TU München
Fakultät für Informatik
PD Dr. Rudolph Triebel
John Chiotellis

Machine Learning for Computer Vision
Winter term 2016

November 25, 2016
K-Means, Expectation-Maximization, Mixture Models

Exercise 1: Expectation-Maximization for GMM

In the standard EM algorithm, we first define the responsibilities γ as

$$\gamma_{nk} = p(z_{nk} = 1|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \quad , z_{nk} \in \{0, 1\}, \sum_{k=1}^{K} z_{nk} = 1$$

a) Find the optimal means, covariances and mixing coefficients that maximize the data likelihood. How can we interpret the results?

We want to maximize the data likelihood, so as usual we minimize the negative log-likelihood:

$$-\mathcal{LL} = -\log p(X|\mu, \mathbf{\Sigma}, \pi) = -\log \prod_n \sum_k \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \tag{1}$$

This time we minimize 3 times independently with respect to the means, the covariances and the mixture coefficients:

$$\mu_k^* = \arg\min_{\mu_k} -\mathcal{LL} \tag{2}$$
$$\Sigma_k^* = \arg\min_{\Sigma_k} -\mathcal{LL} \tag{3}$$
$$\pi_k^* = \arg\min_{\pi_k} -\mathcal{LL} \tag{4}$$

In the following, to avoid confusion of sums and covariances, we denote covariance Σ_k as C_k. To simplify some expressions, let us agree on the following notation:

$$\mathcal{N}_{nk} \equiv \mathcal{N}(x_n|\mu_k, C_k) \tag{5}$$
$$f_k \equiv ((2\pi)^d |C_k|)^{-1/2} \tag{6}$$
$$D_{nk} \equiv (x_n - \mu_k)^T C_k^{-1} (x_n - \mu_k) \tag{7}$$

Thus, we have:

$$\begin{aligned} -\mathcal{L}\mathcal{L} &= -\sum_n \log \sum_k \pi_k \mathcal{N}_{nk} \\ &= -\sum_n \log \sum_k \pi_k f_k \exp(-\frac{1}{2} D_{nk}) \end{aligned}$$

Solving for the means:

$$\begin{aligned} \frac{\partial \mathcal{L}\mathcal{L}}{\partial \mu_k} &= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \frac{\partial \sum_k \pi_k f_k \exp(-\frac{1}{2} D_{nk})}{\partial \mu_k} \tag{8} \\ &= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k f_k \frac{\partial \exp(-\frac{1}{2} D_{nk})}{\partial \mu_k} \tag{9} \\ &= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k f_k \exp(-\frac{1}{2} D_{nk}) C_k^{-1} (x_n - \mu_k) \tag{10} \\ &= \sum_n \frac{\pi_k \mathcal{N}_{nk}}{\sum_j \pi_j \mathcal{N}_{nj}} C_k^{-1} (x_n - \mu_k) \tag{11} \\ &= \sum_n \gamma_{nk} C_k^{-1} (x_n - \mu_k) \tag{12} \\ & \tag{13} \end{aligned}$$

Setting $-\frac{\partial \mathcal{L}\mathcal{L}}{\partial \mu_k} \overset{!}{=} 0$ gives us:

$$\begin{aligned} \sum_n \gamma_{nk} C_k^{-1} \mu_k &= \sum_n \gamma_{nk} C_k^{-1} x_n \tag{14} \\ C_k^{-1} \mu_k \sum_n \gamma_{nk} &= C_k^{-1} \sum_n \gamma_{nk} x_n \tag{15} \\ C_k^{-1} \mu_k \sum_n \gamma_{nk} &= C_k^{-1} \sum_n \gamma_{nk} x_n \tag{16} \\ \mu_k \sum_n \gamma_{nk} &= \sum_n \gamma_{nk} x_n \tag{17} \\ \mu_k &= \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}} \tag{18} \end{aligned}$$

Solving for the covariances:

$$\begin{aligned}
\frac{\partial \mathcal{L}\mathcal{L}}{\partial C_k} &= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \frac{\partial \sum_k \pi_k f_k \exp(-\frac{1}{2}D_{nk})}{\partial C_k} && (19)\\
&= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k \frac{\partial f_k \exp(-\frac{1}{2}D_{nk})}{\partial C_k} && (20)\\
&= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k \left(\frac{\partial f_k}{\partial C_k} \exp(-\frac{1}{2}D_{nk}) + f_k \frac{\partial \exp(-\frac{1}{2}D_{nk})}{\partial C_k} \right) && (21)\\
&= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k \left((-\frac{1}{2} f_k C_k^{-1}) \exp(-\frac{1}{2}D_{nk}) + \frac{1}{2} f_k \exp(-\frac{1}{2}D_{nk}) C_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \right) \\
& && (22)\\
&= (-\frac{1}{2}) \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \pi_k f_k \exp(-\frac{1}{2}D_{nk}) \left(C_k^{-1} - C_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \right) \\
& && (23)\\
&= (-\frac{1}{2}) \sum_n \gamma_{nk} \left(C_k^{-1} - C_k^{-1}(x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} \right) && (24)\\
& && (25)
\end{aligned}$$

Here, we used the derivative of the determinant as follows:

$$\begin{aligned}
\frac{\partial f_k}{\partial C_k} &= \frac{\partial ((2\pi)^d |C_k|)^{-\frac{1}{2}}}{\partial C_k} = ((2\pi)^d)^{-\frac{1}{2}} \frac{\partial (|C_k|)^{-\frac{1}{2}}}{\partial C_k} && (26)\\
&= ((2\pi)^d)^{-\frac{1}{2}} (-\frac{1}{2}) |C_k|^{-\frac{3}{2}} \frac{\partial (|C_k|)}{\partial C_k} = ((2\pi)^d)^{-\frac{1}{2}} (-\frac{1}{2}) |C_k|^{-\frac{3}{2}} |C_k| (C_k^{-1})^T && (27)\\
&= (-\frac{1}{2}) ((2\pi)^d)^{-\frac{1}{2}} |C_k|^{-\frac{1}{2}} C_k^{-1} = -\frac{1}{2} f_k C_k^{-1} && (28)
\end{aligned}$$

and the derivative of the Mahalanobis distance as:

$$\frac{\partial x^T C^{-1} x}{\partial C} = -C^{-T} x x^T C^{-T} = -C^{-1} x x^T C^{-1} \qquad (29)$$

Setting $-\frac{\partial \mathcal{L}\mathcal{L}}{\partial C_k} \overset{!}{=} 0$ gives us:

$$\begin{aligned}
\sum_n \gamma_{nk} C_k^{-1} &= \sum_n \gamma_{nk} C_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} && (30)\\
C_k^{-1} \sum_n \gamma_{nk} &= C_k^{-1} \sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} && (31)\\
\sum_n \gamma_{nk} &= \sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T C_k^{-1} && (32)\\
C_k &= \frac{\sum_n \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_n \gamma_{nk}} && (33)
\end{aligned}$$

Solving for the mixture coefficients: Here we must take into account that $\sum_k \pi_k = 1$. We enforce this constraint with a Lagrange multiplier. Our objective then becomes:

$$\mathcal{LL}' = \mathcal{LL} + \lambda(\sum_k \pi_k - 1) \tag{34}$$

where $\lambda < 0$.

Deriving w.r.t. π_k, we get

$$\begin{aligned}
\frac{\partial \mathcal{LL}'}{\partial \pi_k} &= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \frac{\partial \sum_k \pi_k \mathcal{N}_{nk}}{\partial \pi_k} + \lambda && (35)\\
&= \sum_n \frac{1}{\sum_j \pi_j \mathcal{N}_{nj}} \mathcal{N}_{nk} + \lambda && (36)\\
&= \sum_n \frac{\gamma_{nk}}{\pi_k} + \lambda && (37)
\end{aligned}$$

Setting equal to zero and solving for λ, we get

$$\begin{aligned}
\lambda &= -\sum_n \frac{\gamma_{nk}}{\pi_k} && (38)\\
\lambda \pi_k &= -\sum_n \gamma_{nk} && (39)\\
\sum_k \lambda \pi_k &= -\sum_k \sum_n \gamma_{nk} && (40)\\
\lambda &= -N && (41)
\end{aligned}$$

Now we can plug this back to the objective and actually solve for π_k:

$$\begin{aligned}
\frac{\partial \mathcal{LL}'}{\partial \pi_k} &= \sum_n \frac{\gamma_{nk}}{\pi_k} - N \overset{!}{=} 0 && (42)\\
\frac{1}{\pi_k} \sum_n \gamma_{nk} &= N && (43)\\
\pi_k = \frac{\sum_n \gamma_{nk}}{N} &= \frac{N_k}{N} && (44)
\end{aligned}$$

We can interpret these results as weighted averages of means and covariances, the weights corresponding to the responsibilities γ_{nk}. The mixture coefficients π_k are simply the ratio of data points explained by each component.

b) Define the complete-data-log-likelihood. What is the difference to the standard log-likelihood?

Assuming we observe not only the data but also the binary latent variables Z we define the complete data likelihood as:

$$p(X, Z|\pi, \mu, C) = \prod_n p(z_n|\pi) p(x_n|z_n, \mu, C) \tag{45}$$

where $\quad p(z_n|\pi) = \prod_k \pi_k^{z_{nk}} \quad$ and $\quad p(x_n|z_n, \mu, C) = \prod_k \mathcal{N}(x_n|\mu_k, C_k)^{z_{nk}}$.
Remember that $\sum_k z_{nk} = 1$.

Since now we only have products, we can more easily compute the logarithm:

$$\log p(X, Z|\pi, \mu, C) = \sum_n \sum_k z_{nk} (\log \pi_k + \log \mathcal{N}(x_n|\mu_k, C_k)) \tag{46}$$

Of course in practice, the latent variables are not known, so we maximize the *expectation*:

$$\mathbb{E}[\log p(X, Z|\pi, \mu, C)] = \sum_n \sum_k \mathbb{E}[z_{nk}] (\log \pi_k + \log \mathcal{N}(x_n|\mu_k, C_k)) \tag{47}$$

where we know that $\mathbb{E}[z_{nk}] = \gamma_{nk}$.

The theory says that the log-marginal is also maximized implicitly!

Exercise 2: K-Means Compression and EM for GMM

See code.