CS329 Homework #3

Course: Machine Learning(H)(CS329) - Instructor: Qi Hao

Name: Jianan Xie(谢嘉楠)

SID: 12110714

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

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

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.

Ans:

To find an expression for the solution $\mathbf{w}^*$, we need to set the derivative of $E_D(\mathbf{w})$ to zero. Set $\mathbf{t}' = [\sqrt{r_1}t_1, \sqrt{r_2}t_2, \ldots, \sqrt{r_n}t_n]^{\mathrm{T}}$, and $\Phi(\mathbf{x}) = [\sqrt{r_1}\phi(\mathbf{x_1})^{\mathrm{T}}, \sqrt{r_2}\phi(\mathbf{x_2})^{\mathrm{T}}, \ldots, \sqrt{r_n}\phi(\mathbf{x_n})^{\mathrm{T}}]^{\mathrm{T}}$. Then we rewrite the $E_D(\mathbf{w})$:

$$\begin{aligned}
E_D(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w^T}\phi(\mathbf{x}_n)\}^2. \\
&= \frac{1}{2}\sum_{n=1}^{N}\{\sqrt{r_n}t_n - \sqrt{r_n}\phi(\mathbf{x}_n)^{\mathrm{T}}\mathbf{w}\}^2 \\
&= \frac{1}{2}||\mathbf{t}' - \Phi(\mathbf{x})\mathbf{w}|| \\
&= \frac{1}{2}(\mathbf{t}' - \Phi(\mathbf{x})\mathbf{w})^{\mathrm{T}}(\mathbf{t}' - \Phi(\mathbf{x})\mathbf{w})
\end{aligned}$$

As what we learned before, the solution $\mathbf{w^*}$ to minimize $E(w) = \frac{1}{2}(\mathbf{y} - \mathbf{Xw})^T(\mathbf{y} - \mathbf{Xw})$ is $\hat{\mathbf{w}} = \mathbf{(X^TX)^{-1}X^Ty}$, thus here we find the $\mathbf{w^*}$ for $E_D(\mathbf{w}) = \frac{1}{2}(\mathbf{t}' - \Phi(\mathbf{x})\mathbf{w})^{\mathrm{T}}(\mathbf{t}' - \Phi(\mathbf{x})\mathbf{w})$ is $\mathbf{w^*} = [\Phi(\mathbf{x})^{\mathrm{T}}\Phi(\mathbf{x})]^{-1}\Phi(\mathbf{x})^{\mathrm{T}}\mathbf{t}'$

Two alternative interpretations: (i)if we take data dependent noise variance from β^{-1} to $r_n\beta^{-1}$ then we can get the weighted sum-of-squares error function above. (ii)we can consider r_n as the times $(\mathbf{x_n}, t_n)$ repeatedly occurs.

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,

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

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

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

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

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

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

Ans:

The conjugate prior for $\mathbf{w}$ and β:

$$\begin{aligned} p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0)\mathrm{Gam}(\beta|a_0, b_0) \\ &\propto (\beta\mathbf{S}_0^{-1})^{\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{w}-\mathbf{m_0})^{\mathrm{T}}\beta\mathbf{S_0^{-1}}(\mathbf{w}-\mathbf{m_0})} b_0^{a_0}\beta^{a_o-1}e^{-b_0\beta} \end{aligned}$$

The likelihood function:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X}, \mathrm{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}(t_n|\mathrm{w}^{\mathrm{T}}\phi(\mathrm{x}_n), \beta^{-1}) \\ &\propto \prod_{n=1}^{N} \beta^{\frac{1}{2}} e^{-\frac{\beta}{2}(t_n - \mathrm{w}^{\mathrm{T}}\phi(\mathrm{x_n}))^2} \end{aligned}$$

According to Bayesian Inference $p(\mathbf{w}, \beta|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{X}, w, \beta) \times p(\mathbf{w}, \beta)$, the posterior is also in the form of $p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\mathrm{Gam}(\beta|a_N, b_N)$.

First focus on quadratic term of $\mathbf{w}$:

$$\begin{aligned} \textit{quadratic term} &= -\frac{\beta}{2}\mathbf{w^T S_0^{-1} w} - \frac{\beta}{2}\sum_{n=1}^{N}\mathbf{w^T}\boldsymbol{\phi}(\mathbf{x_n})\boldsymbol{\phi}(\mathbf{x_n})^{\mathbf{T}}\mathbf{w} \\ &= -\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}[\mathbf{S_0^{-1}} + \boldsymbol{\phi}(\mathbf{x_n})\boldsymbol{\phi}(\mathbf{x_n})^{\mathrm{T}}]\mathbf{w} \end{aligned}$$

Then we get $\mathbf{S_N^{-1}} = \mathbf{S_0^{-1}} + \boldsymbol{\phi}(\mathbf{x_n})\boldsymbol{\phi}(\mathbf{x_n})^{\mathrm{T}}$.

Second focus on linear term of $\mathbf{w}$:

$$\begin{aligned} \textit{linear term} &= -\beta\mathbf{m_0^{\mathrm{T}} S_0^{-1} w} - \beta\sum_{n=1}^{N}\mathbf{t_n}\boldsymbol{\phi}(\mathbf{x_n})^{\mathbf{T}}\mathbf{w} \quad (\text{As } S_0 \text{ is symmetric}) \\ &= -\beta[\mathbf{m_0^{\mathrm{T}} S_0^{-1}} + \sum_{n=1}^{N}\mathbf{t_n}\boldsymbol{\phi}(\mathbf{x_n})^{\mathbf{T}}]\mathbf{w} \end{aligned}$$

Then we get $\mathbf{m_N}^{\mathrm{T}}\mathbf{S_N^{-1}} = \mathbf{m_0^{\mathrm{T}} S_0^{-1}} + \sum_{n=1}^{N}\mathbf{t_n}\boldsymbol{\phi}(\mathbf{x_n})^{\mathbf{T}}$, thus $\mathbf{m_N = S_N S_0^{-1} m_0 + S_N \sum_{n=1}^{N} t_n}\boldsymbol{\phi}(\mathbf{x_n})$

Third focus on constant term of $\mathbf{w}$:

$$constant\ term = (-\frac{\beta}{2}\mathbf{m_0^T S_0^{-1} m_0} - b_0\beta) - \frac{\beta}{2}\sum_{n=1}^{N} t_n^2$$

Then we get $-\frac{\beta}{2}\mathbf{m_N^T S_N^{-1} m_N} - b_N\beta = -\frac{\beta}{2}\mathbf{m_0^T S_0^{-1} m_0} - b_0\beta - \frac{\beta}{2}\sum_{n=1}^{N} t_n^2$, thus $b_N = \frac{1}{2}\mathbf{m_0^T S_0^{-1} m_0} + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2 - \frac{1}{2}\mathbf{m_N^T S_N^{-1} m_N}$.

Fourth focus the exponential term of β:

$$\beta's\ exponential\ term = (\frac{1}{2} + a_o - 1) + \frac{N}{2}$$

Then we get $\frac{1}{2} + a_N - 1 = (\frac{1}{2} + a_0 - 1) + \frac{N}{2}$, thus $a_N = a_0 + \frac{N}{2}$.

Question 3

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

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

Hence show that the log marginal likelihood is given by

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

Ans:

According to the definition of $E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(w - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N)$, where $\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$. Thus, what we need to integrate is that:

$$\begin{aligned}\int \exp\{-E(\mathbf{w})\}\mathrm{d}\mathbf{w} &= \int \exp\{-E(\mathbf{m}_N) + \frac{1}{2}(w - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\}\mathrm{d}\mathbf{w} \\ &= \exp\{-E(\mathbf{m}_N)\}\int \exp\{\frac{1}{2}(w - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\}\mathrm{d}\mathbf{w}\end{aligned}$$

As for a multivariate normal distribution, we know:

$$\int \frac{1}{(2\pi)^{\frac{M}{2}}}\frac{1}{|\mathbf{A^{-1}}|^{\frac{1}{2}}}\exp\{\frac{1}{2}(w - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\}\mathrm{d}\mathbf{w} = 1$$

Thus :

$$\begin{aligned}\int \exp\{-E(\mathbf{w})\}\mathrm{d}\mathbf{w} &= \exp\{-E(\mathbf{m}_N)\}\int \exp\{\frac{1}{2}(w - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\}\mathrm{d}\mathbf{w} \\ &= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}\end{aligned}$$

Then the log marginal likelihood is:

$$
\begin{aligned}
\ln p(\mathbf{t}|\alpha,\beta) &= \ln\{(\frac{\beta}{2\pi})^{\frac{N}{2}}(\frac{\alpha}{2\pi})^{\frac{M}{2}}\int \exp\{-E(\mathbf{w})\}d\mathbf{w}\} \\
&= \frac{M}{2}\ln\alpha - \frac{M}{2}\ln 2\pi + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \ln\{\exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}\} \\
&= \frac{M}{2}\ln\alpha - \frac{M}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - E(\mathbf{m}_N) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \\
&= \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi)
\end{aligned}
$$

Question 4

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

$$\epsilon \sim N(0,\sigma^2)$$

$$Y = aX + \epsilon$$

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

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

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:

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

Ans:

Following the hint, we start with the simplest form of the problem, trying to minimize $F(a)$:

$$
\begin{aligned}
\frac{\partial F(a)}{\partial a} &= \frac{\partial}{\partial a}\{\frac{1}{2}\sum_{i=1}^{n}(Y_i - aX_i)^2\} \\
&= \sum_{i=1}^{n}(Y_i - aX_i)(-X_i)
\end{aligned}
$$

set above as zero, then we get the $a^* = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$.

And next we return to the original problem:

$$
\begin{aligned}
a_{ML} &= \underset{a}{argmax}(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2) \\
&= \underset{a}{argmax}(\frac{1}{\sqrt{2\pi}\sigma})^n \exp(\sum_{i=1}^{n} - \frac{1}{2\sigma^2}(Y_i - aX_i)^2) \\
&= argmax(\sum_{i}^{n} - \frac{1}{2\sigma^2}(Y_i - aX_i)^2)
\end{aligned}
$$

$$
\begin{aligned}
&= \underset{a}{argmin}(F(a)) \\
&= a^* && (a^* \text{ derived above}) \\
&= \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}
\end{aligned}
$$

Question 5

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

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

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 .

Ans:

The log-likelihood of the data as a function of θ:

$$
\begin{aligned}
\ln \prod_{i=1}^n p(y_i|\theta) &= \ln \prod_{i=1}^n \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\
&= \sum_{i=1}^n (y_i \ln\theta - \theta - \sum_{k=1}^{y_i} \ln k) \\
&= \ln\theta \sum_{i=1}^n y_i - \sum_{i=1}^n \sum_{k=1}^{y_i} \ln k - n\theta
\end{aligned}
$$

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.

Ans:

Aiming to maximize the log-likelihood $\ln \prod_{i=1}^n f_X(X_i)$:

$$
\begin{aligned}
\ln \prod_{i=1}^{n} f_X(X_i) &= \sum_{i=1}^{n} \ln f_X(X_i) \\
&= \sum_{i=1}^{n} \ln\{\frac{1}{\Gamma(\alpha)} \lambda^\alpha X_i^{\alpha-1} e^{-\lambda X_i}\} \\
&= n\alpha \ln \lambda + (\alpha - 1) \sum_{i=1}^{n} \ln X_i - \lambda \sum_{i=1}^{n} X_i - n \ln \Gamma(\alpha)
\end{aligned}
$$

Then we set the devirative of it as zero to get the λ_{ML}:

$$
\begin{aligned}
\frac{\partial}{\partial \lambda} \ln \prod_{i=1}^{n} f_X(X_i) &= n\alpha \ln \lambda + (\alpha - 1) \sum_{i=1}^{n} \ln X_i - \lambda \sum_{i=1}^{n} X_i - n \ln \Gamma(\alpha) \\
&= \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i
\end{aligned}
$$

So, we get the maximum likelihood estimator for λ is : $\lambda_{ML} = \frac{n\alpha}{\sum_{i=1}^{n} X_i}$.