CS405 Homework #3

12110644 周思呈

Question 1

Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x}_n)\}^2. \tag{1}$$

Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.

Take the derivative of (3.104) with respect to $\boldsymbol{w}$ and set it equal to 0

$$\begin{aligned} \nabla E_D(\boldsymbol{w}) &= \sum_{n=1}^{N} r_n \left\{t_n - \boldsymbol{w}^T \boldsymbol{\Phi}\left(\boldsymbol{x}_n\right)\right\} \boldsymbol{\Phi}(\boldsymbol{x}_n)^T = 0 \\ &= 0 = \sum_{n=1}^{N} r_n t_n \boldsymbol{\Phi}(\boldsymbol{x}_n)^T - \boldsymbol{w}^T \left(\sum_{n=1}^{N} r_n \boldsymbol{\Phi}\left(\boldsymbol{x}_n\right) \boldsymbol{\Phi}(\boldsymbol{x}_n)^T\right) \end{aligned} \tag{2}$$

To achieve a similar form with (3.14), we denote $\sqrt{r_n}\boldsymbol{\phi}\left(\boldsymbol{x_n}\right) = \boldsymbol{\phi}'\left(\boldsymbol{x_n}\right)$ and $\sqrt{r_n} t_n = t_n'$

$$\begin{gathered} 0 = \sum_{n=1}^{N} t_n' \Phi'(\boldsymbol{x}_n)^T - \boldsymbol{w}^T \left(\sum_{n=1}^{N} \boldsymbol{\Phi}'\left(\boldsymbol{x}_n\right) \boldsymbol{\Phi}'(\boldsymbol{x}_n)^T\right) \\ \boldsymbol{w}_{ML} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{t} \end{gathered} \tag{3}$$

where

$$\begin{gathered} \boldsymbol{t} = \left[\sqrt{r_1} t_1, \sqrt{r_2} t_2, \ldots, \sqrt{r_N} t_N\right]^T \\ \boldsymbol{\Phi}(i, j) = \sqrt{r_i} \phi_j\left(x_i\right) \end{gathered} \tag{4}$$

If we substitute β^{-1} by $r_n \cdot \beta^{-1}$ in the summation term, (3.12) will be the same as (3.104).

r_n can be viewed as the observation time of $(\mathbf{x}_n, t_n)$.

Question 2

We saw in Section 2.3.6 that the conjugate prior for a Gaussian distribution with unknown mean and unknown precision (inverse variance) is a normal-gamma distribution. This property also holds for the case of the conditional Gaussian distribution $p(t|\mathbf{x}, \mathbf{w}, \beta)$ of the linear regression model. If we consider the likelihood function,

$$p(\mathbf{t}|\mathbf{X}, \mathrm{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathrm{w}^{\mathrm{T}}\phi(\mathrm{x}_n), \beta^{-1}) \tag{5}$$

then the conjugate prior for $\mathbf{w}$ and β is given by

$$p(\mathbf{w}, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0)\text{Gam}(\beta|a_0, b_0). \tag{6}$$

Show that the corresponding posterior distribution takes the same functional form, so that

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\text{Gam}(\beta|a_N, b_N). \tag{7}$$

and find expressions for the posterior parameters $\mathbf{m}_N$, $\mathbf{S}_N$, a_N, and b_N.

From (3.112) we have

$$\begin{aligned} p(\boldsymbol{w}, \beta) &= \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{m_0}, \beta^{-1}\mathbf{S_0}\right) \text{Gam}\left(\beta \mid a_0, b_0\right) \\ &\propto \left(\frac{\beta}{|\mathbf{S_0}|}\right)^2 \exp\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m_0})^T \beta \mathbf{S_0}^{-1}\left(\boldsymbol{w}-\boldsymbol{m_0}\right)\} b_0^{a_0} \beta^{a_0-1} \exp\{-b_0\beta\} \end{aligned} \tag{8}$$

Because

$$p(\boldsymbol{w}, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \mathbf{X}, \boldsymbol{w}, \beta) \times p(\boldsymbol{w}, \beta) \tag{9}$$

and we have

$$\begin{aligned} p(\mathbf{t} \mid \mathbf{X}, \boldsymbol{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right), \beta^{-1}\right) \\ &\propto \prod_{n=1}^{N} \beta^{1/2} \exp\left[-\frac{\beta}{2}\left(t_n - \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right)\right)^2\right] \end{aligned} \tag{10}$$

$$\begin{aligned} \text{quadratic term } &= -\frac{\beta}{2}\boldsymbol{w}^T \boldsymbol{S}_0^{-1}\boldsymbol{w} + \sum_{n=1}^{N} -\frac{\beta}{2}\boldsymbol{w}^T \phi\left(x_n\right)\phi(x_n)^T \boldsymbol{w} \\ &= -\frac{\beta}{2}\boldsymbol{w}^T \left[\boldsymbol{S}_0{}^{-1} + \sum_{n=1}^{N} \phi\left(x_n\right)\phi(x_n)^T\right]\boldsymbol{w} \\ \Rightarrow \mathbf{S}_N^{-1} &= \boldsymbol{S}_0{}^{-1} + \sum_{n=1}^{N} \phi\left(x_n\right)\phi(x_n)^T \end{aligned} \tag{11}$$

$$\begin{aligned} \text{linear term } &= \beta \boldsymbol{m_0}^T \boldsymbol{S_0}^{-1}\boldsymbol{w} + \sum_{n=1}^{N} \beta t_n \boldsymbol{\phi}(\boldsymbol{x_n})^T \boldsymbol{w} \\ &= \beta\left[\boldsymbol{m_0}^T \boldsymbol{S_0}^{-1} + \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\boldsymbol{x_n})^T\right]\boldsymbol{w} \\ \Rightarrow \boldsymbol{m}_N &= \boldsymbol{S_N}\left[\boldsymbol{S}_0{}^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N} t_n \phi\left(\boldsymbol{x}_n\right)\right] \end{aligned} \tag{12}$$

$$
\begin{aligned}
\text{constant term } &= \left(-\frac{\beta}{2}\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0} - b_0\beta\right) - \frac{\beta}{2}\sum_{n=1}^{N} t_n^2 \\
&= -\beta\left[\frac{1}{2}\boldsymbol{m}_0{}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0} + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2\right] \\
\Rightarrow b_N &= \frac{1}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2 - \frac{1}{2}\boldsymbol{m}_N{}^T\boldsymbol{S_N}^{-1}\boldsymbol{m_N}
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
\text{exponent term } &= (2 + a_0 - 1) + \frac{N}{2} \\
\Rightarrow a_N &= a_0 + \frac{N}{2}
\end{aligned}
\tag{14}
$$

Question 3

Show that the integration over w in the Bayesian linear regression model gives the result

$$
\int \exp\{-E(\mathbf{w})\}\mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}. \tag{15}
$$

Hence show that the log marginal likelihood is given by

$$
\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi) \tag{16}
$$

From multivariate normal distribution, we have

$$
\int \frac{1}{(2\pi)^{M/2}}\frac{1}{|\mathbf{A}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\mathbf{A}\left(\boldsymbol{w}-\boldsymbol{m}_N\right)\right\}d\boldsymbol{w} = 1 \tag{17}
$$

Hence

$$
\int \exp\left\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\mathbf{A}\left(\boldsymbol{w}-\boldsymbol{m}_N\right)\right\}d\boldsymbol{w} = (2\pi)^{M/2}|\mathbf{A}|^{1/2} \tag{18}
$$

Question 4

Consider real-valued variables X and Y. The Y variable is generated, conditional on X, from the following process:

$$
\begin{aligned}
\epsilon &\sim N(0,\sigma^2) \\
Y &= aX + \epsilon
\end{aligned}
\tag{19}
$$

where every ϵ is an independent variable, called a noise term, which is drawn from a Gaussian distribution with mean 0, and standard deviation σ. This is a one-feature linear regression model, where a is the only weight parameter. The conditional probability of Y has distribution $p(Y|X,a) \sim N(aX,\sigma^2)$, so it can be written as

$$
p(Y|X,a) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(Y-aX)^2) \tag{20}
$$

Assume we have a training dataset of n pairs (X_i, Y_i) for $i = 1..n$, and σ is known.

Derive the maximum likelihood estimate of the parameter a in terms of the training example X_i's and Y_i's. We recommend you start with the simplest form of the problem:

$$F(a) = \frac{1}{2}\sum_i (Y_i - aX_i)^2 \tag{21}$$

$$\begin{aligned}\frac{\partial F}{\partial a} &= \sum_i (Y_i - aX_i)(-X_i) \\ &= \sum_i aX_i^2 - X_iY_i \\ \Rightarrow a &= \frac{\sum_i X_iY_i}{\sum X_i^2}\end{aligned} \tag{22}$$

Question 5

If a data point y follows the Poisson distribution with rate parameter θ, then the probability of a single observation y is

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \text{for } y = 0, 1, 2, \ldots \tag{23}$$

You are given data points $y_1, \ldots, y_n$ independently drawn from a Poisson distribution with parameter θ . Write down the log-likelihood of the data as a function of θ .

$$\begin{aligned} log\ p(y|\theta) &= ylog\theta - \theta - \sum_{i=0}^{y} logi \\ \Rightarrow L(\theta) &= \sum_{i=1}^{n}(y_i log\theta - \theta - logy_i!)\end{aligned} \tag{24}$$

Question 6

Suppose you are given n observations, $X_1, \ldots, X_n$, independent and identically distributed with a $Gamma(\alpha, \lambda)$ distribution. The following information might be useful for the problem.

- **If $X \sim Gamma(\alpha, \lambda)$, then $\mathbb{E}[X] = \frac{\alpha}{\lambda}$ and $\mathbb{E}[X^2] = \frac{\alpha(\alpha+1)}{\lambda^2}$**
- **The probability density function of $X \sim Gamma(\alpha, \lambda)$ is $f_X(x) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}$, where the function Γ is only dependent on α and not λ.**

Suppose we are given a known, fixed value for α. Compute the maximum likelihood estimator for λ.

$$\begin{aligned} log\ f_X(x) &= \alpha log\lambda + (\alpha-1)logx - \lambda x - log\Gamma(\alpha) \\ L(\lambda) &= n\alpha log\lambda + (\alpha-1)log\prod_{i=1}^{n} x_i - \lambda\sum_{i=1}^{n} x_i - nlog\Gamma(\alpha) \\ \frac{dL(\lambda)}{d\lambda} &= \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i \\ \Rightarrow \lambda &= \frac{\alpha}{\frac{1}{n}\sum_{i=1}^{n} x_i}\end{aligned} \tag{25}$$