ECE 271A: Homework #4

Due on November 6, 2023 at 11:59pm

Professor Vasconcelos

Ray Tsai
A16848188

Problem 1

Bayesian regression: in last week's problem set we showed that various forms of linear regression by the method of least squares are really just particular cases of ML estimation under the model

$$\mathbf{z} = \mathbf{\Phi}\theta + \epsilon,$$

where $\mathbf{z} = (z_1, \ldots, z_n)^T$, $\theta = (\theta_1, \ldots, \theta_k)^T$

$$\mathbf{\Phi} = \begin{bmatrix} 1 & \ldots & x_1^K \\ \vdots & \ddots & \vdots \\ 1 & \ldots & x_n^K \end{bmatrix}$$

and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is a normal random process $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. It seems only natural to consider the Bayesian extension of this model, an extension that has been the subject of some recent research under the denomination of *Gaussian processes*. For this, we simply extend the model considering a Gaussian prior

$$P_\theta(\theta) = \mathcal{G}(\theta, \mathbf{0}, \mathbf{\Gamma}).$$

Part A

Given a training set $\mathcal{D} = \{(\mathcal{D}_x, \mathcal{D}_z)\} = \{(x_1, z_1), \ldots, (x_n, z_n)\}$, compute the posterior distribution

$$P_{\theta|T}(\theta|\mathcal{D})$$

and the predictive distribution

$$P_{z|T}(z|\mathcal{D}).$$

Solution

We know

$$\begin{aligned} P_{\theta|T}(\theta|\mathcal{D}) &= P_{\theta|T_x,T_z}(\theta|\mathcal{D}_x, \mathcal{D}_z) \\ &= \frac{P_{T_z|\theta,T_x}(\mathcal{D}_z|\theta, \mathcal{D}_x) P_{\theta|T_x}(\theta|\mathcal{D}_x)}{\int P_{T_z|\theta,T_x}(\mathcal{D}_z|\theta, \mathcal{D}_x) P_{\theta|T_x}(\theta|\mathcal{D}_x)\, d\theta} \\ &= \frac{P_{T_z|\theta,T_x}(\mathcal{D}_z|\theta, \mathcal{D}_x) P_\theta(\theta)}{\int P_{T_z|\theta,T_x}(\mathcal{D}_z|\theta, \mathcal{D}_x) P_\theta(\theta)\, d\theta}. \end{aligned}$$

Since $P_{T_z|\theta,T_x}(\mathcal{D}_z|\theta, \mathcal{D}_x) = \mathcal{G}(\mathbf{z}, \mathbf{\Phi}\theta, \mathbf{\Sigma})$

$$\begin{aligned} P_{\theta|T}(\theta|\mathcal{D}) &\propto P_{T_z|\theta,T_x}(\mathcal{D}_z|\theta, \mathcal{D}_x) P_\theta(\theta) \\ &\propto \exp\left\{ -\frac{1}{2} \left[(\mathbf{z} - \mathbf{\Phi}\theta)^T \mathbf{\Sigma}^{-1} (\mathbf{z} - \mathbf{\Phi}\theta) + \theta^T \mathbf{\Gamma}^{-1} \theta \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[\theta^T (\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi} + \mathbf{\Gamma}^{-1}) \theta - 2\theta^T \mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{z} \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left(\theta^T \mathbf{\Sigma}_\theta^{-1} \theta - 2\theta^T \mathbf{\Sigma}_\theta^{-1} \mu_\theta \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[(\theta - \mu_\theta)^T \mathbf{\Sigma}_\theta^{-1} (\theta - \mu_\theta) \right] \right\}, \end{aligned}$$

where $\mathbf{\Sigma}_\theta = (\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1}$ and $\mu_\theta = \mathbf{\Sigma}_\theta \mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{z} = (\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1} \mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{z}$. Therefore,

$$P_{\theta|T}(\theta|\mathcal{D}) = \mathcal{G}(\theta, (\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1} \mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{z}, (\mathbf{\Phi}^T \mathbf{\Sigma}^{-1} \mathbf{\Phi} + \mathbf{\Gamma}^{-1})^{-1}).$$

Problem 1 continued on next page...

Let $\xi = \phi^T\theta$. Then, the predictive distribution

$$\begin{aligned} P_{z|T,x}(z|\mathcal{D},x) &= \int P_{z|\xi,x}(z|\xi,x)P_{\xi|T}(\xi|\mathcal{D})d\xi \\ &= \int \mathcal{G}(z,\phi^T\theta,\sigma(x)^2)\mathcal{G}(\phi^T\theta,\phi^T\mu_\theta,\phi^T\mathbf{\Sigma}_\theta\phi)\,d\xi \\ &= \mathcal{G}(z,0,\sigma(x)^2) * \mathcal{G}(z,\phi^T\mu_\theta,\phi^T\mathbf{\Sigma}_\theta\phi) \\ &= \mathcal{G}(z,\phi^T\mu_\theta,\sigma(x)^2+\phi^T\mathbf{\Sigma}_\theta\phi). \end{aligned}$$

Part B

Consider the MAP estimate

$$\theta_{MAP} = \arg\max_\theta P_{\theta|\mathbf{T}}(\theta|\mathcal{D}).$$

How does it differ from the weighted least squares estimate? What is the role of the terms that were not present in the latter? Is there any advantage in setting them to anything other than zero?

Solution

Since $P_{\theta|T}(\theta|\mathcal{D}) = \mathcal{G}(\theta,(\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi}+\mathbf{\Gamma}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z},(\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi}+\mathbf{\Gamma}^{-1})^{-1})$ is simply a gaussian distribution, the map estimate $\theta_{MAP} = (\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi}+\mathbf{\Gamma}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z}$. From last week, we get that the weighted least equares estimate is $\theta_{ML} = (\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{\Sigma}^{-1}\mathbf{z}$. Notice that the weighted least squares estimate is missing the term $\mathbf{\Gamma}^{-1}$ in the $\mathbf{\Sigma}_\theta$ term of θ_{MAP}. $\mathbf{\Gamma}^{-1}$ plays the role of regularization for the model. An example of the advantage of setting $\mathbf{\Gamma}^{-1}$ to a non-zero value is that it allows us to adjust the importance of each order of the polynomial terms.

Part C

Consider the case in which prior covariance $\mathbf{\Gamma}$ is a diagonal matrix, not necessarily the identity. Suppose that you are told that K, i.e. the number of parameters in θ or the degree of the polynomial $\phi(x)^T\theta$, is somewhere between 1 and 25. How would you set up $\mathbf{\Gamma}$ and why? Discuss the implications of your selection on the bias and variance of your MAP solution

$$z_{MAP} = \mathbf{\Phi}(x)\theta_{MAP}.$$

Solution

Since a polynomial of degree 25 is going to be extremely "wiggly," it is highly likely that it would end up overfitting the training data, so I would heavily bias against the higher orders and favor the lower orders. Therefore, I would set a monotonicity decreasing sequence along the diagonal of $\mathbf{\Gamma}$, where the last few entries are close to 0. In terms of the prior distribution $P_\theta(\theta) = \mathcal{G}(\theta,\mathbf{0},\mathbf{\Gamma})$, doing so it basically say that the first few entries of θ is of a wide range of numbers depending on the data, but the last few are most likely going to be 0. Since we are removing the flexibility of the higher order terms of the polynomial z_{MAP}, the variance would get lower but the bias would get higher.

Problem 2

In this problem we explore the exponential family and conjugate priors. The exponential family is the family of densities of the form

$$P_{\mathbf{X}|\theta} = f(\mathbf{x})g(\theta)e^{\phi(\theta)^T u(\mathbf{x})}$$

with

$$[g(\theta)]^{-1} = \int f(\mathbf{x})e^{\phi(\theta)^T ux`(\mathbf{x})}d\mathbf{x}.$$

Part A

Show that, for a density in this family, the likelihood of a sequence $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is

$$P_{\mathbf{T}|\theta} \propto \prod_{i=1}^{n} f(\mathbf{x}_i) \exp\left\{\phi(\theta)^T \sum_{i=1}^{n} u(\mathbf{x}_i)\right\}.$$

What is the normalization constant?

Solution

Since

$$\begin{aligned} P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) &= \prod_{i=1}^{n} P_{\mathbf{X}|\theta}(\mathbf{x}_i|\theta) \\ &= \prod_{i=1}^{n} f(\mathbf{x}_i)g(\theta) \exp\left\{\phi(\theta)^T u(\mathbf{x}_i)\right\} \\ &= g(\theta)^n \left[\prod_{i=1}^{n} f(\mathbf{x}_i)\right] \exp\left\{\phi(\theta)^T \sum_{i=1}^{n} u(\mathbf{x}_i)\right\}, \end{aligned}$$

we know

$$P_{\mathbf{T}|\theta} \propto \prod_{i=1}^{n} f(\mathbf{x}_i) \exp\left\{\phi(\theta)^T \sum_{i=1}^{n} u(\mathbf{x}_i)\right\},$$

the normalization constant is $g(\theta)^n$.

Part B

It has been shown that, apart from certain irregular cases, the exponential family is the only family of distributions for which there is a conjugate prior. Show that

$$P_\theta(\theta) = \frac{g(\theta)^\eta e^{\phi(\theta)^T \nu}}{\int g(\theta)^\eta e^{\phi(\theta)^T \nu} d\theta}$$

is a conjugate prior for the exponential family and compute the posterior distribution $P_{\theta|\mathbf{T}}(\theta|\mathcal{D})$. Denoting $\mathbf{s} = \sum_{i=1}^{n} u(\mathbf{x}_i)$ as the *sufficient statistic*, compare the posterior with prior density. What is the result of "propagating" the prior through the likelihood function?

Solution

Since

$$\begin{aligned} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &= P_{\mathbf{T}|\theta}(\mathcal{D}|\theta)P_\theta(\theta) \\ &\propto g(\theta)^\eta e^{\phi(\theta)^T\nu} \cdot g(\theta)^n \left[\prod_{i=1}^n f(\mathbf{x}_i)\right] \exp\left\{\phi(\theta)^T \sum_{i=1}^n u(\mathbf{x}_i)\right\} \\ &\propto g(\theta)^{\eta+n} \exp\left\{\phi(\theta)^T \left(\nu + \sum_{i=1}^n u(\mathbf{x}_i)\right)\right\}, \end{aligned}$$

We know

$$P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) = \frac{g(\theta)^{\eta+n} \exp\left\{\phi(\theta)^T(\nu+\mathbf{s})\right\}}{\int g(\theta)^{\eta+n} \exp\left\{\phi(\theta)^T(\nu+\mathbf{s})\right\} d\theta},$$

and so the posterior is of the same form as the prior, with η replaced as $\eta+n$ and ν replaced as $\nu+\mathbf{s}$. Hence, P_θ is indeed a conjugate prior for the exponential family. The result could be viewed as updating the prior with the newly observed data. We can think of η as the virtual sample size and n as the sample size of the training set. Similarly, ν can be think of as the value of the virtual sample and $\mathbf{s}$ as the actual data from the training set.

Part C

Consider table 1. For each row i) show that the likelihood function on the left column belongs to the exponential family, ii) show that the prior on the left column is a conjugate prior for the likelihood function on the right column, iii) compute the posterior $P_{\theta|T}(\theta|\mathcal{D})$, and iv) interpret the meaning of the sufficient statistic and the "propagation" discussed in part B.

(i) We show that each of the following likelihood functions belongs to the exponential family.

Bernoulli

Since

$$\begin{aligned} \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= \exp\left\{\log(\theta)\sum_{i=1}^n x_i + \log(1-\theta)\sum_{i=1}^n 1-x_i\right\} \\ &= \exp\left\{\log(1-\theta)n + \log\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^n x_i\right\} \\ &= (1-\theta)^n \exp\left\{\log\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^n x_i\right\}, \end{aligned}$$

it indeed belongs to the exponential family, with $g(\theta) = (1-\theta)$, $f(x) = 1$, $\phi(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$, and $u(x) = x$.

Poisson

Since

$$\begin{aligned} \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} &= \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \\ &= \frac{e^{-n\theta}}{\prod_{i=1}^n x_i!} \exp\left\{\log\theta \sum_{i=1}^n x_i\right\}, \end{aligned}$$

it indeed belongs to the exponential family, with $g(\theta) = e^{-\theta}$, $f(x) = \frac{1}{x!}$, $\phi(\theta) = \log\theta$, and $u(x) = x$.

Exponential

Since

$$\prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i},$$

it indeed belongs to the exponential family, with $g(\theta) = \theta$, $f(x) = 1$, $\phi(\theta) = -\theta$, and $u(x) = x$.

Normal

Since

$$\prod_{i=1}^{n} \sqrt{\frac{\theta}{2\pi}} \exp\left\{-\frac{\theta}{2}(x_i - \mu)^2\right\} = \sqrt{\frac{\theta}{2\pi}}^{n} \exp\left\{-\frac{\theta}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\},$$

it indeed belongs to the exponential family, with $g(\theta) = \sqrt{\frac{\theta}{2\pi}}$, $f(x) = 1$, $\phi(\theta) = -\frac{\theta}{2}$, and $u(x) = (x-\mu)^2$.

(ii) We now show that the prior on the right column is a conjugate prior for the likelihood function on the left column.

Bernoulli

Since

$$\begin{aligned}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} &\propto \exp\left\{(\alpha-1)\log(\theta) + (\beta-1)\log(1-\theta)\right\} \\ &\propto (1-\theta)^{\alpha+\beta-2}\exp\left\{(\alpha-1)\log\left(\frac{\theta}{1-\theta}\right)\right\} \\ &\propto (1-\theta)^{\eta}\exp\left\{\log\left(\frac{\theta}{1-\theta}\right)\nu\right\},\end{aligned}$$

the Beta function is a conjugate prior for the likelihood function for the Bernoulli distribution, with $\eta = \alpha + \beta - 2$, $\nu = \alpha - 1$, $g(\theta) = 1 - \theta$, and $\phi = \log\frac{\theta}{1-\theta}$.

Poisson

Since

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \propto e^{-\eta\theta}\exp\left\{\log(\theta)\nu\right\},$$

the Gamma function is a conjugate prior for the likelihood function for the Poisson distribution, with $\eta = \beta$, $\nu = \alpha - 1$, $g(\theta) = e^{-\theta}$, and $\phi = \log\theta$.

Exponential

Since

$$\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \propto \theta^{\eta}\exp\left\{-\nu\theta\right\},$$

the Gamma function is a conjugate prior for the likelihood function for the Exponential distribution, with $\eta = \alpha - 1$, $\nu = \beta$, $g(\theta) = \theta$, and $\phi = -\theta$.

Problem 2 continued on next page...

Normal

Since

$$\begin{aligned}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} &\propto \left(\sqrt{\frac{\theta}{2\pi}}\right)^{2\alpha-2}\exp\left\{-\frac{\theta}{2}\cdot 2\beta\right\}\\ &\propto \left(\sqrt{\frac{\theta}{2\pi}}\right)^{\eta}\exp\left\{-\frac{\theta}{2}\cdot\nu\right\},\end{aligned}$$

the Gamma function is a conjugate prior for the likelihood function for the Normal distribution, with $\eta = 2\alpha - 2$, $\nu = 2\beta$, $g(\theta) = \sqrt{\frac{\theta}{2\pi}}$, and $\phi = -\frac{\theta}{2}$.

(iii) We now compute the posterior $P_{\theta|\mathbf{T}}(\theta|\mathcal{D})$ for each distribution.

Bernoulli

$$\begin{aligned}P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}\\ &\propto (1-\theta)^{\eta}\exp\left\{\log\left(\frac{\theta}{1-\theta}\right)\nu\right\}(1-\theta)^{n}\exp\left\{\log\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^{n}x_i\right\}\\ &= (1-\theta)^{\eta+n}\left\{\log\left(\frac{\theta}{1-\theta}\right)\left(\nu+\sum_{i=1}^{n}x_i\right)\right\}\\ &= (1-\theta)^{\alpha+\beta-2+n}\left\{\log\left(\frac{\theta}{1-\theta}\right)\left(\alpha-1+\sum_{i=1}^{n}x_i\right)\right\}.\end{aligned}$$

Poisson

$$\begin{aligned}P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!}\\ &\propto e^{-\eta\theta}\exp\{\log(\theta)\nu\}\frac{e^{-n\theta}}{\prod_{i=1}^{n}x_i!}\exp\left\{\log\theta\sum_{i=1}^{n}x_i\right\}\\ &= \frac{e^{-(\eta+n)\theta}}{\prod_{i=1}^{n}x_i!}\exp\left\{\log\theta\left(\nu+\sum_{i=1}^{n}x_i\right)\right\}\\ &= \frac{e^{-(\beta+n)\theta}}{\prod_{i=1}^{n}x_i!}\exp\left\{\log\theta\left(\alpha-1+\sum_{i=1}^{n}x_i\right)\right\}.\end{aligned}$$

Exponential

$$\begin{aligned}P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!}\\ &\propto \theta^{n}e^{-\theta\sum_{i=1}^{n}x_i}\theta^{\eta}\exp\{-\nu\theta\}\\ &= \theta^{\eta+n}\exp\left\{-\theta\left(\nu+\sum_{i=1}^{n}x_i\right)\right\}\\ &= \theta^{\alpha-1+n}\exp\left\{-\theta\left(\beta+\sum_{i=1}^{n}x_i\right)\right\}.\end{aligned}$$

Problem 2 continued on next page...

Normal

$$\begin{aligned}
P_{\theta|\mathbf{T}}(\theta|\mathcal{D}) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\sqrt{\frac{\theta}{2\pi}}\exp\left\{-\frac{\theta}{2}(x_i-\mu)^2\right\} \\
&\propto \left(\sqrt{\frac{\theta}{2\pi}}\right)^{\eta}\exp\left\{-\frac{\theta}{2}\cdot\nu\right\}\sqrt{\frac{\theta}{2\pi}}^{n}\exp\left\{-\frac{\theta}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
&= \left(\sqrt{\frac{\theta}{2\pi}}\right)^{\eta+n}\exp\left\{-\frac{\theta}{2}\left(\nu+\sum_{i=1}^{n}(x_i-\mu)^2\right)\right\} \\
&= \left(\sqrt{\frac{\theta}{2\pi}}\right)^{2\alpha-2+n}\exp\left\{-\frac{\theta}{2}\left(2\beta+\sum_{i=1}^{n}(x_i-\mu)^2\right)\right\}.
\end{aligned}$$

(iv) We now interpret the meaning of the sufficient statistic and the "propagation" for each distribution.

Bernoulli

The sifficient statistic here represents the number of tosses in $\mathcal{D}$ that results in 1. The propagation in this case represents the addition of virtual tosses we have. Of the total $\alpha+\beta-2$ tosses, $\alpha-1$ of them are 1 and $\beta-1$ of them are 0. Hence, the prior suggests that the chance to toss an 1 should be closer to $\frac{\alpha-1}{\alpha+\beta-2}$.

Poisson

The sifficient statistic here represents the sum of the number of times an event was triggered in $\mathcal{D}$. The propagation in this case represents addition of the virtual experiments that was done. Of the β virtual experiments, the total number of times an event was triggered is $\alpha-1$. Hence, the prior suggests that the rate of trigger should be closer to $\frac{\alpha-1}{\beta}$.

Exponential

The sufficent statistic here represents the sum of the time we waited in $\mathcal{D}$. The propagation in this case represents the addition of the virtual experiments that was done. Of the $\alpha-1$ virtual experiments, the total time we waited is β. Hence, the prior suggests that the wait time should be closer to $\frac{\beta}{\alpha-1}$.

Normal

The sufficent statistic here represents the sum of the variance between each sample in $\mathcal{D}$ and the mean. The propagation in this case represents the addition of the virtual sample. Of the $2\alpha-2$ virtual experiments, the sum of all variances is 2β. Hence, the prior suggests that the variance should be closer to $\frac{\beta}{\alpha-1}$.