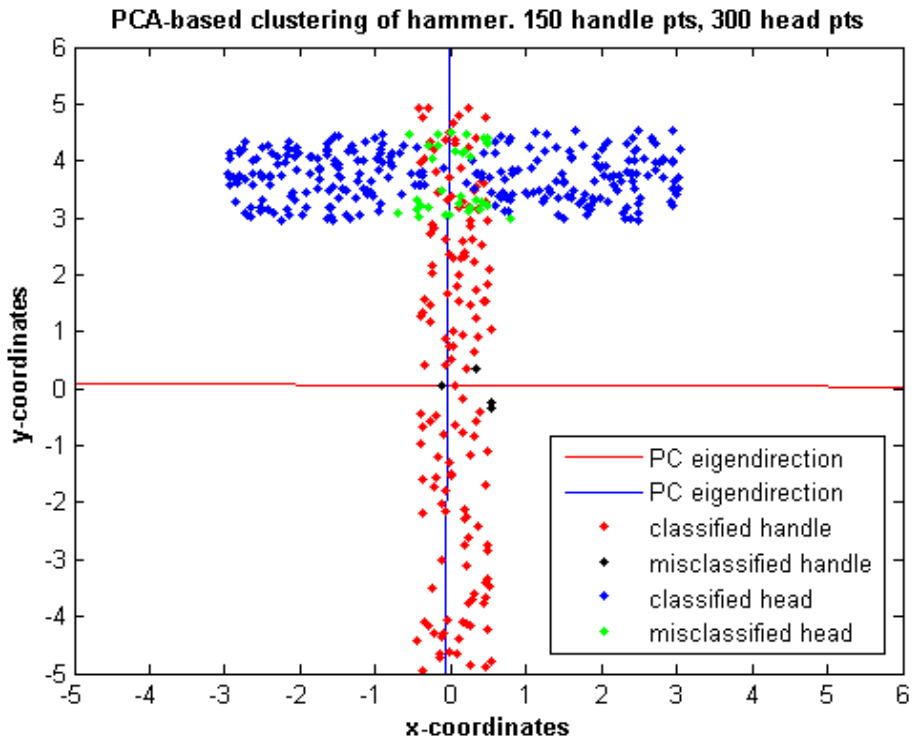
$$\hat{C}(\cdot) := \frac{1}{m-1}\sum_{i=1}^{m} \hat{\Phi}(X_i)(\hat{\Phi}(X_i), \cdot)_K$$

En la gran mayoría de las aplicaciones H_K es un espacio finito-dimensional,
$H_K \approx \mathbb{R}^d$, $d \gg n$ y por lo tanto $\hat{C}$ es una matriz cuadrada simétrica positivo-definida.

4. Algunos resultados de reportes de investigación no publicados

Uso del Análisis de Componentes Principales Lineal (L-PCA) para la determinación de la forma de objetos no convexos.

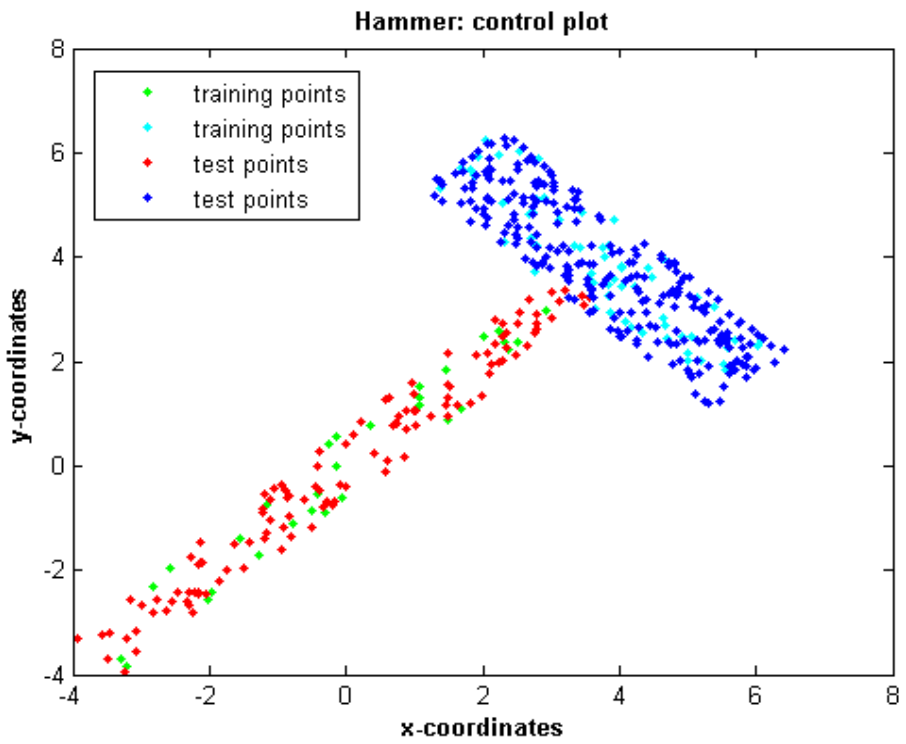


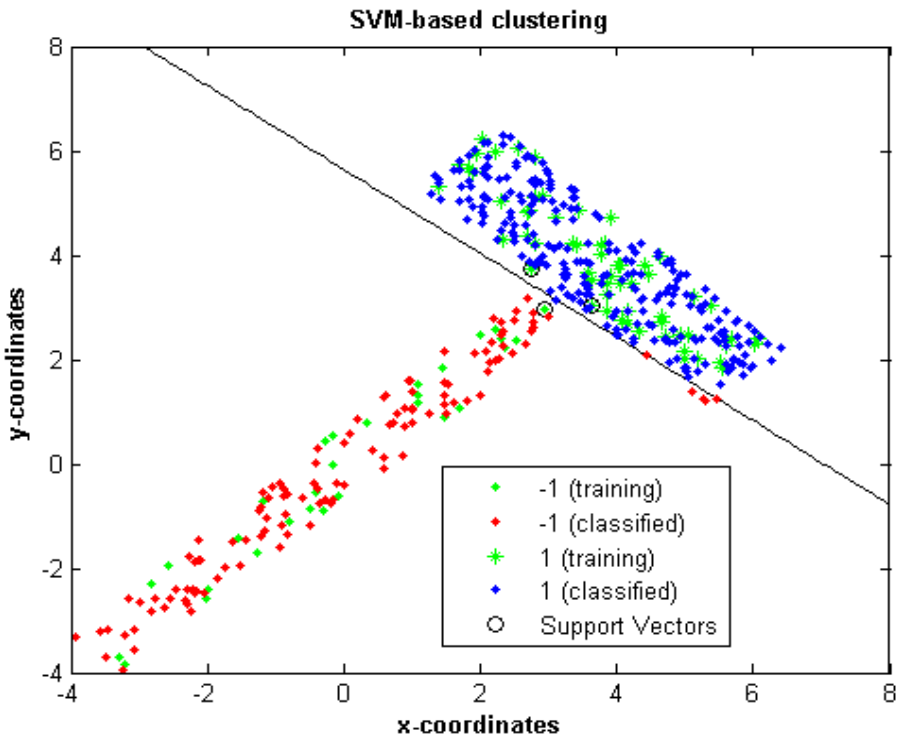


"Martillo." Los puntos de la cabeza y el mango del martillo son generados utilizando una distribución uniforme. Los vectores propios de la matriz de covarianza de los datos son calculados, luego cada punto de la imagen es clasificado de acuerdo al tamaño de su proyección sobre los vectores propios. Un punto es considerado parte del mango si su proyección sobre el vector propio paralelo al eje del mango es mayor que la proyección sobre el vector propio paralelo al eje de la cabeza. Análogamente para los puntos de la cabeza.

4. Algunos resultados de reportes de investigación no publicados

Uso de SVMs para la determinación de las partes mas significativas de una figura parcialmente clasificada.

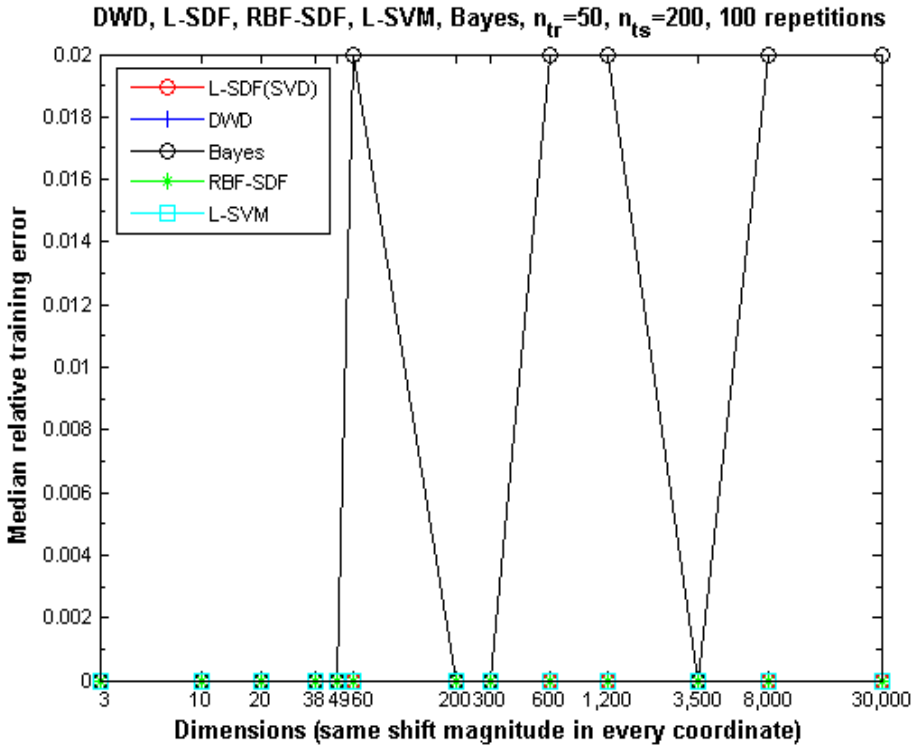




Los puntos en rojo y azul de la fig. de la izq. son los puntos del mango y cabeza, respectivamente, sin clasificar. La figura de la derecha muestra en rojo todos los puntos que fueron clasificados como “mango” y en azul todos los puntos que fueron clasificados como “cabeza.” Basándose en esta clasificación semi-automática, se obtuvo dos elipses, una para los puntos de la cabeza y otra para los puntos del mango (no se muestra).

4. Algunos resultados de reportes de investigación no publicados

A continuación se muestran resultados comparativos de clasificación de datos sintéticos utilizando diversos métodos de clasificación: SVMs, Signed Distance Function (SDF), Distance Weighted Discrimination (DWD), etc. (ver referencias al final para obtener fuentes de información sobre estos métodos).



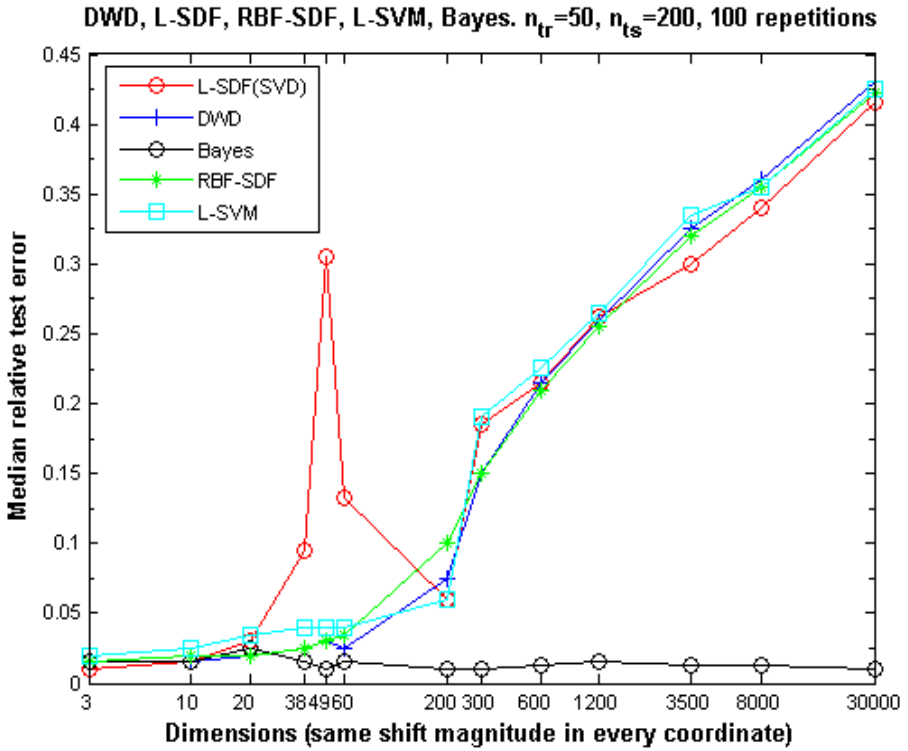


Figura de la izquierda: resultados de clasificación sobre el conjunto de entrenamiento. Figurea de la derecha: resultados de clasificación sobre el conjutnto de prueba (examen Del poder de generalización de los métodos de clasificación –ver página siguiente)

4. Algunos resultados de reportes de investigación no publicados

Los datos son vectores en $\mathbb{R}^n$ cuyas componentes se han generado utilizando una distribución normal. Estos vectores se han dividido en dos clases (positiva y negativa). A los vectores de una clase se les ha trasladado una distancia *d* en una dirección predeterminada, mientras que a los vectores de la otra clase se les ha trasladado la misma distancia pero en la dirección opuesta. De esta manera la clasificación de los datos se puede hacer observando las proyecciones de sus vectores en una dirección apropiada pero desconocida.

Para probar el poder de generalización de los métodos se han utilizado los clasificadores construidos con el conjunto de entrenamiento sobre otros datos generados de la misma forma.

Este proceso se ha repetido incrementando cada vez la dimensión de los datos (cuidando que los centros de masa de cada clase estén siempre a la misma distancia).

El comportamiento de las gráficas de error es un tema que no está completamente entendido.

5. Palabras finales

Los temas anteriores son una parte muy pequeña de un problema cuya complejidad es tan rica como lo es su utilidad científica y tecnológica.

Hemos empleado la mayor parte de la exposición al caso de clasificación binaria lineal, el cual es la base para problemas de clasificación más complejos como lo son el problema de clasificación multiclase no lineal. En las referencias el lector encontrará fuentes de información muy útiles para estudiar otros casos a profundidad.

A lo largo de esta presentación hemos expuesto algunas preguntas las cuales creemos pueden ser las directrices de una tesis de licenciatura muy interesante. Aquellas personas que deseen obtener más información, están cordialmente invitadas a escribirme un correo electrónico.

El tema de clasificación y análisis de datos es un tema muy actual y de intensa actividad hoy en día. La demanda de personas expertas en estos campos perdurará mientras El flujo de información sobrepase nuestra capacidad de análisis del mismo.

6. Recursos en linea

Existe una gran cantidad de información disponible en-línea. **MIT** tiene disponibles en internet cursos enteros, incluidos videos de clases y material de trabajo (tareas, etc).

6. Recursos en linea

Material para cada curso abierto para todo público

Home > Courses > Brain and Cognitive Sciences > Statistical Learning Theory and Applications

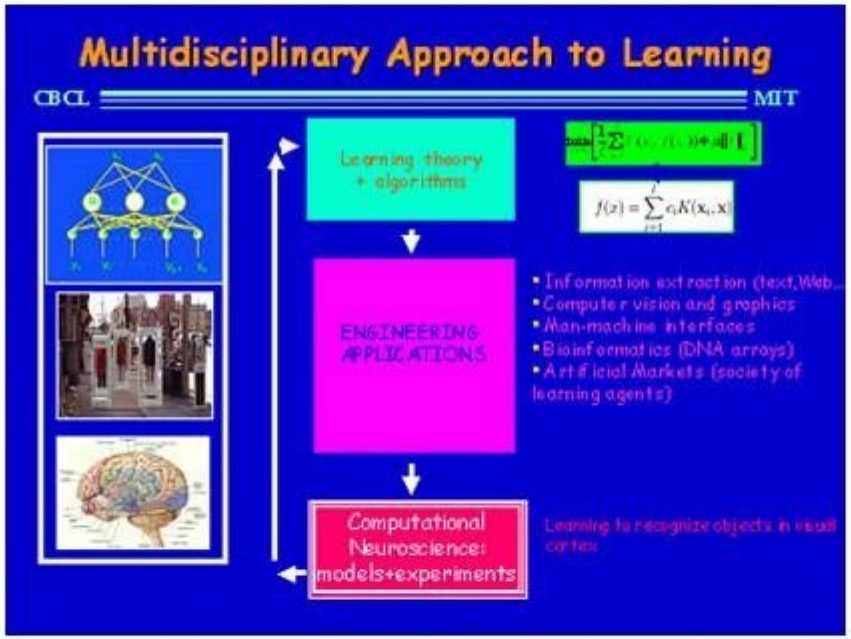
Email this page

› VIEW ALL COURSES

› Course Home
› Syllabus
› Calendar
› Readings
› Lecture Notes
› Assignments
› Download Course Materials

9.520 Statistical Learning Theory and Applications

Spring 2006



Designing and building a system that will function the same way as a human visual system, but without getting bored, and with a greater degree of accuracy. (Image courtesy of Poggio Laboratory, MIT Department of Brain and Cognitive Sciences.)

Course Highlights

This course features extensive lecture notes. The assignments focus on some of the functions needed to make problem-solving more efficient for computer systems.

Staff

Instructor:
Prof. Tomaso Poggio

Course Meeting Times

Lectures:
Two sessions / week
1.5 hours / session

Level

Graduate

***Translations**

› Chinese (Simplified)

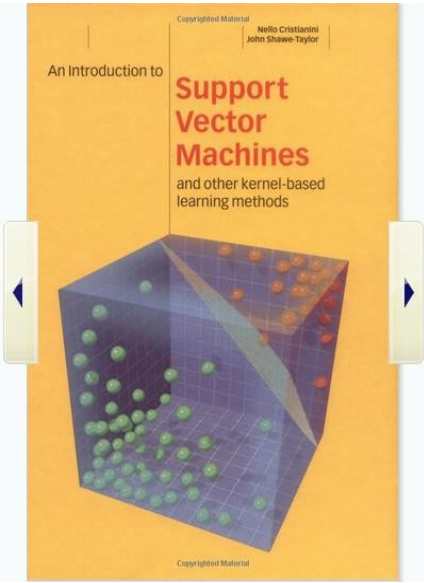
› **Download these course materials**

Feedback

› Send feedback on this course.

› Find out how much your company uses OCW.

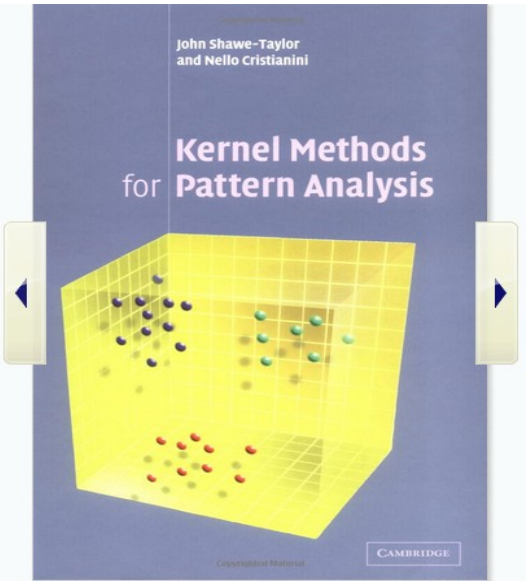
6. Recursos en linea http://www.support-vector.net/



Introducción SVMs y a la teoría del aprendizaje supervisado.

- Comparación con otros métodos de clasificación así como su presentación.
- Referencias (casi) completas a trabajos mas sobresalientes (no incluye últimos 10 años). Incluye también referencias por capítulo.
- **Software libre** (Matlab, java, etc).
- Páginas de web y conferencias sobre teoría del aprendizaje.
- Becas (Estancias 1 año. Requiere experiencia en programación, java y menjando bases de datos).
- Anuncios de escuelas sobre análisis de datos.

- Listas de algoritmos, núcleos, códigos en Matlab y pesudocódigos empleados incluyendo referencias de página.
- Códigos y herramientas de matlab disponibles gratuitamente.
- Tutorial sobre análisis de patrones (powepoint, 1.28 MB).
- Becas (1 año. Requiere experiencia en programación, java y manejo de bancos de datos).
- Anuncios de escuelas de análisis de datos.



6. Recursos en linea

Algunas bases de datos de expresión genética:

Stanford Microarray Database: http://smd.stanford.edu/
Yale Microarray Database: http://info.med.yale.edu/microarray/
National Center for Biotechnology Information (NCBI) – **G**ene **E**xpression Omnibus:
http://www.ncbi.nlm.nih.gov/geo/

Otros sitios de interés:

Machine Learning International Conferences: 2009 / 2010 (IMLS)
Machine Learning Summer Schools
Machine Learning Summer School, Tübingen 2003 (videolectures.net)
Foundations of Data and Visual Analytics (FODAVA) –Georgia Tech.

Páginas personales:

Vladimir Vapnik (NEC Laboratories , Empirical Inference lecture)
Corinna Cortes (Google Research , video lecture ICML'09)
Bernhard Schölkpof (Max Planck Institute for Biological Cybernetics)
Jean-PhilippeVert (Mines ParisTech/Institut Curie)
Susan Dumais (Microsoft Research)

7. Referencias (1/4)

I. PCA y aplicaciones

Wijewickrema, S. N. R.; Paplinski A. P. Principal Component Analysis for the approximation of a fruit as an ellipse. Publicación electrónica (2004). Disponible via CiteSeer-x (Beta).

II. Perceptrons

Freund, Y.; Schapire, R. E. Large margin classification using the preceptron algorithm, *Machine Learning*, 37 (3): 277-296, (1999).

III. SVMs y métodos de clasificación de tipo nuclear.

Cristianini, N.; Shawe-Taylor. *Support Vector Machines and other kernel-based methods*. Cambridge University Press (2000).

Cristianini, N.; Shawe-Taylor. *Kernel methods for pattern analysis*. Cambridge University Press (2004).

Cortes, C.; Vapnik, V. Support-Vector Networks. Machine Learning **20** (1995) pp. 273-297

7. Referencias (2/4)

IV. Aplicaciones en biomedicina y bioinformática.

Schölkopf, B.; Tsuda, K.; Vert, J-P. *Kernel methods in computational Biology*. MIT Press (2004).

V. Aplicaciones diversas en otras ramas de la Ingenieria.

Hearst, M. A. Trends and controversies, Support Vector Machines.
IEEE Intelligent Systems, July/August 1998, pp. 18-28.

Dumais, S. T.; Platt, J.; Heckerman, D.; Shami, M. Inductive learning algorithms and representations for text categorization. *Proceedings of ACM-CIKM98,* Nov. 1998 pp. 148-155.

Osuna, E.; Freund, R.; Girosi, F. Training Support Vector Machines: an application to Face detection. (preprint to appear in *Proc. CVPR'97,* June 17-19, 1997).

Pfister, M; Behnke, S; Rojas, R. Recognition of hand-written zipcodes in a real-world, non-standard-letter sorting system. *Applied Intelligence,* 12 (2000) pp. 95-115

7. Referencias (3/4)

Kutyniok, G. Shearlets: a wavelet-based approach to the detection of directional features. *Oberwolfach report* **36** (2007) 35-38.

VI. Teoría de clasificación:

Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*. Wiley (2001).

VII. Programación lineal (método simplex) e introducción a la programación no lineal

Ferris, M.C.; Mangasarian, O.L.; Wright, S.J. *Linear Programming with Matlab*.
MPS-SIAM Series on Optimization (2007).
Material suplementario disponible en la página de web del libro.

VIII. Signed Distance Function (SDF)

Boczko, E.M.; Young, T.R. The Signed Distance Function: a new tool for binary classification. (2005) (Electronic Preprint)

Boczko, E.M.; Young, T.R.; Xie, M.; Wu, D. Comparison of binary classification based on Signed Distance Functions with Support Vector Machines. (Electronic publication.)

7. Referencias (4/4)

IX. Distance Weighted Discrimination (DWD)

Marron, J.S.; Todd, M.J.; Ahn, J. Distance Weighted Discrimination. ePrints for the Optimization Community (2004)

Benito, M.; Parker, J.; Du, Q.; Wu, J.; Xiang, D.; Perou, C.M.; Marron, J.S. Adjustment of systematic microarray data biases. *Bioinformatics* **20** no. 1 (2004) pp. 105-114.

Software para DWD: NCI / UNC
Plática de J.S. Marron en MSRI sobre DWD.

140 AÑOS DE HISTORIA Y TRADICIÓN

UAEH