



Aprendizaje no Supervisado

Diplomado en Computación Inteligente, UPB 2005

Jerónimo Castrillón Mazo
IEO UPB.
www.geocities.com/jeronimocm



Previa

- Muchos, demasiados campos!
- Abrir el panorama.
- No hay tiempo para verlos todos.
- Se hará énfasis en lo clásico.
- Veremos métodos transversales al ANS, p.e. PCA... muy importante... ver eigenfaces.
- Alentar a que busquen el método que les guste y profundicen en el.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronomocm@yahoo.com

2



Nombres!

- K-MEANS, K-MEDOIDS, K-NEAREST NEIGHBOR, FUZZY C-MEANS, SOFM, K-HARMONIC MEANS, LBG-U, NEURAL GAS, HEBBIAN, ART, FUZZY ART, CMAC, AGG-LANCE/WILLIAMS, DIV-KAUFMAN/ROUSSEAU, EM, CLARANS, FOCUS-CLARANS, ISODATA, DBSCAN, OPTICS (vble Eps), DENCLUE, DBCLASD, STING, SINGLE LINK, COMPLETE LINK, BIRCH (CENTROID LINK), CURE, SOON, PART-FILTERS, SIM-ANNEALING, MRKD TREES, SHORTEST SPANNING PATH, MST-MINIMAL SPANNING TREE, LEADER, KERNEL DENSITY ESTIMATION, WAVE CLUSTER, CATEGORICAL: STIRR, ROCK, CACTUS, PROJECTED: CLIQUE, PROCLUST, ORCLUST, OPTIC GRID, HD-EYE...
- PCA, ICA, KPCA, KL-EXP, IVH, R-TREE, X-TREE, V-A FILE, HILBERT R-TREE, SS-TREE, SR-TREE, DENSITY BASED OUTLIERS, GMM...

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

3





Contenido

- Introducción
- Esquema general del proceso de *clustering*.
- Métodos de representación: PCA.
- Medidas de similitud.
- Principales técnicas de *clustering*.
- Otras técnicas.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

4





Introducción

- Tipos de aprendizaje
- Objetivos del ANS.
- Usos del ANS.
- Diferentes acercamientos al ANS.

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

5



Tipos de Aprendizaje

Imagine un organismo o máquina que experimenta una serie de entradas sensoriales :

$$x_1, x_2, \dots, x_n$$

Dependiendo de la información adicional y de lo que se espera que el agente realice existen 4 tipos de aprendizaje.

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

6



Tipos de Aprendizaje (2)

- **Aprendizaje supervisado:** al agente se le entrega también salidas deseadas:

$$d_1, d_2, \dots, d_n$$

y debe aprender a producir salidas correctas ante nuevas entradas.

- **Aprendizaje no supervisado:** El objetivo del agente es construir representaciones de x que sean útiles para procesos de decisión, predicción, reconocimiento, razonamiento, comunicación, etc.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

7



Tipos de Aprendizaje (3)

- **Aprendizaje por refuerzo:** el agente produce también acciones:

$$a_1, a_2, \dots, a_n$$

que modifican el estado del ambiente y recibe recompensas (o castigos):

$$r_1, r_2, \dots, r_n$$

Su objetivo es aprender a actuar de tal manera que maximice la recompensa a largo plazo.

- **Aprendizaje con índice de desempeño:** El agente recibe una medida cuantitativa por su comportamiento.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

8



Objetivos del ANS

Encontrar representaciones útiles de los datos, por ejemplo:

- *Clusters*.
- Reducción de dimensionalidad.
- Hallar fenómenos o datos escondidos en un *set* de datos.
- Modelado de la densidad de probabilidad de la cual son extraídos los datos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

9



Usos del ANS

Entre muchas otros usos se pueden distinguir:

- Compresión de datos.
- Detección de características generales.
- Clasificación.
- **Facilitar otros procesos de aprendizaje.**
- Teoría y comprensión de los procesos de aprendizaje humanos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

10



Diferentes acercamientos al ANS

- Redes neuronales artificiales:
 - Con ANN se simula el comportamiento observado en los seres vivos.
 - Se basan, por lo general, en el aprendizaje Hebbiano (Hebb, D 1949).
 - “Los pesos de las sinapsis cambian en proporción a la correlación entre señales pre- y post-sinápticas”.
 - Nosotros tampoco sabemos cual es la salida deseada!
- *Clustering*:
 - No necesariamente implica el uso de una ANN.
 - Se verá más adelante.

Existen modelos neuronales para muchos métodos de *clustering* pero usualmente es más fácil entender el proceso desde el *clustering*.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

11



Contenido

- Introducción
- Esquema general del proceso de *clustering*.
- Métodos de representación: PCA.
- Medidas de similitud.
- Principales técnicas de *clustering*.
- Otras técnicas.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

12



Esquema general del proceso de *clustering*.



- Que es *clustering*?
- Esquema general.
- Representación.
- Medidas de similitud.
- Técnicas de *clustering* y su clasificación.
- Abstracción.
- Validación.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

13



Que es *clustering*?



- Es uno de los campos más amplios del ANS.
- Busca agrupar datos similares, según una determinada medida de similitud, en *clusters* o grupos diferentes.
- Es utilizado para análisis de patrones, exploración de datos (*data mining*), recuperación de documentos, segmentación de imágenes* y clasificación de patrones entre otras aplicaciones.
- Ejemplos de aplicaciones: segmentación de mercados, usuarios WWW, **estructuración jerárquica de documentos**, mapas temáticos en imágenes satelitales.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

14



Que es *clustering*? (2)

A.K. Jain *et al*:

“Even though there is an increasing interest in the use of clustering methods in **pattern recognition**, **image processing** and **information retrieval**, clustering has a rich history in other disciplines such as biology, psychiatry, psychology, archaeology, geology, geography, and marketing.

Other terms ... *unsupervised learning*, *numerical taxonomy*, *vector quantization*, and *learning by observation*. The field of spatial analysis of point patterns is also related to cluster analysis. The importance and interdisciplinary nature of clustering is evident through its vast literature”

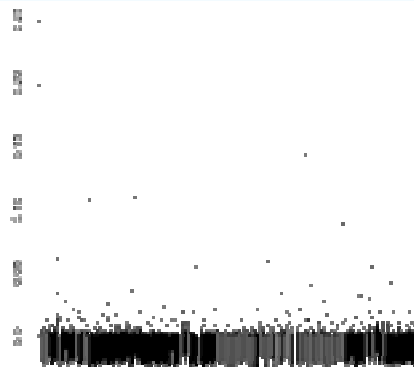
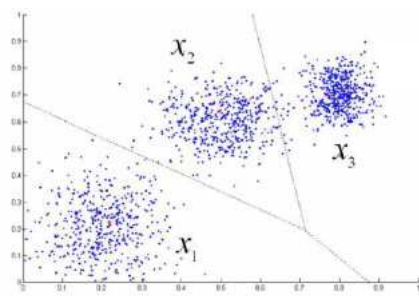
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

15



Que es *clustering*? (3)



Tomado de “Data Clustering: A Review”, JAIN, A.K., MURTY, M.N. y FLYNN, P.J. En: ACM Computing Surveys. Ohio university. Vol. 31, No. 3 (Sep. 1999); p. 306.

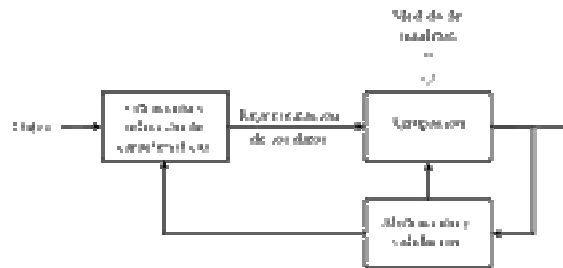
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

16



Esquema General



- Los datos son representados (transformados) de una mejor forma.
- Se agrupan mediante alguna técnica de *clustering* utilizando una medida de similitud.
- Se interpretan y prueban los resultados para proveer una realimentación a los primeros procesos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

17



Representación

- Se refiere a un proceso de extracción y/o selección de características.
- **Selección** de características: toma algunas características que describan los datos de manera aceptable reduciendo la dimensionalidad.
- **Extracción** de características: es una transformación para hacerlas más fáciles de agrupar (ej. una transformación de coordenadas).

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

18



Medida de Similitud

- La medida de similitud indica qué tan similares son un par de datos.
- Generalmente es una medida de distancia.
- Dos datos son similares si están cerca en el espacio de entrada según una métrica determinada.
- Existe mucho trabajo hecho al respecto, sobre todo para datos no numéricos.
- Más adelante, más sobre esto

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

19



Técnicas de *Clustering* y su Clasificación

Existen muchas formas de agrupar los datos. Para elegir una técnica determinada se debe tener en cuenta:

- Cantidad de datos: tamaño de la base de datos.
- Complejidad computacional deseada.
- Tipo de datos: numéricos conmensurables, no conmensurables, forma esperada de los patrones, etc.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

20



Técnicas de *Clustering* y su Clasificación (2)



Tipos de datos (Gowda y Diday, 1992):

- Características cuantitativas:
 - Valores continuos (peso, estatura, velocidad).
 - Valores discretos (número de carros por hora).
- Características cualitativas:
 - Nominales o sin orden: color, textura.
 - Ordinales: medidas cualitativas relativas a una escala de valores, p.e. temperatura (frio - caliente).

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

21



Clasificación *Clustering*



De acuerdo a la partición*:

- *Aglomerativas* (divisivas): cada dato es un *cluster* y éstos se van juntando hasta cumplir un criterio de paro.
- *Unicaracterística* (*multicaracterística*): cuantas características se toman en cuenta en el algoritmo.
- Partición fuerte (difusa): un dato sólo puede pertenecer a uno y sólo un *cluster*.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

22



Clasificación *Clustering* (2)

Según el tipo de algoritmo:

- Jerárquicos:

- *Cluster* anidado que genera un dendograma (**TABLERO**).
- Puede ser aglomerativo o divisivo.
- Provee múltiple resolución.
- La unión o separación de *clusters* se hace de acuerdo a una distancia mínima.
- Dos tipos principales: *single link* y *complete link*.
- SL: la distancia entre dos *clusters* C1, C2 se define:
$$d(C1, C2) = \min_{i,j} d(c_i^1, c_j^2), \forall c_i^1 \in C1, c_j^2 \in C2$$
- CL: la distancia se define a partir del máximo.
- CL se prefiere, es más robusto y menos sensible al ruido.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

23



Clasificación *Clustering* (3)

- Particionales: dividen el espacio en una partición (fuerte o difusa) generalmente mediante división **son de los más utilizados**.
- Basados en modelo: tratan de buscar un modelo (generalmente mezclas de densidades de probabilidad) a partir del cual los datos pudieran haber sido obtenidos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

24



Clasificación *Clustering* (4)

- Basados en análisis de densidad: agrupan los datos según la densidad de los mismos en una determinada región. Permiten formar *clusters* de diversas formas.
- Basados en *grilla*: por lo general hacen una primera división del espacio en una grilla. En las celdas se pueden utilizar otras técnicas de *clustering*.

Feb-Jun, 2005

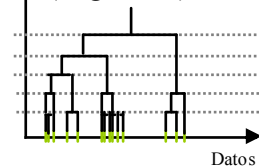
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

25

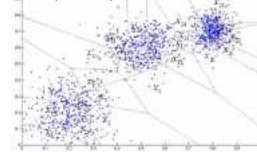


Técnicas de *clustering*

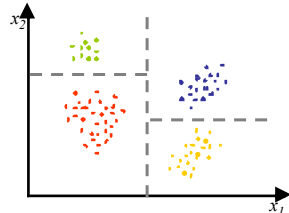
Jerárquico por aglomeración
(*Single Link*)



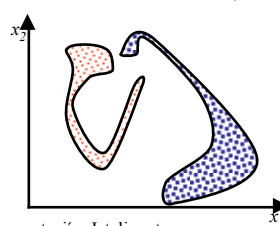
Part. fuerte multicaracterística
(so fm)



Part. fuerte unicaracterística



Basado en densidad (DBSCAN)



Feb-Jun, 2005

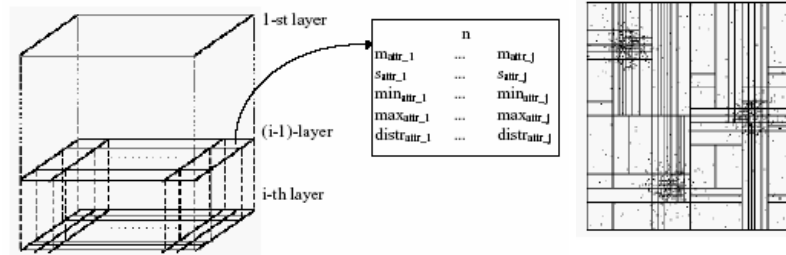
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

26



Técnicas de *clustering* (2)

Basado en Grilla: STING (STatistical INformation Grid based)



Tomados de: “*Clustering for Mining in Large Spatial Databases*”, Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. En: Special Issue on Data Mining, KI-Journal, ScienTec Publishing, Vol. 1, 1998

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

27



Abstracción

- Interpretación de los *clusters* obtenidos (humana o automática).
- Es común el análisis del centroide.
- Ayuda a visualizar el resultado del *clustering* y analizar su pertinencia.

Cualquier método entregará una solución ... La abstracción permite analizar cual es la más apropiada.

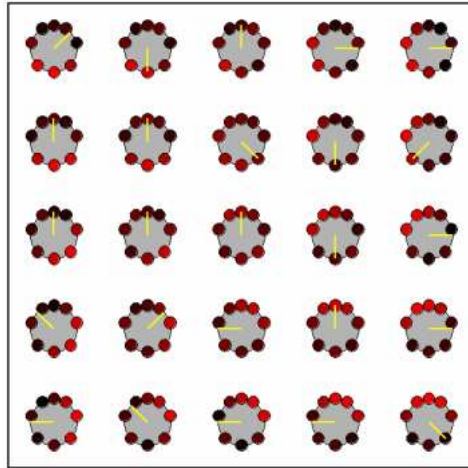
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

28



Ejemplo de abstracción



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

29



Validación

- Es un proceso más riguroso que la validación.
- Se basa en análisis de indicadores de desempeño de los algoritmos.
- Generalmente se torna intuitivo.
- Es más fácil de realizar cuando se utilizan métodos que optimizan alguna cantidad (ej. *Log-likelihood* en EM).
- Mucho por hacer...

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

30



Contenido

- Introducción
- Esquema general del proceso de *clustering*.
- **Métodos de representación: PCA.**
- Medidas de similitud.
- Principales técnicas de *clustering*.
- Otras técnicas.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

31



Métodos de representación: PCA.

- Introducción.
- Preliminares: estadística.
- Preliminares: álgebra lineal.
- Método de análisis de componentes principales (PCA: *principal component analysis*).
- Ejemplo Sencillo.
- Otros métodos: KPCA e ICA.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

32



Introducción

- La representación busca reducir dimensionalidad y/o complejidad en los datos.
- PCA es un método lineal moderno de análisis de datos que logra este objetivo.
- Es muy utilizado en procesamiento de imágenes, reconocimiento de caras, neurociencia, *data-mining* y muchos más.
- Es un método **no paramétrico**.
- Jon Shlens, UCSD: “*PCA is a mainstay of modern data analysis - a black box that is widely used but poorly understood ... has been called one of the most valuable results from applied linear algebra*”

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

33



Introducción (2)

- Notación:
 - K experimentos o mediciones diferentes.
 - N variables diferentes medidas (voltaje, velocidad, ...).
 - Set de datos (mediciones):

$$X \in M_{N \times K}$$

- Se llamará x^n a la n -ésima fila de la matriz X .

Feb-Jun, 2005

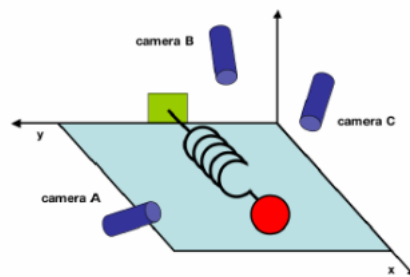
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

34



Introducción (3)

Idea intuitiva:



- Dinámica que depende sólo de x .
- Fenómeno medido a lo largo de 3 ejes arbitrarios mediante tres cámaras.
- Cómo descubrir la dinámica simple unidimensional, a partir de datos 3-dimensionales?

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

35



Introducción (4)

- Ejemplo no trivial.
- Muchas veces un fenómeno se mide como se puede y no de la mejor manera.
- Además las mediciones pueden ser redundantes ... pueden sobrar datos.
- Adicionalmente está el **ruido**.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

36



Preliminares: Estadística

- **Media:** promedio de las mediciones tomadas de una misma variable (p.e. de una fila de X):

$$\bar{x}^n = \frac{\sum_{k=1}^K x_k^n}{K}$$

- **Varianza:** medida de la variación de las mediciones de una misma variable (una fila de la matriz) respecto a su media.

$$\sigma^2 = \frac{\sum_{k=1}^K (x_k^n - \bar{x}^n)^2}{K - 1}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

37



Preliminares: Estadística (2)

- **Covarianza:** mide qué tanto varía cada una de las dimensiones de su media en relación con las otras. La covarianza entre el vector de medidas x^n y x^m está dado por:

$$\text{cov}(x^n, x^m) = \frac{\sum_{k=1}^K (x_k^n - \bar{x}^n) \cdot (x_k^m - \bar{x}^m)}{K - 1}$$

La covarianza de un vector consigo mismo es la varianza.

Esta matriz ayuda a observar la correlación que existe entre los datos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

38



Preliminares: Estadística (3)

Correlación según el signo de la matriz de covarianza:

- Si es positiva, indica que hay una correlación directa entre las variables
- Si es negativa, indica una relación inversa
- Si la covarianza es cero, las dos variables no tienen relación

Gráficamente, en un set de datos correlacionados si se fija una componente se está restringiendo el rango para otras componentes.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

39



Preliminares: Estadística (4)

- **Matriz de covarianza:** almacena la covarianza de cada variable respecto a todas las demás. Por lo tanto es una matriz simétrica con la varianza de cada variable en la diagonal y de dimensión $N \times N$.

$$C = [c_{nm}] = \begin{bmatrix} \text{cov}(x^1, x^1) & \text{cov}(x^1, x^2) & \cdots & \text{cov}(x^1, x^N) \\ \text{cov}(x^2, x^1) & \text{cov}(x^2, x^2) & \cdots & \text{cov}(x^2, x^N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x^N, x^1) & \text{cov}(x^N, x^2) & \cdots & \text{cov}(x^N, x^N) \end{bmatrix}$$

Note que si:

$$\bar{x}^j = 0 \quad \forall j = 1, 2, \dots, K \quad \text{entonces} \quad C = \frac{1}{K-1} \cdot X \cdot X^T$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

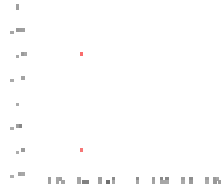
40



Preliminares: Estadística (5)

Ejemplito:

$$X = \begin{bmatrix} 0.6 & 0.4 \\ 0.6 & 0.6 \\ 0.4 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}^T$$



$$\bar{x}^1 = \frac{0.6 + 0.6 + 0.4 + 0.4}{4} = 0.5,$$

$$\bar{x}^2 = \frac{0.4 + 0.6 + 0.6 + 0.4}{4} = 0.5$$

$$\begin{aligned} Cov(x^1, x^2) &= \frac{(0.6 - 0.5)(0.4 - 0.5) + (0.6 - 0.5)(0.6 - 0.5) + \dots}{4 - 1} \\ &= \frac{(0.4 - 0.5)(0.4 - 0.5) + (0.4 - 0.5)(0.6 - 0.5)}{4 - 1} = \frac{-0.1^2 + 0.1^2 + 0.1^2 - 0.1^2}{3} = 0 \end{aligned}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

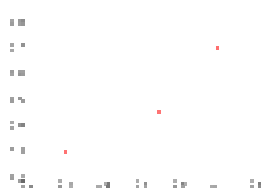
41



Preliminares: Estadística (6)

Ejemplito: (Cont.)

$$X = \begin{bmatrix} 0.3 & 0.3 \\ 0.7 & 0.7 \\ 0.45 & 0.55 \\ 0.55 & 0.45 \end{bmatrix}^T$$



$$\bar{x}^1 = \frac{0.3 + 0.7 + 0.45 + 0.55}{4} = 0.5,$$

$$\bar{x}^2 = \frac{0.3 + 0.7 + 0.55 + 0.45}{4} = 0.5$$

$$\begin{aligned} Cov(x^1, x^2) &= \frac{(0.3 - 0.5)(0.3 - 0.5) + (0.7 - 0.5)(0.7 - 0.5) + \dots}{4 - 1} \\ &= \frac{(0.45 - 0.5)(0.55 - 0.5) + (0.55 - 0.5)(0.45 - 0.5)}{4 - 1} = \frac{0.2^2 + 0.2^2 - 0.05^2 - 0.05^2}{3} \\ Cov(x^1, x^2) &= 0.25 \end{aligned}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

42



Preliminares: Álgebra Lineal

- **Valores propios:** los valores propios de una matriz A son los λ_i que satisfacen la ecuación:

$$|A - \lambda \cdot I| = 0$$

- **Vectores propios:** Son los vectores que generen la solución del sistema asociado a cada λ_i :

$$(A - \lambda_i \cdot I)x = 0$$

- **Base propia*:** Si la dimensión algebraica de cada valor propio es igual a la dimensión geométrica, los vectores propios normalizados forman una base ortonormal conocida como base propia.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

43



Preliminares: Álgebra Lineal (2)

- **Matriz propia:** es la matriz que posee los elementos de la base propia (e.g. los vectores propios) en sus columnas.

Hecho: toda matriz simétrica posee valores propios reales y matriz propia.

- **Diagonalización:** toda matriz A que posea una matriz propia E puede expresarse por medio del producto:

$$A = E \cdot D \cdot E^T$$

donde D es una matriz diagonal en los valores propios de A .

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

44



Preliminares: Álgebra Lineal (2)

- **Matriz de cambio de base:** un espacio vectorial posee infinitas bases. Una matriz de cambio de base permite expresar en una nueva base un vector expresado en otra base. Este procedimiento permite observar mejor los datos.

Si P es una matriz de cambio de base y x es un vector expresado en la base inicial, entonces el vector en la nueva base se determina mediante:

$$y = P \cdot x$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

45



Preliminares: Álgebra Lineal (3)

Ejemplito:

$$T: R^3 \rightarrow R^3, T(x, y, z) = [3x + y, 2x - y, y]$$

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 5 & -2 \\ 2 & -2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \Rightarrow \begin{matrix} |A - \lambda I| = 0 \\ -\lambda^3 + 9\lambda^2 + 14\lambda + 26 = 0 \end{matrix} \Rightarrow \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} 4.2 \\ -1 \\ 5.8 \end{bmatrix}$$

$$E = \begin{bmatrix} -0.87 & -0.48 & -0.08 \\ -0.1 & 0.34 & -0.93 \\ -0.48 & 0.05 & 0.35 \end{bmatrix} \Rightarrow A = E \cdot D \cdot E^T$$

Si los vectores de entrada son expresados en la base propia, la matriz asociada a T es la diagonal D .

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

46



PCA

- Componente principal: vector a lo largo del cual los datos presentan una gran variación.
- PCA: método que halla una matriz de cambio de base por medio de la cual el *set* de datos transformados queda desprovisto de correlación (e.g. matriz de covarianza diagonal).
- **Extrae características:** mediante la rotación se reduce la redundancia.
- **Selecciona características:** una vez transformados los datos, se pueden eliminar las componentes “menos principales”.
- **Suposición: la variación está directamente relacionada con la información.**

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

47



PCA: Explicación

- Se supone que el *set* de datos original (X) se le ha extraído la media. Por lo tanto las filas del *set* de datos transformado (Y) también tendrán media cero (**Ilustrar**).
- El problema: hallar la matriz P de cambio de base

$$Y = P \cdot X$$

para que el nuevo *set* de datos Y posea una matriz de covarianza diagonal:

$$C_Y = \frac{1}{K-1} \cdot Y \cdot Y^T = \frac{1}{K-1} \cdot (P \cdot X) \cdot (P \cdot X)^T$$

$$C_Y = \frac{1}{K-1} \cdot P \cdot X \cdot X^T \cdot P^T = P \cdot C_X \cdot P^T$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

48



PCA: Explicación (2)

- La matriz C_X es simétrica por lo tanto posee una base propia y puede diagonalizarse:

$$C_X = E \cdot D \cdot E^T$$

Observe que si se elige $P = E^T$ y se reemplaza en la expresión para C_Y :

$$C_Y = P \cdot C_X \cdot P^T$$

$$C_Y = P \cdot (P^T \cdot D \cdot P) \cdot P^T = I \cdot D \cdot I$$

$$C_Y = D$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

49



PCA: Procedimiento

- Paso 1: organizar los datos en una matriz X de manera que cada experimento quede registrado en una columna.
- Paso 2: extraer la media de cada fila de X para obtener una nueva matriz X' .
- Paso 3: hallar la matriz de covarianza $C_{X'}$.
- Paso 4: extraer los valores y vectores propios de $C_{X'}$.
- Paso 5: elegir cuántas componentes son relevantes, formar la matriz de cambio de base P y transformar los datos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

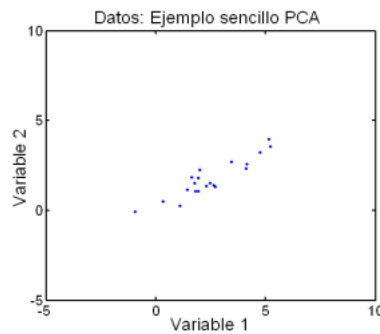
50



Ejemplo Sencillo

- Paso 1- Los datos:

x	y
4.7541	3.1944
1.7953	1.4917
1.819	1.0512
4.112	2.2963
4.1498	2.5465
1.1213	0.25804
0.35205	0.48722
1.9387	1.7739
2.7121	1.2996
1.6484	1.8261
2.3267	1.3347
3.4448	2.6955
5.2408	3.5376
2.4925	1.4849
-0.92949	-0.095649
5.1726	3.9335
2.6566	1.3837
1.456	1.1168
1.9364	1.0708
2.0247	2.2351



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

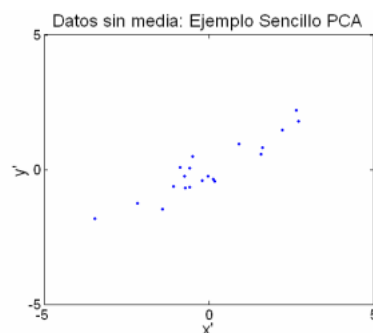
51



Ejemplo Sencillo (2)

- Paso 2- Extracción de la media

x	y	x'	y'
4.7541	3.1944	2.2429	1.4484
1.7953	1.4917	-0.71588	-0.25444
1.819	1.0512	-0.69225	-0.69485
4.112	2.2963	1.6008	0.55023
4.1498	2.5465	1.6386	0.80044
1.1213	0.25804	-1.3899	-1.4881
0.35205	0.48722	-2.1592	-1.2589
1.9387	1.7739	-0.57252	0.027769
2.7121	1.2996	0.20086	-0.44653
1.6484	1.8261	-0.86284	0.079974
2.3267	1.3347	-0.18454	-0.41143
3.4448	2.6955	0.93362	0.94939
5.2408	3.5376	2.7296	1.7915
2.4925	1.4849	-0.018688	-0.2612
-0.92949	-0.095649	-3.4407	-1.8417
5.1726	3.9335	2.6614	2.1874
2.6566	1.3837	0.14541	-0.36243
1.456	1.1168	-1.0553	-0.62927
1.9364	1.0708	-0.57484	-0.67527
2.0247	2.2351	-0.48655	0.48899



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

52



Ejemplo Sencillo (3)

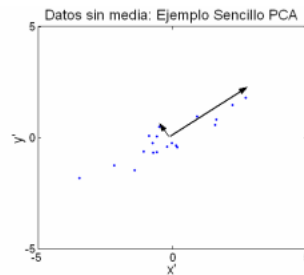
- Paso 3- Matriz de covarianza:

$$C_{X'} = \begin{bmatrix} 2.524 & 1.552 \\ 1.552 & 1.123 \end{bmatrix}$$

- Paso 4- Valores y vectores propios:

$$\lambda = \begin{bmatrix} 3.527 \\ 0.121 \end{bmatrix}$$

$$E = \begin{bmatrix} -0.840 & 0.543 \\ -0.543 & -0.840 \end{bmatrix}$$



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

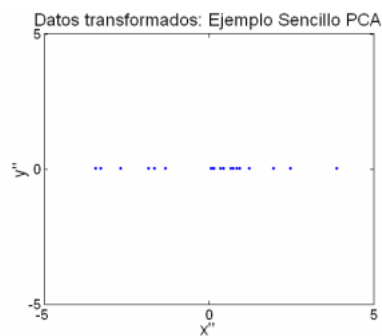
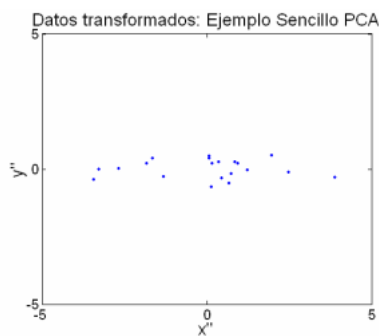
53



Ejemplo Sencillo (4)

- Paso 5: Se puede eliminar la segunda componente.

$$P = \begin{bmatrix} -0.840 & -0.543 \\ 0 & 0 \end{bmatrix}$$



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

54



PCA: Otros Acercamientos

Tarea:

- Consultar implementaciones *on-line* de PCA.
- Consultar implementaciones Neuronales de PCA (Oja, 1985)

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

55



Otros Métodos: KPCA e ICA

- PCA es un método muy poderoso por ser no paramétrico. Pero esto lo hace poco flexible.
- Con los mismos datos siempre arroja el mismo resultado, e.g. es independiente del usuario.
- KPCA: A veces un conocimiento previo de los datos puede ayudar, este conocimiento se puede poner en un *kernel* (transformación no lineal) que mejore el desempeño de PCA.
- ICA: A veces la distribución de las variables no se aproxima a la normal y PCA falla.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

56



Kernel PCA (KPCA)

- Consiste en aplicar un *kernel* o transformación no lineal antes de aplicar el método de PCA.
- Ejemplo: transformación a coordenadas curvilíneas.
- El mayor problema reside en la identificación del *kernel* más adecuado. Generalmente se utiliza el gaussiano.
- Permite obtener más componentes que el número inicial!

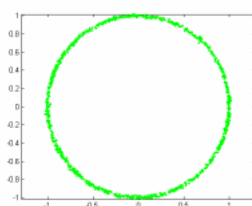
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

57



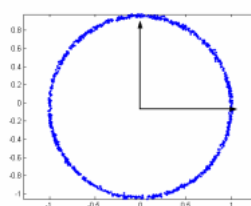
KPCA (2)



Datos originales.

$$K: R^2 \rightarrow R^2$$

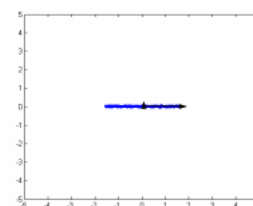
$$\begin{bmatrix} r \\ \theta \end{bmatrix} = K(x, y) = \begin{bmatrix} x^2 + y^2 \\ \tan^{-1}\left(\frac{y}{x}\right) \end{bmatrix}$$



Datos transformados
mediante PCA

$$\lambda_1 = 0.5084$$

$$\lambda_2 = 0.4921$$



Datos transformados
mediante PCA luego
de aplicar el Kernel



$$\lambda_1 = 0.8235$$

$$\lambda_2 = 0.0004$$

Feb-Jun, 2005

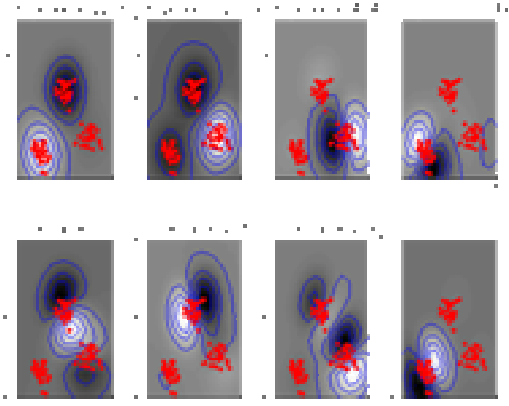
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

58

KPCA: Ejemplo



- **Ejemplo, Código de Schölkopf.**



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

59

ICA



- ICA: *Independent Component Analysis*.
- También busca hallar $Y=PX$ tal que C_Y sea diagonal.
- Forma más formal de reducción de redundancia-
Independencia estadística:

$$P(y_i, y_j) = P(y_i) \cdot P(y_j) \forall i, j \ i \neq j$$
- Independencia estadística implica correlación nula (no al revés).
- Desventajas:
 - No linealidad.
 - Método iterativo de optimización.
 - Varias soluciones posibles.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

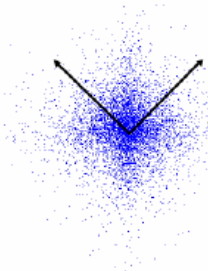
60

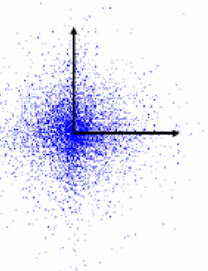
ICA (2)

Ejemplo: distribución exponencial.

PCA





ICA



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

61

ICA (3)

- Surgió para solucionar el “cocktail problem”.
- Relacionado con BSS (*Blind Signal Separation.*)
- Permite obtener señales originales a partir de una mezcla de ellas.
- Se aplica en:
 - Señales de electroencefalograma.
 - Obtención de frecuencia cardíaca de bebés durante embarazo.
 - Extracción de características en procesamiento digital de señales.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

62



ICA (4)

- Método fuertemente basado en análisis estadístico
- No funciona para variables con distribución normal.
- **“no gaussianidad” implica independencia.**
- Kurtosis: medida de la “gaussianidad” de un set de datos.
- Ojo: matlab implementa kurtosis pero no substrahe 3 veces la varianza (3 para normalizadas)
- Existen otras, p.e. *negentropy*.
- Generalmente en el preprocesamiento inicial se aplica PCA con ecualización de varianzas, es decir: $Cov(X) = I$

Feb-Jun, 2005

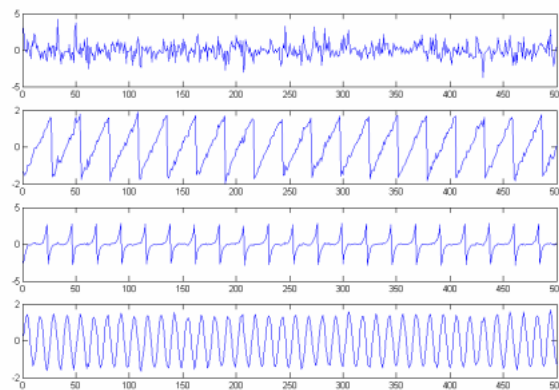
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

63



ICA: Ejemplo BSS



- **Resultados obtenidos con FastICA de Aapo Hyvärinen.**



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

64





Contenido

- Introducción
- Esquema general del proceso de *clustering*.
- Métodos de representación: PCA.
- **Medidas de similitud.**
- Principales técnicas de *clustering*.
- Otras técnicas.

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

65



Medidas de Similitud

- Introducción.
- Espacios métricos.
 - Producto interno.
 - Norma.
 - Distancia.
- Métricas comunes.
- Otras medidas de similitud.

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

66



Introducción

- Depende mucho del tipo de datos que se están trabajando.
- Cuando los datos no son numéricos se deben establecer relaciones heurísticas (similitudes conceptuales).
- Si los datos pueden representarse numéricamente debe revisarse que las magnitudes sean comparables.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

67



Producto Interno

- Sea V un espacio vectorial sobre R^n . Un producto interno en V es una forma bilineal definida por:

$$T: V^2 \rightarrow R$$

$$T(x, y) = \langle x, y \rangle$$

Que cumple las siguientes propiedades:

- Aditividad en la primera componente: $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- Homogeneidad en la primera componente: $\langle \alpha \cdot x, z \rangle = \alpha \cdot \langle x, z \rangle$
- Simetría: $\langle x, z \rangle = \langle z, x \rangle$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

68



Producto Interno (2)

- Si los vectores x y y están expresados en una base determinada de V , se tiene que:

$$\langle x, y \rangle = x_B^T \cdot A \cdot y_B$$

- Donde x_B es el vector expresado en la base B . A es la matriz asociada al producto interno, la cual debe ser simétrica y generalmente es definida no negativa.
- Existen infinitos posibles productos internos en un espacio vectorial.
- La matriz asociada al producto interno convencional en R^n es la matriz identidad $A = I_n$.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

69



Norma

- Sea V un espacio vectorial sobre R^n . Una norma en V es una función definida por:

$$N: V \rightarrow R$$

$$N(x) = \|x\|$$

Que cumple con:

- No negatividad

$$\|x\| \geq 0, \|x\| = 0 \leftrightarrow x = \vec{0}$$

- Desigualdad triangular:

$$\|x + y\| \leq \|x\| + \|y\|$$

- “Homogeneidad”

$$\|\alpha \cdot x\| = |\alpha| \cdot \|x\|$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

70



Norma (2)

- Si en V existe un producto interno, se puede asociar una norma a este producto interno, así:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

$$\|x\| = \sqrt{x^T \cdot A \cdot x}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

71



Métrica (Distancia)

- Sea V un espacio vectorial sobre R^n . Una norma en V es una función definida por:

$$d: V^2 \rightarrow R$$

$$(x, y) \rightarrow d(x, y)$$

Que cumple con:

- No negatividad: $d(x, y) \geq 0, y d(x, y) = 0 \leftrightarrow x = y$
- Simetría: $d(x, y) = d(y, x)$
- Desigualdad triangular: $d(x, z) \leq d(x, y) + d(y, z)$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

72



Métrica (2)

- Si en V existe un producto interno, se puede asociar una norma a este producto interno, y a su vez una métrica a dicha norma, así:

$$d(x, y) = \|x - y\|$$

$$d(x, y) = \sqrt{(x - y)^T \cdot A \cdot (x - y)}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

73



Métricas comunes

- Asociadas con productos internos:
 - Si la matriz asociada es la identidad se llama distancia euclídea:

$$d(x, y) = \sqrt{(x - y)^T \cdot I_n \cdot (x - y)} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Si la matriz asociada es la inversa de la matriz de covarianza se llama distancia de Mahalanobis:

$$d(x, y) = \sqrt{(x - y)^T \cdot \Sigma^{-1} \cdot (x - y)}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

74



Métricas Comunes (2)

- Norma del *Máximo*:

$$d(x, y) = \max_i |x_i - y_i|$$

- Métrica *taxi* (*Manhattan Distance*):

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

- Métrica de Minkowsky:

$$d_p(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$

Observe: d_1 equivale a la distancia de Manhattan, d_2 equivale a la distancia euclídea y d_∞ a la métrica del máximo.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

75



Otras Medidas de Similitud

- Todavía mucho por hacer.
- Algunas no son métricas.
 - Frecuencia de términos en documentos de texto.
 - Medidas de similitud: *cosine similarity*
- Algunos trabajos realizados incluyen términos adicionales en la determinación de la similitud conocido como el “contexto” que toma en consideración los puntos alrededor.

Feb-Jun, 2005

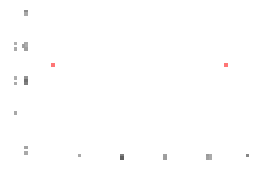
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

76



Análisis Mahalanobis

- Presento un desarrollo desarrollado por mi para mostrar la distancia de Mahanabolis como una rotación variante a la métrica.
- Primero considere los datos:



$$X = \begin{bmatrix} 0.1 & 0.5 \\ 0.9 & 0.5 \\ 0.5 & 0.4 \\ 0.5 & 0.6 \end{bmatrix}^T \Rightarrow \begin{aligned} \text{Cov}(X) &= \begin{bmatrix} 0.1067 & 0 \\ 0 & 0.0067 \end{bmatrix} \\ (\text{Cov}(X))^{-1} &= \begin{bmatrix} 9.375 & 0 \\ 0 & 150 \end{bmatrix} = K \end{aligned}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

77



Análisis Mahalanobis (2)

- Observe la distancia de Mahanabolis al centro de la columna 1 (c_1) y de la columna 3 (c_3) de X :

$$d(c_1, c) = [0.1 - 0.5, 0.5 - 0.5] \cdot \begin{bmatrix} 9.375 & 0 \\ 0 & 150 \end{bmatrix} \cdot \begin{bmatrix} 0.1 - 0.5 \\ 0.5 - 0.5 \end{bmatrix}$$

$$d(c_1, c) = [-0.4, 0] \cdot \begin{bmatrix} -3.75 \\ 0 \end{bmatrix} = 1.5,$$

$$\text{PERO } d(c_3, c) = [0, -0.1] \cdot \begin{bmatrix} 9.375 & 0 \\ 0 & 150 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ -0.1 \end{bmatrix} = 1.5!!$$

➡ **Comparar con la distancia euclidiana**

Ver código Matlab

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

78



Análisis Mahalanobis (3)

- Ahora si ... **La norma de Mahalanobis puede verse como una norma euclidiana en un espacio *diferente*.**

$$\begin{aligned}\|\vec{y}\|_e &= \sqrt{\vec{y}^T \cdot I \cdot \vec{y}} = \sqrt{(M \cdot \vec{x})^T \cdot I \cdot (M \cdot \vec{x})} \\ &= \sqrt{\vec{x}^T \cdot M^T \cdot I \cdot M \cdot \vec{x}} = \sqrt{\vec{x}^T \cdot K \cdot \vec{x}} \\ &= \|\vec{x}\|_{Ma}\end{aligned}$$

- La idea: Dada K será posible hallar la matriz de transformación M ?

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

79



Análisis Mahalanobis (4)

- Problema, factorizar: $K = M^T M$.
- Supuse M triangular superior y se soluciona. Para ello la diagonal de K debe ser positiva y K debe ser simétrica ... no hay problema! Una matriz de covarianza siempre cumple con esto.
- Relacionado con la Expansión KL.

Ver código Matlab

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

80



Contenido

- Introducción
- Esquema general del proceso de *clustering*.
- Métodos de representación: PCA.
- Medidas de similitud.
- Principales técnicas de *clustering*.
- Otras técnicas.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

81



Principales técnicas de *clustering*

- Notación.
- Características deseadas.
- *k-means*.
- *k-means* incremental.
- Vector Quantization.
- SOFM.
- Fuzzy *c-means*.
- Ejemplo con PCA.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

82



Notación

- Patrón (observación) x : simple dato utilizado por el algoritmo de *clustering*.

$$x = [x^1, x^2, \dots, x^m \dots x^M] \quad , \quad x \in R^M$$

- Características (atributos) x^i : son las componentes del patrón.
- Set de patrones X : es la colección de N patrones. Puede verse como una matriz de patrones $N \times M$.

$$X = \begin{pmatrix} [x_1 & x_2 & \dots & x_n & \dots & x_N]^T \end{pmatrix}_{N \times M}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

83



Notación (2)

- Espacio de *medidas*: espacio donde se encuentran los patrones, por simplicidad los reales ($X = R^M$).
Puede ser multidimensional y contener variables aleatorias continuas, discretas o categóricas.
- Clase (*label*) ϕ refleja una posible fuente de patrones que genera una distribución de los mismos en el espacio de medidas.
- Espacio de *significados* Φ (o de características): es un conjunto finito y discreto de *labels* $\phi \in \Phi$.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

84



Notación: Definición

- Definición- *Clustering* (Fraley and Raftery, 1998):

“Con base en unas mediciones de un conjunto de patrones encontrar un método que asigne estos patrones a subclases con algún significado”

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

85



Características Deseadas

- Velocidad y baja complejidad.
- Insensibilidad al ruido.
- Insensibilidad al orden.
- *Clusters* de formas arbitrarias.
- “*Muy no supervisado*”
- Varios tipos de datos: binarios, ordinales, categóricos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

86



k-means (Hartigan HCM, 1975)

- Simple pero todavía muy utilizado.
- Método particional, genera k particiones proveídas por el usuario. $2 \leq k \leq N$
- El i -ésimo *cluster* C_i se describe por el vector $c_i \in R^M$ donde se halla la media de los patrones que pertenecen a él.
- Busca optimizar:

$$J = \sum_{i=1}^k \sum_{x_n \in C_i} \|x_n - c_i\|^2 = \sum_{i=1}^k \sum_{n=1}^N m_{ni} \cdot \|x_n - c_i\|^2$$

$$\text{Matriz de membresía: } m_{ni} = \begin{cases} 1 & \text{si } x_n \in C_i (d_{ni} \leq d_{ki} \forall k \neq i) \\ 0 & \text{de otra forma} \end{cases}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

87



k-means (2)

1. Inicializar los k centros*.
2. Asignar cada uno de los N patrones al *cluster* con menor distancia (bajo alguna métrica).
3. Actualizar los centros.
4. Reasignar patrones a los *clusters*.
5. Recalcular los centros.
6. Revisar criterio de parada sino pasar a 4*.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

88



k-means (3)

- Reasignación de patrones:

- Para cada $x_n \in C_i$ hallar:

$$h = \arg \min_{r, r \neq i} \frac{n_r \cdot \|x_n - c_r\|}{n_r - 1}$$

donde n_r es el número de patrones en el cluster r .

- Pasar el patrón n del cluster i al cluster h si:

$$\frac{n_h \cdot \|x_n - c_h\|}{n_h - 1} < \frac{n_i \cdot \|x_n - c_i\|}{n_i - 1}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

89



k-means: Ventajas y Problemas

- Ventajas:

- Rápido y sencillo.
- Muy utilizado, comprobado y documentado.

- Desventajas:

- Determinación de k .
- El resultado depende de la inicialización de los centros.
- Cluster hipersféricos (y tienden a ser del mismo tamaño).
- Sensibilidad al orden de presentación.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

90



k-means: Mejoras

- Mejorar la inicialización de los centros. Por ejemplo utilizando *clustering* jerárquico.
- División y fusión de *clusters* (cotas de varianza, ISODATA).
- Distancia de Mahanabolis (local) para obtener *clusters* hiperelípticos.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

91



k-means Incremental

- Se inicializan los centros c_i .
- Los centros son actualizados ante un patrón x_n según:

$$c_i = \begin{cases} c_i + \alpha \cdot (x_n - c_i) & \text{si } i = \arg \min_r \|x_n - c_r\| \\ c_i & \text{de otra manera} \end{cases}$$

donde $0 < \alpha < 1$ es la tasa de aprendizaje. Se acostumbra a decrecer de manera exponencial para asegurar convergencia.

- Se revisa algún criterio de parada.
- La ecuación de actualización se parece bastante a la regla delta para **Mencionar esquema neuronal WTA**

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

92



k-means Incremental (2)

GEOMÉTRICAMENTE: **TABLERO**

Forma muy común de actualización ... $\text{taget} - \text{valor actual}$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

93



k-means: Silhoutte-Coefficient

- Kaufman y Rousseeuw, 1990.
- Mide la calidad del *clustering* de *k-means* (y de *k-medoids*).
- Para x_n , perteneciente al cluster de centro c_{j^*} , el coeficiente es:

$$s(x_i) = \frac{d_i^1 - d_i^2}{\max(d_i^1, d_i^2)}, \text{ con } s = \begin{cases} -1 & \text{mala asignación} \\ 0 & \text{indiferente} \\ 1 & \text{buena asignación} \end{cases}$$

$$d_i^1 = \|x_i - c_{j^*}\| \quad \text{y} \quad d_i^2 = \min_{j, j \neq j^*} \|x_i - c_j\|$$

- El coeficiente de Silhoutte se define como:

$$s_c = \frac{\sum_{n=1}^N s(x_n)}{N}$$

$$s_c \geq 0.7 \quad \text{buen clustering}$$

$$s_c \leq 0.3 \quad \text{regular}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

94



Vector Quantization

- Una de las principales aplicaciones de aprendizaje competitivo y de algunas técnicas de *clustering*.
- Utilizado para compresión de datos (e.g. voz e imágenes). Útil tanto en almacenamiento como en transmisión.
- Un *set* X de patrones se representa por l *templates*.
- El espacio de entrada queda dividido (*partido*) en regiones.
- Ante un nuevo vector x el cuantizador ubica la región y devuelve el *template* de dicha región.
- Los l vectores se llaman *codebook*.
- Si se la métrica es la euclidiana se llama **Cuantizador de Voronoi** y el espacio queda dividido en celdas de Voronoi.

Feb-Jun, 2005

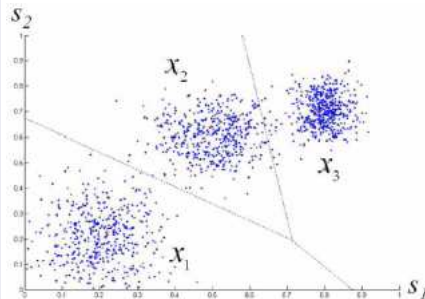
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

95

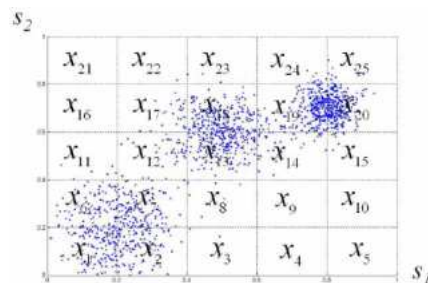


Cuantizador de Voronoi

Particiones del espacio de entrada realizadas por un cuantizador de Voronoi con 3 vectores de reconstrucción



Particiones del espacio de entrada realizadas a mano por un cuantizador 5-ario.



Feb-Jun, 2005

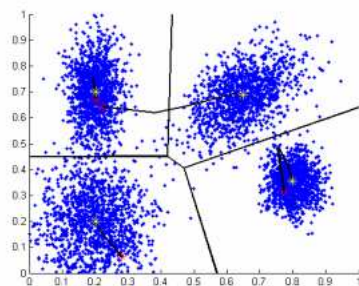
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

96



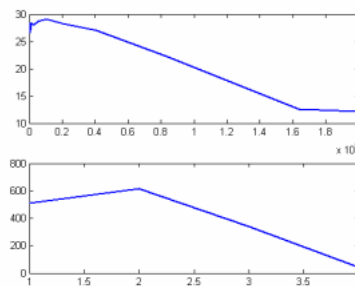
k-means: Ejemplo 1

Partición alcanzada por *k*-means tradicional.



*: punto inicial
.: punto final

Comportamiento del objetivo J y del número de patrones movidos.



Feb-Jun, 2005

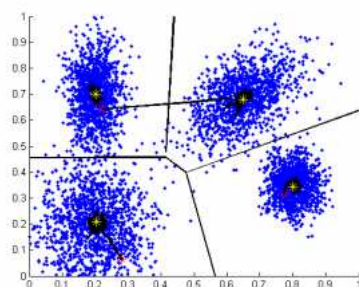
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

97



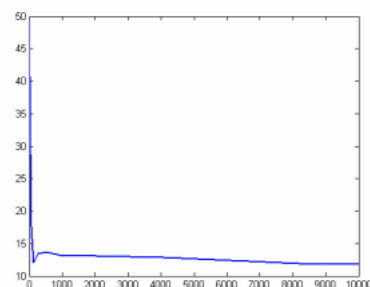
k-means: Ejemplo 2

Partición alcanzada por *k*-means incremental*.



*: punto inicial
.: punto final

Comportamiento de la función objetivo J



*Tasa de aprendizaje con variación exponencial

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

98



Self-Organizing Feature Maps, SOFM

(Kohonen, 1982)



- *Self-organizing*: adaptación no supervisada que genera un mapa de características de los datos a partir de una red donde las unidades (neuronas) interactúan localmente en respuesta a los estímulos recibidos.
- *Topology-preserving*: habilidad de los mapas de características de conservar relaciones topológicas existentes entre los datos en el espacio de entrada en un arreglo de unidades en otro espacio.
- Datos cercanos en R^n activan neuronas cercanas en R^2 y la distancia entre las neuronas tiende a cero a medida que la distancia entre los datos también lo hace.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

99



SOFM (2)



- Los SOFM capturan la distribución de probabilidad de los patrones en el espacio de entrada.
- Generalmente es una red neuronal en dos dimensiones de unidades.
- Cada unidad esta descrita por su centro $c_i \in R^n$.
- Con cada dato el centro de la neurona se mueve en el espacio de entrada, pero permanece fijo en el arreglo 2-dimensional.
- Algunos le llaman *k-means* neuronal.
- Para simular la respuesta local incluye un término de “vecindad”.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

100



SOFM: Algoritmo

- Se inicializan los centros c_i .
- Los centros son actualizados ante un patrón x_n según:

$$c_i = c_i + \alpha_k \cdot \Lambda(r_i, r_{i^*}) \cdot (x_n - c_i)$$

donde:

- r_i es el vector radar a la i -ésima unidad en el arreglo 2-dimensional.
- r_{i^*} es el vector radar a la unidad ganadora en el arreglo 2-dimensional.
- Λ es un *kernel* de vecindad.
- $0 < \alpha_k < 1$ es la tasa de aprendizaje.
- Se revisa algún criterio de parada.

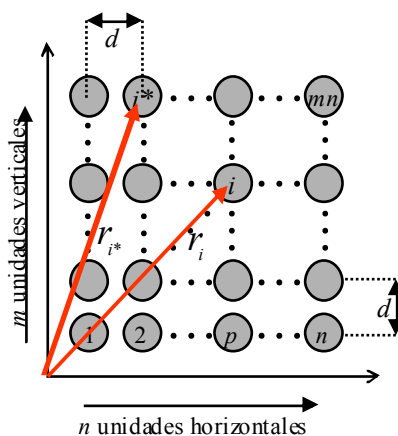
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

101



SOFM: Arreglo 2-Dimensional



- Arreglo $n \times m$ de neuronas.
- Las posiciones en el arreglo dos dimensional son fijas.
- Generalmente d es 1.
- La unidad ganadora se determina al igual que en *k-means* así:

$$i^* = \arg \min_j \|c_j - x_n\|$$

- Las vecinas inmediatas de i son: $i-1, i+1, i+n, i-n$, siempre que i esté en el interior.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

102



SOFM: *Kernel*

- Función estrictamente positiva con simetría radial parametrizada por un “centro” μ y un “ancho” σ
- Valor máximo en el centro y decae, de manera controlada por σ , a medida que la distancia al centro aumenta.
- Formalmente: $K : R^N \rightarrow U \subset R^+$ generalmente: $U = (0,1]$
 - Positiva: $K(x, \mu, \sigma) > 0 \quad \forall x \in R^N$
 - Simetría radial: $K(x, \mu, \sigma) = K(x', \mu, \sigma) \quad \forall x' / \|x' - \mu\| = \|x - \mu\|$
 - Valor máximo: $K(x, \mu, \sigma) < K(\mu, \mu, \sigma) \quad \forall x \neq \mu$
 - Decreciente: $K(x, \mu, \sigma) \geq K(x', \mu, \sigma) \quad si \quad \|x - \mu\| \leq \|x' - \mu\|$
 - Asintótica a cero: $K(x, \mu, \sigma) \rightarrow 0 \quad si \quad \|x - \mu\| \rightarrow \infty$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

103



SOFM: *Kernel (2)*

- Principales *kernels*:
 - Gaussiano: $K(x, \mu, \sigma) = K\left(\frac{\|x - \mu\|}{\sigma}\right) = \frac{1}{(2\pi \cdot \sigma^2)} \exp\left(-\frac{\|x - \mu\|^2}{2 \cdot \sigma^2}\right)$
 - Sigmoidal: $K\left(\frac{\|x - \mu\|}{\sigma}\right) = \left[1 + \exp\left(\frac{\|x - \mu\|^2}{2 \cdot \sigma^2} - \theta\right)\right]^{-1}, \theta: bias$
 - Polinomiales: $K(x, \mu, d) = (\langle x, \mu \rangle + c)^d, c: bias$
 - De Laplace: $K\left(\frac{\|x - \mu\|}{\sigma}\right) = \frac{1}{2 \cdot \sigma} \cdot \exp\left(-\frac{\|x - \mu\|}{\sigma}\right)$
 - Indicator kernel: $K(x, \mu, \sigma) = \frac{1}{\sigma} \left(1 - \frac{1}{\sigma} \|x - \mu\|_1\right) \cdot \Pi\left(\frac{\|x - \mu\|}{2 \cdot \sigma}\right)$

Feb-Jun, 2005

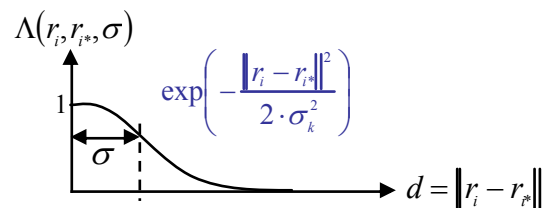
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

104



SOFM: *Kernel* (3)

- Generalmente se emplea el *kernel* gaussiano.
- El centro del *kernel* es el centro de la neurona ganadora: r_{i^*} .
- El dominio del *kernel* son los radios a todas las neuronas: r_i .
- El ancho puede ser constante y determinada por el usuario. Este regula la “localidad” de la perturbación. Generalmente decrece con las iteraciones.



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

105



SOFM: ρ y σ

- Tanto ρ como σ deben decrementarse. ρ para lograr la convergencia y σ para decrecer el efecto de neuronas vecinas a medida que la red se estabiliza.
- Ritter y Shulten en 1988 propusieron:

$$\alpha_k = \alpha_0 \cdot \left(\frac{\alpha_f}{\alpha_0} \right)^{1/k_{\max}} \quad \sigma_k = \sigma_0 \cdot \left(\frac{\sigma_f}{\sigma_0} \right)^{1/k_{\max}}$$

donde ρ_0 , σ_0 , ρ_f y σ_f son valores iniciales y finales de estos valores seleccionados por el usuario y k_{\max} es el número de iteraciones máximas.

- Se recomienda:

$$0 < \alpha_0 \leq 1, \text{ típico } 0.8 \quad \sigma_0 \approx d \cdot m_d / 2$$

$$0 < \alpha_f < 1 \quad \sigma_f = d / 2$$

Con d la distancia entre neuronas vecinas y m_d el número de neuronas en la diagonal de mayor longitud del arreglo

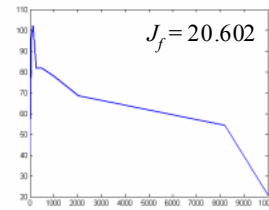
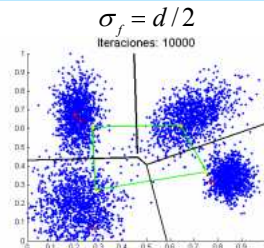
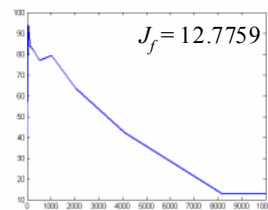
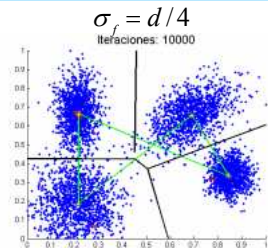
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

106



SOFM: Sobre el Ejemplo Anterior



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

107



SOFM: Sobre el Ejemplo Anterior (2)

- ¿Que pasó?, ¿Son peores los SOFM?
- La característica del SOFM de capturar la **densidad de probabilidad** hace que sea más lento y que el aprendizaje de una unidad influya en las otras.
- Si el ancho del *kernel* se disminuye el máximo el algoritmo funciona igual al de *k-means* incremental.
- Sin embargo en la mayoría de los problemas los SOFM se desempeñan mejor por que utilizan todos los *clusters* que le provea el usuario para cubrir de la mejor manera el espacio de entrada dando mas resolución donde se presentan más casos.

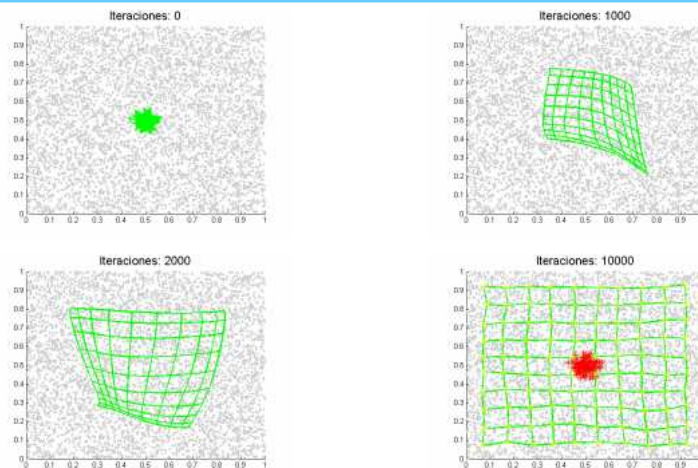
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

108



SOFM: Ejemplo Distribución Uniforme



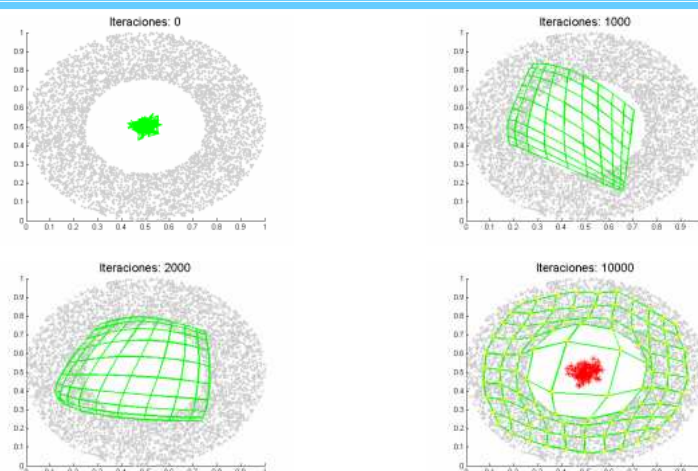
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

109



SOFM: Ejemplo otras Distribuciones



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

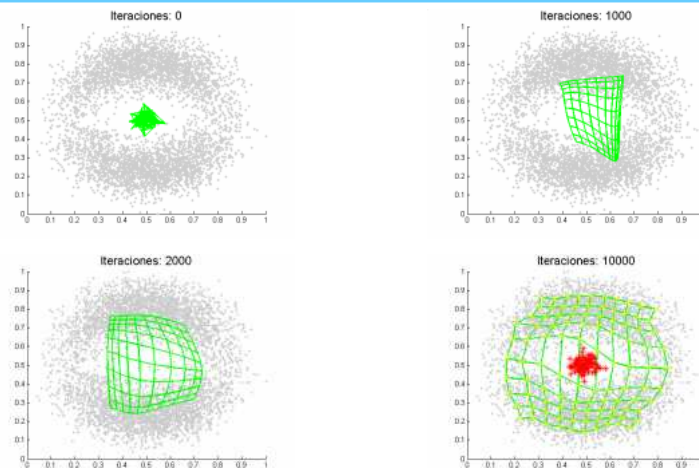
110



SOFM: Ejemplo otras Distribuciones



(2)



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

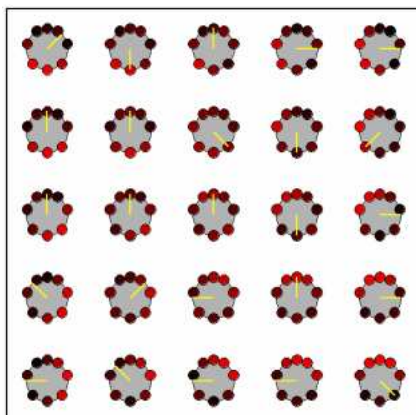
111



SOFM: Ejemplo otras Distribuciones



(3)



- Ejemplo tomado de la tesis: "Aprendizaje por Refuerzo en Espacios Continuos para la Evasión de Obstáculos en un Robot Móvil".
- Mapa de Kohonen aprendido sobre un espacio 8-dimensional de entrada generado por 8 sensores infrarojos.
- En la abstracción entre más encendido está el rojo más cercano se encuentra un obstáculo.
- Se puede observar como se conserva la topología.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

112



SOFM: Características Adicionales

- En la convergencia, la distribución de frecuencia de activación a lo largo de las diferentes unidades debe ser uniforme.
- En la convergencia, si un punto se desplaza en el espacio de entrada a lo largo de una trayectoria suave, la distribución de las transiciones entre neuronas debe ser “normal” en términos de la distancia en el arreglo 2-dimensional.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

113



Fuzzy c-means (FCM)

- Método de partición difusa (suave), genera k particiones proveídas por el usuario, $2 \leq k \leq N$
- Un dato x_n pertenece a cada *cluster* c_i con un determinado grado de certeza m_{ni} .
- Busca optimizar:

$$J = \sum_{i=1}^k \sum_{n=1}^N m_{ni}^r \cdot \|x_n - c_i\|^2; r > 1$$

Sujeto a restricciones. Donde r es el grado de fusificación.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

114



FCM (2)

- Optimización realizada por Bezdek (1981) utilizando *multiplicadores de Lagrange*.

- Restricciones:

$$1) m_{ni} \in [0,1] \forall n, i \quad 2) 0 < \sum_{n=1}^N m_{ni} < N \quad 3) \sum_{i=1}^k m_{ni} = 1$$

- Valor óptimo:

$$m_{ni}^* = \frac{1}{\sum_{h=1}^k \left(\frac{\|x_n - c_i\|}{\|x_n - c_h\|} \right)^{\frac{2}{r-1}}} = \frac{\|x_n - c_i\|^{-\frac{2}{r-1}}}{\sum_{h=1}^k \|x_n - c_h\|^{-\frac{2}{r-1}}}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

115



FCM (3)

1. Inicializar la matriz de membresía m_{ni} .
2. Actualizar los centros de los *clusters* como una suma ponderada.
3. Hallar nuevamente los m_{ni} óptimos.
4. Revisar criterio de parada sino pasar a 2.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

116



FCM: Ventajas y Problemas

- Ventajas:
 - Rápido y sencillo.
 - Grado de certeza de pertenencia
- Desventajas: las mismas de HCM
 - Determinación de k .
 - El resultado depende de la inicialización de los centros.
 - Cluster hipersféricos (y tienden a ser del mismo tamaño).
 - Sensibilidad al orden de presentación.

Feb-Jun, 2005

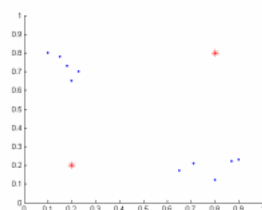
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

117



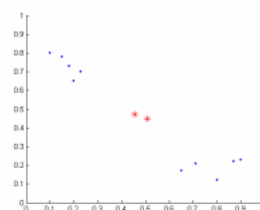
FCM: Ejemplo Sencillo

Iteración: 1



0.5698	0.4302
0.6538	0.3462
0.5551	0.4449
0.5717	0.4283
0.5805	0.4195
0.5579	0.4421
0.6734	0.3266
0.5779	0.4221
0.4055	0.5945
0.4317	0.5683

Iteración: 2

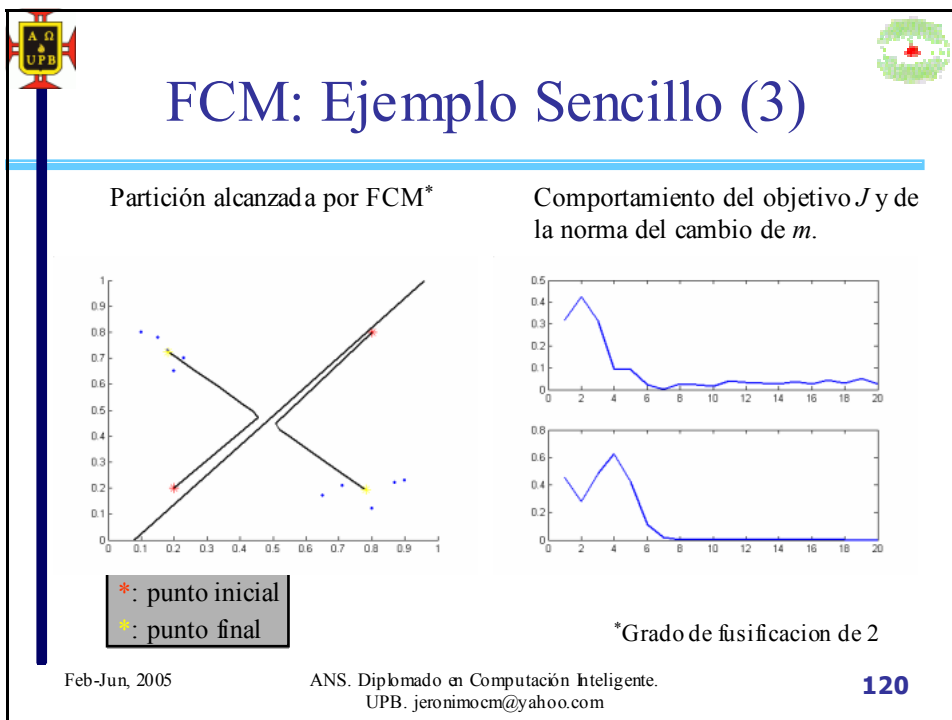
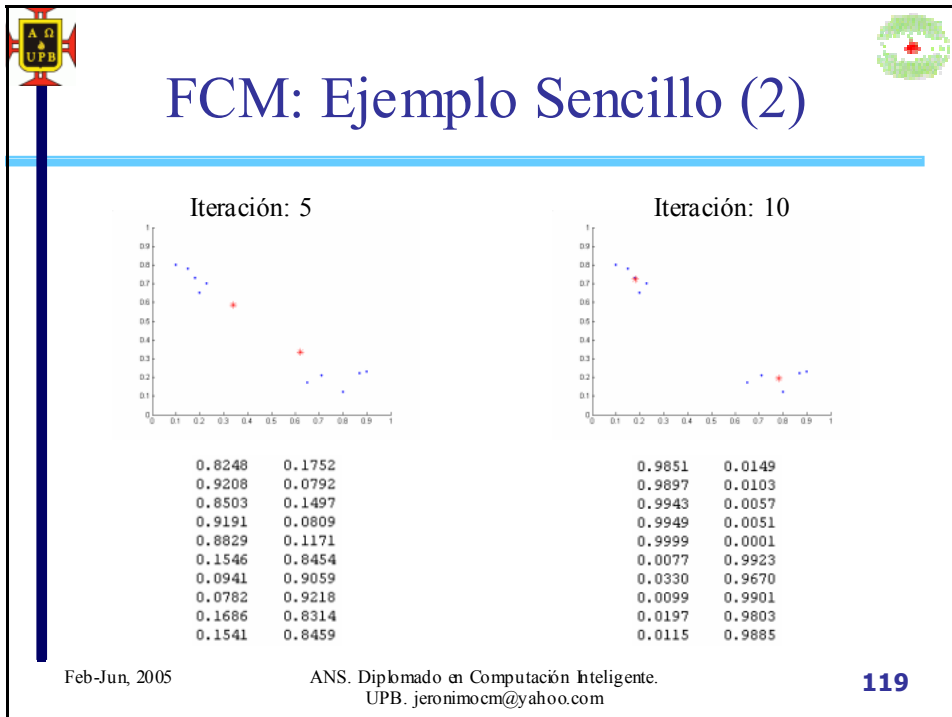


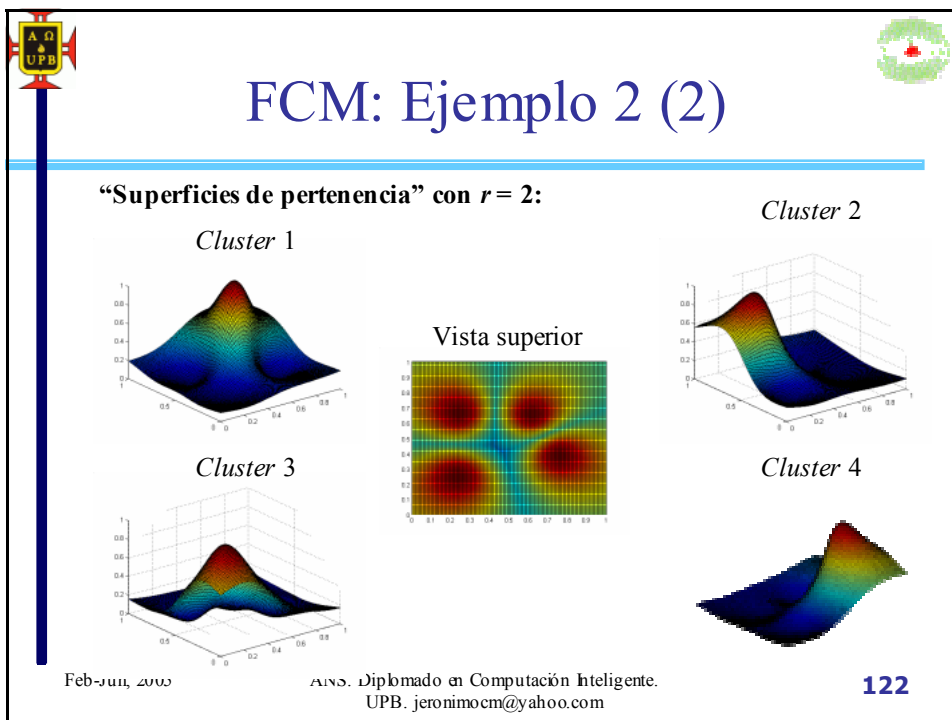
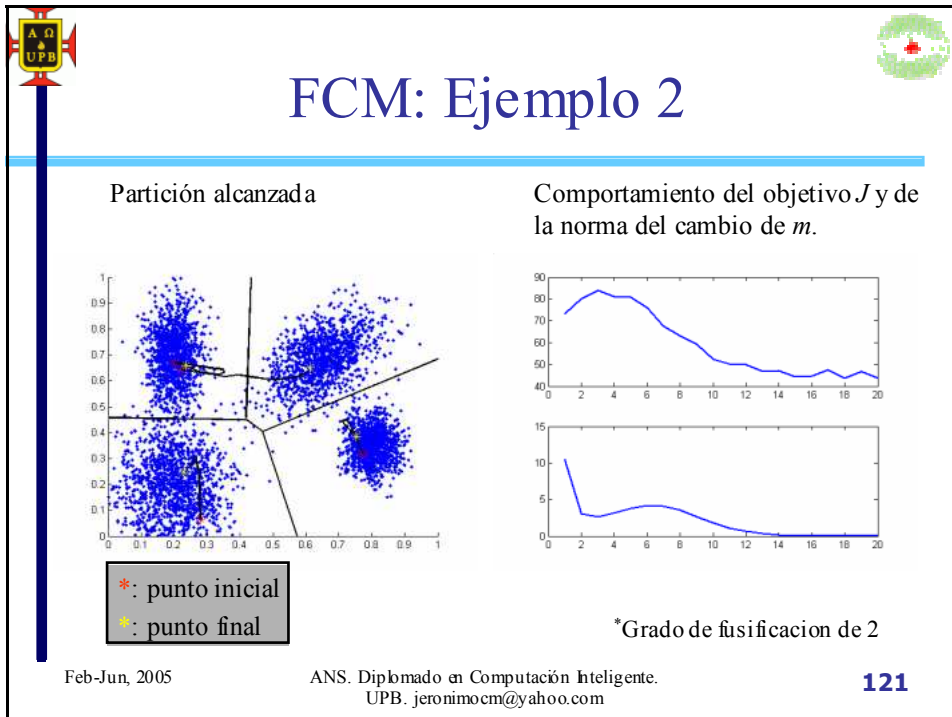
0.5534	0.4466
0.5826	0.4174
0.5584	0.4416
0.5772	0.4228
0.5672	0.4328
0.4433	0.5567
0.4309	0.5691
0.4226	0.5774
0.4401	0.5599
0.4377	0.5623

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

118

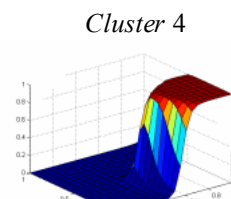
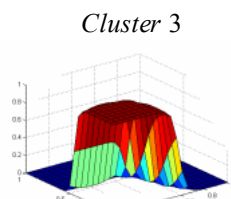
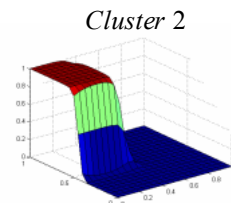
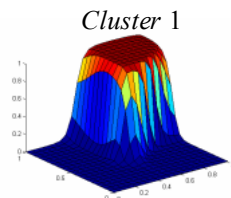






FCM: Ejemplo 2 (3)

“Superficies de pertenencia” con $r = 1.2$:



Feb-Jun, 2005

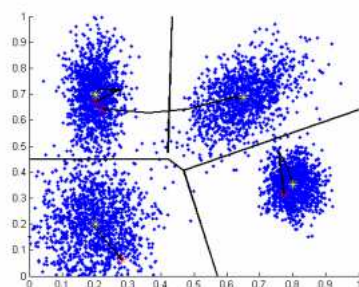
ANS. Diplomado en Computación Inteligente.
UPB. jeronomcm@yahoo.com

123



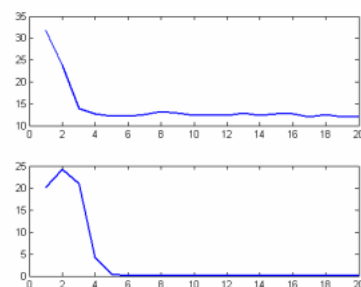
FCM: Ejemplo 2 (4)

Partición alcanzada por FCM*



*: punto inicial
.: punto final

Comportamiento del objetivo J y de la norma del cambio de m .



*Grado de fusificación de 1.2

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronomcm@yahoo.com

124



Ejemplo con PCA

- Base de datos X en 4 dimensiones con 5000 datos:

0.15817	0.5793	0.45856	0.11184	0.22809	0.2952	0.35726	0.16126
0.14093	0.084092	0.18554	0.099651	0.77469	0.2876	0.92155	0.54775
0.79458	0.30714	0.93537	0.56186	0.78969	0.6993	1.155	0.55835
0.77783	0.70398	1.1392	0.55001	0.19903	0.75773	0.57146	0.14074
0.20939	0.64026	0.52382	0.14806	0.15988	0.16252	0.23967	0.11301
0.22727	0.25662	0.35968	0.16071	0.86691	0.25824	0.99845	0.613
0.75936	0.23964	0.87275	0.53695	0.62319	0.76148	1.0547	0.44067
0.50517	0.56421	0.75328	0.35721	0.21925	0.65559	0.56929	0.15501
0.22703	0.60025	0.56645	0.16053	0.28089	0.22265	0.38808	0.19862
0.048693	0.45644	0.30257	0.034431	0.90446	0.35207	1.0791	0.63955
0.76247	0.24916	0.89301	0.53915	0.73765	0.74321	1.1516	0.52155
0.66277	0.63205	0.96228	0.46865	0.16549	0.5763	0.48242	0.11702
0.15951	0.69356	0.53679	0.11279	0.22957	0.0099237	0.23912	0.16231
0.2708	0.12177	0.32275	0.19148	0.76496	0.30899	0.92682	0.54091

Feb-Jun, 2005

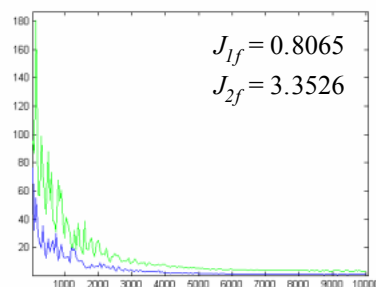
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

125



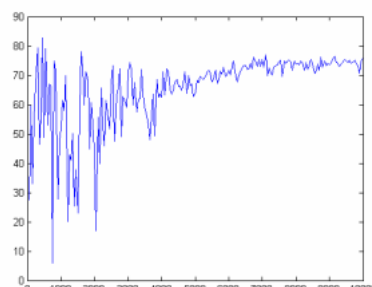
Ejemplo con PCA (2)

- Desempeño de un SOFM en el *clustering* de dicha base de datos:



Comportamiento de J para:

- 2 componentes
- 4 componentes





Diferencia relativa porcentual:

$$r = \frac{J_2 - J_1}{J_2} \times 100\%$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com


126



Contenido

- Introducción
- Esquema general del proceso de *clustering*.
- Métodos de representación: PCA.
- Medidas de similitud.
- Principales técnicas de *clustering*.
- Otras técnicas.

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com **127**



Otras Técnicas

- *Adaptive Resonance Theory* (ART)
- *Fuzzy ART*.
- Basada en Modelo: *Expectation Maximization* (EM).
- Basada en Densidad: *Density Boundary Scan* (DBSCAN).

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com **128**



ART1 (Carpenter y Grossberg, 1987)

- Principios de *adaptive resonance theory*: Gail Carpenter, 1976.
- Biológicamente motivada.
- ART: familia de diferentes arquitecturas neuronales
 - ART1: patrones binarios, Carpenter y Grossberg 1987.
 - ART2: patrones reales (entre cero y uno), Carpenter y Grossberg 1987
 - ART2-A: ART2 mejorado. Dos o tres órdenes de magnitud más rápido, Carpenter *et al* 1991. También ART 2A-C y ART 2A-E.
 - Fuzzy ART. Carpenter, Grossberg y Rosen, 1991.
 - ARTMAP: *real time learning ART model*. Implementa aprendizaje supervisado, Carpenter *et al* 1992.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

129



ART1 (2)

- Resuelve el dilema estabilidad-plasticidad.
- Modela memoria corto y largo plazo.
- Visión biológica compleja, descrito a través de ecuaciones diferenciales.
- Visión algorítmica simple. ART1 modelo simplificado.
- Se emplea para visión artificial.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

130



ART1: Algoritmo

- Estructura *winner take all*, WTA.
- Entradas: patrones $x_n \in \{0,1\}^M$
- Ubicación dinámica de nuevas unidades.
- Cada unidad caracterizada por su centro: $c_j \in \{0,1\}^M$
- Ante una entrada x_n , la j -ésima unidad produce:

$$y_j = \frac{c_j^T \cdot x_n}{\|c_j\|^2}$$

- La red WTA decide la ganadora: $j^* = \arg \max_j \left(\frac{c_j^T \cdot x_n}{\|c_j\|^2} \right)$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

131



ART1: Algoritmo (2)

- Para que $x_n \in C_{j^*}$ se debe satisfacer:

$$i) \frac{c_j^T \cdot x_n}{\|c_j\|^2} > \frac{\|x_n\|^2}{M} \quad \text{Test de cercanía.}$$

$$ii) \frac{c_j^T \cdot x_n}{\|x_n\|^2} \geq \rho \quad \text{Test de similaridad.}$$

donde $0 < \rho < 1$ se llama: parámetro de vigilancia.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

132



ART1: Algoritmo (3)

- **Cercanía:** una fracción suficiente de los bits de x_n concuerden con los de c_j .
- **Similitud:** una significativa fracción de los unos en x_n aparezcan en c_j . La fracción queda determinada por el parámetro de vigilancia.
- Cercanía – Normalización: Evita la selección de *supersets*.
- La ecuación de similitud crea mayor diferenciación en vectores de menor magnitud. A esto se le llama: *automatic scaling, self scaling o noise insensitivity*.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

133



ART1: Algoritmo (4)

- Si se pasan los test, se dice que la red *resuena*. x_n se une al *cluster* j^* y su centro se actualiza según:

$$c_{j^*} = c_{j^*} \wedge x_n$$

- Si j^* no pasa el segundo test, se intenta con la segunda mayor y_j .
- El proceso se repite. Si nunca se llega a pasar el segundo test, se crea una nueva unidad con centro en:

$$c_i = x_n$$

- Si j^* no pasa el primer test, x_n esta muy lejos de cualquiera y también se crea una nueva unidad.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

134



ART1: Algoritmo (5)

Observe, según la ecuación de actualización:

- Los nuevos centros sólo pueden tener menos unos.
- Después de varias actualizaciones, elementos de un cluster pueden dejar de pertenecer a él.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

135





ART1: Comentarios

- ρ controla la granularidad. Valor pequeño crea *clusters* grandes (*set* pequeño de *clusters*).
- Ante un *training set* finito, ART1 **siempre** llega a una configuración **estable***.
- **Plasticidad**, permite aprendizaje continuo. Creación de nuevas unidades.
- La mezcla de estabilidad y plasticidad permite a ART1 seguir distribuciones no estacionarias de entradas.
- ART1 es sensible al orden.

Feb-Jun, 2005

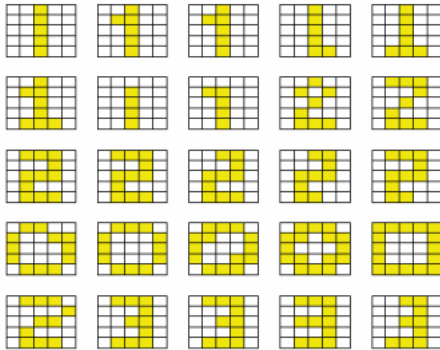
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

136



ART1: Ejemplo

Parte de la base de datos

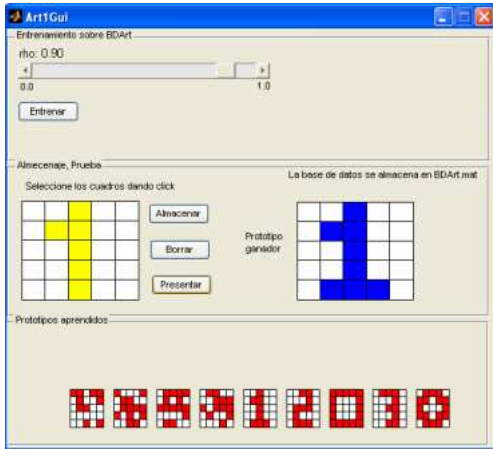


Patrones binarios representados en forma gráfica para mejor visualización

Feb-Jun, 2005
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com
137

ART1: Ejemplo (2)



Feb-Jun, 2005
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com
138



Fuzzy ART. (Carpenter, Grossberg y Rosen, 1991)



- Similar a ART1 pero ampliada para valores continuos.
- Se utilizan operadores difusos (más de esto en lógica difusa).
- Entrada: vectores análogos:

$$\vec{x}' \in U \subset R^{M/2}, \text{ normalmente } U = [0,1]^{M/2}$$

Importante: los datos de entrada son completados con su complemento. Si el mayor valor es x_{\max} (normalmente 1) entonces un vector de entrada x'_n se completa:

$$\vec{x}' = [x'_1, x'_2, \dots, x'_{M/2}] \Rightarrow \vec{x} = [x'_1, \dots, x'_{M/2}, x_{\max} - x'_1, \dots, x_{\max} - x'_{M/2}]$$

$$\text{así: } \sum_{i=1}^M x_i = M \cdot x_{\max}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

139



Fuzzy ART (2)



- Los tests se redefinen de la siguiente manera (para $x_{\max} = 1$):

$$i) \frac{\|c_j \wedge x_n\|_1}{\alpha + \|c_j\|_1} > \frac{\|x_n\|_1}{M} \quad \text{Test de cercanía.}$$

$$ii) \frac{\|c_j \wedge x_n\|_1}{\|x_n\|_1} \geq \rho \quad \text{Test de similitud.$$

Donde x_n es el vector de entrada completado, M es el número de componentes del mismo, ρ es el parámetro de vigilancia y α es un parámetro adicional que favorece unidades con menor área de influencia (**más sobre esto adelante**).

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

140



Fuzzy ART (3)

- En las ecuaciones anteriores: $\|\cdot\|_1$ representa la norma 1 de Minkowsky, y el operador difuso \wedge se define como el mínimo componente a componente, por lo que:

$$\|c_j \wedge x_n\|_1 = \sum_{i=1}^M \min(c_{ji}, x_{ni})$$

- La salida de la unidad j es:

$$y_j = \frac{\|c_j \wedge x_n\|_1}{\alpha + \|c_j\|_1}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

141



Fuzzy ART (4)

- De nuevo... la ganadora es aquella que satisfaga ambos *tests* y con el mayor valor de salida:

$$j^* = \arg \max_j (y_j)$$

- La ganadora se actualiza, según:

$$c_{j^*} = c_{j^*} + \eta \cdot (c_{j^*} \wedge x_n - c_{j^*})$$

Donde η es la tasa de aprendizaje, generalmente $\eta = 1$.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

142



Fuzzy ART: Explicación

- Interpretación para datos originales en 2 dimensiones: los centros de los *clusters* definen “cajas” en el espacio. En 2D un centro $[a,b,c,d]$ define un rectángulo cuyos vértices opuestos están en: (a,b) y $(1-c,1-d)$

Un primer ejemplo:

- Sea la entrada:
 $x'_n = [0.4, 0.4]$, por lo que: $x_n = [0.4, 0.4, 0.6, 0.6]$
- Dos centros:
 $c_1 = [0.1, 0.2, 0.3, 0.4]$ y $c_2 = [0.25, 0.25, 0.5, 0.5]$
- $\alpha = 0.2$.

Feb-Jun, 2005

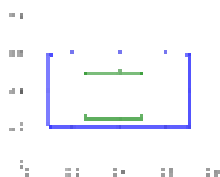
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

143



Fuzzy ART: Explicación (2)

- Graficamente:



$$y_1 = \frac{\|c_1 \wedge x_n\|_1}{0.2 + \|c_1\|_1} = \frac{\|c_1\|_1}{0.2 + \|c_1\|_1} = \frac{1}{0.2 + 1} = 0.833$$

$$y_2 = \frac{\|c_2 \wedge x_n\|_1}{0.2 + \|c_2\|_1} = \frac{\|c_2\|_1}{0.2 + \|c_2\|_1} = \frac{1.5}{0.2 + 1.5} = 0.88$$

Observe:

- α hace que la caja más pequeña gane!
- Si el punto esta dentro de la caja: $c_j \wedge x_n = c_j$
En la ecuación de actualización el centro no cambia!!

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

144



Fuzzy ART: Explicación (3)

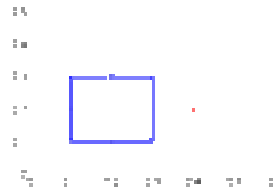
Un segundo ejemplo, entrada por fuera de la caja:

$x'_n = [0.4, 0.2]$, por lo que:

$x_n = [0.4, 0.2, 0.6, 0.8]$

$c_1 = [0.1, 0.1, 0.7, 0.7]$,

$x_n \wedge c_1 = [0.1, 0.1, 0.6, 0.7]$



Observe, el punto esta fuera de la caja y:

$$\|x_n \wedge c_1\|_1 = 1.5 < \|c_1\|_1 = 1.6$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

145



Fuzzy ART: Explicación (4)

- Con $\alpha = 0.2$ y $\rho = 0.75$ y $\eta = 1$:

Test de Cercanía: $y_1 = \frac{1.5}{1.8} = 0.833 \geq \frac{\|x_n\|_1}{M} = \frac{2}{4} = 0.5$

Test de Similaridad: $\frac{\|c_j \wedge x_n\|_1}{\|x_n\|_1} = \frac{1.5}{2} \geq \rho = 0.75$

- Actualización:

$$c_1 = c_1 + \eta \cdot (c_j \wedge x_n - c_j) = [0.1, 0.1, 0.6, 0.7]$$

La caja se agranda para contener el nuevo punto!

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

146



Fuzzy ART: Explicación (5)

- Si ningún centro cumple los *tests* entonces se crea una nueva unidad con centro en x_n .
- Según el análisis el nuevo centro representa una caja de área nula.
- Esta caja irá creciendo con el entrenamiento.
- Fuzzy ART genera una división del espacio en clusters no disjuntos. A estas particiones se les conoce también como *tiles* (similares a las generadas por CMAC).

Feb-Jun, 2005

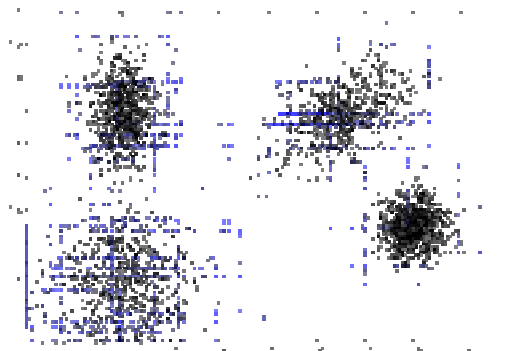
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

147



Fuzzy ART: Ejemplo

- **Ver Código Matlab**



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

148



Fuzzy ART - Tarea

- Consultar implementaciones en Hardware, VLSI.
- Consultar FAST, de Andrés Pérez Uribe, EPFL.
- Consultar CMAC (Albus).

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

149



Expectation Maximization, EM

- Idea original: Hartley 1958, Dempster *et al* 1977. Creo comprensión y aplicaciones ~1995.
- Método basado en modelo.
- Altas dosis de estadística.
- Relacionado con GP (*Gaussian Processes*) y GMM (*Gaussian Mixture Model*).
- Un GMM es una distribución de probabilidad (compuesta por una mezcla). EM es un método para aprenderlos!
- **Comentario acerca de lo que se presentará**

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

150



EM (2)

- Mientras que las ANNs son aproximadores universales de funciones, los GMMs son aproximadores universales de densidades.
- EM es un método general, no necesariamente para modelos basados en funciones gaussianas.
- **Cluster: Es una función de densidad de probabilidad (*probability density function*, pdf), por lo tanto está descrito por los parámetros de la pdf, p.e. para una gaussiana: media y matriz de covarianzas.**
- Se supone que los datos fueron extraídos de una mezcla de funciones de densidad de probabilidad (p.e. un GMM).

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

151



EM (3)

- EM busca aproximar esta mezcla a partir de las observaciones.
- Se considera que cada dato es generado por una y sólo una distribución de la mezcla.
- Generación: para obtener un punto se selecciona al azar una distribución y luego de ella se obtiene un punto.
- EM: *missing data problem*. En este caso, los *labels*!

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

152



EM: Estadística

- Probabilidad del condicional:

$$P(A, B) = P(A \cap B) = P(A/B) \cdot P(B)$$

$P(A, B)$: *joint probability*

$P(A/B)$: probabilidad de que se de A dado B

Si los eventos son independientes entonces la probabilidad de que se de A no depende de B y: $P(A, B) = P(A) \cdot P(B)$

Ejemplitos:

- Probabilidad de que al lanzar dos monedas caiga cara.

Por ser independientes:

$$P(C, C) = P(C) \cdot P(C) = 0.25$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

153



EM: Estadística (2)

- La probabilidad de ganar el diplomado (G) si estudio es de 80%, pero la probabilidad de estudiar en semana santa (E) es de 30%. La probabilidad de ganar y estudiar es:

$$P(G \cap E) = P(G/E) \cdot P(E) = 0.8 \cdot 0.3 = 24\%$$

- **Observe:**

$$A \cap B \equiv B \cap A \Rightarrow P(A, B) = P(B, A)$$

$$\Rightarrow P(A/B) \cdot P(B) = P(A) \cdot P(A)$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

154



EM: Estadística (3)

- *Bayes Rule:*

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Donde:

$$P(B) = \sum_i P(B, A_i) = \sum_i P(B/A_i) \cdot P(A_i)$$

Donde A_i son todos los posibles eventos de los que depende B .

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

155



EM: Estadística (4)

Para el ejemplo,

- Ganar depende de si estudia y si tiene suerte (S).
- Si tiene suerte la probabilidad de ganar es de 50%.
- La probabilidad de tener suerte es del 20%.
- La probabilidad de ganar es:
$$P(G) = 0.8 \cdot 0.3 + 0.5 \cdot 0.2 = 34\%$$
- La probabilidad de que haya estudiado dado que gano será entonces:

$$P(E/G) = \frac{P(G/E) \cdot P(E)}{P(G)} = 70.6\%, \text{ y } P(S/G) = 29.4\%$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

156



EM: Presentación

- Base de datos: N patrones $x \in R^M$
- k clusters.
- Para GMMs, la probabilidad de que se presente el dato x dado que estamos en el cluster (pdf) C_j es:
 - $P_{\mu_j, K_j}(x/C_j) = \frac{1}{\sqrt{(2\pi)^M |\det(K_j)|}} \exp\left(-\frac{1}{2}(x - \mu_j) \cdot K_j^{-1} \cdot (x - \mu_j)^T\right)$
- pdf del clustering: probabilidad de que se presente el patrón x es la suma de las probabilidades de que haya sido producido por cada uno de los clusters:

$$P(x) = \sum_{j=1}^k P_{\mu_j, K_j}(x/C_j) \cdot P(C_j)$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

157



EM: Presentación (2)

- El término $P(C_j)$ representa la probabilidad de que sea el cluster C_j . Esta probabilidad depende del número de elementos que este cluster posea y lo denotaremos W_j .

$$P(C_j) = W_j = \frac{|C_j|}{N}, \text{ con } |\cdot| \text{ la cardinalidad del set } C_j$$

- Note que:

$$\sum_{j=1}^k W_j = 1 \text{ y } W_j \geq 0 \forall j$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

158



EM: Presentación (3)

- Asignación de patrones: probabilidad de que, dado que tengo el patrón x , este haya sido obtenido de la distribución j .

BAYES:

$$P(C_j/x) = \frac{P_{\mu_j, K_j}(x/C_j) \cdot P(C_j)}{P(x)} = \frac{P_{\mu_j, K_j}(x/C_j) \cdot W_j}{\sum_{j=1}^k P_{\mu_j, K_j}(x/C_j) \cdot P(C_j)}$$

- Likelihood:** bajo un modelo dado, es la probabilidad del *training set* (*joint probability* sobre todos los datos) dados los parámetros de dicho modelo. Si los datos se considerarán independientes:

$$P(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N) = P(\vec{x}_1) \cdot P(\vec{x}_2) \dots P(\vec{x}_N)$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

159



EM: Presentación (4)

- Likelihood (Cont.):** si la pdf depende de unos parámetros θ , p.e. k vectores de medias y k matrices de covarianzas, se escribe:

$$L(\theta) = \prod_{i=1}^N P_{\theta}(x)$$



En muchas ocasiones se prefiere trabajar con la función de *log-likelihood*:

$$L(\theta) = \sum_{i=1}^N \log(P_{\theta}(x))$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

160





EM: Presentación (4)

Idea de EM:

maximizar el *likelihood* (o el *log-likelihood*) del modelo a través de los parámetros.


Nota: entre más alto sea el *likelihood* más acercado será el modelo.

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com **161**



EM: Ejemplo *Likelihood*

- Tomamos datos de dos pdfs gaussianas conocidas de medias en -2 y 2 y desviación estándar de 1.5 ambas.



Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com **162**



EM: Ejemplo *Likelihood* (2)

- “Suponemos” un modelo con dos gaussianas de desviaciones 1.5 pero de medias desconocidas.
- A continuación se muestra la función de *likelihood* del *set* barriendo sobre todas las medias posibles entre $[-4, 4]^2$
- Observe que hay dos valores máximos, cerca de $[-2, 2]$ y de $[2, -2]$
no importa cual distribución genera cual conjunto de datos!

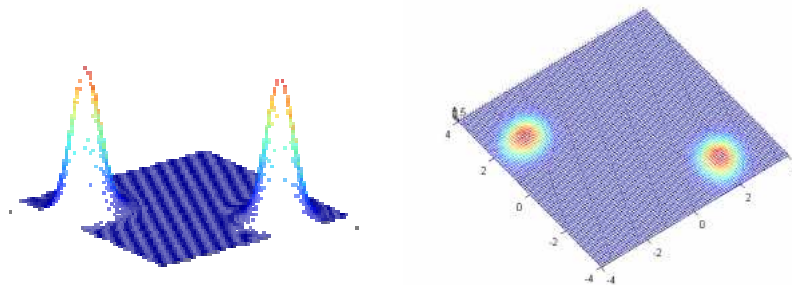
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

163





EM: Ejemplo *Likelihood* (3)



Feb-Jun, 2005

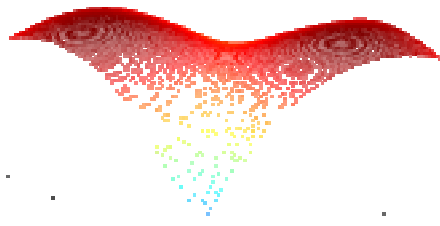
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

164



EM : Ejemplo *Likelihood* (4)

- La función de *log-likelihood* es:



**VER CODIGO
MATLAB**

Feb-Jun, 2005
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com
165

EM: Algoritmo

- Hay muchos algoritmos que implementan EM.
- Tienen en común dos pasos fundamentales:
 - E-Step***: *estimate the distribution of the hidden variable given the data and the current value of the parameters.*
 - M-Step***: *modify the parameters in order to maximize the joint distribution of the data and the hidden variable.*

(Samy Bengio, *An Introduction to Statistical Machine Learning - EM for GMMs* -, 2004.)

Feb-Jun, 2005
ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com
166



EM: Algoritmo (2)

- *E-Step*: consiste en calcular la probabilidad de que los datos hayan pertenecido a un *cluster*, según:

$$P(C_j/x) = \frac{P_{\mu_j, K_j}(x/C_j) \cdot P(C_j)}{P(x)}$$

- *M-Step*: recálculo de W_j y optimización de los parámetros para maximizar la función de *likelihood*.

$$W_j = \frac{\sum_{i=1}^N P(C_j/x_i)}{N}$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

167



EM: Algoritmo (3)

Para optimizar los parámetros existen muchos métodos:

- Búsqueda exhaustiva Imposible para grandes volúmenes de datos. Con cada cluster se añaden $M+M^2$ parámetros. (Recordar ejemplo de likelihood)
- Hallar los parámetros con promedios ponderados.
- Plantear un problema de optimización vectorial de L restringido a que:

$$\sum_{j=1}^k W_j = 1 \text{ y } W_j \geq 0 \forall j$$

Solucionarlo aplicando Multiplicadores de Lagrange (ver Samy Bengio).

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

168



EM: Algoritmo (4)

- La cantidad $P(C_j/x_i)$ puede verse como una membresía z_{ij} : la pertenencia del dato x_i al cluster C_j .
- Si se emplean actualizaciones promediadas el funcionamiento de EM es similar al de *fuzzy c-means*.

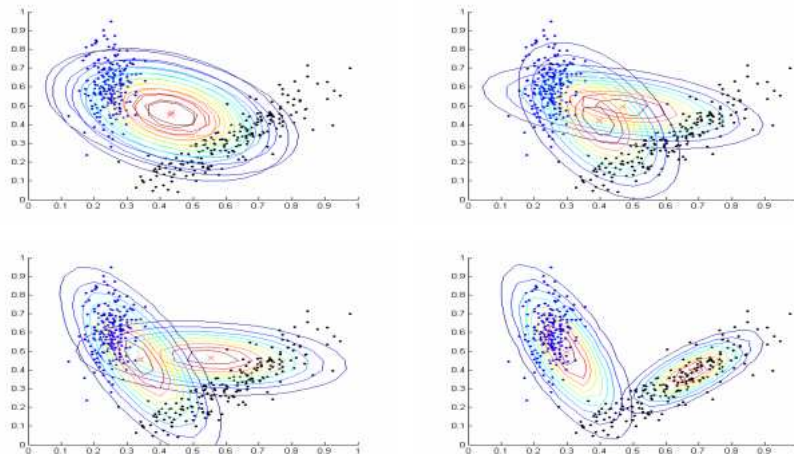
Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

169



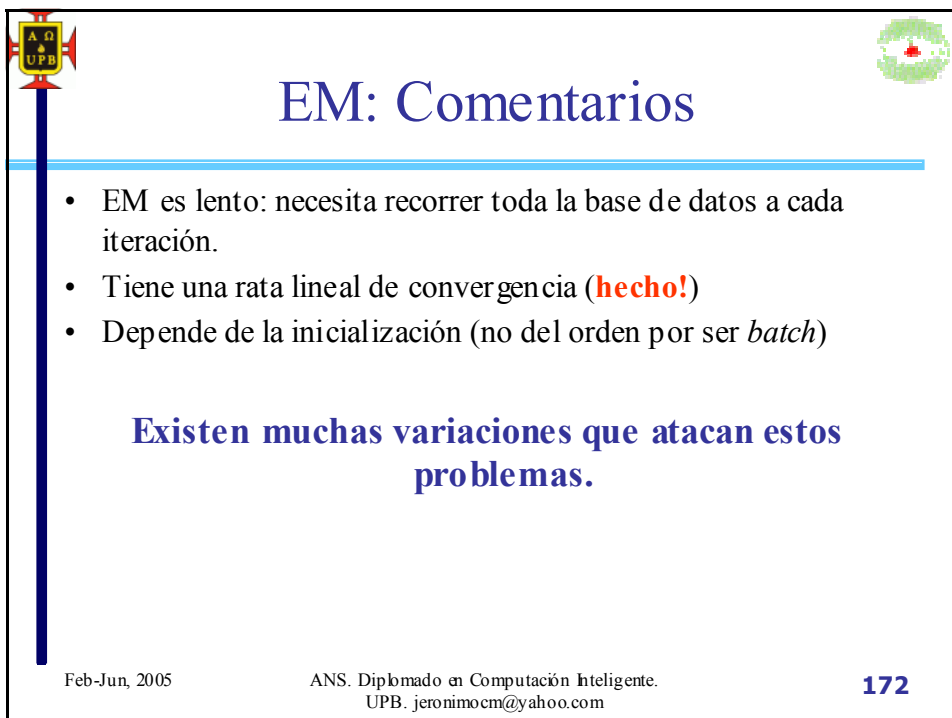
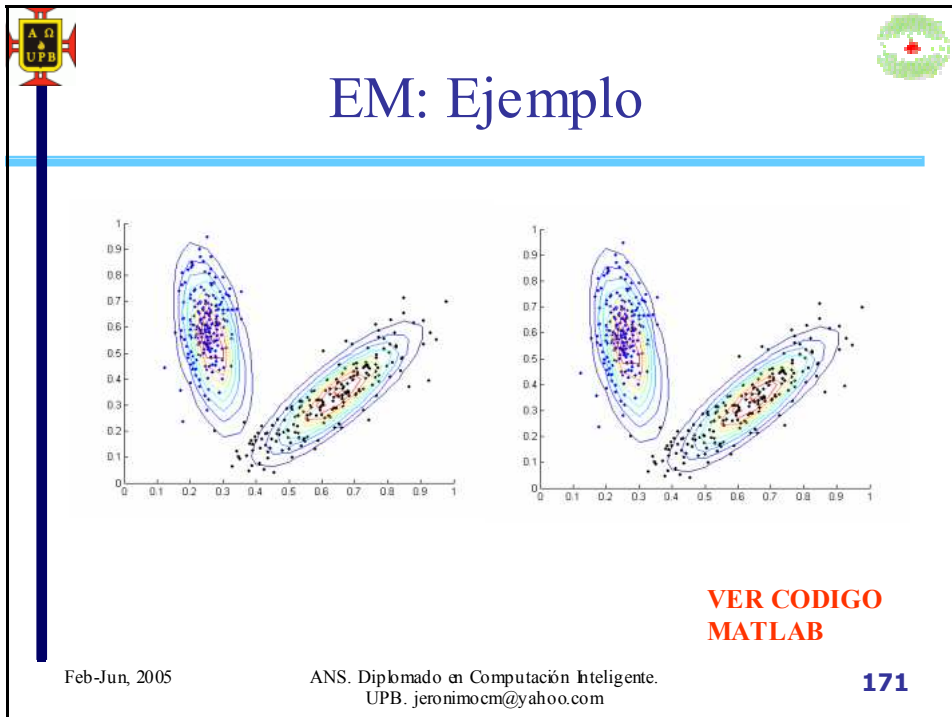
EM: Ejemplo



Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

170



The slide features a title 'EM: Comentarios' in blue. Below it is a bulleted list of three points about the EM algorithm. A bold blue statement follows the list. The footer contains the date 'Feb-Jun, 2005', the course name 'ANS. Diplomado en Computación Inteligente.', the email 'UPB. jeronimocm@yahoo.com', and the slide number '172'.

EM: Comentarios

- EM es lento: necesita recorrer toda la base de datos a cada iteración.
- Tiene una rata lineal de convergencia (**hecho!**)
- Depende de la inicialización (no del orden por ser *batch*)

Existen muchas variaciones que atacan estos problemas.

Feb-Jun, 2005 ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com 172



DBSCAN (Ester *et al* 1996)

- Presentado por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiawei Xu en München en la *2nd International Conference on Knowledge Discovery and Data Mining (KDD)* en 1996.
- DBSCAN: *Density Boundary SCAN*.
- Solución para bases de datos muy grandes: imágenes satelitales, cristalografía de rayos X.
- **Idea:** Si alrededor de un punto a una distancia fija hay un número mayor a un valor de umbral, entonces estos puntos pertenecen al mismo *cluster*.

Que dentro de una distancia haya un determinado numero de puntos equivale a evaluar una densidad mínima de puntos en dicha región

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

173



DBSCAN (2)

Ventajas:

- Halla clusters con formas arbitrarias.
- No hay que definir el número de clusters.
- Baja sensibilidad al ruido.
- Baja dependencia del orden.
- Eficiente computacionalmente.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

174



DBSCAN: Definiciones

- **ε -neighborhood:** la ε -neighborhood de un punto x es el conjunto definido por:

$$N_{\varepsilon}(x) = \{x \in X / \|x - x'\| < \varepsilon\}$$

Note que se puede utilizar cualquier norma (cualquier métrica).

- **Directly density-reachable:** un punto y es directamente alcanzable por densidad desde un punto x (sujeto a m y a ε) si:

$$y \in N_{\varepsilon}(x), \text{ con } |N_{\varepsilon}(x)| \geq m$$

A $|N_{\varepsilon}(x)| \geq m$ se le llama *core-point condition* y garantiza que el número de puntos en la vecindad de x sea mayor a un valor mínimo de umbral m .

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

175



DBSCAN: Definiciones (2)

- **Idea:** definir un cluster como el conjunto de todos los puntos con una densidad mayor a un valor determinado. Este criterio falla porque se debe tener en cuenta que el cluster tiene tanto puntos **interiores** como puntos de **frontera**. En estos últimos es de esperarse que haya una densidad menor.
- **Observación:** ser *directly density-reachable* no es una relación de equivalencia porque no satisface la simetría. Observe que un punto y puede ser DDR de otro x pero si y es un punto frontera entonces la *core-condition* no se satisface para que x sea DDR desde y .

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

176



DBSCAN: Definiciones (3)

- **Density reachable:** un punto y es alcanzable por densidad desde un punto x si existe una cadena $z_1, z_2, \dots, z_p, \dots, z_n$ con $z_1 = x$ y $z_n = y$ tal que z_{i+1} sea directamente alcanzable por densidad desde z_i .
- **Density-connectivity:** dos puntos x y y están conectados por densidad si existe un punto z tal que x y y sean alcanzables por densidad desde z .
- **Cluster:** es un conjunto de puntos conectados por densidad donde su cardinalidad es maximizada según la alcanzabilidad por densidad.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

177



DBSCAN: Definiciones (4)

- Todas las definiciones están sujetas a los parámetros ε y a m .
- DR es una relación transitiva pero tampoco es simétrica por la misma razón que DDR.
- ¿Todo par de puntos conectados por densidad son DR
No. Dos puntos de frontera están conectados por densidad pero no son alcanzables por densidad ya que ninguno de ellos satisfacen la *core-condition*.
- La relación de conectividad si es una relación de equivalencia: reflexiva, simétrica, transitiva.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

178



DBSCAN: Definiciones (5)

- **Cluster:** Formalmente...

Un *cluster* C_k sujeto a ε y a m es un subconjunto no vacío de X que satisface:

– *Maximality:*

$\forall i, j$; si $x_i \in C_k$ y x_j es $DR_{\varepsilon, m}$ desde $x_i \Rightarrow x_j \in C_k$
donde $DR_{\varepsilon, m}$ quiere decir: alcanzables por densidad sujeto a ε y a m .

– *Connectivity:*

$\forall i, j$; si $x_i, x_j \in C_k \Rightarrow x_i$ y x_j están conectados por densidad

Note que esta definición implica que todo cluster debe tener al menos m puntos

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

179



DBSCAN: Definiciones (6)

- **Ruido:** desde el punto de vista de DBSCAN el ruido es un conjunto N con patrones que no pertenecen a ningún *cluster*.

$$N = \{x \in X / \forall i, x \notin C_i\}$$

Observe que los *clusters* generados mediante DBSCAN no forman una partición de la base de datos:

$$\bigcup_i C_i \neq X$$

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

180



DBSCAN: Algoritmo

1. Seleccionar un punto x arbitrario no perteneciente a ningún *cluster*.
2. Verificar que x satisfaga la *core-point condition*. Si no volver a 1.
3. Formar el *cluster* hallando todos los puntos alcanzables por densidad ($DR_{\varepsilon m}$) desde x .
4. Si no se han recorrido todos los puntos de la base de datos volver al paso 1.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

181



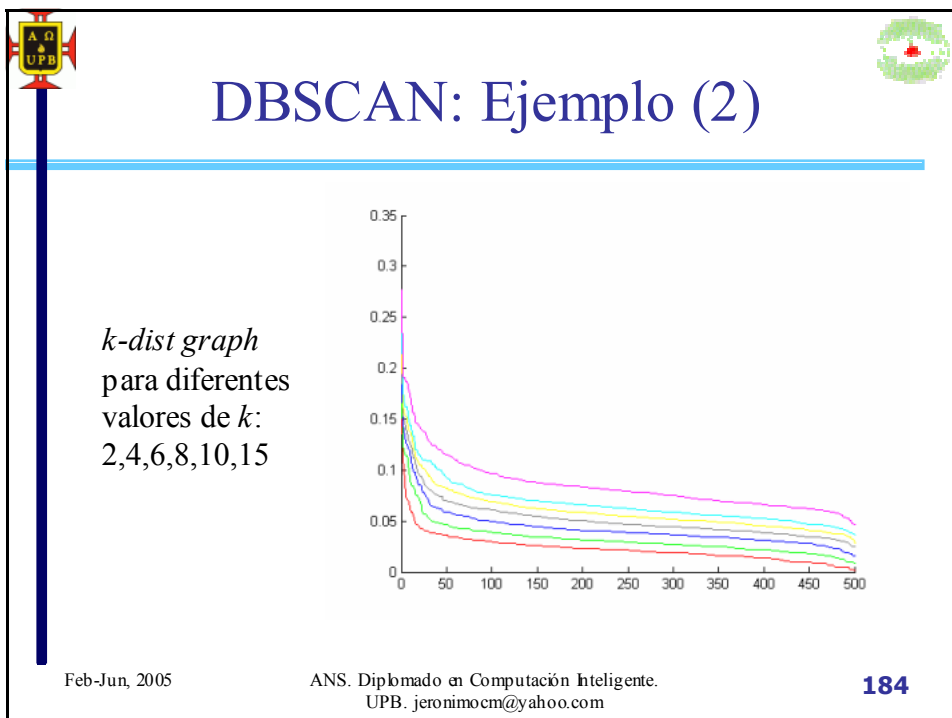
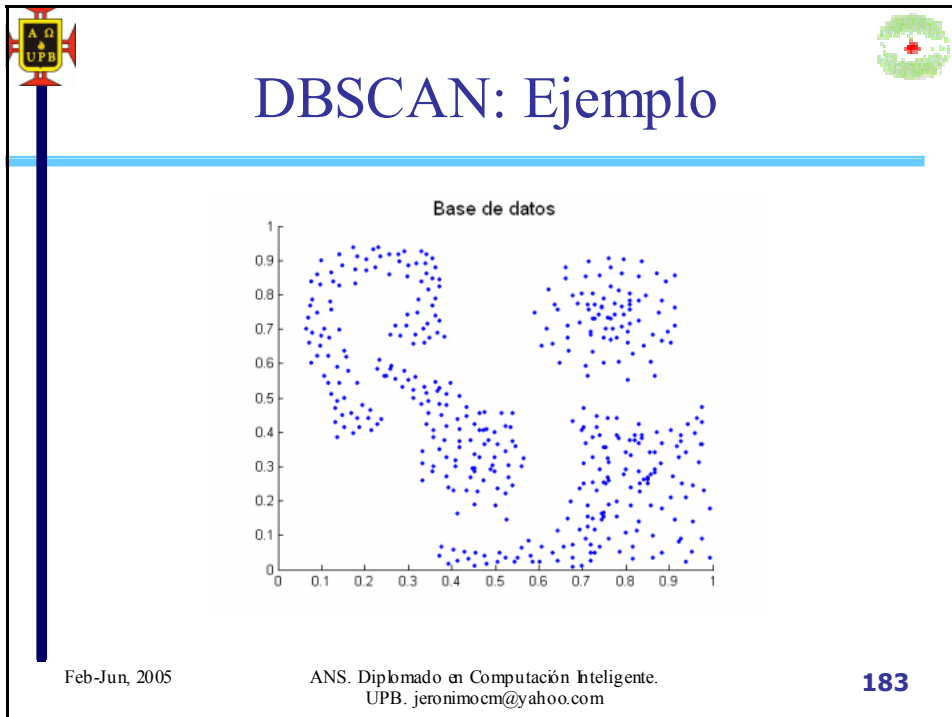
DBSCAN: ε y m

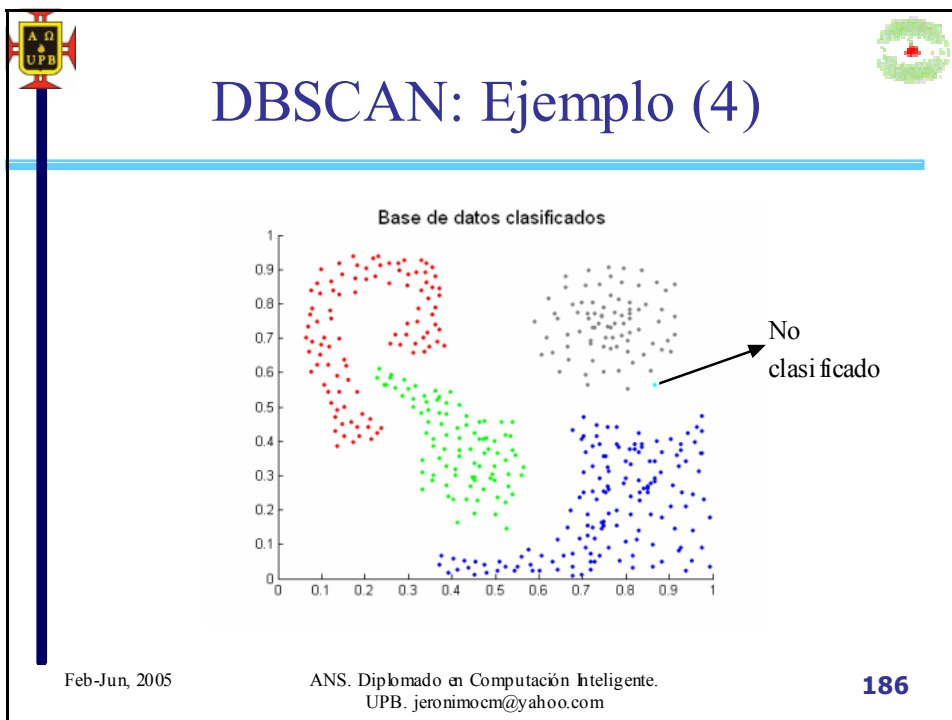
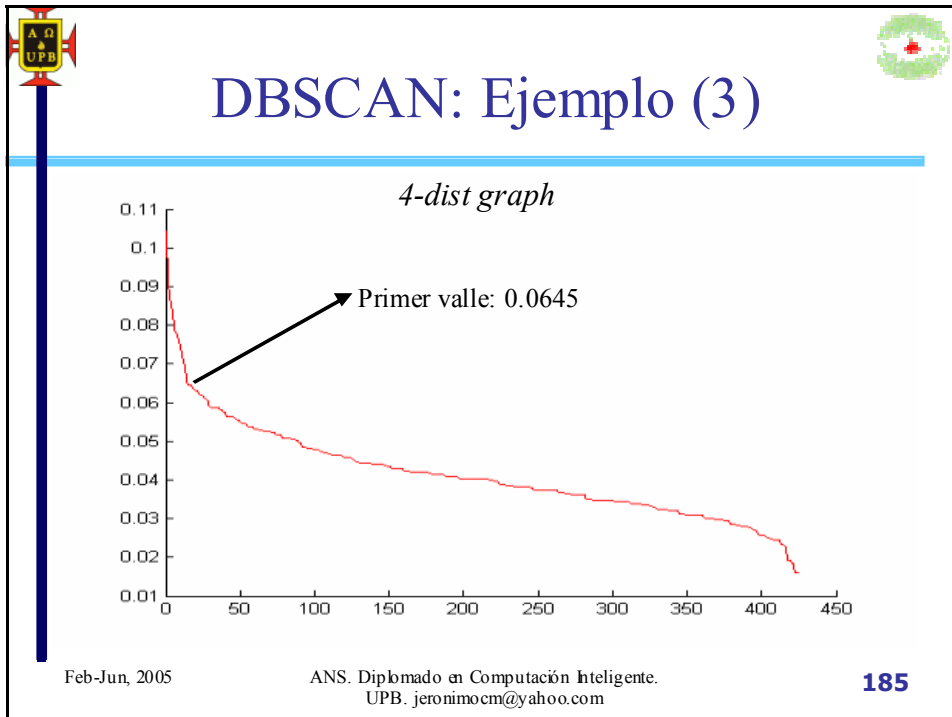
- Los autores sugieren un método heurístico para determinarlos.
- k -dist: distancia desde un punto dado a su k -th nearest neighbor.
- Si $\varepsilon = k - dist \rightarrow |N_\varepsilon| \geq k + 1$
- A medida que k aumenta k -dist no varía mucho.
- Los autores sugieren $k=4=m$ para bases 2-dimensionales.
- k -dist graph: grafica descendente del k -dist de cada punto en la base de datos.
- Fijar ε al valor de k -dist correspondiente al primer valle en la k -dist graph.
- Puntos a la izquierda de este valor son ruido.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

182







Bibliografía

- CASTRILLÓN, Jerónimo, GIRALDO, Daniel y PEÑA, Jorge A. Aprendizaje por Refuerzo en Espacios Continuos para la Evasión de Obstáculos en un Robot Móvil". Tesis (Ingeniería Electrónica). UPB, Medellín 2004, 365p.
- HASSOUN, Mohamad H. Fundamental of artificial neural networks. Boston, Massachusetts- USA. A Bradford Book, MIT Press, 1995. 511p.
- JAIN, A.K., MURTY, M.N. y FLYNN, P.J. DataClustering: A Review. En: ACM Computing Surveys. Ohio university. Vol. 31, No. 3 (Sep. 1999); p. 264-323.
- HYVÄRINEN, Aapo y OJA, Erkki. Independent Component Analysis: A Tutorial. 1999.
- ESTER, Martin et al. A Density-Based Algorithm for discovering Clusters in Large Spatial Databases with Noise. En: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland, USA. AAAI Press. Vol 2. (Ago. 1996); 6p.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

187



Bibliografía (2)

- AN INTRODUCTION TO PATTERN CLASSIFICATION. Yom-Tov, ELAD. (8.:2003. Tübingen). Memorias del MACHINE LEARNING SUMMER SCHOOL. Tübingen (Alemania): Max Planck Institute for Biological Cybernetics, 2003. 432 p.
- SCHÖLKOPF, Bernhard, SMOLA, Alexander y MÜLLER, Klaus-Robert. Kernel PCA, Nonlinear component analysis as a kernel eigenvalue problem. En: B. Schölkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods. MIT Press, 1998, p. 327-352.
- SHLENS Jon. A tutorial on Principal Components Analysis: Derivation, Discussion and Singular Value Decomposition. revisado en el 2003.
- SMITH, Lindsay. A tutorial on Principal Components Analysis. [pdf] revisado en 2002.
- MERCER, D.P. Clustering Large Datasets. [pdf] Linacre College, revised Oct 2003.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

188



Bibliografía (3)

- HINNEBURG, A., KEIM D. Clustering Technics for Large Datasets. [pdf] University of Halle. revised Oct 2000
- Unsupervised learning. Rassmussen. MLS2003
- ART, Carpenter, Grossberg
- LUBKIN J., CAUWENBERGHS, G. VLSI Implementation of Fuzzy Adaptive Resonance and Learning Vector Quantization. Analog Integrated Circuits and Signal Processing, 23, 1–10 (2001). Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.
- ESTER, Martin. Cluster and Outlier Analysis.
- BORGELT, Christian y KRUSE, Rudolf Shape and Size Regularization in Expectation Maximization and Fuzzy Clustering. Knowledge Processing and Language Engineering Otto-von-Guericke-University of Magdeburg.
- BENGIO, Samy. An introduction to statistical learning – EM for GMMs. IDIAP (Dalle Molle Institute for perceptual artificial intelligence), Septiembre 2004.

Feb-Jun, 2005

ANS. Diplomado en Computación Inteligente.
UPB. jeronimocm@yahoo.com

189