CS329 Machine Learning

Homework #3

Fan Site
fanst2021@mail.sustech.edu.cn

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

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

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.

Solution

Let the derivation of $\mathbf{w}$ to be 0:

$$\frac{\partial}{\partial\, \mathbf{w}} E_D(\mathbf{w}) = \sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^{\mathrm{T}}\phi(x_n)\}\Phi(x_n)^{\mathrm{T}} = 0$$

Solving this equation we obtain:

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

where

- Φ is the design matrix with elements $\Phi_{ij} = \phi_j(x_i)$,
- $\mathbf{t} = [t_1, \cdots, t_n]^{\mathrm{T}}$ is the target vector,
- R is a diagonal matrix with r_i as the i-th diagonal element.

1. **Data Dependent Noise Variance:**

   The weighting factors r_n in the error function can be interpreted as representing the inverse of the variance of the noise associated with each data point.

   The larger the r_n, the smaller the associated variance, meaning that the data point has less noise. So, by assigning different weights to different data points, we are effectively modeling data-dependent noise variances.

2. **Replicated Data Points:**

   If a data point is replicated r_n times in the dataset, it can be viewed as if we have r_n identical copies of that data point. The error term for each replicated point is then scaled by r_n.

   This implies that the model is more influenced by the replicated data points with higher weights, effectively giving them more importance in the fitting process. Replicating data points can be a way to emphasize certain observations in the dataset.

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,

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

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.

Solution

$$\mathrm{Gam}(\beta|a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}\beta^{a_0-1}\exp\{-b_0\beta\}$$

By Bayesian Inference,

$$p(\mathbf{w}, \beta|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \times p(\mathbf{w}, \beta)$$

where the likelihood:

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

And the prior:

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

Quadratic part of the exponent:

$$\sum_{n=1}^{N} -\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\Phi(x_n)\Phi(x_n)^{\mathrm{T}}\mathbf{w} - \frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} = -\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\left(\sum_{n=1}^{N}\Phi(x_n)\Phi(x_n)^{\mathrm{T}} + \mathbf{S}_0^{-1}\right)\mathbf{w}$$

Linear part of the exponent:

$$\beta \boldsymbol{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} + \sum_{n=1}^{N}\beta t_n\Phi(x_n)^{\mathrm{T}}\mathbf{w} = \beta\left(\boldsymbol{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1} + \sum_{n=1}^{N} t_n\Phi(x_n)^{\mathrm{T}}\right)\mathbf{w}$$

So we have the posterior parameters for Gaussian part:

$$\mathbf{S}_N = \left(\sum_{n=1}^{N}\Phi(x_n)\Phi(x_n)^{\mathrm{T}} + \mathbf{S}_0^{-1}\right)^{-1}$$

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\boldsymbol{m}_0 + \sum_{n=1}^{N} t_n\Phi(x_n)\right)$$

Constant part of the exponent:

$$p(\mathbf{w}, \beta|\mathbf{t}) \propto -\frac{\beta}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0\mathbf{m}_0 - b_0\beta - \frac{\beta}{2}\sum_{m=1}^{N} t_n^2$$
$$= -\beta\left(\frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0\mathbf{m}_0 + b_0 + \frac{1}{2}\sum_{m=1}^{N} t_n^2\right)$$
$$= -\beta\left(\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + b_N\right)$$

Exponent of β:

$$p(\mathbf{w}, \beta|\mathbf{t}) \propto \beta^{\frac{N}{2}}\beta^{\frac{1}{2}}\beta^{a_0-1} = \beta^{\frac{N}{2}+a_0-\frac{1}{2}}$$

So we obtain the posterior parameters of Gamma part:

$$a_N = a_0 + \frac{N}{2}$$
$$b_N = \frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2 - \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N$$

Question 3

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

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

Hence show that the log marginal likelihood is given by

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

Solution

According to (3.80),

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

where $\mathbf{A} = \alpha\ \mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} = S_N^{-1}$.

Perform total integral over multivariate Gaussian distribution,

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

As $E(\mathbf{m}_N)$ is independent of $\mathbf{w}$, we have

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

Substitute this back,

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

Question 4

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

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

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

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

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:

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

Solution

The log-likelihood function:

$$
L(a) = \ln\prod_{i=1}^{n} p(Y_i|X_i,a) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \sum_{i=1}^{n}\frac{1}{2\sigma^2}(Y_i - aX_i)^2
$$

Maximize $L(a)$ with respect to a,

$$
\frac{\partial}{\partial a}L(a) = -\frac{1}{\sigma^2}\frac{\partial}{\partial a}F(a) = -\sum_{i=1}^{n}\frac{1}{2\sigma^2}(2aX_i^2 - 2X_iY_i) = 0
$$

So we have

$$
a_{\mathrm{ML}} = \frac{\sum_{i=1}^{n} X_iY_i}{\sum_{i=1}^{n} X_i^2}
$$

Question 5

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

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

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

Solution

The log-likelihood function

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

Question 6

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

- If $X \sim \text{Gamma}(\alpha, \lambda)$, then $\mathbb{E}[X] = \frac{\alpha}{\lambda}$ and $\mathbb{E}[X^2] = \frac{\alpha(\alpha+1)}{\lambda^2}$
- The probability density function of $X \sim \text{Gamma}(\alpha, \lambda)$ 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.

Solution

The log-likelihood function:

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

Maximize $L(\alpha, \lambda)$ with respect to λ,

$$\frac{\partial}{\partial \lambda} L(\alpha, \lambda) = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i = 0$$

We obtain

$$\lambda_{\text{ML}} = \frac{n\alpha}{\sum_{i=1}^{n} X_i}$$