9. Simplifcació de la barreja de Gaussianes 1 [G]

Josep de Cid Albert Ribes Kerstin Winter

18 d'octubre de 2017

Considereu el model de barreja de Gaussianes:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)$$

Preneu el cas que totes les matrius de covariança són iguals i diagonals, és a dir, $\Sigma_1 = \ldots, \Sigma_K = \Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$

1. **Enraoneu en quin sentit representa una simplifcació respecte al cas general (amb matrius de covariança generals), des dels punts de vista estadístic i geomètric.**

   Significa que las dimensiones son independientes entre ellas.

   Geométricamente esto significa que cada uno de los clusters generará instancias formando una probabilidad con forma elipsoide alargada en la cual el eje mayor es paralelo a alguno de los ejes del sistema de coordenadas.

2. **Expresseu la funció de densitat de probabilitat** $\mathcal{N}(x; \mu_k, \Sigma_k)$ **que en resulta.**

   En el caso genérico la función de densidad de la distribución gaussiana multivariada con D variables es:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{(\frac{D}{2})}} \cdot \frac{1}{(|\Sigma|)^{\frac{1}{2}}} exp\Bigg(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \Bigg)$$

   Pero se puede simplificar con las siguientes propiedades:

   - $det(diag(\alpha_1, \alpha_2, \ldots, \alpha_n)) = \prod_{i=1}^{n} \alpha_i$
   - $diag(\alpha_1, \alpha_2, \ldots, \alpha_n)^{-1} = diag(\alpha_1^{-1}, \alpha_2^{-1}, \ldots, \alpha_n^{-1})$
   - $(\alpha_1, \alpha_2, \ldots, \alpha_n)^T \cdot diag(\beta_1, \beta_2, \ldots, beta_n) = (\alpha_1\beta_1, \alpha_2\beta_2, \ldots, \alpha_n\beta_n)^T$

   Y entonces la fórmula anterior queda:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{(\frac{D}{2})}} \cdot \frac{1}{(\prod_{i=1}^{n} \sigma_i)} exp\Bigg(-\frac{1}{2} \sum_{i=1}^{d} \Big(\frac{(x_i - \mu_i)^2}{\sigma_i^2} \Big) \Bigg)$$

3. **Construïu la funció de log-versemblança negativa.**

   En el caso de que se disponga de M datos $D = \{x_1, x_2, \ldots, x_M\}$ la función log-verosimilitud negativa es:

$$l(\pi, \mu, \Sigma; D) = -\ln \prod_{j=1}^{M} p(x_j)$$

$$l(\pi, \mu, \Sigma; D) = -\sum_{j=1}^{M} \ln(p(x_j))$$

$$l(\pi, \mu, \Sigma; D) = -\sum_{j=1}^{M} \left[\ln \Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_j; \mu_k, \Sigma_k) \Big) \right]$$

4. **Deriveu les equacions de l'algorisme E-M que en resulta i escriviu l'algorisme de clustering complet.**

   Antes de seguir avanzando vamos a definir la función $\gamma_k(x)$, que es la probabilidad de que sea la distribución k la que ha generado el dato x. Formalmente:

$$\gamma_k(x) = p(z_k = 1|x)$$

$$\gamma_k(x) = \frac{p(x_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)}$$

$$\gamma_k(x) = \frac{\pi_k \mathcal{N}(x; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x; \mu_j, \Sigma_j)}$$

   Esta función nos será útil más adelante.

   Para encontrar los extremos vamos a tener que derivar respecto a los valores de π, μ y Σ, y además debe cumplirse la condición de que $\sum_{k=1}^{K} \pi_k = 1$. Para resolver este problema usaremos los multiplicadores de Lagrange.

   La función de Lagrange es:

$$\mathcal{L}(\pi, \mu, \Sigma, \lambda) = l(\pi, \mu, \Sigma; D) - \lambda \cdot \Big[(\sum_{k=1}^{K} \pi_k) - 1 \Big]$$

$$\mathcal{L}(\pi, \mu, \Sigma, \lambda) = -\sum_{j=1}^{M} \left[\ln \Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_j; \mu_k, \Sigma_k) \Big) \right] - \lambda \cdot \Big[(\sum_{k=1}^{K} \pi_k) - 1 \Big]$$

   Como más adelante hará falta, calculamos previamente algunas derivadas intermedias:

$$\frac{\partial \mathcal{N}(x;\mu,\Sigma)}{\partial \mu_q} = \mathcal{N}(x;\mu,\Sigma)\cdot\Big[-\frac{1}{2}\cdot\frac{2(x_q-\mu_q)}{\sigma_q^2}(-1)\Big]$$

$$\frac{\partial \mathcal{N}(x;\mu,\Sigma)}{\partial \mu_q} = \mathcal{N}(x;\mu,\Sigma)\frac{(x_q-\mu_q)}{\sigma_q^2} \tag{1}$$

Hay que derivar respecto a los valores de π, μ y Σ y λ e igualar a 0.

$$\frac{\partial \mathcal{L}(\pi,\mu,\Sigma,\lambda)}{\partial \pi_q} = \frac{\partial l(\pi,\mu,\Sigma;D)}{\partial \pi_q} - \frac{\partial \lambda\Big[\Big(\sum_{k=1}^{K}\pi_k\Big)-1\Big]}{\partial \pi_q}$$

Y si lo hacemos por partes:

$$\frac{\partial l(\pi,\mu,\Sigma;D)}{\partial \pi_q} = -\sum_{j=1}^{M}\left[\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}\cdot\frac{\partial\Big(\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)\Big)}{\partial \pi_q}\right]$$

$$\frac{\partial l}{\partial \pi_q} = -\sum_{j=1}^{M}\left[\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}\cdot\mathcal{N}(x_j,\mu_q,\Sigma_q)\right]$$

$$\frac{\partial l}{\partial \pi_q} = -\sum_{j=1}^{M}\left[\frac{\mathcal{N}(x_j,\mu_q,\Sigma_q)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}\right]$$

Y esto se puede expresar usando la función γ_k que hemos definido previamente:

$$\frac{\partial l(\pi,\mu,\Sigma;D)}{\partial \pi_q} = -\sum_{j=1}^{M}\left[\frac{\gamma_q(x_j)}{\pi_q}\right]$$

Y

$$\frac{\partial \lambda\Big[\Big(\sum_{k=1}^{K}\pi_k\Big)-1\Big]}{\partial \pi_q} = \lambda$$

Y por lo tanto:

$$\frac{\partial \mathcal{L}(\pi,\mu,\Sigma,\lambda)}{\partial \pi_q} = -\sum_{j=1}^{M}\left[\frac{\gamma_q(x_j)}{\pi_q}\right] - \lambda$$

La derivada respecto a cada una de las μ_q es:

$$l(\pi,\mu,\Sigma;D) = -\sum_{j=1}^{M}\left[\ln\Big(\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)\Big)\right]$$

$$\frac{\partial l}{\partial \mu_q} = -\sum_{j=1}^{M}\left[\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}\cdot\Big(\sum_{k=1}^{K}\pi_k\frac{\partial\mathcal{N}(x_j;\mu_k,\Sigma_k)}{\partial\mu_q}\Big)\right]$$

Y si usamos la ecuación 1 tenemos:

$$\frac{\partial l}{\partial \mu_q} = -\sum_{j=1}^{M}\left[\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}\cdot\Big(\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)\frac{(x_q-\mu_q)}{\sigma_q^2}\Big)\right]$$

$$\frac{\partial l}{\partial \mu_q} = -\sum_{j=1}^{M}\left[\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}\cdot\Big(\frac{x_q-\mu_q}{\sigma_q^2}\Big)\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)\right]$$

$$\frac{\partial l}{\partial \mu_q} = -\sum_{j=1}^{M}\left[\frac{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)}\cdot\Big(\frac{x_q-\mu_q}{\sigma_q^2}\Big)\right]$$

Tengo que corregir a partir de aquí

$$\frac{\partial l}{\partial \mu_q} = -\sum_{j=1}^{M}\left[\frac{1}{p(x_j)}\cdot\Big(\frac{x_q-\mu_q}{\sigma_q^2}\Big)p(x_j)\right]$$

$$\frac{\partial l}{\partial \mu_q} = -\sum_{j=1}^{M}\frac{x_q-\mu_q}{\sigma_q^2}$$

$$\frac{\partial l}{\partial \mu_q} = \sum_{j=1}^{M}\frac{\mu_q-x_q}{\sigma_q^2}$$

$$\frac{\partial l}{\partial \mu_q} = \frac{M(\mu_q-x_q)}{\sigma_q^2}$$

$$\frac{\partial l}{\partial \mu_q} = \frac{M\mu_q}{\sigma_q^2} - \frac{Mx_q}{\sigma_q^2}$$

La derivada respecto a σ_q^2 es:

$$l(\pi,\mu,\Sigma;D) = -\sum_{j=1}^{M}\left[\ln\Big(\sum_{k=1}^{K}\pi_k\mathcal{N}(x_j;\mu_k,\Sigma_k)\Big)\right]$$

5. **Enraoneu sobre les implicacions (possibles avantatges/inconvenients) que representa la simplifcació respecte el cas general des del punt de vista del clustering.**