Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 1

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

$$p(x|\gamma) = \int p(x|\eta)p(n|\gamma)d\eta$$

$$p(\eta|\gamma) = \frac{\gamma^2}{2}\exp(-\frac{\eta\gamma^2}{2}) \quad [= \text{Exp}(\eta|\frac{\gamma^2}{2})]$$

$$p(x|\eta) = \frac{1}{\sqrt{2\pi\eta}}\exp(-\frac{x^2}{2\eta}) \quad [= \mathcal{N}(x|0,\eta)]$$

$$\implies p(x|\gamma) = \frac{\gamma^2}{2\sqrt{2\pi}}\int_0^\infty \frac{\exp(-\frac{x^2}{2\eta} - \frac{\eta\gamma^2}{2})}{\sqrt{\eta}}d\eta$$

M.G.F of the marginal p.d.f :

$$\mathbb{E}_{x|\gamma}\left[e^{tx}\right] = \frac{\gamma^2}{2\sqrt{2\pi}}\int_{-\infty}^\infty\int_0^\infty \frac{\exp(tx - \frac{x^2}{2\eta} - \frac{\eta\gamma^2}{2})}{\sqrt{\eta}}d\eta dx$$

Exchanging order of integrals and integrating out x first, we can couple terms to complete the whole square in the exponential to obtain a normal distribution constant $\sqrt{2\pi\eta}$ to be multiplied.

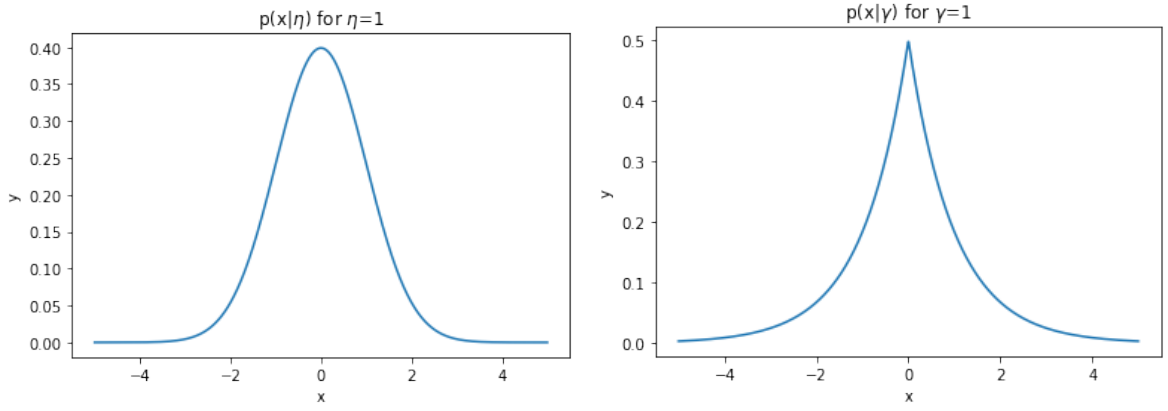
$$\implies M_{x|\gamma}(t) = \frac{\gamma^2}{2}\int_0^\infty \exp(\frac{\eta}{2}(t^2 - \gamma^2))d\eta$$

$$M_{x|\gamma}(t) = \begin{cases} \frac{1}{1-\frac{t^2}{\gamma^2}}, & \text{if } \gamma^2 > t^2 \\ \infty, & \text{otherwise} \end{cases}$$

By observation, the M.G.F obtained is similar to that of a Laplace distribution $L(\mu, b)$ with $\mu = 0$ and $b = \gamma^{-1}$. Therefore, the marginal distribution is the Laplace distribution $L(0, \gamma^{-1})$,

$$p(x|\gamma) = \frac{\gamma}{2}\exp(-\gamma|x|)$$

The plots for the Gaussian and Laplace (marginal) distributions, for $\gamma = \eta = 1$, are as follows:



One obvious difference between the two plots is that the Laplace distribution has a much sharper peak at the mean value ($x = 0$) and thus has more mass at near zero values than the Gaussian distribution.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION

2

Sherman Morrison Formula:

$$(M + uv^T)^{-1} = M^{-1} - \frac{(M^{-1}u)(v^T M^{-1})}{1 + v^T M^{-1} u}$$

Where u and v are column vectors and M is a square matrix

$$\sum_N = (\beta \sum_{n=1}^{N} x_n x_n^T + \lambda \mathbf{I})^{-1}$$

$$= (\beta \sum_{n=1}^{N-1} x_n x_n^T + \lambda \mathbf{I} + \beta x_N x_N^T)^{-1}$$

Taking $u = \beta x_n$ and $v = x_N^T$ in the formula, we get

$$\sum_N = (\beta \sum_{n=1}^{N-1} x_n x_n^T + \lambda \mathbf{I})^{-1} - \frac{((\beta \sum_{n=1}^{N-1} x_n x_n^T + \lambda \mathbf{I})^{-1} \beta x_N)(x_N^T (\beta \sum_{n=1}^{N-1} x_n x_n^T + \lambda \mathbf{I})^{-1}}{1 + x_N^T (\beta \sum_{n=1}^{N-1} x_n x_n^T + \lambda \mathbf{I})^{-1} \beta x_N}$$

It is clear by inspection that the second term on the R.H.S of the above equation is a positive definite matrix since $x_n x_n^T, \beta$(inverse of a covariance matrix) and $\lambda \mathbf{I}$ are all positive definite. If we keep expanding the new M ($\sum_{N-1}$) in a similar manner, we end up with the following:

$$\sum_N = \lambda^{-1}\mathbf{I} - (\text{N positive definite terms})$$

Clearly, $\sum_N$ decreases as N is increased. The variance of the predictive posterior is defined as:

$$\sigma_N^2(x_*) = \beta^{-1} + x_*^T \sum_N x_*$$

Since β^{-1} is constant and independent of N, for a given x_*, the variance decreases with the decrease in $\sum_N$ as the training set size N increases. This also makes sense intuitively because if the model has more training data, it will be able to make better and more accurate predictions.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION 3

Since all x_n's are drawn i.i.d from $\mathcal{N}(\mu, \sigma^2)$,

$$p(\bar{x}) = p(\frac{1}{N}\sum_{n=1}^{N} x_n) = \frac{1}{N}\sum_{n=1}^{N} p(x_n)$$

We know that the sum of two independent normally distributed random variables also follows a normal distribution, where its mean is the sum of the two means and its variance is the sum of the two variances. That is if,

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$
$$\implies Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Therefore,

$$p(\bar{x}) = \frac{1}{N}\mathcal{N}(N\mu, N\sigma^2)$$
$$\implies \bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

This probability distribution for the empirical mean of Gaussian observations makes intuitive sense because,

1. The mean of the distribution of the empirical mean should be the mean of the original normal distribution itself since the empirical mean will ideally be equal to it for $N \to \infty$.
2. The variance of the empirical mean should decrease with increase in N since the empirical mean of multiple observations will be equal to the mean itself as $N \to \infty$ (where $\sigma^2 \to 0$). The probability distribution gets more and more weighted at and around the mean, and the variance decreases.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

QUESTION

4

1. Taking the data for one school as a single observation (empirical mean $\bar{x}^{(m)}$) and using the result of question 3, we can write,

$$p(\bar{x}^{(m)}|\mu_m) = \mathcal{N}(\mu_m, \frac{\sigma^2}{N_m})$$

Since the prior of μ_m is also a Gaussian, the posterior distribution of μ_m will also be a Gaussian,

$$p(\mu_m|\bar{x}^{(m)}, \mu_0, \sigma_0^2) = \mathcal{N}(\mu_w, \sigma_w^2) \text{ where,}$$

$$\frac{1}{\sigma_w^2} = \frac{N_m}{\sigma^2} + \frac{1}{\sigma_0^2} \quad \text{and} \quad \mu_w = \frac{\sigma^2}{N_m\sigma_0^2 + \sigma^2}\mu_0 + \frac{N_m\sigma_0^2}{N_m\sigma_0^2 + \sigma^2}\bar{x}^{(m)}$$

2. Since all $\bar{x}^{(m)}$ are independent,

$$p(\mathbf{x}|\mu_0, \sigma^2, \sigma_0^2) = \int p(\mathbf{x}|\mu, \mu_0, \sigma^2, \sigma_0^2)p(\mu|\mu_0, \sigma^2, \sigma_0^2)d\mu$$

$$= \prod_{m=1}^{M} \int p(\bar{x}^{(m)}|\mu_m)p(\mu_m|\mu_0, \sigma^2, \sigma_0^2)d\mu_m$$

Note that the term under the integral is simply the marginal likelihood with respect to the posterior we derived in part 1 of this question. Since,

$$\text{marginal} = \frac{\text{prior} \times \text{likelihood}}{\text{posterior}}$$

$$p(\mathbf{x}|\mu_0, \sigma^2, \sigma_0^2) = \prod_{m=1}^{M} \frac{\mathcal{N}(\mu_m|\mu_0, \sigma_0^2)\mathcal{N}(\bar{x}^{(m)}|\mu_m, \frac{\sigma^2}{N_m})}{\mathcal{N}(\mu_m|\mu_w, \sigma_w^2)}$$

M.L.E-II estimate should comes out to be,

$$\hat{\mu_0} = \arg\max_{\mu} p(\mathbf{x}|\mu_0, \sigma^2, \sigma_0^2)$$

$$\implies \hat{\mu_0} = \frac{\sum_{m=1}^{M} \frac{N_m\sigma_0^2}{N_m\sigma_0^2+\sigma^2}\bar{x}^{(m)}}{\sum_{m=1}^{M} \frac{N_m\sigma_0^2}{N_m\sigma_0^2+\sigma^2}}$$

3. Using the MLE-II estimate is beneficial because it is able to incorporate the data we have from all the schools and use that to tune the value of μ_0 to best fit our data and make better predictions. The mean of posterior derived previously can be seen as a convex combination of MLE estimate and mean of prior. Hence, if our prior mean is more accurate then our posterior mean will be more accurate as well.

$$p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m) = \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}\mathbf{w}_m, \beta^{-1}\mathbf{I}_N) \text{ and } p(\mathbf{w}_m) = \mathcal{N}(\mathbf{w}_m|\mathbf{w}_0, \lambda^{-1}\mathbf{I}_D)$$

The marginal likelihood can be written as,

$$p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}) = \int p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}, \mathbf{w}_m)p(\mathbf{w}_m)d\mathbf{w}_m = \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}\mathbf{w}_0, \mathbf{X}\lambda^{-1}\mathbf{X}^T + \beta^{-1})$$

Considering all M schools, we have to maximize the objective function $\prod_{m=1}^{M} p(\mathbf{y}^{(m)}|\mathbf{X}^{(m)})$ w.r.t $\mathbf{w}_0$. Taking log,

$$\hat{\mathbf{w}}_0 = \arg\max_{\mathbf{w}_0} \sum_{m=1}^{M} \log \mathcal{N}(\mathbf{y}^{(m)}|\mathbf{X}\mathbf{w}_0, \mathbf{X}\lambda^{-1}\mathbf{X}^T + \beta^{-1}\mathbf{I}_N)$$

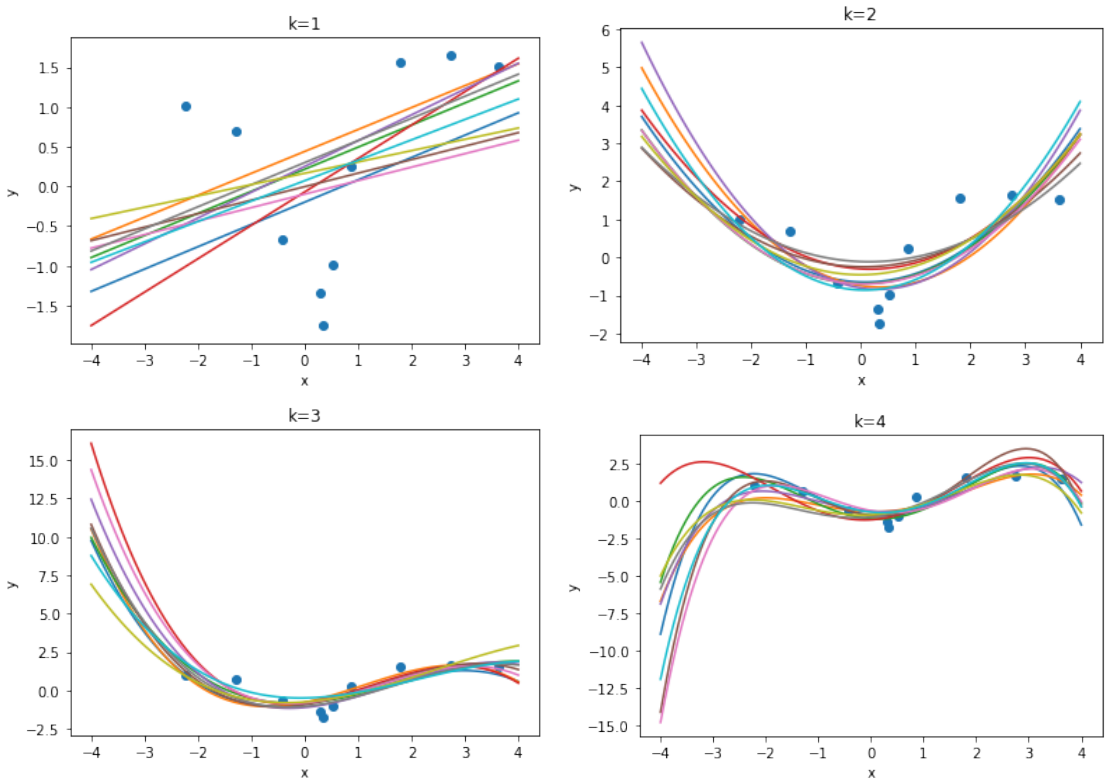
One benefit of this approach is that by maximizing the marginal likelihood, we are tuning the hyperparameter $\mathbf{w}_0$ such that it best fits and explains the data we have. It will give more accurate draws of $\mathbf{w}_m$ for our given data and thus make better predictions.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

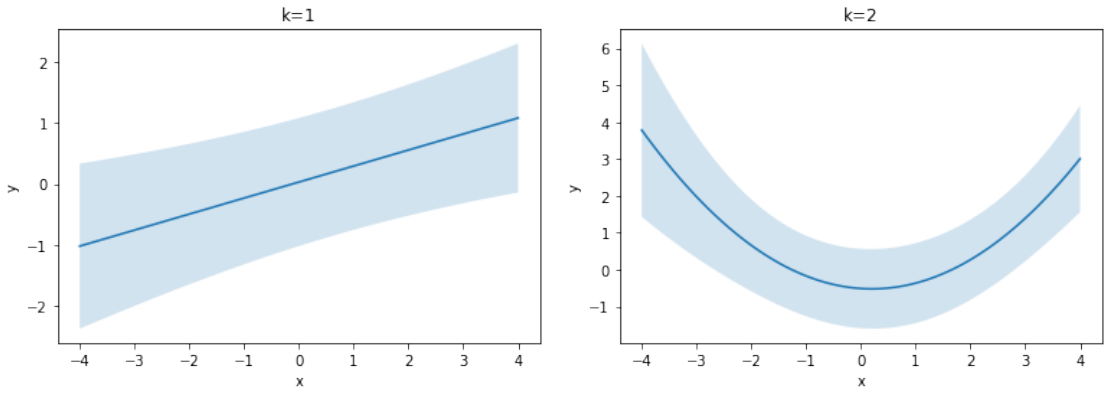
Student Name: Ankur Banga
Roll Number: 180108
Date: June 9, 2021

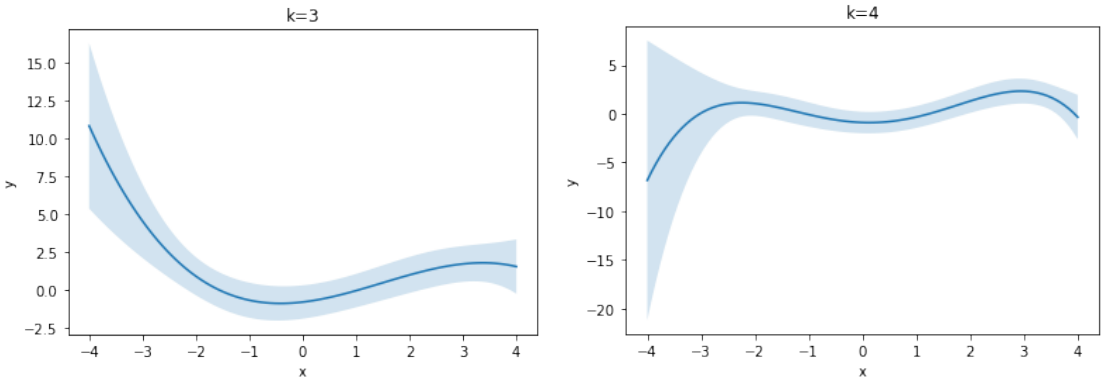
QUESTION 6

1. The plots for each k are as follows:



2. The plots for each k are as follows:





3. The log marginal likelihood of the training data, $\log p(y|\phi_k(x), \beta)$:
   k=1: -32.352
   k=2: -22.772
   k=3: -22.079
   k=4: -22.387

   Model 3 seems to explain the data best since it has the highest log likelihood.

4. The log likelihood of the training data using the MAP estimate, $\log p(y|\boldsymbol{w}_{MAP}, \phi_k(x), \beta)$:
   k=1: -28.094
   k=2: -15.36
   k=3: -10.936
   k=4: -7.225

   Model 3 has the highest log likelihood according to part 3 and model 4 has the highest log likelihood according to part 4. The marginal log likelihood is a better criteria to select the best model since it it averages over all possible values of the parameter provided by the prior. The MAP estimate fits the best w according to the model selected and may overfit the data. Marginal likelihood provides a reasonable explanation for the data itself irrespective of the parameter.

5. For our best model i.e. model 3, we can improve the model by taking an additional training input at $x' \approx -4$ since it is clear from the plot in part 2 that the uncertainty or variance in our prediction is much higher in that region as compared to the other regions in [-4,4].