

Aprendizaje automatizado

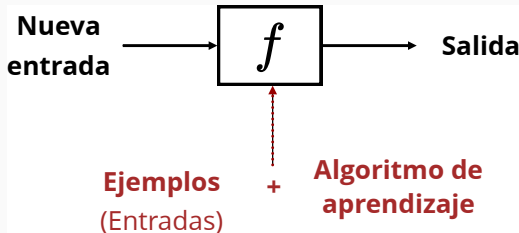
APRENDIZAJE SIN SUPERVISIÓN Y MODELOS DE VARIABLES
LATENTES

Gibran Fuentes Pineda

Mayo 2021

Aprendizaje sin supervisión

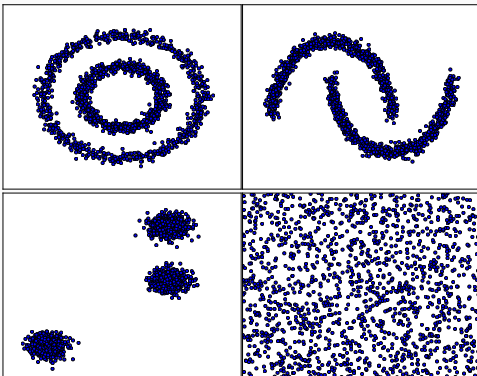
- Ejemplos sólo contienen entradas sin salidas deseadas
- Algunas tareas: el agrupamiento y el descubrimiento de patrones



- Busca encontrar la estructura escondida de los datos sin necesitar etiquetas

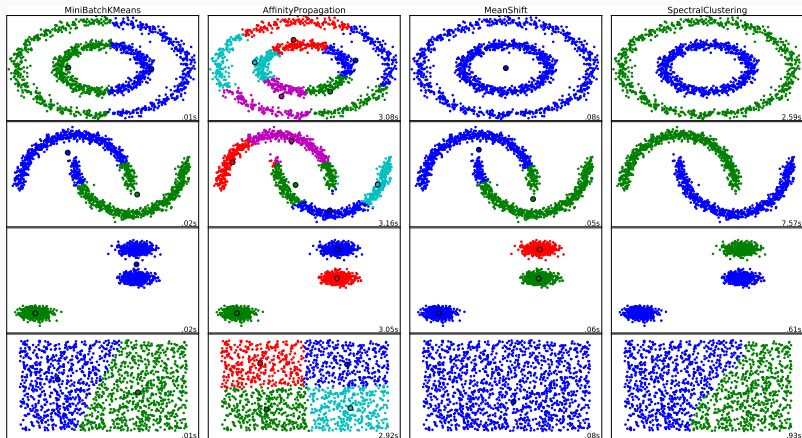
Agrupamiento

- Objetivo: agrupar ejemplos en base a su proximidad
- Criterios: por conectividad y por compacidad



Ejemplo de <http://scikit-learn.org>

Diferentes algoritmos de agrupamiento



Ejemplo de <http://scikit-learn.org>

Agrupamiento por K-medias

- Divide ejemplos en K grupos, asignando cada ejemplo al grupo con el centroide más cercano
- Busca los K centroides que minimicen

$$E[\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K] = \sum_{i=1}^n \sum_{k=1}^K r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

donde $r_{ik} = 1$ si $\boldsymbol{\mu}_k$ es el centroide más cercano a \mathbf{x}_i y $r_{ik} = 0$ en caso contrario

1. Elige K ejemplos aleatoriamente como centroides iniciales

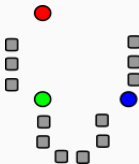


Imagen tomada de Wikipedia (K-means clustering)

Algoritmo de K-medias

1. Elige K ejemplos aleatoriamente como centroides iniciales
2. Asigna cada ejemplo al centroide más próximo

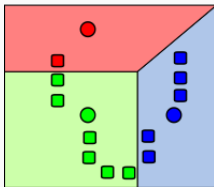


Imagen tomada de Wikipedia (K-means clustering)

Algoritmo de K-medias

1. Elige K ejemplos aleatoriamente como centroides iniciales
2. Asigna cada ejemplo al centroide más próximo
3. Re-calcula los centroides a partir de las asignaciones

$$\mu_k = \frac{1}{n_k} \sum_{i=1}^n r_{ik} \mathbf{x}_i, n_k = \sum_{i=1}^n r_{ik}, k = 1, \dots, K$$

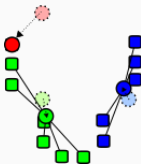


Imagen tomada de Wikipedia (K-means clustering)

Algoritmo de K-Means

1. Elige K ejemplos aleatoriamente como centroides iniciales
2. Asigna cada ejemplo al centroide más próximo
3. Re-calcula los centroides a partir de las asignaciones
4. Repite hasta cumplir criterio de convergencia (por ej. que E no disminuya)

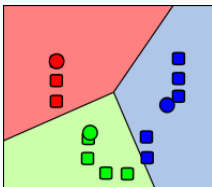
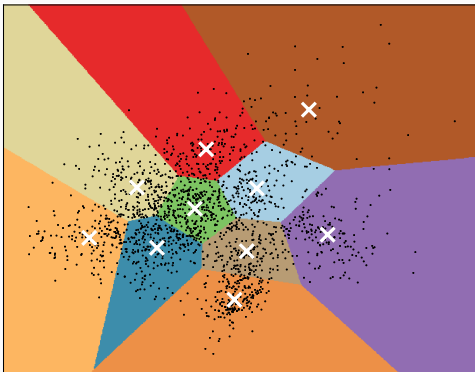


Imagen tomada de Wikipedia (K-means clustering)

Agrupamiento de imágenes de dígitos con K-medias

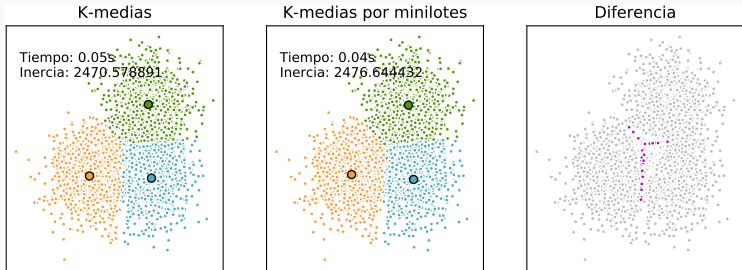
- El algoritmo de K-medias genera una partición del espacio representado por el diagrama de Voronoi en el que cada punto está asociado al centroide más próximo.



Ejemplo de <http://scikit-learn.org>

K-medias por minilotes¹

- Actualiza centroides y asignaciones usando un ejemplo o un subconjunto pequeño de ejemplos a la vez



Ejemplo de <http://scikit-learn.org>

¹D. Sculley. *Web-Scale K-Means Clustering*, 2010.

Agrupamiento jerárquico

- Construye de forma gradual una jerarquía de grupos siguiendo un criterio dado
 - **Aglomerativo:** empieza considerando cada dato como un grupo y va mezclando grupos
 - **Divisivo:** empieza considerando todos los datos como un solo grupo y lo va dividiendo
- Dados dos grupos $\{\mathcal{G}_a, \mathcal{G}_b\}$, algunos criterios son
 - Mínimo o simple: $\min \{dist(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) : \mathbf{x}^{(i)} \in \mathcal{G}_a, \mathbf{x}^{(j)} \in \mathcal{G}_b\}$
 - Completo o máximo: $\max \{dist(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) : \mathbf{x}^{(i)} \in \mathcal{G}_a, \mathbf{x}^{(j)} \in \mathcal{G}_b\}$
 - Promedio

$$\frac{1}{|\mathcal{G}_a| |\mathcal{G}_b|} \sum_{\mathbf{x}^{(i)} \in \mathcal{G}_a} \sum_{\mathbf{x}^{(j)} \in \mathcal{G}_b} dist(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$$

- Mínima varianza: elige el par de grupos en el que la varianza intergrupar se incremente menos

Dendogramas

- Diagrama jerárquico que muestra los agrupamientos en distintos niveles

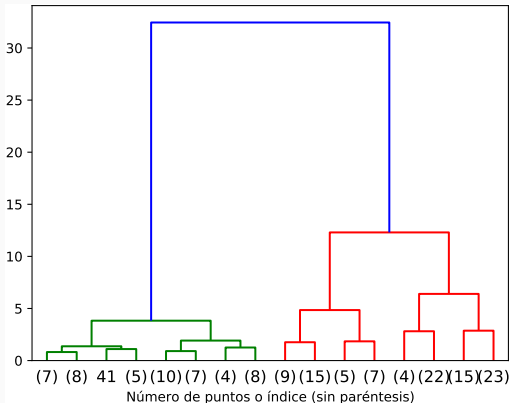


Imagen generada por ejemplo de scikit-learn

Agrupamiento espectral

- Se calcula la matriz laplaciana L a partir de la matriz de adyacencia o afinidad A de la siguiente manera

- Sin normalizar

$$L = D - A$$

- Normalizada simétrica

$$L_{\text{sim}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}$$

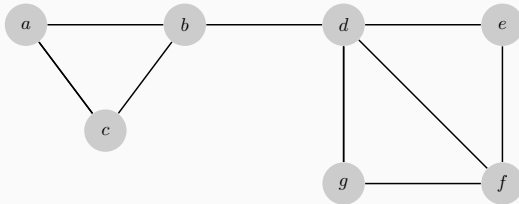
- Caminata aleatoria

$$L_{\text{ca}} = D^{-1} L = I - D^{-1} A$$

- Se realiza el agrupamiento usando K-medias sobre los puntos representados por los K eigenvectores con mayores eigenvalores de la matriz laplaciana L

Corte de grafos

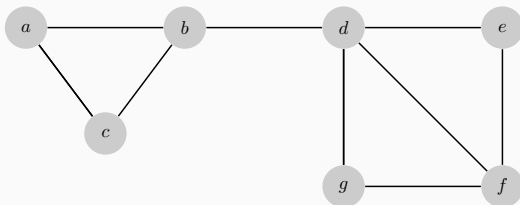
- Consideremos una red como grafo no dirigido $\mathcal{G}(V, E)$
- Objetivo: dividir los vertices en grupos disjuntos tal que
 - Se maximice el número de conexiones entre vértices del mismo grupo
 - Se minimice el número de conexiones entre vértices de grupos distintos



Representación de grafos: matriz de adyacencia

- Matriz de adyacencia A

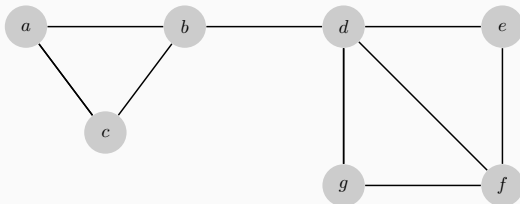
$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$



Representación de grafos: matriz de grado

- Matriz de grado G

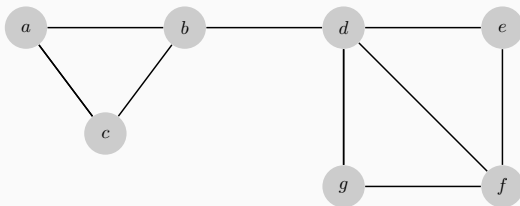
$$G = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$



Representación de grafos: matriz laplaciana

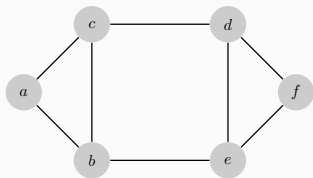
- Laplaciana $L = G - A$

$$L = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 \\ -1 & -1 & 2 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 4 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{bmatrix}$$



1. Construir la matriz laplaciana L .
2. Calcular los eigenvalores y eigenvectores de la matriz laplaciana
3. Mapear los vértices a los elementos del eigenvector \mathbf{e} asociado al segundo eigenvalor
4. Ordenar los elementos del vector \mathbf{e}
5. Agrupar vértices dividiendo el vector ordenado en dos
 - Por ejemplo usando el 0 o la mediana.

Partición espectral: otro ejemplo con 2 grupos



$$L = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & -0 & 0 & 0 \\ -1 & -1 & 3 & 0 & 0 & -1 \\ -1 & 0 & 0 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & -1 & -1 & 3 \end{bmatrix}$$

λ	0	1	2	3	4	5
\mathbf{x}	1	1	-5	-1	-1	-1
	1	2	4	-2	1	0
	1	1	1	3	-1	1
	1	-1	-5	-1	1	1
	1	-2	4	-2	-1	0
	1	-1	1	3	1	-1

Partición espectral: ejemplo con 2 grupos

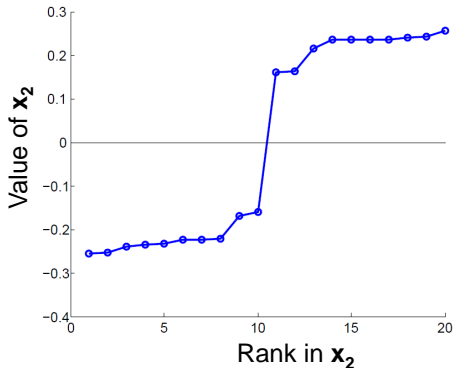
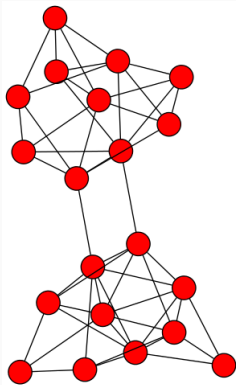


Figura tomada de diapositivas por J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmms.org>

Partición espectral: ejemplo con 4 grupos (1)

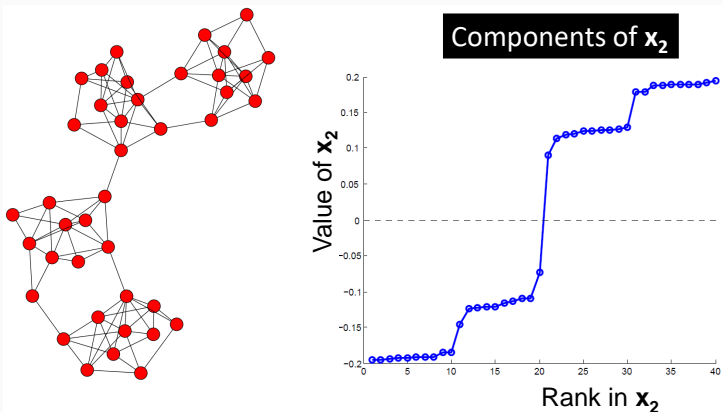


Figura tomada de diapositivas por J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmms.org>

Partición espectral: ejemplo con 4 grupos (2)

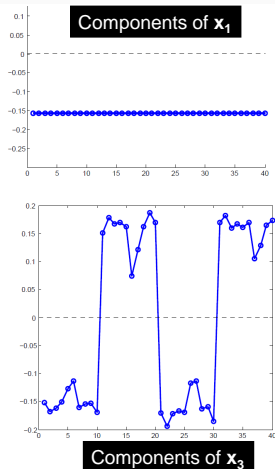
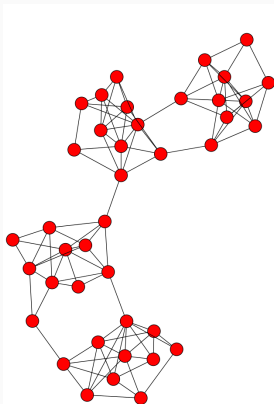


Figura tomada de diapositivas

por J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, <http://www.mmds.org>

Partición de grafos en K grupos

- Partición recursiva
 - Divide el grafo en dos grupos, posteriormente se aplica el algoritmo en cada grupo generando a su vez dos subgrupos cada uno y así sucesivamente.
- Múltiples eigenvectores
 - Se representa cada vértice usando los elementos de múltiples eigenvectores (asociados a los eigenvalores $\lambda_2, \lambda_3, \dots, \lambda_K$).
 - Se agrupan las representaciones usando algún algoritmo de agrupamiento (por ej. K-medias).

Modelando con variables latentes

- En muchos fenómenos las observaciones (variables observadas) dependen de variables no directamente visibles (variables latentes)
- Un modelo con variables no visibles se conoce como **modelo de variable latente (MVL)**
- Ventajas
 1. Son modelos más compactos en general
 2. Es posible aprender ciertas estructuras en los datos sin supervisión

Dependencia local en MVLs

- Suposición: relación entre variables observadas se da únicamente a través de variables latentes
- Ejemplo (de Lazarsfeld and Henry): 1000 personas fueron encuestadas sobre si leen la revista A y B.

	Leyó A	No leyó A	Total
Leyó B	260	140	400
No leyó B	240	360	600
Total	500	500	1000

Dependencia local en MVLs

- Suposición: relación entre variables observadas se da únicamente a través de variables latentes
- Ejemplo (de Lazarsfeld and Henry): 1000 personas fueron encuestadas sobre si leen la revista A y B.

High education	Read A	Did not read A	Total
Read B	240	60	300
Did not read B	160	40	200
Total	400	100	500
Low education	Read A	Did not read A	Total
Read B	20	80	100
Did not read B	80	320	400
Total	100	400	500

- Los modelos de variables latentes se pueden clasificar por la naturaleza de sus variables latentes y observadas

	V. observadas	
V. latentes	Continua	Categórica
Continua	Análisis de factores	Teoría de la respuesta al reactivo
Discreta	Modelo de mezclas	Análisis de clases latentes

- Variable latente discreta $z \in \{1, \dots, K\}$

$$z \sim \text{Cat}(\boldsymbol{\pi})$$

- K distribuciones base $P(\mathbf{x}|z = k) = f_k(\mathbf{x})$

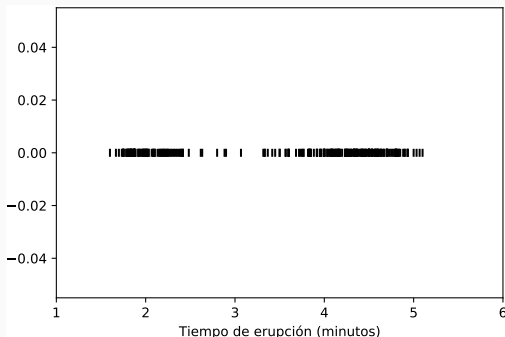
$$\mathbf{x}|z \sim f_k(\mathbf{x})$$

- Distribución de \mathbf{x} se puede expresar como

$$P(\mathbf{x}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x})$$

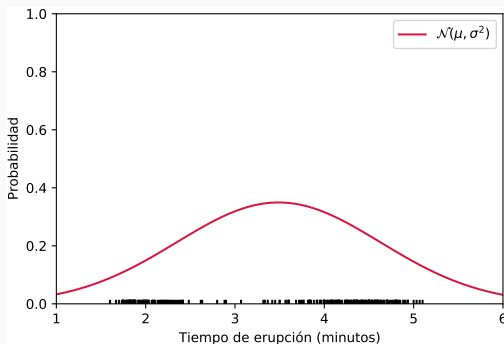
Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



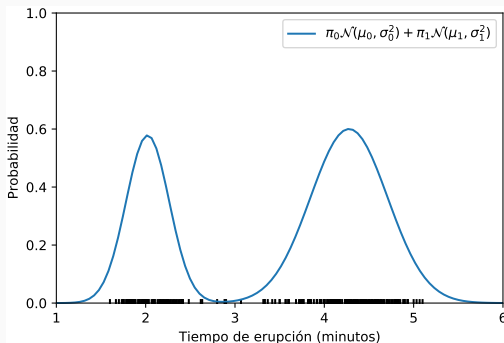
Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



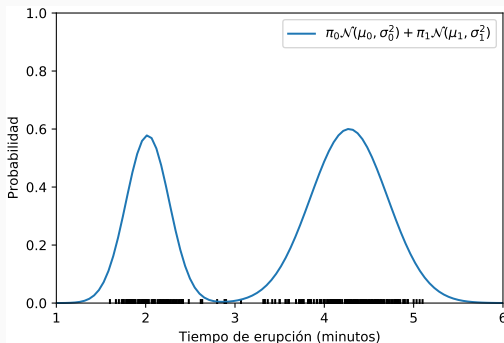
Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



Modelos de mezclas: representando distribuciones complejas

- ¿Qué distribución podríamos presuponer para los siguientes datos?



- ¿Cómo estimamos los parámetros?

- K distribuciones base $f_k(\mathbf{x})$ gaussianas

$$\mathbf{x}|z \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Rightarrow P(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- La verosimilitud logarítmica está dada por

$$\log \{P(\mathcal{D}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\} = \sum_{i=1}^N \log \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- No hay solución cerrada analítica, necesitamos usar algoritmos de optimización iterativa.

- Algoritmo para estimar parámetros por máxima verosimilitud o máximo a posteriori en problemas con datos faltantes y modelos de variables latentes
- Procedimiento general
 1. **Paso E:** inferir valores faltantes o de variables latentes
 2. **Paso M:** optimizar parámetros usando datos inferidos

EM para estimación por máxima verosimilitud

- Considera que el conjunto de ejemplos está dado por los valores tanto de las variables observadas como las variables latentes $\{\mathcal{D}, \mathbf{Z}\}$
- Busca encontrar los valores de los parámetros θ que maximicen la verosimilitud logarítmica de $\{\mathcal{D}, \mathbf{Z}\}$

$$\theta = \arg \max_{\theta} \log \left\{ \sum_{\mathbf{Z}} P(\mathcal{D}, \mathbf{Z} | \theta) \right\}$$

- Como los valores de las variables latentes \mathbf{Z} no se conocen, se calcula la distribución a posteriori $P(\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo})$ con los parámetros actuales $\boldsymbol{\theta}^{viejo}$

$$P(\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo}) = \frac{P(\mathcal{D}|\mathbf{Z}, \boldsymbol{\theta}^{viejo})P(\mathbf{Z}|\boldsymbol{\theta}^{viejo})}{P(\mathcal{D}|\boldsymbol{\theta}^{viejo})}$$

- Se asignan parámetros que maximizan la esperanza de la verosimilitud logarítmica usando $P(\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo})$

$$\begin{aligned}\boldsymbol{\theta}^{nuevo} &= \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{viejo}) \\ &= \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo}} \left[\log \left\{ \sum_{\mathbf{Z}} P(\mathcal{D}, \mathbf{Z}|\boldsymbol{\theta}) \right\} \right] \\ &= \sum_{\mathbf{Z}} P(\mathbf{Z}|\mathcal{D}, \boldsymbol{\theta}^{viejo}) \log \{P(\mathcal{D}, \mathbf{Z}|\boldsymbol{\theta})\}\end{aligned}$$

1. Inicializa parámetros θ
2. **Paso E:** Evaluar $P(\mathbf{Z}|\mathcal{D}, \theta^{viejo})$
3. **Paso M:** Re-estimar parámetros

$$\theta^{nuevo} = \arg \max_{\theta} Q(\theta, \theta^{viejo})$$

4. Repetir 2 y 3 hasta que se cumpla el criterio de convergencia

Distribución a posteriori para modelo de mezclas gaussianas

- Probabilidad a posteriori $P(z = k|\mathbf{x}^{(i)})$ (responsabilidad) está dada por

$$P(z^{(i)} = k|\mathbf{x}^{(i)}) = \gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

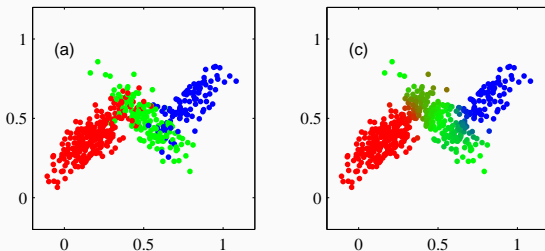


Imagen tomada de Bishop, PRML 2007

EM para modelos de mezclas gaussianas

1. Inicializa $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ y π_k
2. **Paso E:** Evalúa responsabilidades con parámetros actuales

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(i)} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(i)} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

3. **Paso M:** Recalcula parámetros $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ y π_k a partir de $\gamma(z_{nk})$

$$n_k = \sum_{i=1}^n \gamma(z_{ik})$$

$$\boldsymbol{\mu}_k^{\text{nuevo}} = \frac{1}{n_k} \sum_{i=1}^n \gamma(z_{ik}) \cdot \mathbf{x}^{(i)}$$

$$\boldsymbol{\Sigma}_k^{\text{nuevo}} = \frac{1}{n_k} \sum_{i=1}^n \gamma(z_{ik}) \cdot (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{\text{nuevo}})(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{\text{nuevo}})^T$$

$$\pi_k^{\text{nuevo}} = \frac{n_k}{n}$$

4. Evalúa verosimilitud logarítmica

$$\log \{P(\mathcal{D} | \boldsymbol{\mu}^{\text{nuevo}}, \boldsymbol{\Sigma}^{\text{nuevo}}, \boldsymbol{\pi}^{\text{nuevo}})\}$$

Modelo de mezclas gaussianas y EM en acción

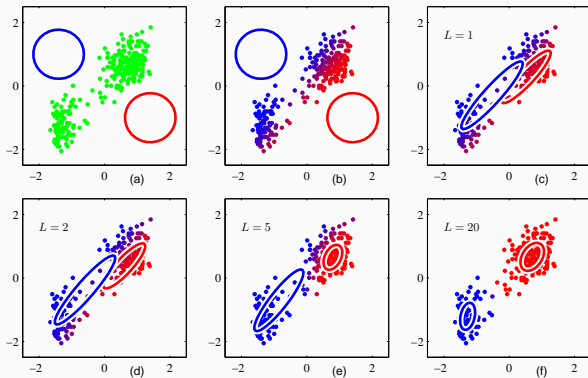


Imagen tomada de Bishop, PRML 2007

Modelo de mezclas de Bernoulli (análisis de clases latentes)

- Ejemplos con d variables binarias $\mathbf{x}^{(i)} = \{x_1, \dots, x_d\}$

$$P(\mathbf{x}^{(i)}|\mathbf{q}) = \prod_{j=1}^d q_j^{x_j^{(i)}} (1 - q_j^{x_j^{(i)}})^{(1-x_j^{(i)})}$$

- Mezcla de K de estas distribuciones

$$P(\mathbf{x}^{(i)}|\mathbf{Q}, \boldsymbol{\pi}) = \sum_{k=1}^K \pi_k \cdot \left[\prod_{j=1}^d q_{kj}^{x_j^{(i)}} (1 - q_{kj}^{x_j^{(i)}})^{(1-x_j^{(i)})} \right]$$

donde $\mathbf{Q} = \{\mathbf{q}_1, \dots, \mathbf{q}_K\}$ y $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$

EM para modelo de mezclas de Bernoulli

1. Inicializa \mathbf{q}_k y $\boldsymbol{\pi}_k$
2. **Paso E:** Evalúa responsabilidades con parámetros actuales

$$\gamma(z_{ik}) = \frac{\pi_k \cdot \left[\prod_{j=1}^d q_{kj}^{x_j^{(i)}} \left(1 - q_{kj}^{x_j^{(i)}}\right)^{(1-x_j^{(i)})} \right]}{\sum_{l=1}^K \pi_l \cdot \left[\prod_{j=1}^d q_{lj}^{x_j^{(i)}} \left(1 - \mu_{lj}^{x_j^{(i)}}\right)^{(1-x_j^{(i)})} \right]}$$

3. **Paso M:** Re-estima parámetros $\boldsymbol{\mu}_k$ y $\boldsymbol{\pi}_k$ a partir de $\gamma(z_{nk})$

$$\boldsymbol{\mu}_k = \sum_{i=1}^n \gamma(z_{ik}) \mathbf{x}^{(i)}$$

$$\pi_k = \frac{n_k}{n}$$

$$n_k = \sum_{i=1}^n \gamma(z_{ik})$$

4. Evalúa verosimilitud logarítmica $\log P(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\pi})$

Desventajas de EMV en MMG

- Singularidades: cuando una media es igual a un ejemplo $\mathbf{x}^{(i)} = \boldsymbol{\mu}_k$, la verosimilitud logarítmica se vuelve infinito ya que $\sigma_k \rightarrow 0$.

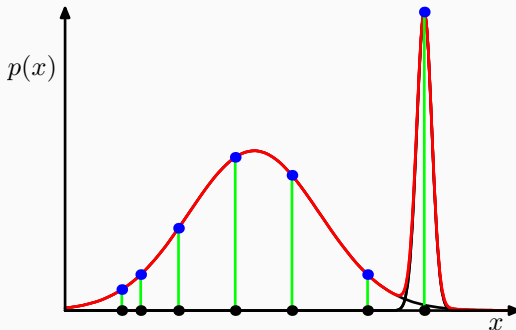
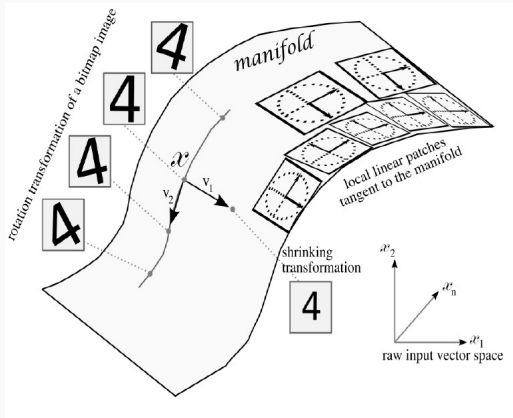


Figura tomada de Bishop, PRML 2007

- Singularidades: cuando una de las medias μ_k es exactamente igual a un dato.
- No identificabilidad: existen $K!$ soluciones equivalentes

La hipótesis de la variedad

- Ejemplos pueden vivir en una variedad de muchas menores dimensiones que el espacio original



Análisis de componentes principales (PCA)

- Busca subespacio de m dimensiones que maximiza varianza (o minimiza error) de los ejemplos
 - Definido por eigenvectores $\mathbf{u}_1, \dots, \mathbf{u}_m$ con eigenvalores más grandes $\lambda_1, \dots, \lambda_m$ de la matriz de covarianza

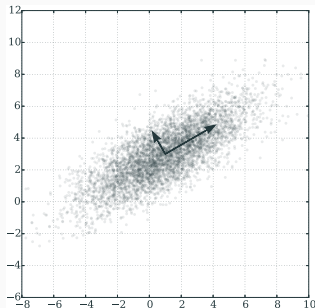
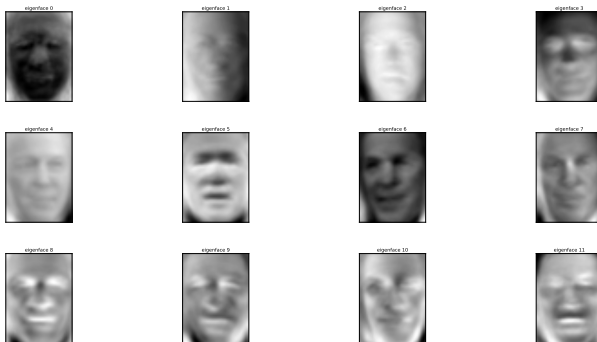


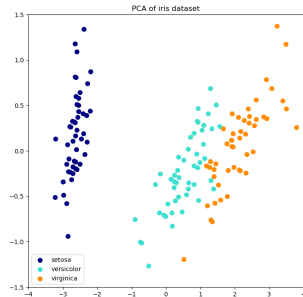
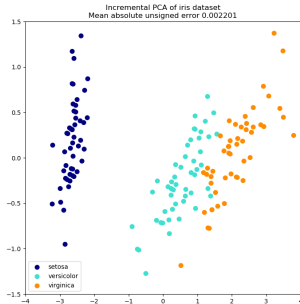
Figura tomada de Wikipedia (*Principal Component Analysis*)

PCA aplicado a imágenes de rostros

- Componentes principales se toman como base (eigenfaces)
- Nuevos rostros se proyectan en subespacio encontrado para ser comparados



PCA incremental



Ejemplo de <http://scikit-learn.org>

Análisis de factores: variables continuas (1)

- Variables latentes continuas $\mathbf{z} \in \mathbb{R}^K$, con a priori gaussiana

$$\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

- Variables observadas continuas $\mathbf{x} \in \mathbb{R}^d$ con²

$$\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

- Distribución sobre \mathbf{x} está dada por

$$P(\mathbf{x}) = \int P(\mathbf{x}|\mathbf{z})P(\mathbf{z})d\mathbf{z} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$$

$$\text{donde } \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \boldsymbol{\Psi}$$

²Cuando $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$, $\boldsymbol{\mu}_0 = \mathbf{0}$ y $\boldsymbol{\Sigma}_0 = \mathbf{I}$, se conoce como *análisis de componentes principales probabilista* (PPCA).

Proceso generativo de PPCA

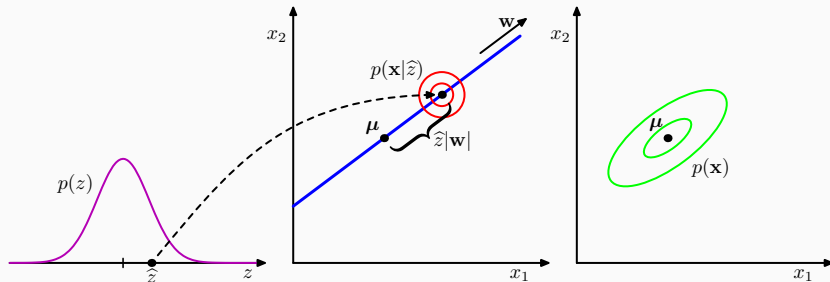


Imagen tomada de Bishop, PRML 2007

- Presuponiendo $\sigma^2 = 0$, se pueden encontrar parámetros de PCA por máxima verosimilitud usando el algoritmo EM

1. Paso E: $\tilde{\mathbf{Z}} = (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \tilde{\mathbf{X}}$

2. Paso M: $\mathbf{W} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{Z}}^\top (\tilde{\mathbf{Z}} \tilde{\mathbf{Z}}^\top)^{-1}$

donde $\tilde{\mathbf{X}} = \mathbf{X}^\top$

Análisis de componentes independientes (ICA)

- ICA considera que variables latentes no siguen una distribución gaussiana pero son independientes

$$P(\mathbf{z}) = \prod_{j=1}^d P(z_j)$$

- Aplicación: separación de fuentes ciega

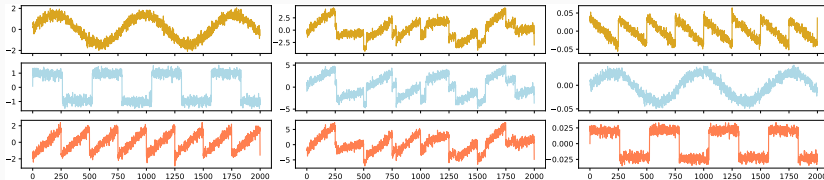


Imagen generado usando ejemplo de <http://scikit-learn.org>

PCA vs ICA

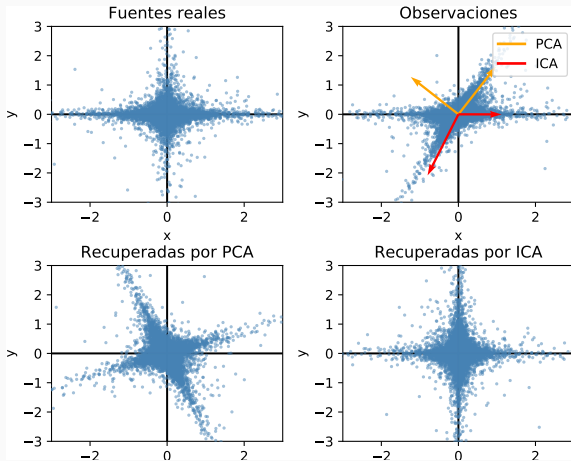


Imagen generada usando ejemplo de <http://scikit-learn.org>

Codificación dispersa

- \mathbf{z} tiene más dimensiones que \mathbf{x}
- Apriori de \mathbf{z} viene de distribución que favorece dispersidad
- \mathbf{x} se aproxima como combinación dispersa de columnas de \mathbf{W}

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \epsilon$$

