2413, Machine Learning, Tutorial 9
Universität Bern

Simon Jenni (jenni@inf.unibe.ch)

Expectation Maximization (EM-Algorithm)

1. Explain the differences between the Mixtures of Gaussian model (MoG) and the Gaussian Discriminant Analysis model (GDA).

   Solution.

   GDA is a supervised generative model in which we assume $p(x|y_c) \sim \mathcal{N}(\mu_c, \Sigma_c)$. Labels (i.e., the assignment of each training example $x^{(i)}$ to the corresponding Gaussian) are given for the training set.

   The MoG on the other hand is an unsupervised model in which we assume that each data point is sampled from a Gaussian distribution. The assignment of each $x^{(i)}$ to one of the Gaussians is unknown in this case (i.e., we treat it as a latent variable) and has to be learnt.

2. Derive the update rule for Σ_l in the Maximization step (M-step) of the EM algorithm for the Mixture of Gaussian model.

Solution.

We need to calculate the gradient of the $J(Q, \theta)$ with respect to Σ_l and set it to zero.

$$J(Q,\theta) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma_j|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\Big)}{w_j^{(i)}}$$

Then we have:

$$\begin{aligned}\nabla_{\Sigma_l} J(Q,\theta) &= -\nabla_{\Sigma_l}\Bigg[\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j) + \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\frac{1}{2}\log|\Sigma_j|\Bigg] \\ &= -\nabla_{\Sigma_l}\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\Big[(x^{(i)}-\mu_l)^T\Sigma_l^{-1}(x^{(i)}-\mu_l) + \log|\Sigma_l|\Big]\end{aligned}$$

We set $\Lambda_l = \Sigma_l^{-1}$ and solve for Λ_l.

$$=\nabla_{\Lambda_l}\Big[-\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}(x^{(i)}-\mu_l)^T\Lambda_l(x^{(i)}-\mu_l) + \frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\log|\Lambda_l|\Big] \tag{1}$$

$$=\nabla_{\Lambda_l}\Big[-\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)} tr\big((x^{(i)}-\mu_l)^T\Lambda_l(x^{(i)}-\mu_l)\big) + \frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\log|\Lambda_l|\Big] \tag{2}$$

$$=\nabla_{\Lambda_l}\Big[-\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)} tr\big((x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T\Lambda_l\big) + \frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\log|\Lambda_l|\Big] \tag{3}$$

$$=-\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}(x^{(i)}-\mu_l)(x^{(i)}-\mu_l)^T + \frac{1}{2}\Lambda_l^{-T}\sum_{i=1}^{m} w_l^{(i)} \tag{4}$$

Where

- (2) follows form (1) by $a = tr(a), \quad \forall a \in \mathcal{R}$. Note that $(x^{(i)}-\mu_j)^T\Lambda_l(x^{(i)}-\mu_j) \in \mathcal{R}$.
- (3) follows from (2) by $tr(ABC) = tr(CAB) = tr(BCA)$.
- (4) follows from (3) by $\nabla_A tr(BA) = B^T$ and $\nabla_A \log|A| = A^{-T}$. Note that $((x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T)^T = (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T$.

By setting to zero we have:

$$\Lambda_l^{-T} = \Lambda_l^{-1} = \Sigma_l = \frac{\sum_{i=1}^{m} w_l^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{m} w_l^{(i)}}$$

Factor Analysis

3. Assume that $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ are sampled i.i.d. from a distribution described by the factor analysis model

$$z \sim \mathcal{N}(0, I) \tag{5}$$
$$\epsilon \sim \mathcal{N}(0, \Psi) \tag{6}$$
$$x = \mu + \Lambda z + \epsilon. \tag{7}$$

What is the optimal μ? Use Maximum-Likelihood estimation.

Solution.

The samples are drawn from the distribution $x \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$. The log-likelihood function according to the ML estimate is

$$\begin{aligned}
l(\mu) &= \log \prod_{i=1}^{m} \frac{\exp(-\frac{1}{2}(x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu))}{(2\pi)^{n/2} |\Lambda\Lambda^T + \Psi|^{1/2}} && (8) \\
&= \sum_{i=1}^{m} -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Lambda\Lambda^T + \Psi|) + && (9) \\
&\quad \sum_{i=1}^{m} -\frac{1}{2} (x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu) && (10)
\end{aligned}$$

Note that the negative log-likelihood is a convex quadratic function in μ, therefore we can find the optimal μ if we set the gradient to 0. The gradient of the log-likelihood w.r.t. μ is

$$\begin{aligned}
\nabla_\mu l(\mu) &= \nabla_\mu \sum_{i=1}^{m} -\frac{1}{2} (x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu) && (11) \\
&= \sum_{i=1}^{m} -(\Lambda\Lambda^T + \Psi)^{-1} \mu + (\Lambda\Lambda^T + \Psi)^{-1} x^{(i)}. && (12)
\end{aligned}$$

From here, the solution is not very surprisingly,

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}. \tag{13}$$