CS181 Practice Questions: Linear Regression, Continued

1. **Posterior Weight Distribution Using Bayes' Rule for Linear Gaussian Systems**

Some background: In section (2.3.3), Bishop derives the following facts about linear Gaussian systems: assuming we have a marginal distribution on $\boldsymbol{x}$ and a conditional distribution on $\boldsymbol{y}$ given by

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \tag{2.114}$$

then

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\Sigma}(\boldsymbol{A}^\top \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}), \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^\top \boldsymbol{L} \boldsymbol{A})^{-1}. \tag{2.117}$$

Now, we know from (3.10) in Bishop that the regression likelihood can be written as

$$p(\boldsymbol{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^\top \phi(\boldsymbol{x}_n), \beta^{-1}),$$

where $\beta = \frac{1}{\sigma^2}$. If the prior distribution on $\boldsymbol{w}$ is given by $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \boldsymbol{S}_0)$, derive that the posterior distribution $p(\boldsymbol{w} \mid \boldsymbol{t})$ is given by

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta \boldsymbol{\Phi}^\top \boldsymbol{t})$$

$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$$

in the following way:

(a) Write down the likelihood in the form of a multivariate Gaussian.

(b) Explain what we need to substitute for $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{L}$ (respectively) in equations (2.113)-(2.117) to derive the posterior.

2. **Posterior Weight Distribution By Completing the Square (Bishop 3.7)**

We know from (3.10) in Bishop that the regression likelihood can be written as

$$\begin{aligned} p(\boldsymbol{t} \mid \boldsymbol{w}) &= \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^\top \phi(\boldsymbol{x}_n), \beta^{-1}) \\ &\propto \exp\left(-\frac{\beta}{2}(\boldsymbol{t} - \boldsymbol{\Phi w})^\top (\boldsymbol{t} - \boldsymbol{\Phi w})\right), \end{aligned}$$

where $\beta = \frac{1}{\sigma^2}$ and in the second line above we have ignored the Gaussian normalizing constants. By completing the square, show that with a prior distribution on $\boldsymbol{w}$ given by $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \boldsymbol{S}_0)$, the posterior distribution $p(\boldsymbol{w} \mid \boldsymbol{t})$ is given by

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\begin{aligned} \boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\top \boldsymbol{t}) \\ \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi} \end{aligned}$$

3. **Predictive Distribution**

Bishop notes in section (3.2.2) that if our prior distribution on $\boldsymbol{w}$ is

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\boldsymbol{I}),$$

and if we assume again that our likelihood involves an inverse variance parameter β, then the predictive distribution of t for a new datapoint $\boldsymbol{x}$ is given by

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta) \, d\boldsymbol{w} \tag{3.57}$$

Does the equation in (3.57) make any independence assumptions about the variables involved? If so, which?

4. **Deriving Lasso Regularization with Lagrange Multipliers**

Show that minimization of the unregularized sum-of-squares error function given by

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^\mathsf{T}\phi(\boldsymbol{x}_n))^2,$$

subject to the constraint

$$\sum_{j=1}^{M}|w_j| \leq \eta,$$

is equivalent to minimizing the regularized error function

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^\mathsf{T}\phi(\boldsymbol{x}_n))^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|$$

5. **Connection between Priors and Regularization**

Consider the Bayesian linear regression model given in Bishop 3.3.1. The prior is given by

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}),$$

where α is the precision parameter that controls the variance of the Gaussian prior. The likelihood can be written as

$$p(\boldsymbol{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n), \beta^{-1}),$$

Using the fact that the posterior is the product of the prior and the likelihood, show that maximizing the log posterior (i.e. $\ln p(\boldsymbol{w} \mid \boldsymbol{t}) = \ln p(\boldsymbol{w}|\alpha) + \ln p(\boldsymbol{t} \mid \boldsymbol{w})$) is equivalent to minimizing the regularized error term given by $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$ with

$$\begin{aligned} E_D(\boldsymbol{w}) &= \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}_n))^2 \\ E_W(\boldsymbol{w}) &= \frac{\lambda}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} \end{aligned}$$

Do this by writing $\ln p(\boldsymbol{w} \mid \boldsymbol{t})$ as a function of $E_D(\boldsymbol{w})$ and $E_W(\boldsymbol{w})$, dropping constant terms if necessary.
Conclude that maximizing this posterior is equivalent to minimizing the regularized error term given by $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$.
(Hint: take $\lambda = \alpha/\beta$)

6. **Bayesian Updates in Linear Regression, Bishop 3.8**

Suppose we have the standard Bayesian linear regression model and we have already observed N data points, so the posterior distribution is the same as the one derived in problem 2,

$$p(\boldsymbol{w} \mid \boldsymbol{t}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N)$$

where

$$\begin{aligned} \boldsymbol{m}_N &= \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{t}) \\ \boldsymbol{S}_N^{-1} &= \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} \end{aligned}$$

Suppose we observe a new data point $(\boldsymbol{x}_{N+1}, t_{N+1})$. Show that the resulting posterior distribution is of the same form with $\boldsymbol{m}_{N+1}$ and $\boldsymbol{S}_{N+1}$.