
Define the complete data for this problem to be $D = \{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^n$. Write out the complete-data (negative) log likelihood.

$$\mathcal{L}(\boldsymbol{\theta}, \{\mu_k, \Sigma_k\}_{k=1}^c) = -\ln p(D \,|\, \boldsymbol{\theta}, \{\mu_k, \Sigma_k\}_{k=1}^c).$$

- **Expectation Step** Our next step is to introduce a mathematical expression for $\mathbf{q}_i$, the posterior over the hidden topic variables $\mathbf{z}_i$ conditioned on the observed data $\mathbf{x}_i$ with fixed parameters, i.e $p(\mathbf{z}_i|\mathbf{x}_i; \boldsymbol{\theta}, \{\mu_k, \Sigma_k\}_{k=1}^c)$.
  - Write down and simplify the expression for $\mathbf{q}_i$.

  - Find an expression for $\boldsymbol{\theta}$ that maximizes this expected complete-data log likelihood. You may find it helpful to use Lagrange multipliers in order to force the constraint $\sum \theta_k = 1$. Why does this optimized $\boldsymbol{\theta}$ make intuitive sense?

    To maximize given a constraint, we will use the method of Lagrangian multipliers. As we are maximizing with respect to θ_k, we may also drop the terms (on the right) not including θ. Previously, in homework 2, we had z_{ik} as our indicator variable. However, we now treat z as a latent variable, and instead we have a "soft estimate" q variable.

$$\mathcal{L}(\theta_{\mathbf{k}}, \lambda) = \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik} \ln \theta_k + \lambda(\sum_{k=1}^{c} \theta_k - 1) \tag{1}$$

Take the partial derivative of $\mathcal{L}$ with respect to θ_k and set it equal to zero, and noting that

$$\sum_i q_{ik} = n_k \tag{2}$$

(see homework 2 problem 2.2 for more details)

$$0 = \sum_{i=1}^{n} \frac{q_{ik}}{\theta_k} - \lambda \tag{3}$$

$$\theta_k = \frac{n_k}{\lambda} \tag{4}$$

Take the partial derivative with respect to λ and set it equal to zero to get

$$0 = \sum_{k=1}^{c} \theta_k - 1 \tag{5}$$

$$\tag{6}$$

Now let's solve from lambda by combining **??** and **??**

$$0 = \sum_{k=1}^{c} \frac{n_k}{\lambda} - 1 \tag{7}$$

$$\lambda = n_k \tag{8}$$

Returning to **??** we now see that

$$\theta_k = \frac{n_k}{n} \tag{9}$$

(Note that since we are actually estimating θ_k, it would be clearer to write $\hat{\theta_k}$)

This solution for $\hat{\theta_k}$makes sense: the optimal (prior) probability of a given x_i belong to a class k is equal to the proportion of observations (we've estimated in this iteration) that come from class k.

– Apply a similar argument to find the value of the (μ_k, Σ_k)'s that maximizes the expected complete-data log likelihood. For $\boldsymbol{\mu}_k$ case.

$$= \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik} \left(\ln \theta_k + \ln \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{10}$$

To solve for optimal $\boldsymbol{\mu}_k$, taking the derivative of **??** with respect to $\boldsymbol{\mu}_k$ and set to zero. Note that the left hand terms drop out. Furthermore, we remove terms in the log gaussian without $\boldsymbol{\mu}_k$

$$= \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik} \left(\ln \theta_k + \ln \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik} \ln \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{?? revisited}$$

The natural log of the Gaussian is equal to

$$= -\frac{1}{2} \Big(D \ln 2\pi + \ln |\boldsymbol{\Sigma}| + (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \Big) \tag{11}$$

$$= -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \tag{12}$$

Thus we get that we are solving for $\boldsymbol{\mu}_k$ s.t.

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = -\frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) + const \tag{13}$$

$$= 0 \tag{14}$$

Carrying the derivative out:

$$0 = -\frac{1}{2}(\mathbf{\Sigma}^{-1} + \mathbf{\Sigma}^{-T}) \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k) \tag{15}$$

$$0 = \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k) \tag{16}$$

$$0 = \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik}\mathbf{x}_i - q_{ik}\boldsymbol{\mu}_k \tag{17}$$

$$\sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik}\boldsymbol{\mu}_k = \sum_{i=1}^{n} \sum_{k=1}^{c} q_{ik}\mathbf{x}_i \tag{18}$$

$$\tag{19}$$

As the μ_k is the same for all points in the class, $\sum_{k=1}^{c} q_{ik}\boldsymbol{\mu}_k$ is simply $n_k\boldsymbol{\mu}_k$. Thus we get that

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} q_{ik}\mathbf{x}_i \tag{20}$$

– Apply a similar argument to find the value of the (μ_k, Σ_k)'s that maximizes the expected complete-data log likelihood. $\mathbf{\Sigma}_k$ case.

$$\hat{\mathbf{\Sigma}}_k = \frac{1}{n_k} \sum_{i=1}^{n} q_{ik}(x_i - \hat{\boldsymbol{\mu}}_k)(x_i - \hat{\boldsymbol{\mu}}_k)^T \tag{21}$$

Dropping terms without $\mathbf{\Sigma}$

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{k=1}^{c} \left[-\frac{1}{2}(\ln|\mathbf{\Sigma}| + (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)) \right] \tag{22}$$

Taking the partial derivative and collapsing sums, we get

$$\frac{\partial \mathcal{L}}{\partial \mathbf{\Sigma}_k} = \frac{n_k}{\mathbf{\Sigma}2} - \frac{1}{2} \sum_{i=1}^{n} q_{ik}(x_i - \hat{\boldsymbol{\mu}}_k)(x_i - \hat{\boldsymbol{\mu}}_k)^T \tag{23}$$

$$\mathbf{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n} q_{ik}(x_i - \boldsymbol{\mu}_k)(x_i - \boldsymbol{\mu}_k)^T \tag{24}$$