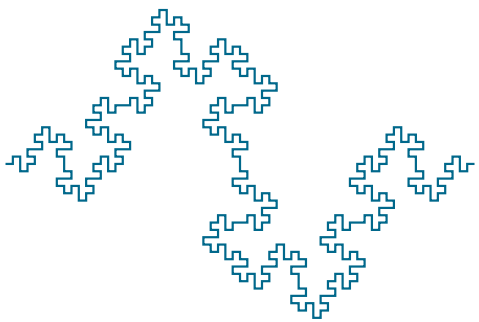
Modelo de mezclas Gaussianas (GMM)

Rafael Pérez Torres

Tópicos Selectos en Reconocimiento de Patrones
Profesor: Dr. Wilfrido Gómez Flores

LTI Cinvestav



Agenda

INTRODUCCIÓN

- Normalmente, la naturaleza de una distribución refleja características intrínsecas en los datos.
- Las técnicas de agrupamiento basadas en distribución intentan crear un modelo parametrizado que describa a los datos.
- Con el modelo es posible realizar inferencias que sean útiles para labores de identificación - agrupamiento

INTRODUCCIÓN

- El modelo que describe a una distribución puede ser conocido a través de una *PDF* (Probability Density Function).
- Normalmente, los datos que provienen del mundo real obedecen a una distribución normal o gaussiana, definida como

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{(l/2)}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \quad (1)$$

Esta PDF es parametrizada por las medias μ y covarianzas Σ de los datos.

INTRODUCCIÓN

- En general, las técnicas de agrupamiento asumen que los datos han sido generados por una mezcla de distribuciones (PDF).
- Cada componente de la mezcla define a cada uno de los grupos.

EL MODELO DE MEZCLAS GAUSSIANAS (GMM)

- El GMM es una suma ponderada de M PDFs Gaussianas, expresado como

$$p(x|\omega_i, \Sigma_i) = \sum_{i=1}^{M} \omega_i g(x|\mu_i, \Sigma_i) \tag{2}$$

Donde $g(x|\mu_i, \Sigma_i)$ es la PDF gaussiana de cada grupo y ω_i es el vector de coeficientes mixtos que se encuentra sujeto a la restricción $\sum_{i=1}^{M} \omega_i = 1$.

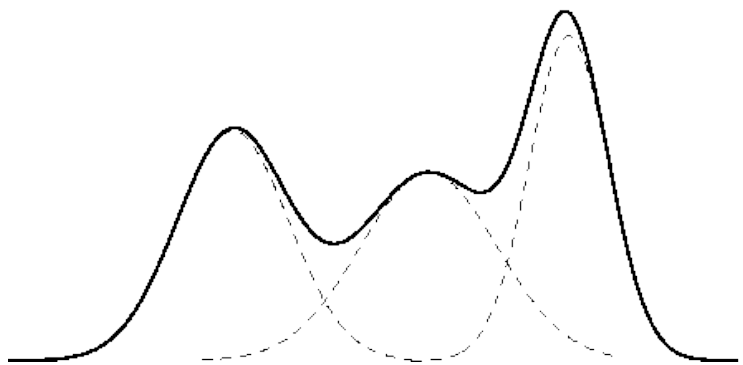


Figura: Una combinación de PDFs Gaussianas

EL MODELO DE MEZCLAS GAUSSIANAS (GMM)

- Entonces, el GMM es parametrizado por los vectores de medias, covarianzas y coeficientes de cada grupo:

$$\lambda = \{\omega_i, \mu_i, \Sigma_i\}$$

- El resto del trabajo consiste en encontrar los valores adecuados para dichos parámetros.

ESTIMACIÓN DE LOS PARÁMETROS DEL GMM

- Existen varios métoos para estimar λ de un GMM.
- El más utilizado es el método de estimación de máxima verosimilitud (ML).
- El objetivo del método de ML es maximizar la verosimilitud del GMM dado el conjunto de datos de entrenamiento.

EL MÉTODO DE MÁXIMA VEROSIMILITUD

Sea $\mathcal{X}$ una variable aleatoria con función de probabilidad $p(\mathcal{X};\theta)$ donde θ es un parámetro desconocido.
Además, sean $x_1, x_2, \ldots, x_N$ los valores observados en una muestra aleatoria de tamaño N de la misma variable.
La función de verosimilitud de la muestra (o función de densidad conjunta) es:

$$\mathcal{L}(\theta) = p(\mathcal{X};\theta) = p(x_1, x_2, \ldots, x_n;\theta) = \prod_{i=1}^{N} p(x_i;\theta) \tag{3}$$

EL MÉTODO DE MÁXIMA VEROSIMILITUD, EJEMPLO

- Se desea estimar la probabilidad p de que salga cara en el lanzamiento de una moneda (no necesariamente regular).
- Se lanza cinco veces la moneda y se obtiene: $C + CC+$
- $p(C + CC+) = p * (1 - p) * p * p * (1 - p) = p^3(1 - p)^2$

Valor de p	**Probabilidad de la muestra observada**
0.0	0.0000
0.1	0.0008
0.2	0.0051
0.3	0.0132
0.4	0.0230
0.5	0.0313
0.6	0.0346
0.7	0.0309
0.8	0.0205
0.9	0.0073
1.0	0.0000

Cuadro: Valores obtenidos para distintos valores de p

EL MÉTODO DE MÁXIMA VEROSIMILITUD, EJEMPLO

$$p(k \text{ caras en} n \text{ lanzamientos}) = \binom{n}{k} p^k (1-p)^{n-k} = \mathcal{L}(p) \tag{4}$$

Derivando e igualando a 0

$$\frac{\partial \ln\big(\mathcal{L}(p)\big)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} \tag{5}$$

$$\frac{k}{\hat{p}} - \frac{n-k}{1-\hat{p}} = 0 \rightarrow k(1-\hat{p}) - \hat{p}(n-k) = 0 \rightarrow \hat{p} = \frac{k}{n} \tag{6}$$

El *estimador máximo verosimil* de la probabilidad de un suceso es la frecuencia relativa.

ML EN DISTRIBUCIONES NORMALES

Al aplicar ML a distribuciones Gaussianas, se observa que los estimadores máximos verosímiles coinciden con la μ y la Σ.

$$\mathcal{L}(\mu, \sigma^2) = \ln \prod_{i=1}^{N} p(x_i; \theta) \tag{7}$$

$$= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{8}$$

$$= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x - \mu)^2 \tag{9}$$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 \tag{10}$$

EL ALGORITMO EM

- El algoritmo Expectation-Maximization se basa en el método de estimación de máxima verosimilitud para determinar los valores *óptimos* de los parámetros (λ) de un GMM.
- Se deben calcular las derivadas de los parámetros de la función de log-verosimilitud $\left(\lambda = \{\mu_k, \Sigma_k, \omega_k\}\right)$:

$$\ln p(x|\lambda) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{M} \omega_k g\left(x_i|\mu_k, \Sigma_k\right) \right\}$$

EL ALGORITMO EM

Pasos del algoritmo EM

1. Selección de valores iniciales para λ.
2. Paso E: Evaluar probabilidades a posteriori utilizando el actual conjunto de valores λ.
3. Paso M: Reestimar λ utilizando las probabilidades a posteriori obtenidas.
4. Criterio de parada atendiendo a los resultados de la función de verosimilitud de la iteración t y $t-1$.

EM: PASO 1

Paso 1: Selección de valores iniciales

- El algoritmo EM requiere la cantidad de M componentes Gaussianas a considerar en el GMM y el valor inicial de λ
- Normalmente, se utiliza k-means para obtener el valor de las medias iniciales.
- Σ_k se calcula en base a todos las instancias pertenecientes al grupo k.
- Los coeficientes se calculan como $\omega_k = N_k/N$ donde N_k es la cantidad de *instancias* del grupo y N es la cantidad total de instancias.

EM: PASO 2, E

Paso 2: Expectation

Para la i-ésima muestra se calculan las probabilidades a posteriori para la k-ésima componente, o en otras palabras, se calcula la probabilidad de que cada muestra pertenezca a cada clúster, mediante:

$$p(k|x_i, \lambda) = \frac{w_k g(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{M} w_k g(x_i|\mu_k, \Sigma_k)} \tag{11}$$

donde $g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{(l/2)}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right)$.
Este paso puede ser considerado similar al paso de asignación de clúster en el algoritmo k-means.

EM: PASO 3, M

Paso 3: Maximization

Se reestiman los valores del parámetro λ, es decir, $\mu_k, \Sigma_k, \omega_k$, a través de:

$$\hat{\omega_k}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} p(k|x_i, \lambda) \tag{12}$$

$$\hat{\mu_k}^{t+1} = \frac{\sum_{i=1}^{N} p(k|x_i, \lambda) x_i}{\sum_{i=1}^{N} p(k|x_i, \lambda)} \tag{13}$$

$$\hat{\Sigma_k}^{t+1} = \frac{\sum_{i=1}^{N} p\left(k|x_i, \lambda\right) \left(x_i - \hat{\mu}_k^{t+1}\right) \left(x_i - \hat{\mu}_k^{t+1}\right)^T}{\sum_{i=1}^{N} p\left(k|x_i, \lambda\right)} \tag{14}$$

Este paso puede ser considerado similar al paso de recálculo de clústers en el algoritmo k-means.

EM: PASO 4

Paso 4: Criterio de parada

Se evalúa la función de log-verosimilitud utilizando los valores estimados de λ:

$$\mathcal{L}(\lambda) = \sum_{i=1}^{N}\sum_{k=1}^{M} p\left(k|x_i,\lambda\right)\left(-\frac{1}{2}(x_i-\hat{\mu}_k)^T\Sigma_k^{-1}(x_i-\hat{\mu}_k) + \ \ln P(\hat{w}_k) + c_k\right) \tag{15}$$

La convergencia puede ser medida por la diferencia entre las evaluaciones de la función de verosimilitud que debería ser menor a la de un umbral determinado:

$$|\mathcal{L}(\lambda)_{t-1} - \mathcal{L}(\lambda)_t| \leq \epsilon \tag{16}$$

Otro mecanismo consiste en observar μ y detener la ejecución ante la ausencia de cambios o variaciones menores a un umbral.