Expectation-Maximization from ml_stanford_cs229/CS229_Lecture_Notes.pdf

Mixture of Gaussians model is an example of EM for missing data

$\{x^{(1)}, \ldots, x^{(n)}\}$ training set — **Observed, incomplete-data**

$p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$. — the model is the joint probability that each x^(i) was generated by randomly choosing z^(i) from {1…k}.

$z^{(i)} \sim \text{Multinomial}(\phi)$ — **Latent, unobserved cluster assignments, needed for complete-data**

$\phi_j \geq 0, \sum_{j=1}^k \phi_j = 1,$

ϕ_j gives $p(z^{(i)} = j)$). — **Cluster (gaussian) weights**

p(x^(i) | z^(i) = j) ~ N(μ_j, Σ_j) — we adopt the gaussian distribution to represent x

ϕ, μ and Σ are the model parameters.
k is a hyper-parameter — **Number of clusters (number of gaussians)**

we need the likelihood of x, **p(x; ϕ, μ, Σ)** to estimate the parameters:

$$
\begin{aligned}
\ell(\phi, \mu, \Sigma) &= \sum_{i=1}^{n} \log p(x^{(i)}; \phi, \mu, \Sigma) \\
&= \sum_{i=1}^{n} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)} | z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi) = \sum_{i=1}^{n} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}, z^{(i)})
\end{aligned}
$$

If we knew z^(i)'s, the problem would be solvable using:

$$
\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \log p(x^{(i)} | z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi).
$$

then from setting derivatives to 0 to find the maxima, we would have:

$$
\begin{aligned}
\phi_j &= \frac{1}{n} \sum_{i=1}^{n} 1\{z^{(i)} = j\}, \\
\mu_j &= \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}, \\
\Sigma_j &= \frac{\sum_{i=1}^{n} 1\{z^{(i)} = j\} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} 1\{z^{(i)} = j\}}.
\end{aligned}
$$

since we do NOT know the z^(i)'s, we use expectation-maximization:

in Expectation step: guess the z^(i)'s

in Maximization step: update the model parameters, based upon the z^(i) estimates

---

```
step 0: init parameters k, and ϕ_j, μ_j, Σ_j) for k=1 to k

prev_log_l = 1E-11;  tol=1E-4

while (true):

  #E-step:
    for each j=1,n, for each i=1,k:  #estimate the posterior
      #set Q_i(z^(i) = j) := p(z^(i) | x^(i); θ) so that ELBO(x;Q,θ) = log p(x; θ) for x and the current θ

      Q_i(z^(i) == j) = w^(i)_j = p(z^(i) = j | x^(i); ϕ, μ, Σ) from Bayes Rule = p(x|z)*p(z)/p(x)
                      = N(x^(i) | μ_j, Σ_j) * ϕ_j / sum_over ell = 1 to_k( ϕ_ell * N(x^(i) | μ_ell, Σ_ell) )
      #note that these should sum to 1 for a single data point x^(i)

    #calculate log likelihood too as it is used to define convergence.
    log_l = l(θ) = sum over j=1,n ( log p(x; θ) ) = sum over j=1,n ( log (sum over i=1,k (w^(i)_j)))
    if (log_l - prev_log_l <= tol):
      converged = true
      break
```

#M-step:

estimate ϕ, μ, Σ by maximizing ELBO:

EQN 11.12 θ = argmax ELBO(x^(i); Q_i, θ) w.r.t θ while fixing the choice of Q_i

$$\sum_{i=1}^{n}\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} Q_i(z^{(i)} = j) \log \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

hold w^(i)_j fixed while taking the gradient of the rest of the ELBO. set gradient to zero and solve for parameters.

note that : w^(i)_j =Q_i(j)

$$\phi_j := \frac{1}{n}\sum_{i=1}^{n} w_j^{(i)},$$

$$\mu_j := \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)}},$$

$$\Sigma_j := \frac{\sum_{i=1}^{n} w_j^{(i)}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)}}$$

Generalization of details of **p(z|x), choosing** Q as estimate and the ELBO:

choose Q(z) proportional to the joint distribution as the model. the posterior follows.

$$Q(z) \propto p(x, z; \theta) \quad = \quad p(z|x; \theta) \qquad \boxed{\text{posterior of z's}} \tag{11.8}$$

$$Q_i(z^{(i)}) \quad = \quad p(z^{(i)}|x^{(i)}; \theta)$$

$$\log p(x^{(i)}; \theta) \geq \mathrm{ELBO}(x^{(i)}; Q_i, \theta) = \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

$$\begin{aligned} \ell(\theta) &\geq \sum_i \mathrm{ELBO}(x^{(i)}; Q_i, \theta) && (11.11) \\ &= \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \end{aligned}$$