Assignment 1

Gurpreet Singh - 150259

Question 1

Part A

MLE Estimate

Likelihood function on the distribution (X) is given as

$$\begin{aligned} p(X|\lambda) &= \prod_{i=1}^{N} p(x_i|\lambda) \\ &= \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= e^{-N\lambda} \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!} \end{aligned}$$

We can find the MLE of λ by maximizing the Log Likelihood Function $(log(p(X|\lambda))))$

$$\begin{aligned} \lambda_{MLE} &= \arg\max_{\lambda} p(X|\lambda) \\ &= \arg\max_{\lambda} log(p(X|\lambda)) \end{aligned}$$

In order to maximize, we differentiate wrt λ

$$\begin{aligned} \frac{d(log(p(X|\lambda)))}{d\lambda} &= \frac{d(-N\lambda + log(\lambda)\sum_{i=1}^{N} x_i - log(\prod_{i=1}^{N} x_i!))}{d\lambda} = 0 \\ &= -N + \frac{\sum_{i=1}^{N} x_i}{\lambda} = 0 \end{aligned}$$

$$\implies \lambda = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{X}$$

Therefore, the $\lambda_{MLE} = \bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$

MAP Estimate

The prior on λ is given by $p(\lambda) = Gamma(\lambda; \alpha, \beta)$. The posterior on λ using Bayes rule can be written as

$$p(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{\int p(X|\lambda)p(\lambda)d\lambda}$$

MAP estimate of λ is given by

$$\begin{aligned}\lambda_{MAP} &= \arg\max_{\lambda} p(\lambda|X) \\ &= \arg\max_{\lambda} \frac{p(X|\lambda)p(\lambda)}{p(X)} \\ &= \arg\max_{\lambda} p(X|\lambda)p(\lambda) \\ &= \arg\max_{\lambda} log(p(X|\lambda)p(\lambda))\end{aligned}$$

In order to compute $\arg\max_{\lambda} log(p(X|\lambda)p(\lambda))$, we differentiate the quantity wrt λ and equate it to zero

$$\begin{aligned}\frac{d(log(p(X|\lambda)p(\lambda)))}{d\lambda} &= \frac{d(log(p(X|\lambda))}{d\lambda} + \frac{d(log(p(\lambda)))}{d\lambda} = 0 \\ &= -N + \frac{\sum_{i=1}^{N} x_i}{\lambda} - \beta + \frac{\alpha - 1}{\lambda} = 0\end{aligned}$$

$$\implies \lambda = \frac{\alpha - 1 + \sum_{i=1}^{N} x_i}{N + \beta}$$

Therefore, $\lambda_{MAP} = \frac{\alpha - 1 + \sum_{i=1}^{N} x_i}{N+\beta}$

Part B

The posterior distribution of λ is given by $p(\lambda|X)$

$$p(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{\int p(X|\lambda)p(\lambda)d\lambda}$$

From the previous part, we know

$$p(X|\lambda) = e^{-N\lambda} \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N} x_i!}$$

Therefore

$$\begin{aligned}p(X|\lambda)p(\lambda) &= \frac{\beta^{\alpha}}{\Gamma(\lambda; \alpha)\prod_{i=1}^{N} x_i!} e^{-(N+\beta)\lambda} \lambda^{\alpha - 1 + \sum_{i=1}^{N} x_i} \\ &= \frac{\beta^{\alpha}\Gamma(\lambda; \alpha + \sum_{i=1}^{N} x_i)}{(N+\beta)^{\alpha}\Gamma(\lambda; \alpha)\prod_{i=1}^{N} x_i!} Gamma(\lambda; \alpha + \sum_{i=1}^{N} x_i, N + \beta)\end{aligned}$$

Hence, we can find out the integral

$$\int p(X|\lambda)p(\lambda)d\lambda = \frac{\beta^\alpha \Gamma(\lambda; \alpha + \sum_{i=1}^N x_i)}{(N+\beta)^\alpha \Gamma(\lambda;\alpha) \prod_{i=1}^N x_i!}$$
$$\implies p(\lambda|X) = Gamma(\lambda; \alpha + \sum_{i=1}^N x_i, N+\beta)$$

Hence the posterior distribution of λ is $Gamma(\lambda; \alpha + \sum_{i=1}^N x_i, N+\beta)$

Part C

From the previous parts, we have the posterior and estimates

$$\begin{aligned} p(\lambda|X) &= Gamma(\lambda; \alpha + \sum_{i=1}^N x_i, N+\beta) \\ \lambda_{MLE} &= \frac{\sum_{i=1}^N x_i}{N} \\ \lambda_{MAP} &= \frac{\alpha - 1 + \sum_{i=1}^N x_i}{N+\beta} \end{aligned}$$

Mode of the posterior

$$\begin{aligned} \arg\max_\lambda p(\lambda|X) &= \lambda_{MAP} \\ &= \frac{\alpha - 1 + \sum_{i=1}^N x_i}{N+\beta} \\ &= \frac{\alpha - 1 + N\lambda_{MLE}}{N+\beta} \\ &= \frac{N}{N+\beta}\lambda_{MLE} + \frac{\alpha-1}{N+\beta} \\ &= \frac{N}{N+\beta}\lambda_{MLE} + \frac{\beta}{N+\beta} mode(p(\lambda)) \end{aligned}$$

Mean of the posterior

$$\begin{aligned} E[\lambda|X] &= \int \lambda \cdot p(\lambda|X) d\lambda \\ &= \int \lambda \cdot Gamma(\lambda; \alpha + \sum_{i=1}^N x_i, N+\beta) d\lambda \\ &= \int \lambda \cdot Gamma(\lambda; \alpha + N\lambda_{MLE}, N+\beta) d\lambda \\ &= \frac{\alpha + N\lambda_{MLE}}{N+\beta} \\ &= \frac{N}{N+\beta}\lambda_{MLE} + \frac{\alpha}{N+\beta} \\ &= \frac{N}{N+\beta}\lambda_{MLE} + \frac{\beta}{N+\beta} mean(p(\lambda)) \end{aligned}$$

Hence, both Posterior mode and mean can be represented as a weighted sum of the MLE estimate and the Prior's mode and mean respectively.

Question 2

Looking at the variance of the posterior predictive distribution

$$
\begin{aligned}
\sigma_N^2 &= \beta^{-1} + x_*^T \Sigma_N x_* \\
\sigma_{N+1}^2 &= \beta^{-1} + x_*^T \Sigma_{N+1} x_* \\
\\
\sigma_{N+1}^2 - \sigma_N^2 &= x_*^T (\Sigma_{N+1} - \Sigma_N) x_*
\end{aligned}
$$

Therefore, the trend of σ^2 with N can be observed by observing the above quantity

$$
\begin{aligned}
\Sigma_N &= (\beta \sum_{i=1}^{N} x_n x_n^T + \lambda I)^{-1} \\
\Sigma_{N+1} &= (\beta \sum_{i=1}^{N+1} x_n x_n^T + \lambda I)^{-1} \\
&= ((\beta \sum_{i=1}^{N} x_n x_n^T + \lambda I) + x_{N+1} x_{N+1}^T)^{-1}
\end{aligned}
$$

Let $M = (\beta \sum_{i=1}^{N} x_n x_n^T + \lambda I)$ and $v = x_{N+1}$

$$
\begin{aligned}
\implies \Sigma_{N+1} &= (M + vv^T)^{-1} \\
&= M^{-1} - \frac{(M^{-1}v)(v^T M^{-1})}{1 + v^T M^{-1} v} \\
&= M^{-1} - \frac{(M^{-1}v)(v^T M^{-1})}{1 + v^T M^{-1} v}
\end{aligned}
$$

According to the definition of M, $M^{-1} = \Sigma_N$

$$
\implies \Sigma_{N+1} = \Sigma_N - \frac{(\Sigma_N v)(v^T \Sigma_N)}{1 + v^T \Sigma_N v}
$$

Also, $\Sigma_N = \Sigma_N^T$, hence we can rewrite the above as

$$
\implies \Sigma_{N+1} = \Sigma_N - \frac{(\Sigma_N v)(\Sigma_N v)^T}{1 + v^T \Sigma_N v}
$$

Since Σ_N is positive semi-definite, $v^T \Sigma_N v$ will be positive

Cleary $(\Sigma_N v)(\Sigma_N v)^T$ is positive semi-definite, and hence $\frac{(\Sigma_N v)(\Sigma_N v)^T}{1+v^T \Sigma_N v}$ is positive semi-definite

$$
\begin{aligned}
\implies \sigma_{N+1}^2 - \sigma_N^2 &= x_*^T (-\frac{(\Sigma_N v)(\Sigma_N v)^T}{1 + v^T \Sigma_N v} x_* \\
&= -x_*^T (\frac{(\Sigma_N v)(\Sigma_N v)^T}{1 + v^T \Sigma_N v} x_* \\
\\
&< 0 \text{ if } x_* \neq \mathbf{0} \\
&= 0 \text{ if } x_* = \mathbf{0}
\end{aligned}
$$

Therefore, on increasing the number of observations, the variance decreases if x_* is non-zero, otherwise remains the same.

Question 3

Part A

Note: We consider x_{ni} for $i = 1...D$ to be i.i.d.

Case 1: $x_n \in R^D$

We can take $x|y = k$ to be a multivariate Gaussian Distribution, i.e.

$$
\begin{aligned}
p(x|y = k) &= \mathcal{N}(x|\mu_k, \Sigma_k) \\
\therefore p(x|y = k) &= \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} exp(-\frac{1}{2}(x - \mu_k)^T\Sigma^{-1}(x - \mu_k))
\end{aligned}
$$

Case 2: $x_n \in \{0, 1\}^D$

We can take $x|y = k$ to be a multivariate Bernoulli Distribution, i.e.

$$
\begin{aligned}
p(x|y = k) &= \prod_{i=1}^{D} Bernoulli(x_i; \boldsymbol{p}_{kd}) \\
&= \prod_{i=1}^{D} \boldsymbol{p}_{kd}^{x_i}(1 - \boldsymbol{p}_{kd})^{x_i}
\end{aligned}
$$

Case 3: $x_n \in \{1, 2...V\}^D$

We can take $x|y = k$ to be a multivariate Multinoulli Distribution, i.e.

$$
\begin{aligned}
p(x|y = k) &= \prod_{i=1}^{D} Multinoulli(x_i; \boldsymbol{p}_k) \\
\boldsymbol{p}_k \in [0, 1]^D &\text{ such that } \sum_{i=1}^{D} \boldsymbol{p}_{ki} = 1
\end{aligned}
$$

Part B

We have already defined the class conditional distributions in the previous part. We also assume the class prior to be multinoulli distribution i.e.

$$
p(y|\boldsymbol{\pi}) = \prod_{k=1}^{K} \boldsymbol{\pi}_k^{\mathbb{I}[y=k]}
$$

Using these distributions, from Bayes rule, we can find the generative classification model $p(y = k|x, \theta, \boldsymbol{\pi})$. Here, θ is the set of parameters in the class conditional distributions.

$$
\begin{aligned}
p(y=k|x,\theta,\boldsymbol{\pi}) &= \frac{p(y=k)p(x|y=k)}{p(x)} \\
&= \frac{p(y=k)p(x|y=k)}{\sum_{i=1}^{K} p(y=i)p(x|y=i)}
\end{aligned}
$$

We can find the MLE estimates of the parameters θ and $\boldsymbol{\pi}$ by maximizing the log $p(\mathcal{D}|\theta,\boldsymbol{\pi})$

$$
\begin{aligned}
p(\mathcal{D}|\theta,\boldsymbol{\pi}) &= \prod_{n=1}^{N} p(y_n, x_n|\theta,\boldsymbol{\pi}) \\
&= \prod_{n=1}^{N} p(x_n|y_n,\theta)p(y_n|\boldsymbol{\pi}) \\
log(p(\mathcal{D},\theta,\boldsymbol{\pi})) &= \sum_{n=1}^{N} log(p(x_n|y_n,\theta)p(y_n|\boldsymbol{\pi})) \\
&= \sum_{n=1}^{N} log(p(x_n|y_n,\theta)) + log(p(y_n|\boldsymbol{\pi})) \\
\implies log(p(\mathcal{D},\theta,\boldsymbol{\pi})) &= \sum_{k=1}^{K}\sum_{n:y_n=k} log(p(x_n|y_n,\theta)) + \sum_{n=1}^{N}\sum_{k=1}^{K} \mathrm{II}[y_n=k] log(\boldsymbol{\pi}_k)
\end{aligned}
$$

Since in the above equation, θ and $\boldsymbol{\pi}$ are independent, we can find the MLE estimate of $\boldsymbol{\pi}$ independent of the class conditional

Using the constraint on $\boldsymbol{\pi}$, we can write

$$
\boldsymbol{\pi}_K = 1 - \sum_{k=1}^{K-1} \boldsymbol{\pi}_k
$$

For the MLE estimate of $\boldsymbol{\pi}_k$ for $k \in 1...K-1$

$$
\frac{d(log(p(\mathcal{D}|\theta,\boldsymbol{\pi})))}{d\boldsymbol{\pi}_k} = \sum_{n=1}^{N} \mathrm{II}[y_n=k]\frac{1}{\boldsymbol{\pi}_k} - \sum_{n=1}^{N} \mathrm{II}[y_n=K]\frac{1}{1-\sum_{k=1}^{K-1}\boldsymbol{\pi}_k} = 0
$$

Let N_k be the number of points such that $y_n = k$, then

$$
\frac{d(log(p(\mathcal{D}|\theta,\boldsymbol{\pi})))}{d\boldsymbol{\pi}_k} = \frac{N_k}{\boldsymbol{\pi}_k} - \frac{N_K}{\boldsymbol{\pi}_K} = 0
$$

From this, we can say

$$\boldsymbol{\pi} = \{\frac{N_k}{N}\}_{k=1}^{K}$$

Now, we can do case wise MLE analysis of θ

Case 1

Here, $\theta = \{\mu_k, \Sigma_k\}_{k=1}^{K}$ and $p(x|y=k) = \mathcal{N}(x|\mu_k, \Sigma_k)$.

In order to find the MLE estimate of θ, we need to maximize $p(\mathcal{D}|\theta, \boldsymbol{\pi})$ or maximize $log(p(\mathcal{D}|\theta, \boldsymbol{\pi}))$

$$\begin{aligned}
log(p(\mathcal{D}|\theta, \boldsymbol{\pi})) &= \sum_{k=1}^{K} \sum_{n:y_n=k} log(p(x_n|y_n, \theta)) + \sum_{n=1}^{N} \sum_{k=1}^{K} \mathrm{II}[y_n = k] log(\boldsymbol{\pi}_k) \\
&= \sum_{k=1}^{K} \sum_{n:y_n=k} log(\frac{1}{(\sqrt{(2\pi)^D \Sigma_k})} exp(-\frac{1}{2}(x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k))) + \sum_{n=1}^{N} \sum_{k=1}^{K} \mathrm{II}[y_n = k] log(\boldsymbol{\pi}_k)
\end{aligned}$$

MLE estimate of $\boldsymbol{\mu_k}$

$$\begin{aligned}
\frac{d(log(p(\mathcal{D}|\theta, \boldsymbol{\pi})))}{d\mu_k} &= \sum_{n:y_n=k} \frac{d(-\frac{1}{2}(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{d\mu_k} = 0 \\
&\implies \sum_{n:y_n=k} \Sigma^{-1}(x_n - \mu_k) = 0 \\
&\implies \mu_k = \frac{1}{N_k} \sum_{n:y_n=k} x_n
\end{aligned}$$

MLE estimate of $\boldsymbol{\Sigma_k}$

$$\begin{aligned}
\frac{d(log(p(\mathcal{D}|\theta, \boldsymbol{\pi})))}{d\Sigma_k} &= \sum_{n:y_n=k} \frac{d(-\frac{1}{2}(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{d\mu_k} = 0 \\
&\implies -\frac{1}{2}\Sigma_k^{-1} \sum_{n:y_n=k} (\boldsymbol{I} - (x_n - \mu_k)(x_n - \mu_k)_k^{T-1}) = 0 \\
&\implies \Sigma_k = \frac{1}{N_k} \sum_{n:y_n=k} (x_n - \mu_k)(x_n - \mu_k)^T
\end{aligned}$$

Case 2

Here, $\theta = \{\boldsymbol{p}_k\}_{k=1}^{K}$ and $p(x|y=k) = \prod_{d=1}^{D} Bernoulli(x_i; \boldsymbol{p}_{kd})$. In this case, we assume x_{nd} to be a binary vector representing 0 or 1. Therefore, any element x will be a matrix of dimensions $D \times 2$. Similarly, $\boldsymbol{p}_k$ will be a matrix of Dimension $D \times 2$, and $\boldsymbol{p}_{kij} = P(x_{nij} = 1 \text{ and } x_{nik} = 0)(k \neq j)$

In order to find the MLE estimate of θ, we need to maximize $p(\mathcal{D}|\theta, \boldsymbol{\pi})$ or maximize $log(p(\mathcal{D}|\theta, \boldsymbol{\pi}))$

$$\begin{aligned}
log(p(\mathcal{D}|\theta, \boldsymbol{\pi})) &= \sum_{k=1}^{K} \sum_{n:y_n=k} log(p(x_n|y_n, \theta)) + \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] log(\boldsymbol{\pi}_k) \\
&= \sum_{k=1}^{K} \sum_{n:y_n=k} \sum_{d=1}^{D} (x_{nd} log(\boldsymbol{p}_{kd}) + (1 - x_{nd}) log(1 - \boldsymbol{p}_{kd})) + \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] log(\boldsymbol{\pi}_k)
\end{aligned}$$

MLE estimate of $\boldsymbol{p}_k$

$$\begin{aligned}
\frac{d(log(p(\mathcal{D}|\theta, \boldsymbol{\pi})))}{d\boldsymbol{p}_{kd}} &= \sum_{n:y_n=k} \frac{d(\sum_{d=1}^{D} (x_{nd} log(\boldsymbol{p}_k) + (1 - x_{nd}) log(1 - \boldsymbol{p}_k)))}{d\mu_k} = 0 \\
&\implies \sum_{n:y_n=k} \frac{x_{nd} - \boldsymbol{p}_{kd}}{\boldsymbol{p}_{kd}(1 - \boldsymbol{p}_{kd})} = 0 \\
&\implies \boldsymbol{p}_{kd} = \frac{1}{N_k} \sum_{n:y_n=k} x_{nd} \\
&\implies \boldsymbol{p}_k = \frac{1}{N_k} \sum_{n:y_n=k} x_n
\end{aligned}$$

Case 3

Here, $\theta = \{\boldsymbol{p}_k\}_{k=1}^{K}$ and $\prod_{i=1}^{D} Multinoulli(x_i; \boldsymbol{p}_k)$. In this case, we consider x and $\boldsymbol{p}_k$ to be $D \times V$ dimensional matrices with each row of x as the transpose of a binary vector and $\sum_{i=1}^{V} \boldsymbol{p}_{kdi} = 1$ (for all $d \in \{1...D\}$)

MLE estimate of $\boldsymbol{p}_k$

If we follow the same steps as we followed in case 2 (combined with the MLE analysis of $\boldsymbol{\pi}$), we can get the MLE estimate of p_k as

$$\boldsymbol{p}_k = \frac{1}{N_k} \sum_{n:y_n=k} x_n$$

Question 4

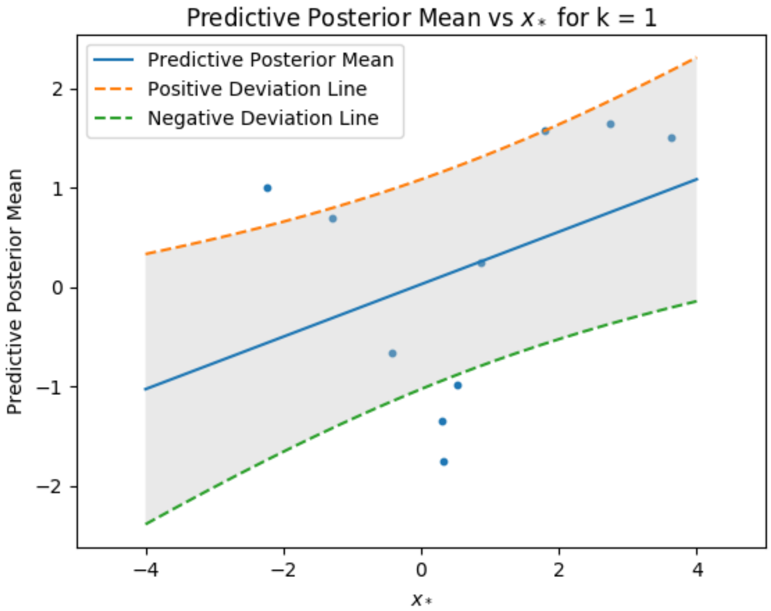
Part A

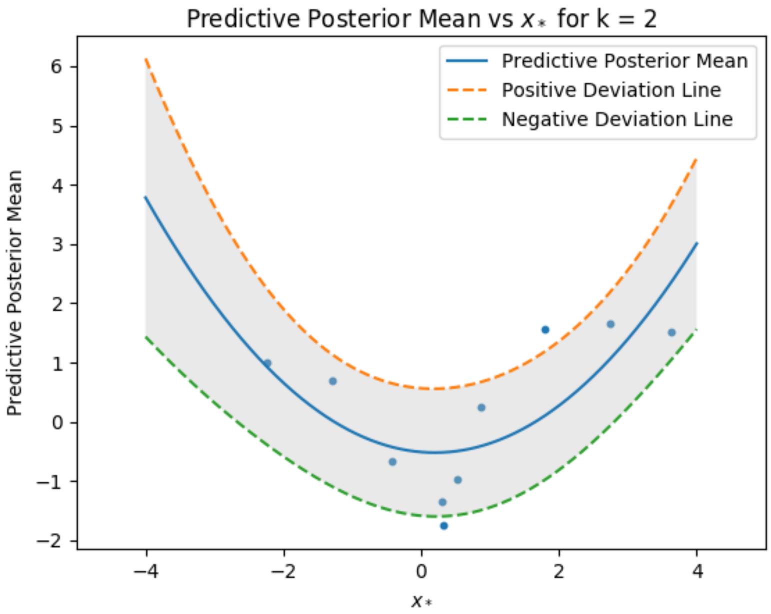
We know that the posterior predictive distributions is of the form

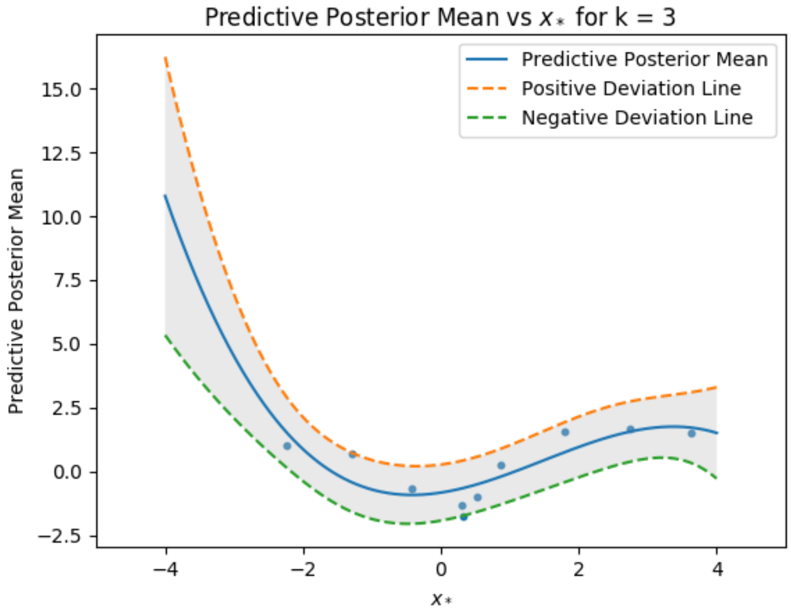
$$
\begin{aligned}
p(y_*|\phi(x)_*, \phi(\boldsymbol{X}), \boldsymbol{y}, \beta, \lambda) = \mathcal{N}(x_*|\mu^T x_*, \beta^{-1} + x_*^T \Sigma x_*) \\
\mu = (\phi(\boldsymbol{X})^T \phi(\boldsymbol{X}) + \frac{\lambda}{\beta}\boldsymbol{I_D})^{-1} \phi(\boldsymbol{X})^T \boldsymbol{y} \\
\Sigma = (\beta \phi(\boldsymbol{X})^T \phi(\boldsymbol{X}) + \lambda \boldsymbol{I_D})^{-1}
\end{aligned}
$$

$\beta \boldsymbol{I_D}$ and $\lambda \boldsymbol{I_D}$ are the covariance matrices of the likelihood and the prior distributions

We need to plot $\mu^T x_*$ for all $x_* \in [-4, 4]$







Part B

Marginal Likelihood is defined as

$$\begin{aligned} p(\boldsymbol{y}|\phi(\boldsymbol{X}), \beta, \lambda) &= \int p(\boldsymbol{y}|\phi(\boldsymbol{X}), \boldsymbol{w}, \beta, \lambda) p(\boldsymbol{w}|\lambda) d\boldsymbol{w} \\ &= \mathcal{N}(\boldsymbol{y}|0, \beta^{-1}\boldsymbol{I}_D + \lambda^{-1}\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T) \end{aligned}$$

Putting in the values, we get

Marginal Likelihood for $k = 1 :$ $8.90631979852e^{-15}$

Marginal Likelihood for $k = 2 :$ $1.28878221246e^{-10}$

Marginal Likelihood for $k = 3 :$ $2.57739780619e^{-10}$

Clearly, k = 3 is the best fit model for this data distribution. This is also evident from the plot between Predictive Posterior Mean and x_*

Part C

We need to improve the fit, that is reduce the variance in the posterior distribution. Hence we need to minimize Σ. Since $\Sigma = (\lambda \boldsymbol{I} +^T X)^{-1}$, our aim is to increase the value of $X^T X$. Now looking at the posterior predictive distribution, the standard deviation is directly proportional to $X^T X$. Hence the point where the standard deviation is highest should be taken as the new point. Therefore we take the point where the standard deviation is highest as the next training example. The point is $x' = 4$ for all ks.