CS772: Assignment 2

Gurpreet Singh — 150259

Question 1

We can write $P(\theta)$ as a convex combination of K priors as follows

$$P(\theta) = \sum_{k=1}^{K} P(z = k)P(\theta|z = k) \tag{1}$$

Using Chain Rule of Probability, we can write the posterior in the following manner

$$P(\theta|\mathcal{D}) = \sum_{k=1}^{K} P(\theta|\mathcal{D}, z = k)P(z = k|\mathcal{D}) \tag{2}$$

Now, using Bayes rule, we can write

$$P(\theta|\mathcal{D}, z = k) = \frac{P(\mathcal{D}|\theta, z = k)P(\theta|z = k)}{P(\mathcal{D}|z = k)} \tag{3}$$

Note: The denominator is independent of θ.

$$P(\mathcal{D}|z = k) = \int_{\theta} P(\mathcal{D}|\theta, z = k)P(\theta|z = k) \tag{4}$$

Now, since we are also given that $P(\theta|z = k)$ is conjugate to $P(\mathcal{D}|\theta|z = k)$, hence $P(\theta|\mathcal{D}, z = k)$ is tractable and can be computed. Also

$$\sum_{k=1}^{K} P(z = k|\mathcal{D}) = 1$$

Therefore, using equation 2 we can say that $P(\theta|\mathcal{D})$ can be written as a convex combination of conjugate distributions. Precisely

$$P(\theta|\mathcal{D}) = \sum_{k=1}^{K} \frac{P(\mathcal{D}|\theta, z = k)P(\theta|z = k)P(z = k)}{P(\mathcal{D}|z = k)} \tag{5}$$

Question 2

Part A

We know $\boldsymbol{x} \in \mathbb{R}^D$ and $y \in \mathbb{R}$. This suggests that for the joint distribution of (x, y), we should use Gaussian Distribution. Therefore

$$(\boldsymbol{x}, y) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

This distribution can be justified as we are essentially modelling y as a function of $\boldsymbol{x}$ alongside modelling $\boldsymbol{x}$ as a part of the generative model. This can be later seen when we compute the MLE estimate of $\boldsymbol{x}$ and Σ.

Part B

Using the Chain Rule of Probability, we can say

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}, y)}{P(\boldsymbol{x})}$$

Also, we can write

$$P(\boldsymbol{x}) = \int_{y \in \mathbb{R}} P(\boldsymbol{x}, y) dy$$

Using properties of Gaussian distributions, we can easily marginalize over y to get $P(\boldsymbol{x}, y)$

$$P(y|\boldsymbol{x}) = \mathcal{N}(\mu', \Sigma')$$

where μ' and Σ' are defined as follows

$$\mu' = \boldsymbol{\mu}_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})$$
$$\Sigma' = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

Note: Here $\Sigma' \in \mathbb{R}$

Part C

For MLE estimate, we need to maximize $P(\mathcal{D})$. Assuming all points to iid, we can write

$$P(\mathcal{D}) = \prod_{n=1}^{N} p(\boldsymbol{x}^n, y^n)$$

We can also minimize the NLL of $P(\mathcal{D})$

$$\begin{aligned} NLL &= \sum_{n=1}^{N} -\log(P(\boldsymbol{x}^n, y^n)) \\ &= \sum_{n=1}^{N} \frac{1}{2}(\boldsymbol{p}^n - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{p}^n - \boldsymbol{\mu}) \end{aligned}$$

where

$$\boldsymbol{p}^n = \begin{bmatrix} \boldsymbol{x}^n \\ y^n \end{bmatrix}$$

Since this is the NLL of a multivariate gaussian distribution, we can directly write the MLE estimates of μ' and Σ'

$$\begin{aligned}
\boldsymbol{\mu} &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{p}^n \\
\implies \boldsymbol{\mu}_x &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}^n \qquad \mu_y = \frac{1}{N}\sum_{n=1}^{N} y^n \\
\Sigma &= \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{p}^n - \boldsymbol{\mu})(\boldsymbol{p}^n - \boldsymbol{\mu})^T \\
\implies \Sigma &= \frac{1}{N}\sum_{n=1}^{N} \begin{matrix} (\boldsymbol{x}^n - \boldsymbol{\mu}_x)(\boldsymbol{x}^n - \boldsymbol{\mu}_x)^T & (\boldsymbol{x}^n - \boldsymbol{\mu}_x)(y^n - \mu_y) \\ (y^n - \mu_y)(\boldsymbol{x}^n - \boldsymbol{\mu})^T & (y^n - \mu_y)(y^n - \mu_y) \end{matrix}
\end{aligned}$$

Question 3

For a one dimensional GMM, $x \in R$ and $z \in \{1, 2 \dots K\}$. Hence, we model $x|z = k$ as a univariate gaussian, and z as a multinoulli.

$$\begin{aligned} x|z = k &\sim \mathcal{N}(\mu_k, \sigma_k^2) \\ z &\sim Multinoulli(\boldsymbol{\pi}) \end{aligned}$$

Assuming that z is observed, we can write the joint distribution, x, z as follows

$$P(x, z) = \frac{1}{\sqrt{2\pi}\sigma_z} \boldsymbol{\pi}_z e^{-\frac{1}{2\sigma_z^2}(x-\mu_z)^2}$$

This can be written in the exponential form as follows

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma_z^2}(x-\mu_z)^2 - \log \sigma_z + \log \boldsymbol{\pi}_z}$$

Looking at only the part in the exponent

$$\begin{aligned} exp &= -\frac{1}{2\sigma_z^2}(x - \mu_z)^2 - \log \sigma_z + \log \boldsymbol{\pi}_z \\ &= -\left[\frac{x^2}{2\sigma_z^2} - \frac{x\mu_z}{\sigma_z^2} + \frac{\mu_z^2}{2\sigma_z^2} + \log \sigma_z - \log \boldsymbol{\pi}_z\right] \end{aligned}$$

If we represent z in the one-hot representation, i.e. $\boldsymbol{z}_k = \mathbb{I}\big[z = k\big]$

$$\begin{aligned} exp &= -\sum_{k=1}^{K} \boldsymbol{z}_k \big(\frac{x^2}{2\sigma_k^2} - \frac{x\mu_k}{\sigma_k^2} + \frac{\mu_k^2}{2\sigma_k^2} + \log \sigma_k - \log \boldsymbol{\pi}_k\big) \\ &= \theta^T \phi(x, z) - A(\theta) \end{aligned}$$

where

$$\theta = \begin{bmatrix} -\frac{1}{2\sigma_k^2} \\ \frac{\mu_k}{\sigma_k^2} \\ \log \boldsymbol{\pi}_k \end{bmatrix}_{k=1}^{K} \qquad \phi(x, z) = \begin{bmatrix} \boldsymbol{z}_k x^2 \\ \boldsymbol{z}_k x \\ \boldsymbol{z}_k \end{bmatrix}_{k=1}^{K}$$

Since we can write the joint distribution in the exponential form, we can say that x, z is an exponential distribution if z is observed.

Question 4

Since each data point is a binary vector of D dimensions (assuming the features to be independent), and we have K components, we can model $\boldsymbol{x}|z$ as a Multivariate Bernoulli Distribution, and z as a Multinoulli Distribution.

$$P(\boldsymbol{x}|z=k) = \prod_{d=1}^{D} \boldsymbol{\rho_k}_d^{\boldsymbol{x}_d} (1-\boldsymbol{\rho_k}_d)^{(1-\boldsymbol{x}_d)}$$
$$P(z) = \prod_{k=1}^{K} \boldsymbol{\pi}_k^{\mathbb{I}\left[z=k\right]}$$

Similar to the GMM case, we compute the Complete Log Likelihood ($\log P(\boldsymbol{X}, z)$) it is not easy to compute the MAP estimate directly. It is easier to represent z in a one hot representation, such that $\boldsymbol{z}_k = \mathbb{I}\left[z=k\right]$

$$P(\boldsymbol{x}, \boldsymbol{z}) = \prod_{k=1}^{K} \prod_{d=1}^{D} \left(\boldsymbol{\rho_k}_d^{\boldsymbol{x}_d} (1-\boldsymbol{\rho_k}_d)^{(1-\boldsymbol{x}_d)} \boldsymbol{\pi}_k \right)^{\boldsymbol{z}_k}$$
$$P(\boldsymbol{X}, \boldsymbol{z}) = \prod_{n=1}^{N} P(\boldsymbol{x}^n, \boldsymbol{z}^n)$$

Therefore, CLL can be written as

$$CLL = \sum_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{z}_k^n \log \left(\boldsymbol{\rho_k}_d^{\boldsymbol{x}_d^n} (1-\boldsymbol{\rho_k}_d)^{(1-\boldsymbol{x}_d^n)} \right)$$

Here, we are assuming that z or $\boldsymbol{z}$ is observed, however, that is not the case. It is also hard to simulateously model both $\boldsymbol{x}$ and $\boldsymbol{z}$. Therefore, we alternately model them, using EM algorithm.

Note: I am using z to represent the integer value of z where as $\boldsymbol{z}$ represents the one-hot representation

Expectation Step

We first start with an initial value of all the parameters, namely $\left\{\boldsymbol{\rho}_k\right\}_{k=1}^{K}$ and $\boldsymbol{\pi}$. Then, we find the expected values of $\boldsymbol{z}_k^n$ using these parameters. As discussed in class, it is most optimal to calculate the expectation wrt the posterior distribution of the latent variable.

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{z}_k^n\right] &= 0 \cdot P(z^n \neq k|\boldsymbol{x}^n, \left\{\boldsymbol{\rho}_k\right\}_{k=1}^{K}, \boldsymbol{\pi}) + 1 \cdot P(z^n = k|\boldsymbol{x}^n, \left\{\boldsymbol{\rho}_k\right\}_{k=1}^{K}, \boldsymbol{\pi}) \\
&= P(z^n = k|\boldsymbol{x}^n, \left\{\boldsymbol{\rho}_k\right\}_{k=1}^{K}, \boldsymbol{\pi}) \\
&= \frac{\boldsymbol{\pi}_k P(\boldsymbol{x}|z=k, \rho_k)}{\sum_{k'=1}^{K} \boldsymbol{\pi}_{k'} P(\boldsymbol{x}|z=k', \rho_{k'})}
\end{aligned}$$

Now using these values, we will find out the best MAP estimates for the parameters mentioned above.

Maximization Step

In order to find the MAP estimate, we require the posterior distributions of the parameters.

MAP estimate of $\boldsymbol{\pi}$

Since $\boldsymbol{\pi}$ defines a parameter for a Multinoulli distribution, it is only logical to assume Dirichlet Distribution, i.e $\boldsymbol{\pi} \sim Dir(\boldsymbol{\gamma})$.

$$P(\boldsymbol{\pi}) = \frac{1}{B(\boldsymbol{\gamma})} \prod_{k=1}^{K} \boldsymbol{\pi}_k^{\boldsymbol{\gamma}_k}$$
$$P(\boldsymbol{\pi}|\boldsymbol{X}, \boldsymbol{z}) \propto P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\pi})P(\boldsymbol{\pi})$$

$$\boldsymbol{\pi}_{MAP} = \arg\max_{\boldsymbol{\pi}} \log P(\boldsymbol{\pi}|\boldsymbol{X}, \boldsymbol{z})$$
$$\text{s.t.} \sum_{k=1}^{K} \boldsymbol{\pi} = 1$$

This reduces to the following optimization function

$$\boldsymbol{\pi}_{MAP} = \arg\max_{\boldsymbol{\pi},\lambda} \sum_{n=1}^{N}\sum_{k=1}^{K} \boldsymbol{z}_k^n \log \boldsymbol{\pi}_k + \sum_{k=1}^{K} \boldsymbol{\pi}_k(\boldsymbol{\gamma}_k - 1) + \lambda(1 - \sum_{k=1}^{K} \boldsymbol{\pi}_k)$$

where λ is the Lagrange Multiplier. We need to replace $\boldsymbol{z}_k^n$ with its expected value, as calculated in the Expectation Step. After solving, we finally get

$$\boldsymbol{\pi}_{MAP} = \left\{ \frac{\sum_{n=1}^{N} \mathbb{E}\big[\boldsymbol{z}_k^n\big] + \boldsymbol{\gamma}_k - 1}{N + \sum_{k'=1} \boldsymbol{\gamma}_{k'} - K} \right\}_{k=1}^{K}$$

MAP estimate of $\{\boldsymbol{\rho}\}_{k=1}^{K}$

Since we are using $\boldsymbol{\rho}_k$ to model multivariate independent binary data, hence it makes sense to take $\boldsymbol{\rho}_{kd}$ as a Beta distribution.

$$P(\boldsymbol{\rho}_{kd}) = \boldsymbol{\rho}_{kd}^{(\boldsymbol{\alpha}_{kd}-1)}(1 - \boldsymbol{\rho}_{kd})^{(\boldsymbol{\beta}_{kd}-1)}$$
$$P(\boldsymbol{\rho}_{kd}|\boldsymbol{X}, \boldsymbol{z}) \propto P(\boldsymbol{X}, \boldsymbol{z}|\boldsymbol{\rho}_{kd})P(\boldsymbol{\rho}_{kd})$$

$$\boldsymbol{\rho}_{kd,MAP} = \arg\max_{\boldsymbol{\rho}_{kd}} P(\boldsymbol{\rho}_{kd}|\boldsymbol{X}, \boldsymbol{z})$$

The above optimization can be easily solved by differentiating the objective function partially wrt $\boldsymbol{\rho}_{kd}$

$$\boldsymbol{\pi}_{MAP} = \frac{\sum_{n=1}^{N} \mathbb{E}\big[\boldsymbol{z}_k^n\big]\boldsymbol{x}_d^n + \boldsymbol{\gamma}_k - 1}{\sum_{n=1}^{N} \mathbb{E}\big[\boldsymbol{z}_k^n\big] + \boldsymbol{\alpha}_{kd} + \boldsymbol{\beta}_{kd} - 2}$$

We can alternate between these steps until convergence.

Question 5

We have the following distributions

$$\begin{aligned}\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta &\sim \prod_{n=1}^{N}\mathcal{N}(y^n|\boldsymbol{x}^n,\boldsymbol{w},\beta^{-1})\\ \boldsymbol{w}|\lambda &\sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\lambda^{-1}\mathbb{I}_D)\\ \boldsymbol{w}|\boldsymbol{y},\boldsymbol{X},\lambda,\beta &\sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\Sigma)\end{aligned}$$

where $\boldsymbol{\mu}$ and Σ are defined as follows

$$\Sigma = (\beta\boldsymbol{X}^T\boldsymbol{X}+\lambda\mathbb{I}_D)^{-1} \qquad \boldsymbol{\mu} = (\boldsymbol{X}^T\boldsymbol{X}+\frac{\lambda}{\beta}\mathbb{I}_D)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Using these distributions, we can find the CLL (Complete Log Likelihood) i.e. $(\log P(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{X},\beta,\lambda))$

$$\begin{aligned}P(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{X},\beta,\lambda) &= P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta)P(\boldsymbol{w}|\lambda)\\ \log P(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{X},\beta,\lambda) &= \log P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta)+\log P(\boldsymbol{w}|\lambda)\\ \implies CLL &= \frac{1}{2}\Big[N\log\beta+D\log\lambda-\beta(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})\\ &\qquad \lambda\boldsymbol{w}^T\boldsymbol{w}-(N+D)\log 2\pi\Big]\end{aligned}$$

We have computed the CLL assuming we have observed the value of $\boldsymbol{w}$, however, that is not the case. Therefore, we will use EM Alogrithm to alternatively model the parameters and the hyperparameters.

Expectation Step

We first start with an initial value of all the hyper parameters, namely β and λ. Then, we find the expected values of $\boldsymbol{z}_k^n$ using these parameters. As discussed in class, it is most optimal to calculate the expectation wrt the posterior distribution of the latent variable. Here, the latent variable is $\boldsymbol{w}$.

$$\mathbb{E}\big[\boldsymbol{w}\big] = \int_{\boldsymbol{w}\in\mathbb{R}^D}\boldsymbol{w}P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X},\lambda,\beta)d\boldsymbol{w}$$

Since the posterior of $\boldsymbol{w}$ is a normal distribution, we can directly write the expectation of $\boldsymbol{w}$

$$= \boldsymbol{\mu}$$

Since we are to maximize the expectation of CLL, we also need to compute $\mathbb{E}\big[CLL\big]$

$$\begin{aligned}\mathbb{E}\big[CLL\big] &= \mathbb{E}\big[\frac{1}{2}\Big[N\log\beta+D\log\lambda-\beta(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})-\lambda\boldsymbol{w}^T\boldsymbol{w}-(N+D)\log 2\pi\Big]\big]\\ i &= \frac{1}{2}\Big[N\log\beta+D\log\lambda-\beta\big(\boldsymbol{y}^T\boldsymbol{y}-\mathbb{E}\big[\boldsymbol{w}^T\big]\boldsymbol{X}^T\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{X}\mathbb{E}\big[\boldsymbol{w}\big]+\mathbb{E}\big[\boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X}+\frac{\lambda}{\beta}\mathbb{I}_D)\boldsymbol{w}\big]\big)\\ &\qquad -(N+D)\log 2\pi\Big]\end{aligned}$$

We can compute the expectations as follows

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{w}^T\big] &= \mathbb{E}\big[\boldsymbol{w}\big]^T \\ &= \boldsymbol{\mu}\end{aligned}$$

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{w}\boldsymbol{w}^T\big] &= Cov(w) + \mathbb{E}\big[\boldsymbol{w}\big]\mathbb{E}\big[\boldsymbol{w}\big]^T \\ &= \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{w}^T R\boldsymbol{w}\big] &= Tr(R\mathbb{E}\big[\boldsymbol{w}\boldsymbol{w}^T\big]) \\ &= Tr(R(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T))\end{aligned}$$

Hence, we can write the expected value of CLL

$$\mathbb{E}\big[CLL\big] = \frac{1}{2}\Bigg[N\log\beta + D\log\lambda - \beta\big(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\mu}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\mu} + Tr((\boldsymbol{X}^T\boldsymbol{X} + \frac{\lambda}{\beta}\mathbb{I}_D)\boldsymbol{\mu}\boldsymbol{\mu}^T)\big) - (N+D)\log 2\pi$$

Maximization Step

MLE Estimate of β

We simply require to maximize the Expectation of CLL wrt β

$$\begin{aligned}\beta_{MLE} &= \arg\max_{\beta}\mathbb{E}\big[CLL\big] \\ \beta^{-1}_{MLE} &= \frac{1}{N}\big(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\mu}^T\boldsymbol{X}^T\boldsymbol{y} + Tr((\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\mu}\boldsymbol{\mu}^T)\big)\end{aligned}$$

MLE Estimate of λ

We simply require to maximize the Expectation of CLL wrt λ

$$\begin{aligned}\lambda_{MLE} &= \arg\max_{\lambda}\mathbb{E}\big[CLL\big] \\ \lambda^{-1}_{MLE} &= \frac{Tr(\boldsymbol{\mu}\boldsymbol{\mu}^T)}{D}\end{aligned}$$