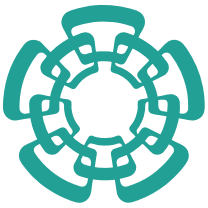

Centro de Investigación y de Estudios Avanzados del IPN
Laboratorio de Tecnologías de Información

Mezclas Gaussianas

Rafael Pérez Torres

Análisis de grupos basado en distribución
Tópicos Selectos en Reconocimiento de Patrones

Profesor Dr. Wilfrido Gómez Flores

28 de mayo de 2015

Resumen

Las técnicas de análisis de grupos basadas en distribución tienen como objetivo la creación de un modelo que describa a los mismos. Este modelo es creado a partir de propiedades y características descritas por los mismos datos, para posteriormente utilizarlo en las tareas de clasificación. En este documento se describe el modelo de mezclas gaussianas (GMM, Gaussian Mixture Model) como una técnica de agrupación basada en distribucón, así como el algoritmo EM (Expectation - Maximization) como una vía para ajustar los parámetros de un GMM determinado.

Índice

1. Introducción

Es posible que la naturaleza de los datos bajo análisis describa una distribución que sea de antemano conocida. Dicha distribución se encuentra caracterizada por una función de la que únicamente hace falta estimar los mejores valores para sus parámetros y así obtener un modelo optimizado de los datos.
La anterior característica es aprovechada por las técnicas de agrupamiento basadas en distribución, obteniendo dicho modelo probabilístico que permite describir de forma ajustada a los datos, para luego utilizarlo con fines de identificación - agrupación. En general, estas técnicas de agrupamiento asumen que los datos han sido generados por una mezcla de distribuciones, donde cada distribución distinta define a uno de los grupos.

2. El modelo de mezclas Gaussianas

La naturaleza de los datos puede reflejar características intrínsecas que permiten describirlos a través de un modelo con una estructura determinada. Este modelo es también conocido como una *PDF* (Probability Density Function, Función de densidad de probabilidad) que al igual que cualquier otra función matemática admite una serie de parámetros específicos.
Por ejemplo, si se pudiera analizar una cantidad de datos provenientes del mundo real que tienda a infinito, se observará que mantienen una distribución normal también conocida como gaussiana. La distribución gaussiana de un dato x se expresa como:

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

Donde $g(x|\mu_i,\Sigma_i)$ predice el valor que tomará una variable aleatoria desconocida a partir de parámetros conocidos (o fijos).

$$\mu_i = \frac{1}{n_i} \sum_{x\in\omega_i} x$$

y

$$\Sigma_i = \frac{1}{n_i - 1} \sum_{x\in\omega_i} (x-\mu_i)$$

indican los valores de media y covarianza de los datos para cada uno de los grupos.
Como puede observarse, la PDF Gaussiana es definida en términos de los parámetros μ y Σ.
El modelo de mezclas Gaussianas (GMM - Gaussian Mixture Model) es una suma ponderada de M PDFs Gaussianas, donde cada PDF se refiere a un grupo de datos (clúster) específico. Un ejemplo de esta combinación de PDFs es mostrado en la Figura 1.

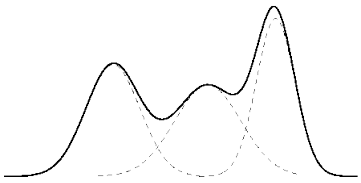

Figura 1: Una combinación de PDFs Gaussianas

Matemáticamente, un GMM es definido 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 $\sum_{i=1}^{M} \omega_i = 1$.
De lo anterior se desprende que el GMM es parametrizado por los vectores de medias, covarianzas y coeficientes de cada uno de los grupos, siendo expresado como:

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

Usando un número adecuado de PDF's Gaussianas y ajustando sus parámetros (μ_i y Σ_i) así como sus coeficientes (ω_i) en una combinación lineal, se obtiene una densidad muy aproximada a la distribución de los datos observados.

3. Estimación de los parámetros del GMM

3.1. El método de estimación de máxima verosimilitud

Existen varios métodos para realizar la estimación de los parámetros λ de un GMM, sin embargo 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.
Sea $\mathscr{X}$ una variable aleatoria con función de probabilidad $p(\mathscr{X};\theta)$ 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:

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

La función de verosimilitud es una función de los parámetros de un modelo estadístico que permite realizar inferencias acerca de su valor a partir de un conjunto de observaciones. El criterio de ML entonces busca elegir el valor adecuado de los parámetros θ que maximice a $\mathscr{L}$.

Por ejemplo, imaginar que se quiere estimar la probabilidad p de que salga cara en el lanzamiento de una moneda (no necesariamente regular). Inicialmente, se lanza cinco veces la moneda y se obtiene: $C+CC+$. Una forma de estimar p sería evaluar la probabilidad de obtener esta muestra para diferentes valores de p y elegir el valor que maximice dicha probabilidad. Esto puede ser calculado como: $p(C+CC+) = p*(1-p)*p*p*(1-p) = p^3(1-p)^2$ para todo valor real de P en el intervalo $[0,1]$. En la Tabla 1 se observan los valores obtenidos por distintos valores de p.

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 1: Valores obtenidos para distintos valores de p

El valor que obtiene la máxima probabilidad es 0,6. Suponiendo que se hayan efectuado n lanzamientos de la moneda obteniendo k caras sin importar el orden en el que han salido, la probabilidad de dicho suceso es dada por:

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

Si asumimos también que los valores de n y k son conocidos, entonces la verosimilitud queda definida exclusivamente en términos del parámetro p. Como en el caso de cualquier función matemática, la función de verosimilitud puede ser maximizada utilizando las técnicas de cálculo: derivando e igualando a cero y resolviendo las ecuaciones resultantes.

Para el ejemplo los cálculos se simplifican si se representa a $\mathscr{L}$ en términos de logaritmos. Así a través de la sucesión de ecuaciones 5, 6, 7 y 8 es posible indicar que la frecuencia relativa (ecuación 8) es el *estimador máximo verosimil* de la probabilidad de un determinado suceso. A la serie de pasos efectuados se le conoce como el método de la máxima verosimilitud.

$$\ln\big(\mathscr{L}(p)\big) = k\ln(p) + (n-k)\ln(1-p) \tag{5}$$

$$\frac{\partial \ln\big(\mathscr{L}(p)\big)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} \tag{6}$$

$$\frac{k}{\hat{p}} - \frac{n-k}{1-\hat{p}} = 0 \quad \rightarrow \quad k(1-\hat{p}) - \hat{p}(n-k) = 0 \tag{7}$$

$$\rightarrow \quad \hat{p} = \frac{k}{n} \tag{8}$$

Al aplicar este método a muestras que observan una distribución Gaussiana (ecuaciones 9, 10, 11), se observa que los estimadores máximos verosímiles coinciden con la media (ecuación 12) y la covarianza (ecuación 13).

$$\begin{aligned}
\mathscr{L}(\mu,\sigma^2) &= \ln \prod_{i=1}^{N} p(x_i;\theta) && (9)\\
&= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) && (10)\\
&= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x-\mu)^2 && (11)
\end{aligned}$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{12}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 \tag{13}$$

3.2. El algoritmo EM

El algoritmo EM (Expectation-Maximization) se basa en el método de estimación de máxima verosimilitud para determinar los valores *óptimos* de los parámetros (λ) de un GMM.

Para ello se deben calcular las derivadas de los parámetros de la función de log-verosimilitud $\big(\lambda = \{\mu_k, \Sigma_k, \omega_k\}\big)$:

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

Este algoritmo define cuatro pasos básicos:

1. Selección de valores iniciales para el parámetro $\big(\lambda = \{\mu_k, \Sigma_k, \omega_k\}\big)$.
2. **Paso Expectation (E)**: Evaluar las probabilidades *a posteriori* utilizando los valores actuales de λ .
3. **Paso Maximization (M)**: Reestimar los valores de λ utilizado las probabilidades *a posteriori* obtenidas.

4. Evaluar un criterio de parada basado en los valores de las medias o bien en la función de verosimilitud utilizando los nuevos valores del parámetro y compararlos con la iteración anterior.

3.2.1. Paso 1

El algoritmo EM requiere la cantidad de M componentes Gaussianas que se desea generar en el GMM. Esto puede equipararse con la cantidad de grupos que se desea reconocer en el agrupamiento.
Además, requiere el establecimiento de valores iniciales $(\mu_0, \Sigma_0, \omega_0)$ en los parámetros. Para dicha tarea es posible utilizar un algoritmo de agrupamiento como el *k-means*. Para el caso de los coeficientes ω_k sus valores son calculados a través de $\omega_k = N_k/N$ donde N_k es la cantidad de *instancias* del grupo y N es la cantidad total de instancias.

3.2.2. Paso 2: E

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:

$$p(k|x_i, \lambda) = \frac{w_k g(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{M} w_k g(x_i|\mu_k, \Sigma_k)} \tag{14}$$

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.

3.2.3. Paso 3: M

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

$$\hat{\omega_k}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} p(k|x_i, \lambda) \tag{15}$$

$$\hat{\mu_k}^{t+1} = \frac{\sum_{i=1}^{N} p(k|x_i, \lambda) x_i}{\sum_{i=1}^{N} p(k|x_i, \lambda)} \tag{16}$$

$$\hat{\Sigma_k}^{t+1} = \frac{\sum_{i=1}^{N} p\left(k|x_i, \lambda\right)\left(x_i - \hat{\mu}_k^{t+1}\right)\left(x_i - \hat{\mu}_k^{t+1}\right)^T}{\sum_{i=1}^{N} p\left(k|x_i, \lambda\right)} \tag{17}$$

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

3.2.4. Paso 4

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

$$\mathscr{L}(\lambda) = \sum_{i=1}^{N} \sum_{k=1}^{M} p(k|x_i, \lambda) \left(-\frac{1}{2}(x_i - \hat{\mu}_k)^T \Sigma_k^{-1} (x_i - \hat{\mu}_k) + \ln P(\hat{w}_k) + c_k \right) \tag{18}$$

donde

$$c_k = -\frac{l}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma_k| \tag{19}$$

Esta evaluación es comparada con la de la iteración anterior y se finaliza la ejecución con los valores actuales si se observa convergencia en la solución.
La convergencia puede ser definida en términos de la diferencia entre las evaluaciones de la función de verosimilitud que debería ser menor a la de un umbral determinado:

$$|\mathscr{L}(\lambda)_{t-1} - \mathscr{L}(\lambda)_t| \leq \epsilon \tag{20}$$

Otro mecanismo consiste en observar a los cambios existentes en los valores de μ, deteniendo la ejecución ante la ausencia de cambios o variaciones menores a un umbral.

4. Conclusiones

Este documento ha presentado una introducción a las técnicas de agrupación basadas en distribución. Se ha descrito el funcionamiento del GMM como un exponente de esta categoría de técnicas, así como el algoritmo EM como mecanismo para ajustar los valores de un GMM para así encontrar la pertenencia de cada dato a los clústers participantes en el proceso de agrupación.

Referencias

[Gan u. a. 2007] GAN, Guojun ; MA, Chaoqun ; WU, Jianhong: *Data clustering: theory, algorithms, and applications*. Bd. 20. Siam, 2007

[Gomez Flores 2015] GOMEZ FLORES, Wilfrido: *Diapositivas de clase Clasificación Bayesiana III: Mezcla de Gaussianas*. 2015

[McCormick 2014] MCCORMICK, Chris: *Gaussian Mixture Models Tutorial and MATLAB Code*. https://chrisjmccormick.wordpress.com/2014/08/04/gaussian-mixture-models-tutorial-and-matlab-code/. Version: August 2014

[Reynolds 2007] REYNOLDS, Douglas: Gaussian Mixture Models / MIT Lincoln Laboratory. 2007. – Forschungsbericht

[Theodoridis u. Koutroumbas 2009] THEODORIDIS, Sergios ; KOUTROUMBAS, Konstantinos ; THEODORIDIS, Sergios (Hrsg.) ; KOUTROUMBAS, Konstantinos (Hrsg.): *Pattern Recognition*. Fourth Edition. Boston : Academic Press, 2009. – 1 – 12 S. http://dx.doi.org/http://dx.doi.org/10.1016/B978-1-59749-272-0.50003-7. http://dx.doi.org/http://dx.doi.org/10.1016/B978-1-59749-272-0.50003-7. – ISBN 978–1–59749–272–0