VECTORIZED EM ALGORITHM

It is possible to fully vectorize the EM algorithm for the mixture of gaussians model.

E-STEP

$$\gamma_{nk} = p(\text{class}=k \mid \underline{x}^{(n)})$$

$$= \frac{p(\underline{x}^{(n)} \mid k) \cdot p(k)}{\sum_j p(\underline{x}^{(n)} \mid j) \cdot p(j)}$$

$$= \frac{\pi_k \mathcal{N}(\underline{x}^{(n)}; \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\underline{x}^{(n)}; \mu_j, \Sigma_j)}$$

M-STEP

$$\pi_k^* = \frac{N_k}{N}$$

$$\mu_k^* = \frac{1}{N_k} \sum_n \gamma_{nk} \underline{x}^{(n)}$$

$$\Sigma_k^* = \frac{1}{N_k} \sum_n \gamma_{nk} (\underline{x}^{(n)} - \mu_k^*)(\underline{x}^{(n)} - \mu_k^*)^T$$

First we define the parameters

$$X = \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_N^T - \end{bmatrix} \qquad \mu = \begin{bmatrix} - \mu_1^T - \\ \vdots \\ - \mu_K^T - \end{bmatrix}$$

$$\Pi = [\pi_1 \cdots \pi_K] \qquad \Sigma = [\Sigma_1 \cdots \Sigma_K]$$

One of the first things we can vectorize is the likelihood $p(X \mid Z) = \mathcal{N}(X; \mu, \Sigma)$ into a matrix with dimensions (N, K)

$$p(X \mid Z) = \begin{bmatrix} p(\underline{x}^{(1)} \mid z=1) & \cdots & p(\underline{x}^{(1)} \mid z=k) \\ \vdots & & \vdots \\ p(\underline{x}^{(N)} \mid z=1) & \cdots & p(\underline{x}^{(N)} \mid z=k) \end{bmatrix}$$

Once we have a vectorization of the likelihood we can easily compute the E-step using broadcasting vectorizing the likelihood requires us to calculate the gaussian pdf of every point-cluster pair

$$p(\underline{x}^{(n)} | z=k) = \frac{1}{(2\pi)^{D/2}} |\Sigma_k|^{-1/2} e^{-\frac{1}{2}(\underline{x}^{(n)}-\mu_k)^T \Sigma_k^{-1} (\underline{x}^{(n)}-\mu_k)}$$

Rather than using a loop to individually calculate elements of the likelihood we can vectorize the pdf by vectorizing the <u>Mahalanobis distance</u>.

$$\begin{aligned}
\Delta^2 &= (\underline{x}^{(n)}-\mu_k)^T \Sigma_k^{-1} (\underline{x}^{(n)}-\mu_k) \\
&= \mathrm{Tr}\{ (\underline{x}^{(n)}-\mu_k)^T \Sigma_k^{-1} (\underline{x}^{(n)}-\mu_k) \} \\
&= \mathrm{Tr}\{ (\underline{x}^{(n)}-\mu_k)(\underline{x}^{(n)}-\mu_k)^T \Sigma_k^{-1} \} \\
&= \mathrm{vec}[\underbrace{(\underline{x}^{(n)}-\mu_k)(\underline{x}^{(n)}-\mu_k)^T}_{\triangleq\, \delta_{k,n}}]^T \mathrm{vec}[\Sigma_k^{-1}]
\end{aligned}$$

→ The vectorization of the precision involves flattening the last 2 dimensions of the tensor $\Sigma \in (K,D,D)$

→ The vectorization of the co-deviation involves flattening the last 2 dimensions of the tensor $\delta \in (K,N,D,D)$

Consequently to vectorize the Mahalanobis distance we must get the co-deviation tensor δ, and to get the co-deviation tensor we must get the deviation tensor $\omega \in (K,N,D)$ where.

$$\omega_{k,n} \triangleq \underline{x}^{(n)} - \mu_k$$

and then the co-deviation is

* The parenthesis (x) indicates an added dimension of size x.

* where the subscripts indicate the shape of the tensor

$$\boxed{\delta_{K,N,D,D} = \omega_{K,N,D,(1)} @ \omega_{K,N,(1),D}}$$

notice $\omega_{\ldots D,1} \times \omega_{\ldots 1,D} = \delta_{\ldots D,D}$
as expected in matrix multiplication.

The deviation tensor can also be vectorized using broadcasting

$$\boxed{\omega_{K,N,D} = X_{N,D} - \mu_{K,(N),D}}$$

*subscripts indicate the shape of the tensor (not index!)

where we have added a dimension to $\mu_{K,D}$ by stacking/tiling along axis = 1 for N times.

<u>LIKELIHOOD SUMMARY</u>

1. First we calculate the deviation tensor:

$$\omega_{K,N,D} = X_{N,D} - \mu_{K,(N),D}$$

2. Then we calculate the co-deviations

$$\delta_{K,N,D,D} = \omega_{K,N,D,(1)} \,@\, \omega_{K,N,(1),D}$$

3. Next the Mahalanobis distance can be calculated

$$\Delta^2_{K,N,1} = \text{vec}[\,\delta_{K,N,D,D}\,] \,@\, \text{vec}[\,\Sigma^{-1}_{K,D,D}\,]_{(1)}$$

we add a dimension so we can use matrix multiplication instead of broadcasting $(N, D^2) \times (D^2, 1) = (N, 1)$

$$K\left\{\begin{array}{ccc}
\begin{bmatrix}[-\ \text{vec}(\delta)_{0,0}\ -]\\ [-\ \text{vec}(\delta)_{0,1}\ -]\\ \vdots \\ [-\ \text{vec}(\delta)_{0,N}\ -]\end{bmatrix} & \xleftrightarrow{\text{matmul}} & \begin{bmatrix}\begin{bmatrix} | \\ \text{vec}(\Sigma^{-1})_0 \\ | \end{bmatrix}\end{bmatrix} \\
\vdots & @ & \vdots \\
N\left\{\underbrace{\begin{bmatrix}[-\ \text{vec}(\delta)_{K,0}\ -]\\ [-\ \text{vec}(\delta)_{K,1}\ -]\\ \vdots \\ [-\ \text{vec}(\delta)_{K,N}\ -]\end{bmatrix}}_{D^2}\right. & \xleftrightarrow{\text{matmul}} & \left.\underbrace{\begin{bmatrix}\begin{bmatrix} \\ \text{vec}(\Sigma^{-1})_K \\ \\ \end{bmatrix}\end{bmatrix}}_{1}\right\} D^2
\end{array}\right.$$

4. We use the squeezed Mahalanobis distance $(\Delta^2_{K,N})$ to determine the likelihood tensor $p(X \mid Z)$

E-STEP VECTORIZATION

Once we have the vectorized likelihood tensor we can easily calculate the E-Step responsibility matrix

$$\gamma_{N,K}$$

M-STEP VECTORIZATION

The class priors can be easily calculated, however the mean and covariance need to be vectorized.

$$\mu_K = \frac{1}{N_K} \sum_n \gamma_{nk} \underline{x}^{(n)}$$

$$\Rightarrow \mu_{ki} = \frac{1}{N_k} \sum_n \gamma_{nk} x_{ni}$$

$$= \frac{1}{N_k} \underbrace{\sum_n \gamma^T_{kn} x_{ni}}_{\text{matrix multiplication}} \qquad \Rightarrow \boxed{\mu_{K,D} = \frac{1}{N_k} \cdot \gamma^T_{N,K} @ X_{N,D}}$$

subscripts indicate tensor shape

$$\Sigma_k = \frac{1}{N_k} \sum_n \gamma_{nk} \underbrace{(\underline{x}^{(n)} - \mu_n)(\underline{x}^{(n)} - \mu_n)^T}_{\text{co-deviation}}$$

$$= \frac{1}{N_k} \sum_n \gamma_{nk} \delta_{kn}$$

$$\Rightarrow \Sigma_{k,i,j} = \frac{1}{N_k} \sum_n \gamma_{nk} \delta_{knij}$$

$$= \frac{1}{N_k} \underbrace{\sum_n \delta^{T(2,3,0,1)}_{i,j,k,n} \gamma_{nk}}_{\text{matrix multiplication}}$$

$$= \frac{1}{N_k} (\delta^{T(2,3,0,1)} @ \gamma)_{i,j,k,k}$$

$$= \frac{1}{N_k} [\mathrm{diag}(\delta^{T(2,3,0,1)} @ \gamma)]_{i,j,k}$$

$$= \frac{1}{N_k} [\mathrm{diag}(\delta^{T(2,3,0,1)} @ \gamma)]^{T(2,0,1)}_{k,i,j}$$

subscripts indicate tensor shape

$$\Rightarrow \boxed{\Sigma_{K,D,D} = \frac{1}{N_k} [\mathrm{diag}(\delta^{T(2,3,0,1)}_{K,N,D,D} @ \gamma_{N,K})]^{T(2,0,1)}}$$