EM algorithm for Gaussian mixtures

Kitov Victor

1 Lemmas

Lemma 1. *For any square non-degenerate matrix* B $\frac{\partial |B|}{\partial B} = |B| B^{-1}$

Proof. Row expansion of determinant:

$$|B| = \sum_{j=1}^{D} b_{ij} A_{ij}$$

where A_{ij} is algebraic complement (also known as cofactor) of B for b_{ij}. So

$$\frac{\partial |B|}{\partial b_{ij}} = A_{ij}$$

Combining this with the fact that $B^{-1} = \frac{1}{|B|} A^T$, where $\{A\}_{ij} = A_{ij}$, obtain that $\frac{\partial |B|}{\partial B} = |B| B^{-T}$. □

Lemma 2. *For any square symmetric non-degenerate matrix* B $\frac{\partial \ln |B|}{\partial B} = B^{-1}$

Proof. $\frac{\partial \ln |B|}{\partial B} = \frac{1}{|B|} |B| B^{-T} = B^{-T} = B^{-1}$ since B is symmetric. □

Lemma 3. *For any square matrix* U $\frac{\partial \left\{a^T U b\right\}}{\partial U} = ab^T$.

Proof.

$$a^T U b = \sum_{i,j} a_i u_{ij} b_j$$

$$\frac{\partial \left\{a^T U b\right\}}{\partial u_{ij}} = a_i b_j$$

$$\frac{\partial \left\{a^T U b\right\}}{\partial U} = ab^T$$

□

2 Definitions

Training set $x_1, x_2, ...x_N$.
If component is known

$$p(x|z, \theta) = N(x|\mu_z, \Sigma_z) = \frac{1}{(2\pi)^{D/2} |\Sigma_z|^{1/2}} e^{-\frac{1}{2}(x-\mu_z)^T \Sigma_z^{-1} (x-\mu_z)} \tag{1}$$

Priors

$$p(z|\theta) = \phi_z \tag{2}$$

Each object is modelled with mixture of Z Gaussians with parameters $\{\mu_z, \Sigma_z\}$ and priors ϕ_z, $z = 1, 2, ...Z$.

$$p(x|\theta) = \sum_{z=1}^{Z} \phi_z N(x|\mu_z, \Sigma_z) = \sum_{z=1}^{Z} \phi_z \frac{1}{(2\pi)^{D/2} |\Sigma_z|^{1/2}} e^{-\frac{1}{2}(x-\mu_z)^T \Sigma_z^{-1} (x-\mu_z)} \tag{3}$$

Parameters of mixture: $\theta = \{\phi_z, \mu_z, \Sigma_z\}_{z=1}^{Z}$.

3 Expectation step

$$q_n(z) = w_{nz} = P(z|x_n) = \frac{P(z,x_n)}{P(x_n)} = \frac{P(z,x_n)}{\sum_k P(k,x_n)} = \frac{P(z)P(x_n|z)}{\sum_k P(k)P(x_n|k)} = \frac{\widehat{\phi}_z N(x_n|\widehat{\mu},\widehat{\Sigma}_z)}{\sum_k \widehat{\phi}_k N(x_n|\widehat{\mu}_k,\widehat{\Sigma}_k)}$$

4 Maximization step

4.1 Optimization task

Given, that $\phi_z = p(z)$, be need to ensure

$$\sum_{z=1}^{Z} \phi_z = 1 \tag{4}$$

Given (4) we need to solve

$$\sum_{n=1}^{N}\sum_{z=1}^{Z} w_{nz} \ln \frac{p(x_n,z|\theta)}{w_{nz}} \to \max_\theta$$

This is equivalent to

$$\sum_{n=1}^{N}\sum_{z=1}^{Z} w_{nz} \ln p(x_n,z|\theta) = \sum_{n=1}^{N}\sum_{z=1}^{Z} w_{nz} \ln p(z|\theta)p(x_n|z,\theta) \to \max_\theta \tag{5}$$

Substituting (2) and (1) into (5) and using (4) we obtain final optimization task for the M-step:

$$\begin{cases} \sum\limits_{n=1}^{N}\sum\limits_{z=1}^{Z} w_{nz} \ln \phi_z + \sum\limits_{n=1}^{N}\sum\limits_{z=1}^{Z} w_{nz} \ln(2\pi)^{-D/2} + \frac{1}{2}\sum\limits_{n=1}^{N}\sum\limits_{z=1}^{Z} w_{nz} \ln\left|\Sigma_z^{-1}\right| \\ \qquad\qquad - \frac{1}{2}\sum\limits_{n=1}^{N}\sum\limits_{z=1}^{Z} w_{nz} \left(x_n-\mu_z\right)^T \Sigma_z^{-1} \left(x_n-\mu_z\right) \to \max\limits_\theta \\ \sum_{z=1}^{Z} \phi_z = 1 \end{cases} \tag{6}$$

4.2 Optimization for ϕ_z.

Optimization for ϕ_z leads to Lagrangian

$$\mathcal{L} = \sum_{n=1}^{N}\sum_{z=1}^{Z} w_{nz} \ln \phi_z + \lambda(1 - \sum_{z=1}^{Z} \phi_z) \to \text{extr}$$

Optimal values satisfy

$$\frac{\partial \mathcal{L}}{\partial \phi_z} = \sum_{n=1}^{N} w_{nz} \frac{1}{\phi_z} - \lambda = 0$$

$$\phi_z = \frac{1}{\lambda}\sum_{n=1}^{N} w_{nz}$$

Using (4) we find λ:

$$1 = \sum_z \phi_z = \frac{1}{\lambda}\sum_n\sum_z w_{nz}$$

$$\lambda = \sum_n\sum_z w_{nz} = \sum_{n=1}^{N} 1 = N$$

It follows that

$$\phi_z = \frac{1}{N}\sum_{n=1}^{N} w_{nz}$$

Interpretation: ϕ_z is sample average of component correspondences, when each object may partly belong to each component.

4.3 Optimization for μ_z.

Optimization for μ_z leads to

$$\mathcal{L} = \sum_{n=1}^{N}\sum_{z=1}^{Z} w_{nz}\left(x_n - \mu_z\right)^T \Sigma_z^{-1}\left(x_n - \mu_z\right) \to \max_{\mu_z}$$

$$\frac{\partial \mathcal{L}}{\partial \mu_z} = 2\sum_n w_{nz}\Sigma_z^{-1}(x_n - \mu_z)(-1) = 0$$

$$\sum_n w_{nz}\Sigma_z^{-1}(x_n - \mu_z) = \Sigma_z^{-1}\sum_n \{w_{nz}(x_n - \mu_z)\} = 0$$

$$\sum_n \{w_{nz}(x_n - \mu_z)\} = \sum_n w_{nz}x_n - \sum_n w_{nz}\mu_z = 0$$

$$\mu_z = \frac{\sum_n w_{nz}x_n}{\sum_n w_{nz}}$$

Interpretation: μ_z equals to weighted average of objects with weights=correspondences to corresponding component.

4.4 Optimization for Σ

From (6) we obtain the following optimization problem to find covariance matrices:

$$\sum_{n=1}^{N}\sum_{z=1}^{Z} w_{nz}\ln\left|\Sigma_z^{-1}\right| - \sum_{n=1}^{N}\sum_{z=1}^{Z} w_{nz}\left(x_n - \mu_z\right)^T \Sigma_z^{-1}\left(x_n - \mu_z\right) \to \max_{\Sigma_1,\ldots\Sigma_Z} \tag{7}$$

4.4.1 Unconstrained Σ_z.

For each individual Σ_z we get

$$\mathcal{L} = \sum_{n=1}^{N} w_{nz}\ln\left|\Sigma_z^{-1}\right| - \sum_{n=1}^{N} w_{nz}\left(x_n - \mu_z\right)^T \Sigma_z^{-1}\left(x_n - \mu_z\right) \to \max_{\Sigma_z} \tag{8}$$

Using lemmas (3), (2)

$$\frac{\partial \mathcal{L}}{\partial \Sigma_z^{-1}} = \sum_{n=1}^{N}\{w_{nz}\Sigma_z\} - \sum_{n=1}^{N} w_{nz}(x_n - \mu_z)(x_n - \mu_z)^T = 0$$

$$\Sigma_z = \frac{\sum_{n=1}^{N} w_{nz}(x_n - \mu_z)(x_n - \mu_z)^T}{\sum_{n=1}^{N} w_{nz}}$$

Interpretation: Σ_z equals to weighted sample covariance matrix with weights=correspondences of objects to corresponding component.

4.4.2 Constraint: common covariance $\Sigma_z = \Sigma\,\forall z$.

4.4.3 Constraint: diagonal covariance $\Sigma_z = \mathrm{diag}\{\sigma_{z,1}^2, \ldots\sigma_{z,D}^2\}\,\forall z$.

4.4.4 Constraint: spherical covariance $\Sigma_z = \sigma_z^2 I\,\forall z$.