SUSTech

MACHINE-LEARNING

Fall 2020

Homework 3

刘禹熙 2020 年 10 月 27 日

3.1. Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 9$, so that the sum-of-squares error function becomes

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

Find an expression for the solution 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.

$$\begin{aligned}
E_D(w) &= \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - w^T\phi(x_n)\}^2 \\
&= \frac{1}{2}(\phi w - t)^T R(\phi w - t) \\
&= \frac{1}{2}(w^T\phi^T R\phi w - w^T\phi^T Rt - t^T R\phi w + t^T Rt) \\
&= \frac{1}{2}(w^T\phi^T R\phi w - 2t^T R\phi w + t^T Rt)
\end{aligned}$$

and we defined $R = diag(r_1, r_2, ...r_N)$. Taking the gradient of the error function

$$\nabla E_D(w) = \phi^T R\phi w - t^T R\phi$$

$$\begin{aligned}
w^* &= (\phi^T R\phi)^{-1}t^T R\phi \\
&= (\phi^T R\phi)^{-1}\phi^T Rt
\end{aligned}$$

If R = I, we get the standard solution$w^* = (\phi^T\phi)^{-1}\phi^T t$, and r_n can be seen as a precision parameter.r_n can also be regarded as an effective number of replicatesd observations of data point(x_n, t_n)

3.2. We saw in Section 2.3.6 that the conjugate prior for a Gauussian 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 $p(t|x, w, \beta)$ of the linear regression model. If we consider the likelihood function,

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

then the conjugate prior for w and β is given by

$$p(w, \beta) = \mathcal{N}(w|m_0, \beta^{-1}S_0)Gam(\beta|a_0, b_0)$$

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

$$p(w, \beta|t) = \mathcal{N}(w|m_N, \beta^{-1}S_N)Gam(\beta|a_N, b_N)$$

and find expressions for the posterior parameters m_N, S_N, a_N, and b_N.

Solution.

$$\begin{aligned}\ln p(w, \beta|t) &= \ln p(w, \beta) + \sum_{n=1}^{N} \ln p(t_n|w^T\phi(x_n), \beta^{-1}) \\ &= \frac{M}{2}\ln\beta - \frac{1}{2}\ln|S_0| - \frac{\beta}{2}(w - m_0)^T S_0^{-1}(w - m_0) \\ &- b_0\beta + (a_0 - 1)\ln\beta \\ &+ \frac{N}{2}\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N}\{w^T\phi(x_n) - t_n\}^2 + const\end{aligned}$$

$$p(w, \beta|t) = p(w|\beta, t)p(\beta|t)$$

$$\ln(w|\beta, t) = -\frac{\beta}{2}w^T[\phi^T\phi + S_0^{-1}]w + w^T[\beta S_0^{-1}m_0 + \beta\phi^T t] + const$$

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

$$\begin{aligned}m_N &= S_N[S_0^{-1}m_0 + \phi^T t] \\ \beta S_N^{-1} &= \beta(S_0^{-1} + \phi^T\phi)\end{aligned}$$

$$\begin{aligned}\ln p(\beta|t) = &-\frac{\beta}{2}m_0^T S_0^{-1}m_0 + \frac{\beta}{2}m_N^T S_N^{-1}m_N \\ &+ \frac{N}{2}\ln\beta - b_0\beta + (a_0 - 1)\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N} t_n^2 + const\end{aligned}$$

We recognize this as the log of a Gamma distribution. And we have

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

$$b_N = b_0 + \frac{1}{2}(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + \sum_{n=1}^{N} t_n^2)$$

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

$$\int \exp\{-E(w)\}dw = \exp\{E(m_N)\}(2\pi)^{M/2}|A|^{-1/2}$$

Hence show that the log marginal likelihood is given by

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

Solution. Using $p(t|\alpha,\beta) = (\frac{\beta}{2\pi})^{N/2}(\frac{\alpha}{2\pi})^{M/2}\int \exp\{-E(w)\}dw$ we have

$$\begin{aligned}\ln p(t|\alpha,\beta) &= \frac{M}{2}(\ln\alpha - \ln(2*\pi)) + \frac{N}{2}(\ln\beta - \ln(2\pi)) + ln\int exp\{-E(w)\}dw \\ &= \frac{M}{2}(\ln a - \ln(2\pi)) + \frac{N}{2}(\ln\beta - \ln(2\pi)) - E(m_N) - \frac{1}{2}\ln|A| + \frac{M}{2}\ln(2\pi)\end{aligned}$$

which equals $\ln p(t|\alpha,\beta) = \frac{M}{2} + \frac{N}{2}\ln\beta - E(m_N) - \frac{1}{2}\ln|A| - \frac{N}{2}(2\pi)$.

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

$$\varepsilon \sim \mathrm{N}(0,\sigma^2)$$
$$Y = aX + \varepsilon$$

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 $p(Y|X,a) \sim \mathrm{N}(aX,\sigma^2)$, 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:

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

Solution.

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

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

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

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

$$\hat{a} = \frac{\sum_{i=1}^{n} X_iY_i}{\sum_{i=1}^{n} X_i^2}$$

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

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

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.

$$L(y_1, y_2, ..., y_n|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = e^{-n\theta} \prod_{u=1}^{n} \frac{\theta^{y_i}}{y_i!}$$

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

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

$$\theta = \frac{1}{n}\sum_{i=1}^{n} y_i$$

3.6. Suppose you are given n obserbations, $X_1, ...X_n$, independent and identically distributed with a Gamma(α, λ) distribution. The following informaztion might be useful for the problem.

(a) If $X \sim$ Gamma(α, λ), then $\mathbb{E}$ [X]=$\frac{\alpha}{\lambda}$ and $\mathbb{E}$ $[X^2]$=$\frac{\alpha(\alpha+1)}{\lambda^2}$

(b) The probabiliuty 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.

Solution.

$$p(x_n) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x_n^{\alpha-1}e^{-\lambda x_n}$$

$$\ln p(x_n) = \alpha \ln \lambda - \ln \Gamma(\alpha) + (\alpha - 1)\ln x_n - \lambda x_n$$

$$L(x; \alpha, \lambda) = \sum_{i=1}^{n} \ln p(x_i) = n\alpha \ln \lambda - n \ln \Gamma(\alpha) + (\alpha - 1) \ln \sum_{i=1}^{n} x_i - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial \lambda} L(x; \alpha, \lambda) = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{n\alpha}{\sum_{i=1}^{n} x_i}$$