SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Answer #3

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *October 23th, 2019*

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.

Answer. If we define $\mathbf{R} = \text{diag}(r_1, \ldots, r_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_D(\mathbf{w}) = \frac{1}{2}(\mathbf{t} - \Phi\mathbf{w})^T\mathbf{R}(\mathbf{t} - \Phi\mathbf{w})$$

Setting the derivative with respect to $\mathbf{w}$ to zero, and re-arranging, then gives

$$\mathbf{w}^* = (\Phi^T\mathbf{R}\Phi)^{-1}\Phi^T\mathbf{R}\mathbf{t}$$

which reduces to the standard solution (3.15) for the case $\mathbf{R} = \mathbf{I}$.
If we compare (3.104) with (3.10) - (3.12), we see that r_n can be regarded as a precision (inverse variance) parameter, particular to the data point (x_n, t_n), that either replaces or scales β.
Alternatively, r_n can be regarded as an effective number of replicated observations of data point (x_n, t_n); this becomes particularly clear if we consider (3.104) with r_n taking positive integer values, although it is valid for any $r_n > 0$.

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

$$p(\mathbf{t}|\mathbf{X}, \mathrm{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathrm{w}^{\mathrm{T}}\phi(\mathrm{x}_n), \beta^{-1}) \tag{2}$$

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

$$p(\mathbf{w}, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0)\mathrm{Gam}(\beta|a_0, b_0). \tag{3}$$

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

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\mathrm{Gam}(\beta|a_N, b_N). \tag{4}$$

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

Answer. The log of the posterior distribution is given by

$$\begin{aligned}
\ln p(\mathbf{w}, \beta|\mathbf{t}) &= \ln p(\mathbf{w}, \beta) + \sum_{n=1}^{N} \ln p(t_n|\mathbf{w}^{\mathrm{T}}\phi(\mathrm{x}_n), \beta^{-1}) \\
&= \frac{M}{2}\ln\beta - \frac{1}{2}\ln|\mathbf{S}_0| - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) - b_0\beta \\
&+ (a_0-1)\ln\beta + \frac{N}{2}\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N}\{\mathbf{w}^{\mathrm{T}}\phi(\mathrm{x}_n) - t_n\}^2 + const.
\end{aligned}$$

Using the product rule, the posterior distribution can be written as $p(\mathbf{w}, \beta|\mathbf{t}) = p(\mathbf{w}|\beta, \mathbf{t})p(\beta|\mathbf{t})$. Consider first the dependence on $\mathbf{w}$. We have

$$\ln p(\mathbf{w}, \beta|\mathbf{t}) = -\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}[\Phi^{\mathrm{T}}\Phi + \mathbf{S}_0^{-1}]\mathbf{w} + \mathbf{w}^{\mathrm{T}}[\beta\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^{\mathrm{T}}\mathbf{t}] + const.$$

Thus we see that $p(\mathbf{w}, \beta|\mathbf{t})$ is a Gaussian distribution with mean and covariance given by

$$\mathbf{m}_N = \mathbf{S}_N[\mathbf{S}_0^{-1}\mathbf{m}_0 + \Phi^{\mathrm{T}}\mathbf{t}]$$
$$\beta\mathbf{S}_N^{-1} = \beta(\mathbf{S}_0^{-1} + \Phi^{\mathrm{T}}\Phi)$$

To find $p(\beta|\mathbf{t})$ we first need to complete the square over $\mathbf{w}$ to ensure that we pick up all terms involving β (any terms independent of β may be discarded since these will be absorbed into the normalization coefficient which itself will be found by inspection at the end). We also need to remember that a factor of $(M/2)\ln\beta$ will be absorbed by the normalisation factor of $p(\mathbf{w}|\beta, \mathbf{t})$. Thus

$$\ln p(\mathbf{w}, \beta|\mathbf{t}) = -\frac{\beta}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{\beta}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \frac{N}{2}\ln\beta - b_0\beta + (a_0-1)\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N}t_n^2 + const.$$

We recognize this as the log of a Gamma distribution. Reading off the coefficients of β and $\ln \beta$ we then have

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \sum_{n=1}^{N} t_n^2)$$

Question 3

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

$$\int \exp\{-E(\mathbf{w})\}\mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}. \tag{5}$$

Hence show that the log marginal likelihood is given by

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi) \tag{6}$$

Answer. From

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^T\mathbf{A}(\mathbf{w}-\mathbf{m}_N)$$

we see that the integrand of

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

is an unnormalized Gaussian and hence integrates to the inverse of the corresponding normalizing constant, which can be read off from the r.h.s. of

$$\mathcal{N}(\mathbf{x}|\mu,\sum) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\sum|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\sum^{-1}(x-\mu)\}$$

as

$$(2\pi)^{M/2}|\mathbf{A}^{-1}|^{1/2}$$

Using

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\mathrm{d}\mathbf{w}$$

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

and the properties of the logarithm, we get

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

which equals (6)

Question 4

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

$$\epsilon \sim N(0, \sigma^2) \tag{7}$$

$$Y = aX + \epsilon \tag{8}$$

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

$$p(Y|X, a) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(Y - aX)^2) \tag{9}$$

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:

$$F(a) = \frac{1}{2}\sum_i (Y_i - aX_i)^2 \tag{10}$$

Answer. Use $F(a) = \frac{1}{2}\sum_i (Y_i - aX_i)^2$ and minimizi F. Then

$$\begin{aligned} 0 &= \frac{\partial}{\partial a}\left[\frac{1}{2}\sum_i (Y_i - aX_i)^2\right] \\ &= \sum_i (Y_i - aX_i)(-X_i) \\ &= \sum_i aX_i^2 - X_iY_i \end{aligned}$$

$$a = \frac{\sum_i X_iY_i}{\sum_i X_i^2}$$

Question 5

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

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \text{for } y = 0, 1, 2, \ldots \tag{11}$$

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 .

Answer.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} P(y_i|\theta) \\ &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= e^{-n\theta} \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} \end{aligned}$$

The log-likelihood function of θ

$$\ln L(\theta) = -n\theta + \sum_{i=1}^{n} (y_i \ln \theta - \ln y_i)$$

Question 6

Suppose you are given n observations, $X_1, \ldots, X_n$, independent and identically distributed with a Gamma(α, λ) distribution. The following information might be useful for the problem.
1.If X $\sim$ Gamma(α, λ), then $\mathbb{E}[X] = \frac{\alpha}{\lambda}$ and $\mathbb{E}[X^2] = \frac{\alpha(\alpha+1)}{\lambda^2}$
2.The probability density function of X $\sim$ Gamma(α, λ) 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.

Answer. We first write the likelihood function.

$$L(\lambda|X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \lambda^\alpha X_i^{\alpha-1} e^{-\lambda X_i}$$

The log-likelihood function is given as follows.

$$l(\lambda|X_1, \ldots, X_n) = -n \log(\Gamma(\alpha)) + n\alpha \log \lambda + \sum_{i=1}^{n} (\alpha - 1) \log X_i - \lambda X_i$$

Next, we take the gradient with respect to λ and set it equal to 0.

$$\nabla_\lambda l(\lambda|X_1, \ldots, X_n) = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i = 0$$

Solving for λ, we get

$$\hat{\lambda}_{\text{MLE}} = \frac{n\alpha}{\sum_{i=1}^{n} X_i} = \frac{\alpha}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{\alpha}{\overline{X}_1}$$