latent variable — category of x_i

EM for (Gaussian Mixture Model) (GMM)

$\theta^{(0)} = \{c_y, \mu_y, \Sigma_y\}$

E-step: Given the current estimates of the parameters θ_t^t we can compute the posterior $= \dfrac{\text{prior} \times \text{likelihood}}{\text{Normalizer}}$

$$Q_{i,y}^{(t+1)} \equiv \underbrace{P[Y=y \mid X=x_i]}_{\text{posterior}} = \underbrace{c_y}_{\text{prior (weight)}} \times \underbrace{\frac{1}{\sqrt{(2\pi)^d |\Sigma_y|}} \exp\left(-\frac{1}{2}(x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y)\right)}_{\text{Gaussian}}$$

M-step: Estimating $\theta^{(t+1)} = \arg\max_{\theta} \underbrace{F(Q^{(t+1)}, \theta)}$

For simplicity, we assume $\Sigma_y = I$ (identity matrix) $\forall y$.

$$F(Q^{(t+1)}, \theta) = \sum_i \sum_y \underbrace{Q_{iy}^{(t+1)}} \log\left(P[X=x_i, Y=y]\right)$$

$$= \sum_i \sum_y \underbrace{P_{\theta^t}[Y=y \mid X=x_i]} \log\left(P[X=x_i, Y=y]\right)$$

$$= \sum_i \sum_y P_{\theta^t}[Y=y \mid X=x_i] \left(\underbrace{\log(c_y)}_{\text{prior}} + \underbrace{\log \exp\left(-\frac{1}{2}\|x_i - \mu_y\|^2\right)}_{\text{likelihood}}\right)$$

$$= \sum_i \sum_y P_{\theta^t}[Y=y \mid X=x_i] \left(\log(c_y) - \underbrace{\frac{1}{2}\|x_i - \mu_y\|^2}\right)$$

To maximize $F(Q^{(t+1)}, \theta)$ w.r.t. μ_y we take derivative w.r.t. μ_y and equate it to zero.

$\|x\|^2 = x^T x$

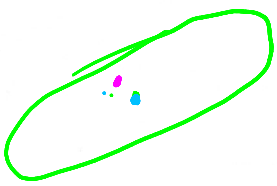
$-\frac{1}{2}\, 2x$

$$\sum_i P_{\theta^t}[Y=y \mid X=x_i] \underbrace{(x_i - \mu_y)} = 0$$

$$\mu_y = \frac{\sum_i \boxed{P_{\theta^t}[Y=y \mid X=x_i]}\, x_i}{\sum_i P_{\theta^{(t)}}[Y=y \mid X=x_i]}$$

responsibility

Posterior: Measure of resp that y^{th} cluster takes for x_i



When optimizing $F(Q^{(t+1)}, \theta)$ w.r.t. c_y w.r.t. c_y we need to satisfy the constraint that $\boxed{\sum_y c_y = 1.}$

We formulate the Lagrangian

$$\sum_i \sum_y P_{\theta^t}[Y=y \mid X=x_i]\left(\log c_y - \frac{1}{2}\|x_i - \mu_y\|^2\right) + \lambda\left(\sum_y c_y - 1\right)$$

Taking derivative w.r.t. c_y and equating it to zero

$$\sum_i P_{\theta^t}[Y=y \mid X=x_i] \times \left(\frac{1}{c_y}\right) + \lambda = 0$$

Multiplying by c_y and summing out y gives

$$\underbrace{\sum_i \sum_y P_{\theta^t} P[Y=y \mid X=x_i]}_{1} + \lambda \sum_y c_y = 0$$

$$\cancel{N + \lambda} \qquad \sum_{i=1}^{m} (1) + \lambda (1) = 0$$

$$\therefore \boxed{\lambda = -m}$$

$m + \lambda = 0$

prior (weight) of yth cluster $\quad c_y = \left(\frac{1}{m}\right) \sum_i P_{\theta^t}[Y=y \mid X=x_i]$

If the covariance matrix is not assumed as identity, then its updated value is

$$\Sigma_y = \frac{1}{m\, c_y} \sum_i P_{\theta^t}[Y=y \mid X=x_i](x_i - \mu_y)(x_i - \mu_y)^T$$

$$\sum_{i=1}^{m} x_i x_i^T$$

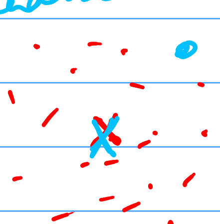
Homework

Read K-means clustering

K-means	GMM
	MV Gaussians (Hyper ellipse)
My Hyper spheres.	
Hard assignment of a data point to a cluster	Q. Soft assignment of data point to clusters.
K-means algorithms: Iterations	Iterations
Converge.	

Objective function
Distortion measure

K: hyperparameter
clusters



(centroid) Mean of a cluster is the prototype (mean). to represent the points

K as large: large # prototypes.