Assignment 2

190070033,190070068,19D070003

Answer 3

If prior information on the parameters θ in the form of probability distribution $P(\theta)$ is known then instead of maximizing over likelihood, we can maximize over posterior function that is

$$arg \max_{\theta} P(\theta|y) \tag{1}$$

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{2}$$

$$\max_{\theta} log(P(\theta|y)) = \max_{\theta} [log(P(y|\theta)) + log(P(\theta)) - log(P(y))]$$

Let $F(q, \theta) = log(P(\theta|y)) - KL(q||p(x|y, \theta))$

E step

Design q(.) to maximize F(q,θ_i)

$F(q, \theta) = log(P(\theta|y)) - KL(q||p(x|y, \theta))$
$\therefore q(x) = P(x|y, \theta_i)$
F(.) is lower bound of the log posterior funciton and satisfies
F(q,θ_i) = log(P($y|\theta_i$)) at $\theta = \theta_i$

M step

Choose θ to maximise $F(q, \theta)$

$F(q, \theta) = log(P(y|\theta)) + log(P(\theta)) - log(P(y)) - KL(q||P(x|y, \theta))$

We also know that,

$$log(P(y|\theta)) - KL(q||p(x|y, \theta)) = E_{q(.)}[log(P(x|y, \theta)] + H(q)$$

where $H(q) = E[-log(q)]$

$$\therefore F(q, \theta) = E_{q(.)}[log(P(x|y, \theta)] + H(q) + log(P(\theta)) - log(P(y))$$

Let $Q(\theta; \theta_i) = E_{q(.)}[log(P(x|y, \theta)]$

$\therefore$ We maximise $Q(\theta; \theta_i) + log(P(\theta))$

i) For the weights, we want $\sum_k w_k = 1$ with each $0 < w_k < 1$
$\therefore$ We can design the prior for weights to be the Dirichlet distribution of the order k with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3..., \alpha_k)$ given by

$$f(w_1, w_2, .., w_k; \alpha_1, \alpha_3..., \alpha_k) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} w_i^{\alpha_i - 1}$$

satisfying $\sum_k w_k = 1$ and $0 < w_k < 1$

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_i^K \alpha_i)}$$

Note that if $\boldsymbol{\alpha} = (1, 1, 1..., 1)$ then the prior reduces to a uniform distribution. We can choose to tune the parameter looking at the histogram of the image intensities or start with a uniform distribution as well.

The prior for mean of the cluster distributions can be modelled as $\mu_i \sim N(\mu_0, \lambda\sigma_i^2)$ and for variance as $\sigma^2 \sim$ Inverse-Gamma(ν, σ_0^2) or inverse

matrix gamma distribution for covariance matrix $\sim \text{IMG}(\alpha, \beta, \Psi)$
$\mu_0, \sigma_0, \nu, \lambda$ are all hyperparameters.

$$f(C_i) = K|C_i|^{-\alpha-(n+1)/2} exp(-\frac{1}{\beta} tr(\Psi C_i^{-1}))$$

Now, M step is:

$$G(\theta) = \underset{\theta}{argmax}\ Q(\theta; \theta_i) + log(P(\theta))$$

$$= \underset{\theta}{argmax} \sum_n \sum_k \gamma_{nk}(-0.5log|C_k| - 0.5(y_n - \mu_k)' C_k^{-1}(y_n - \mu_k) + logw_k)$$
$$- (n + \alpha + (n+1)/2)log(|C_i|) - \frac{1}{\beta}(tr(\Psi C_i^{-1})) - \frac{(\mu_i - \mu_0)^2}{2\sigma_i^2}$$

Now take $\frac{\partial G}{\partial \mu_i} = 0$ and $\frac{\partial G}{\partial C_i} = 0$ will give solution to μ_i and for C_i using cholesky decomposition.