<u>Machine Learning WS 2018</u>

Solution to Assignment 3

Submitted By : Vindhya Singh **Enrollment Number :** 03693296
Collaborated By : Wasiq Rumaney **Enrollment Number :** 03694978

<u>Solution 2 :</u>

$$E_{weighted}(w)=\frac{1}{2}\sum_{i=1}^{N} t_i[w^T \phi(x_i)-y_i]^2 \qquad (1)$$

For finding minima, taking the gradient of 1 and equating it to zero,

$$\frac{(\partial E_{weighted}(w))}{(\partial w)}=0 \qquad (2)$$

Taking the matrix notations in consideration, let W be matrix of weighted coefficients,

$$E_{weighted}(w)=\frac{1}{2}W(\Phi w-y)^T(\Phi w-y)$$ (matrix notation taken form slide 15, Lecture 03)

$$= \frac{1}{2}(w^T\Phi^T W\Phi w-w^T\Phi^T Wy-y^T W\Phi w+y^T y W)$$

$$= \frac{1}{2}(w^T\Phi^T W\Phi w-2y^T W\Phi w+y^T y W)$$

Thus, equation 2 becomes

$$\nabla E_{weighted}(w)=\Phi^T W\Phi w-y^T W\Phi$$

$$w=(\Phi^T W\Phi)^{-1}\Phi^T W\, y$$

Taking W = I in the above equation,

$$\boxed{w=(\Phi^T\Phi)^{-1}\Phi^T y \quad \text{for} \quad w=w_{ML}}$$

$$E_{LS}(w)=\frac{1}{2}\sum_{i=1}^{N}(w^T\phi(x_i)-y\,i)^2 \quad \text{and} \quad E_{weighted}(w)=\frac{1}{2}\sum_{i=1}^{N} t_i[w^T\phi(x_i)-y_i]^2$$

Thus for T_i , it can be regarded as precision for the given data set points.

Solution 3 :

Since for ordinary least square regression,

$$(y-\Phi w)^T(y-\Phi w)$$

$$= \quad (y^T-\Phi^T w^T)(y-\Phi w)$$

$$= \quad y^T y - y^T \Phi w - \Phi^T w^T y + \Phi^T \Phi w^T w \tag{1}$$

For Ridge Regression,

$$(y-\Phi w)^T(y-\Phi w)+\lambda w^T w$$

$$= \quad (y^T-\Phi^T w^T)(y-\Phi w)+\lambda w^T w$$

$$= \quad y^T y - y^T \Phi w - \Phi^T w^T y + \Phi^T \Phi w^T w + \lambda w^T w \tag{2}$$

Given that, we need to "Augment the design matrix $\Phi \in R^{N\times M}$ with M additional rows $\sqrt{\lambda} I_{M\times M}$ and augment y with M zeros."

Therefore, let the augmented matrices be

$$\Phi_{new}=\begin{pmatrix}\Phi \\ \lambda I\end{pmatrix} \quad \text{and} \quad y_{new}=\begin{pmatrix} y \\ 0_{MxM}\end{pmatrix}$$

Now, taking the RSS expression for Ordinary Least Square Regression, we get,

$$\sum_{1=1}^{N+M}\left(y_i-\sum_{j=1}^{M}x_{ij}\beta_j\right)^2$$

$$= \quad \sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{M}(x_{ij}\beta_j)\right)^2+\sum_{i=N+1}^{N+M}\left(y_i+\sum_{j=1}^{M}(x_{ij}\beta_j)\right)^2$$

$$= \quad \sum_{j=1}^{m}\lambda\beta_j^2+\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{M}(x_{ij}\beta j)\right)^2$$

which is the Ridge function's objective function.

Thus, we can say that Ridge Regression estimates can be obtained from Ordinary Least Square Regression with an augmented data set.

Solution 4 :

Given : likelihood as $p(y|\Phi, w, \beta) = \prod_{i=1}^{N} N(y_i|w^T \phi(x_i), \beta^{-1})$ (1)

and conjugate prior as, $p(w, \beta) = N(w|m_o, \beta^{-1} S_o) . Gamma(\beta|a_o, b_o)$ (2)

We know that,

$$Gamma(\lambda|a, b) = \frac{1}{(\tau(a))} b^a \lambda^{(a-1)} \exp(-b\lambda)$$ (from Bishop Pg 100) (3)

Therefore in (2),

$$Gamma(\beta|a_o, b_o) = \frac{1}{(\tau(a_o))} b_o^{(a_o)} \beta^{(a_o-1)} \exp(-b_o \beta) \quad (4)$$

Also, using the definition of Gaussian Distribution,

the Gaussian part of eq. 1 and 2 respectively can be written as

$$N(y_i|w^T \phi(x_i), \beta^{-1}) = \frac{1}{(2\pi\beta^{-1})^{(1/2)}} \exp \frac{-1}{(2\beta^{-1})} (y_i - w^T \phi(x_i)^2) \quad (5)$$

$$N(w|m_o, \beta^{-1} S_o) = \frac{1}{(2\pi\beta^{-1} S_o)^{(1/2)}} \exp \frac{-1}{(2\beta^{-1} S_o)} (w - m_o) \quad (6)$$

In a Bayesian Approach we know that,

$$p(w \mid D) = \frac{p(D \mid w). p(w)}{p(D)} \quad (A)$$

Thus,

$$p(w, \beta|D) = \prod_{i=1}^{N} N(y_i|w^T \phi(x_i), \beta^{-1}) \; . \; p(w, \beta) = N(w|m_o, \beta^{-1} S_o) . Gamma(\beta|a_o, b_o) \quad (7)$$

Putting the values of (4),(5) and (6) in (7) and taking log,

we get, (8) as below :

$$\frac{-\beta}{2} \sum_{i=1}^{N} (w^T \phi(x_i) - y_i^2) + \frac{N}{2} \ln \beta + \frac{M}{2} \ln \beta - \frac{\beta}{2} (w - m_o)^T S_o^{-1} (w - m_o) - b_o \beta + (a_o - 1) \ln \beta - \frac{1}{2} \ln S_o + constant$$

Using Product rule, ln p(w | β,D) = p(w | β, D) . p(β | D)

we get,

p(w | β , D) = (8) / p(β | D)

Given a marginal Gaussian distribution for x and a conditional Gaussian distribution for y given x in the form :

p(x) = N(x| μ , Λ^{-1})

p(y) = N(y | Ax + b, L^{-1}) [Bishop 2.3]

The marginal distribution of y and the conditional distribution of x given y are given by

p(y) = N(y | Aμ +b, $L^{-1} + A\Lambda^{-1}A^T$)
p(x | y) = N(x | ∑{$A^T L$(y-b)+ Λμ}, ∑)

where ∑ = $(\Lambda+A^T LA)^{-1}$

Since the conjugate prior is a Gaussian, the posterior will also be a Gaussian.
where,

$$\boldsymbol{m_N}=\boldsymbol{S_N}\left(\boldsymbol{S_0^{-1}}\boldsymbol{m_0}+\phi^T y\right) \quad \text{is mean}$$

$$\boldsymbol{S_N^{-1}}=\beta\left(\boldsymbol{S_0^{-1}}+\phi^T\phi\right) \quad \text{is covariance}$$
