Practice Set 3 (Problem 1)

$X = \{x_1, x_2, \ldots x_N\}$

$y = \{y_1, y_2, \ldots y_N\}$

The likelihood (both X and y are modeled here)

$$P(X, y \mid \theta) = \prod_{n=1}^{N} P(x_n, y_n \mid \theta)$$

Indicator function: $\mathbb{I}[y_n = k]$

θ (All the params) $= \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^{K}$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} \left[P(x_n, y_n = k \mid \theta) \right]$$

Note that only one term in the product will be selected based on the value of y_n

Thus log-likelihood will be

$$\log P(X, y \mid \theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \left[\log P(x_n, y_n = k \mid \theta) \right]$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \left[\log P(x_n \mid y_n = k, \theta) P(y_n = k \mid \theta) \right]$$

$P(x_n \mid y_n = k, \theta) \rightarrow \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$

$P(y_n = k \mid \theta) \rightarrow \pi_k$

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \left[\log \mathcal{N}(x_n \mid \mu_k, \Sigma_k) + \log \pi_k \right]$$

To optimize w.r.t. π_k, we take partial derivatives w.r.t. π_k <u>BUT</u> need to use the constraint $\sum_{k=1}^{K} \pi_k = 1$. Also note that for $\{\pi_k\}_{k=1}^{K}$, the part of the objective that matters is

$$\mathcal{L}(\pi_1, \ldots \pi_K) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \log \pi_k$$

(Lagrangian)

Need to min. over π, max over λ

$$= \sum_{k=1}^{K} N_k \log \pi_k$$

N_k = # of points with $y_n = k$

↑ The Constrained formulation will be

$$\mathcal{L}(\pi, \lambda) = \sum_{k=1}^{K} N_k \log \pi_k + \lambda \left[\sum_{k=1}^{K} \pi_k - 1 \right]$$

Taking partial derivative w.r.t. π_k,

$$\frac{N_k}{\pi_k} + \lambda = 0$$

$$\pi_k = -\frac{N_k}{\lambda}$$

Also, using $\sum_{k=1}^{K} \pi_k = 1$ gives $\lambda = -N$

thus $\boxed{\pi_k = \frac{N_k}{N}}$

Now, for μ_k and Σ_k :

The relevant part of the objective function is

$$\mathcal{L}\left(\{\mu_k, \Sigma_k\}_{k=1}^{K}\right) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \log \mathcal{N}(x_n | \mu_k, \Sigma_k)$$

MLE for μ_k, Σ_k is exactly the same as the MLE procedure for a multivariate Gaussian, but only using inputs x_n from class k.

(August 21 lecture)

I have provided additional slides on the class webpage showing how to do it for multivariate Gaussian. The procedure for this problem will be exactly the same (but only using x_n with $y_n = k$)

Practice Set 3 (Problem 2)

MAP estimation for $[\pi_1, \ldots, \pi_K]$ is identical to the MLE procedure, except now we maximize $\log p(y|\pi) + \log p(\pi)$

($\log p(\pi)$: Dirichlet)

$$\log p(y|\pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}[y_n = k] \log \pi_k$$

$$\log p(\pi) = \log \prod_{k=1}^{K} \pi_k^{\alpha-1} = \sum_{k=1}^{K} (\alpha-1) \log \pi_k$$

(Dirichlet without prop. const.)

we also have the sum to 1 constraint.

The full objective (Lagrangian will be)

$$L(\pi, \lambda) = \sum_{k=1}^{N} N_k \log \pi_k + \sum_{k=1}^{K} (\alpha-1) \log \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$\sum N_k \log \pi_k$: log lik

$\sum (\alpha-1) \log \pi_k$: log prior (wasn't there in Prob. 1)

We will use the same procedure as used in prob. 1.

The solution will be

$$\hat{\pi}_k = \frac{N_k + \alpha - 1}{\sum_{k=1}^{K} (N_k + \alpha - 1)}$$

$$= \frac{N_k + \alpha - 1}{N + N\alpha - K}$$

Alternatively, we can easily get the posterior, which will be Dirichlet due to conjugacy and the form will be (multiply the multinoulli and Dirichlet and verify)

$$P(\pi|y) = Dir[\alpha + N_1, \alpha + N_2, \ldots, \alpha + N_K]$$

The MAP estimate will be the mode of this distribution.

$$\hat{\pi}_k = \frac{\alpha + N_k - 1}{N + K\alpha - K}$$

Practice set-3 (Problem 3)

Using the expression of Gaussian conditional from Gaussian joint, we get

$p(y|x) = N(\mu', \sigma'^2)$ where μ' and σ'^2 are

$$\mu' = \mu_y + V^T \Sigma_{xx}^{-1}(x - \mu_x)$$

$$\sigma'^2 = \sigma_y^2 - V^T \Sigma_{xx}^{-1} V$$

Note that if Σ_{xx} is diagonal then ("$w^T x$")

$$\mu' = V'^T(x - \mu_x) + \mu_y$$ (Similar to "$w^T x$")

where V' is V with each of its entries divided by the corresponding diagonal entry of Σ_{xx}.

Similarity/difference with linear regression:

if μ_y and $\mu_x = 0$ then

$$\mu' = V'^T x$$

(Probabilistic ↑ similar to a linear regression model's mean)

and

$$\sigma'^2 = \sigma_y^2 - V^T \Sigma_{xx}^{-1} V$$ (Slightly different from prob. linear regression's variance $P(y|x) = N(w^T x, \beta^{-1})$)