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

QUESTION 1

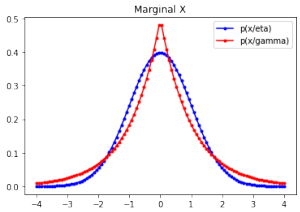
Student Name: Ajita Shree
Roll Number: 20111262
Date: February 26, 2021

When You Integrate Out..

Consider x is a scalar random variable drawn from a uni-variate Gaussian $p(x/\eta) = N(x/0, \eta)$, η is drawn from an exponential distribution $p(\eta/\gamma) = Exp(\eta|\gamma^2/2)$. Exponential distribution is defined as Exp(x—λ) = $\lambda exp(\lambda x)$

Derivation of $p(x/\eta) = \int p(x/\eta)p(\eta/\gamma)d\eta$

- $p(x/\eta) = \int N(x/0, \eta)Exp(\eta/\gamma^2/2)d\eta$
- $= p(x/\eta) = \int \frac{1}{\sqrt{2\pi\eta}} exp(\frac{-x^2}{2\eta})\frac{\gamma^2}{2}exp(\frac{-\gamma^2\eta}{2}))d\eta$
- $M^t(x) = \int_{-\infty}^{\infty} e^{tx}p(x/\eta)dx$
- $M^t(x) = \int_{-\infty}^{\infty} e^{tx} \int \frac{1}{\sqrt{2\pi\eta}} exp(\frac{-x^2}{2\eta}) + \frac{\gamma^2}{2}exp(\frac{-\gamma^2\eta}{2})d\eta dx$
- $M_x(t) = \int \frac{1}{\sqrt{2\pi\eta}}\frac{\gamma^2}{2}\int_{-\infty}^{\infty} exp(tx - x^2/2\eta - \gamma^2\eta/2)dxd\eta$
- Using identity $\int_{-\infty}^{\infty} exp(-ax^2 + bx + c)dx = \sqrt{(\frac{\pi}{a})}exp(\frac{b^2}{4a} + c)$, $M_x(t)$ can be written as
- $M^t(x) = \int \frac{1}{\sqrt{(2\pi\eta)}}\frac{\gamma^2}{2}\sqrt{(2\pi\eta)}exp(\frac{-\gamma^2\eta}{2} + \frac{t^2}{2})d\eta$
- $M^t(x) = -exp(\frac{t^2}{2} - \frac{\gamma^2\eta}{2})$ The form is similar to the Laplace distributions's, $L(\mu, b)$, $M^t(x)$, which is $\frac{e^{t\mu}}{1 - b^2t^2}$



- **The marginal distribution**, $p(x/\eta)$ here mean that probability distribution of x given conditioned on η parameter; when the parameters of x distribution are also drawn from an alternate distribution, marginal distribution gives the probability of x given the last parameters without referring to other in between parameters.Above diagram shows the plot of $p(x/\eta)$ and $p(x/\gamma)$;

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

QUESTION 2

Student Name: Ajita Shree
Roll Number: 20111262
Date: February 26, 2021

It Gets Better

Consider a Bayesian Linear Regression model with likelihood $p(y/x, w) = N(W^T x, \beta^{-1})$ and prior $N(0, \lambda^{-1}I)$, the predictive posterior is $p(y*/x*) = N(\mu_N^T x_*, \beta^{-1} + x^T \Sigma_N x_*) = N(\mu_N^T x_*, \sigma_N^2)$ where $\sigma_N^2 = \beta^{-1} + x^T \Sigma_N x_*$

As the training set size N increases, what happens to the variance of the predictive posterior? Does it increase or decrease or remain the same? μ_N, Σ_N are mean and variance of Gaussian posterior where $\Sigma_N = ({}^T X+)^{-1}$ and $\mu_N = \Sigma_N \beta X^T Y$

Solution

- Eq1) PPD Variance = $\beta^{-1} + x_*^T(\beta X^T X + \lambda I_D)^{-1} x_*$
- Using identity, $(I_D + A_{D*N} B_{N*D})^{-1} = I_D - A_{D*N}(I_N + B_{N*D} A_{D*N})^{-1} B_{N*D}$
- Eq 1 can be written as follows
- $\beta^{-1} + x_*^T(\lambda I_D - \beta X^T(\lambda I_N + \beta X X^T)^{-1} X) x_*$
- Expanding the above term, we have
- $\beta^{-1} + x_*^T(\lambda I_D) x_* - x_*^T \beta X^T(\lambda I_N + \beta X X^T)^{-1} X x_*$
- First two terms i.e. $\beta^{-1} + x_*^T(\lambda I_D) x_*$ will remain constant w.r.t. N, hence no effect of increase in N here
- In last term, $X^T(\lambda I_N + \beta X X^T)^{-1} X$ (Eq2) term will increase will an increase in N, no of training samples but this is in negative
- How? Using identity $(I + AB)^{-1} A = A(I + BA)^{-1}$, we can rewrite, eq 2 as follows
- $\beta X^T X(\lambda I_D + \beta X^T X)^{-1}$; Now using $(I + A)^{-1} = A(A + I)^{-1}$, we have
- $\beta(X^T X)(\beta X^T X)^{-1}({}^T X)^{-1} + \lambda I_D)^{-1} \longrightarrow((\beta X^T X)^{-1} + \lambda I_D)^{-1}$
- Using identity $(A^{-1} + B^{-1})^{-1} = A(A + B)^{-1} B$, we will finally have
- $\lambda^{-1} I_D(\beta X^T X + \lambda^{-1} I_D)^{-1}(\beta X^T X)$
- The above term is positive semi-definite, hence, with an increase in example will grow further and will further decrease the PPD variance.
- Hence, it has been seen formally that the **PPD's variance will decrease with an increase in N** training samples.

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

Student Name: Ajita Shree
Roll Number: 20111262
Date: February 26, 2021

QUESTION
3

Distribution of Empirical Mean of Gaussian Observations

Consider N scalar valued observations $x_1, x_2, ...x_n$, drawn from i.i.d from $N(\mu, \sigma^2)$
Representing empirical mean $\bar{x} = \frac{1}{N}\sum_{n \in N} x_n$ as the linear transformation of a random variable derive the probability distribution (mean, sigma) of $\bar{x}$

Solution

The **mean** for $\bar{x}$ is derived as follows:

1. $E[\bar{x}] = \frac{1}{N}E[x_1 + x_2 + ...x_n]$
2. Using property, $E[\bar{x}] = \frac{1}{N}[E[x_1] + ...E[x_n]]$
3. $E[\bar{x}] = \frac{\mu.N}{N} = \mu$

The **variance** for $\bar{x}$ is derived as follows:

1. $Cov[\bar{x}] = \frac{1}{N}Cov[x_1 + x_2 + ...x_n]$
2. Using property, $Cov[\bar{x}] = [Cov[\frac{x_1}{N}] + ...Cov[\frac{x_n}{N}]]$
3. $Cov[\bar{x}] = \frac{\sigma^2}{N^2} + ...\frac{\sigma^2}{N^2}$
4. $Cov[\bar{x}] = \frac{\sigma^2}{N}$

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

QUESTION 4

Student Name: Ajita Shree
Roll Number: 20111262
Date: February 26, 2021

Benefits of probabilistic Joint modeling-1

Consider a dataset of test-scores of students from M schools in a district: $x = x_{(m)}M = x^m, ..., x^{(m)}M$, where N_m denotes the number of students in school m. Assume the scores of students in school m are drawn independently as x(m): $N(\mu_m, \sigma^2)$ where the Gaussian's mean μ_m is unknown and the variance σ^2 is same for all schools and known (for simplicity). Assume the means $\mu_1, ..., \mu_M$ of the M Gaussians to also be Gaussian distributed $\mu_M : N(\mu_0, \sigma 0^2)$ where μ_0 and $\sigma 0^2$ are hyperparameters.

Solution a)

Derive the posterior distribution of m and write down the mean and variance of this posterior distribution, $p(\mu_m / X^{(m)}, \sigma^2)$

- For a particular school m, the Likelihood, $p(x^{(m)}/\mu_m, \sigma^2) = \prod_{n \in N_m} p(x_n^{(m)}/\mu_m, \sigma^2)$
- Posterior $= \frac{Likelihood.Prior}{Marginal} \propto Likelihood * Prior = \prod_{n \in N_m} N(x_n^{(m)}/\mu_m, \sigma^2) N(\mu_m/\mu_0, \sigma^2)$
- $p(\mu_m / X^{(m)}, \sigma^2) = \prod_{n \in N_m} exp(\frac{-(x_n^{(m)} - \mu_m)^2}{2\sigma^2}) exp(\frac{-(\mu_m - \mu_0)^2}{2\sigma_0^2})$
- Using completing the sqaures trick, it can written as $N(\mu_M, \sigma_M^2)$, where μ_M and σ_M are paramters of posterior.
- $\mu_M = \frac{\sigma^2 \mu_0}{\sigma^2 + N_m \sigma_0^2} + \frac{N_m \sigma_0^2 \bar{x}^{(m)}}{\sigma^2 + N_m \sigma_0^2}$, where $\mathrm{x}^{(m)} = \frac{\sum_{n \in N_m} x_n^{(m)}}{N_m}$
- $\frac{1}{\sigma_M^2} = \frac{N_m}{\sigma} + \frac{1}{\sigma_0^2}$

Solution b)

Assume the hyper-parameter μ_0 to be unknown. Derive the marginal likelihood $p(x/\mu_0, \sigma_2, \sigma_0^2)$ and use MLE-II to estimate μ_0

- $p(x/\mu_0, \sigma_2, \sigma_0^2) = \int p(x/\mu_m, \sigma^2) p(\mu_m/\mu_0, \sigma_0^2) d\mu_m$
- Alternatively, we also know, Marginal likelihood $= \frac{Likelihood * Prior}{Posterior}$
- Likelihood term, $p(x/\mu, \sigma^2) = \prod_{m \in M} \prod_{n \in N_M} P(x_n^{(m)}/\mu_m, \sigma^2)$, Prior term $= p(\mu_m/\mu_0, \sigma_0^2)$
- $p(x/\mu_0, \sigma_2, \sigma_0^2) = \prod_{m \in M} \frac{\prod_{n \in N_M} P(x_n^{(m)}/\mu_m, \sigma^2) * p(\mu_m/\mu_0, \sigma_0^2)}{p(\mu_m / X^{(m)}, \sigma^2)}$

- MLE-2 , $\mu_0 = argmax_{\mu_0} p(x/\mu_0, \sigma, \sigma_0^2) \implies \mu_0 = argmin_{\mu_0} - logp(x/\mu_0, \sigma, \sigma_0^2)$
- $argmin_{\mu_0} - log \prod_{m \in M} \frac{\prod_{n \in N_M} Exp(-(x_n^{(m)} - \mu_m)^2/2\sigma^2) * Exp(-(\mu_m - \mu_0)^2, 2\sigma_0^2)}{Exp(-(\mu_m - X^{(m)})^2, 2\sigma^2)}$
- $argmin_{\mu_0} \sum_{m \in M} \left[\frac{\sum_{n \in N_m} (x_n^{(m)} - \mu_m)^2}{2\sigma^2} + \frac{(\mu_m - \mu_0)^2}{2\sigma_0^2} - \frac{(\mu_m - \mu_M)^2}{2\sigma_M^2} \right]$
- differentiating wrt μ_0 and equating it with 0, we have
- $\sum_{m \in M} \mu_0 = \sum_{m \in M} \mu_M$; Plugging the value of μ_M
- $\sum_{m \in M} \mu_0 = \sum_{m \in M} \left(\frac{\sigma^2 \mu_0}{\sigma^2 + N_m \sigma_0^2} + \frac{N_m \sigma_0^2 \bar{x}_{(m)}}{\sigma^2 + N_m \sigma_0^2} \right)$
- Rearranging the terms and solving it
- Now, $\mu_0 = \frac{\sum_{m \in M} \bar{x}^{(m)}}{M}$; after plugging the value of $\bar{x}^{(m)} as \frac{\sum_{n \in N_m} x_n^{(m)}}{N}$,
- $\mu_0 = \frac{\sum_{m \in M} \frac{\sum_{n \in N_m} x_m^{(m)}}{N_m}}{M}$

Solution c)

After substituting MLE-2 μ_0 in μ_M, we have

- $\mu_M = \frac{\sigma^2 \left[\frac{\sum_{m \in M} \bar{x}^{(m)}}{M} \right]}{\sigma^2 + N_m \sigma_0^2} + \frac{N_m \sigma_0^2 \bar{x}^{(m)}}{\sigma^2 + N_m \sigma_0^2}$, where $\bar{x}^{(m)} = \frac{\sum_{n \in N_m} x_n^{(m)}}{N_m}$
- **Benefit** of using MLE-II estimate as opposed to using a known value is that we are able to use the data to learn the hyper-parameter. It can be seen in the above example that μ_0 term has turned out to be the empirical mean over all the training samples available across schools and over all the students. Prior can have some bias and may not account for scenarios if we have representation of variety of schools (say, music, dance, education etc in the training data).

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

QUESTION 5

Student Name: Ajita Shree
Roll Number: 20111262
Date: February 26, 2021

Benefits of probabilistic Joint modeling-2

Assume a linear regression model for these scores, i.e., $p(y^(m)|x^(m), w) = N(y^(m)|w^{x^(m),\beta^1)}$, where $w_m \in R^D$ denotes the regression weight vector for school m, and β is known. Note that this can also be denoted as $p(y^{(m)}|X^{(m)}, w_m) = N(y^{(m)}|X^{(m)}w_m, \beta^1 I_N)$; Assume a $priorp(w_m) = N(w_m|w_0, \lambda I_D)$, to be known and w_0 to be unknown.

Solution

Derive the expression for the log of the MLE-II objective for estimating $w_0, p(y/x, \beta, \lambda, w_0)$

- $p(y/x, \beta, \lambda, w_0) = \int p(y/x, w_m, \beta)p(w_m/\lambda w_0)dw_m$
- Alternatively, we also know, Marginal likelihood $= \frac{Likelihood * Prior}{Posterior}$, Eq1)
- Likelihood term, $p(x/\mu, \sigma^2) = \prod_{m\in M}\prod_{n\in N_M} N(y_n^{(m)}/x_n^{(m)}, w_m, \beta)$
- Prior term $= N(w_m/w_0, \lambda^{-1}I_D)$
- Posterior term $= p(w_m/x^m, \beta, \lambda) = \prod_{m\in M} N(w_m/\mu_M, \Sigma_M)$
- where $\Sigma_M = (\beta X^T X + \lambda I_D)^{-1}, \mu_M = (X^T X + \lambda/\beta I_D)^{-1}X^T y$
- Substituting values in equation 1 for doing MLE-2
- $p(y/x, \beta, \lambda, w_0) = argmax_{w_0} \prod_{m\in M} \frac{\prod_{n\in N_m} N\left(\frac{y_n^{(m)}}{w_m^T x_n^{(m)}, \beta^{-1}}\right) N\left(\frac{w_m}{w_0, \lambda^{-1}I_D}\right)}{N\left(\frac{w_m}{\mu_M, \Sigma_M}\right)}$
- MLE-2 , $w_0 = argmax_0 p(y/x, \beta, \lambda, w_0) \Longrightarrow w_0 = argmin_0 - logp(y/x, \beta, \lambda, w_0)$
- $argmin_{\mu_0} \sum_{m\in M}\left[\frac{\beta\sum_{n\in N_m}(y_n^{(m)} - w_m^T x_m^{(m)})^2}{2} + \frac{\lambda(w_m - w_0)^2}{2} - \frac{\Sigma_M(w_m - \mu_M)^2}{2}\right]$
- **Benefit** of this approach as opposed to fixing w_0 to some value, if our goal is to learn the school-specific weight vectors $w1, ..., wM$? **Ans** Here w_0 value is not dependent on any prior belief, one of the biggest benefit would be w_0 will be learned based on the trends/patterns in the training set consisting on m schools which is effective when we are learning school specific weight vector

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

Student Name: Ajita Shree
Roll Number: 20111262
Date: February 26, 2021

QUESTION
6

Solution

The Bayesian linear regression model is learned on $\phi_k(x) = [1, x, x^2, ..., x^k]^T$

- **Compute Posterior of w / Plot of 10 random functions from the inferred posterior**

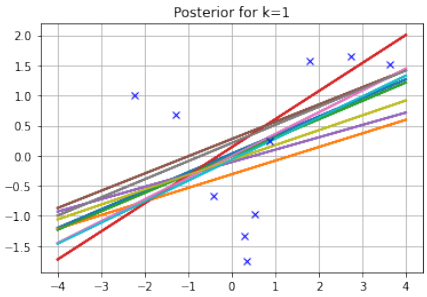


Figure 1: *
K = 1

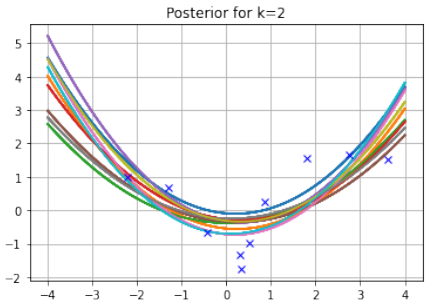


Figure 2: *
K = 2

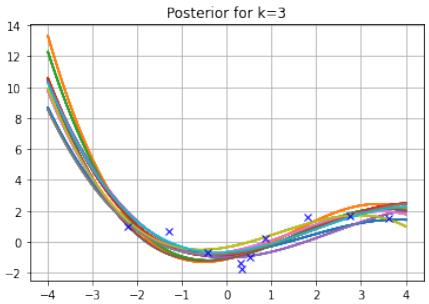


Figure 3: *
K = 3

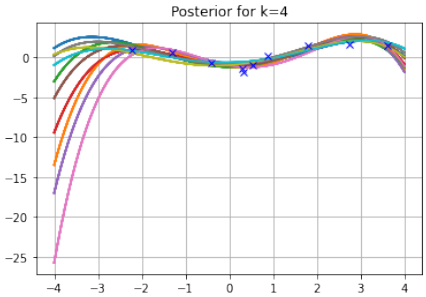


Figure 4: *
K = 4

- **Compute and plot the mean of posterior predictive on $\in -4, 4$ and a dotted line with mean +- 2 times standard deviation**

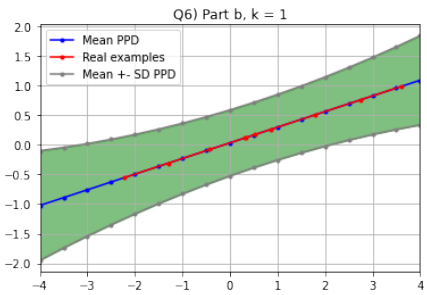


Figure 5: *
K = 1

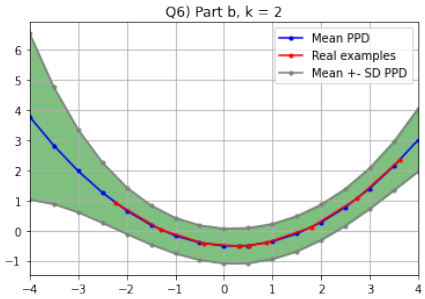


Figure 6: *
K = 2

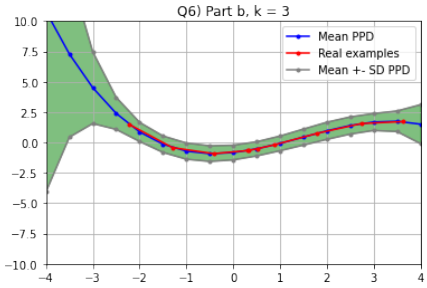


Figure 7: *
K = 3

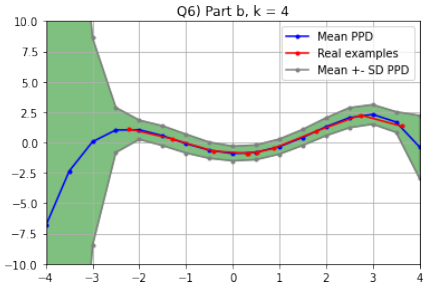


Figure 8: *
K = 4

- **c)Log marginal likelihood of the training data**
  - K = 1: -32.35 and K = 2: -22.77
  - K = 3: -22.07 and K = 4: -22.38
- **d)Log likelihood of the MAP estimate**
  - K = 1: -28.09 and K = 2: -15.35
  - K = 3: -10.9 and K = 4: -7.22
  - K = 4 has the highest log likelihood
  - Is answer same as part3? The answer is different as shown in the table above; Log-likelihood results see an increase with increase in K
  - The criteria to select the best model is highest log marginal likelihood because as we go to k = 4 from k = 3 (part b), we see that variance in the prediction has slightly gone up, this is indicative of uncertainty); Marginal likelihood is capturing this whereas log likelihood does not seem to capture this and give the max for k = 4.
- **e)Region to choose for adding new data point** is [-4, -3] as we can clearly see in plots above because of low data presence, we have a large variance leading to more uncertainty.