**Problem 2:** Let's assume we have a dataset where each datapoint, $(\boldsymbol{x}_i, y_i)$ is weighted by a scalar factor which we will call $t_i$. We will assume that $t_i > 0$ for all $i$. This makes the sum of squares error function look like the following:

$$E_{\text{weighted}}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N} t_i \left[\boldsymbol{w}^T \phi(\boldsymbol{x}_i) - y_i\right]^2$$

Find the equation for the value of $\boldsymbol{w}$ that minimizes this error function.

Furthermore, explain how this weighting factor, $t_i$, can be interpreted in terms of
1) the variance of the noise on the data and
2) data points for which there are exact copies in the dataset.

$$E = \frac{1}{2}(\Phi w - y)^T T (\Phi w - y) \qquad T = \begin{bmatrix} t_1 & & & 0 \\ & t_2 & & \\ & & \ddots & \\ 0 & & & t_N \end{bmatrix}$$

$$\frac{\partial}{\partial w} w^T A w = 2Aw$$

$$E = \frac{1}{2}\left[(w^T\Phi^T - y^T)\, T\, (\Phi w - y)\right]$$

$$= \frac{1}{2}\left[w^T \Phi^T T \Phi w - 2 w^T \Phi^T T y + y^T T y\right]$$

$$\frac{\partial E}{\partial w} = \frac{1}{2}\left[2\Phi^T T \Phi w - 2\Phi^T T y\right] \stackrel{!}{=} 0$$

$$\Phi^T T \Phi w_* = \Phi^T T y \implies w_* = (\Phi^T T \Phi)^{-1} \Phi^T T y$$

if $T = I_{N \times N} \implies w_* = (\Phi^T \Phi)^{-1} \Phi^T y$ (standard formula)

Lemma:

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} A \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \sum_{i,j} A_{ij} x_i x_j$$

$$\left[ w^T\Phi(x_1)-y_1, \; \dots, \; w^T\Phi(x_N)-y_N \right] \boxed{\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}} \begin{bmatrix} w^T\Phi(x_1)-y_1 \\ \vdots \\ w^T\Phi(x_N)-y_N \end{bmatrix}$$

---

$$y_i \sim \mathcal{N}(y_i \mid w^T x_i, \beta^{-1})$$

Approach 1 $\quad y_i \sim \mathcal{N}(y_i \mid w^T x_i, t_i^{-1})$

$$\log P(D|w) = \prod_i \log \mathcal{N}(y_i \mid w^T x_i, t_i^{-1})$$

$$= \sum_i - \frac{t_i}{2}(y_i - w^T x_i)^2$$

$$\Rightarrow \quad w_{MLE} = \operatorname*{argmax}_w \left[ - \sum_i \frac{t_i}{2}(y_i - w^T x_i)^2 \right]$$

$$= \operatorname*{argmin}_w \left[ \sum_i t_i (y_i - w^T x_i)^2 \right]$$

Approach 2 $\quad y_i \sim \mathcal{N}(y_i \mid w^T x_i, t_i \beta^{-1})$

---

Intrepretation 2:

$$\begin{matrix} x_1 \rightsquigarrow y_1 \\ x_2 \rightsquigarrow y_2 \\ x_3 \rightsquigarrow y_3 \end{matrix} \quad \Longrightarrow \quad \left\{ \begin{matrix} x_1 \rightsquigarrow y_1 \\ x_1 \rightsquigarrow y_1 \\ x_2 \rightsquigarrow y_2 \\ x_2 \rightsquigarrow y_2 \\ x_3 \rightsquigarrow y_3 \\ x_3 \rightsquigarrow y_3 \\ \vdots \end{matrix} \right.$$

$$\Rightarrow E = \sum_i t_i (y_i - w^T x_i)^2$$

Works only if $t_i \in \mathbb{N}$

$x_3 \rightsquigarrow y_3$

$\vdots$

$x_3 \longrightarrow y_3$

**Problem 3:** Show that the following holds: The ridge regression estimates can be obtained by ordinary least squares regression on an augmented dataset: Augment the design matrix $\mathbf{\Phi} \in \mathbb{R}^{N\times M}$ with $M$ additional rows $\sqrt{\lambda}\boldsymbol{I}_{M\times M}$ and augment $\boldsymbol{y}$ with $M$ zeros.

$$E = \frac{1}{2}\Big(\sum_i (y_i - \Phi(x_i)^T w)^2 + \underline{\frac{\lambda}{2}\|w\|_2^2} \Rightarrow \text{Ridge Regression}$$

$$\hookrightarrow \quad + \dots + w_M^2$$

extra sample $1 \Rightarrow [\sqrt{\lambda}, 0 \dots, 0]$ $0$ $\rightsquigarrow \lambda w_1^2$

$\vdots$

extra sample $N \Rightarrow [\dots\dots, \sqrt{\lambda}]$ $0$ $\rightsquigarrow \lambda w_M^2$

$$\begin{bmatrix} \Phi(x_1) \\ \vdots \\ \Phi(x_N) \\ \begin{matrix} \sqrt{\lambda} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda} \end{matrix} \end{bmatrix} w - \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} (\Phi(x_1)^T w - y_1) \\ \vdots \\ (\Phi(x_N)^T - y_N) \\ w_1\sqrt{\lambda} \\ \vdots \\ w_M\sqrt{\lambda} \end{bmatrix}$$

$\Rightarrow E =$ cost function Ridge Regression

$$= \sum_i (\Phi(x_i)^T w - y_i)^2 + \lambda w_1^2 + \lambda w_2^2 + \dots + w_M^2 \lambda$$

**Problem 4:** It turns out that the conjugate prior for the situation when we have an unknown mean and unknown precision is a normal-gamma distribution (See section 2.3.6 in Bishop). This is also true when we have a conditional Gaussian distribution of the linear regression model. This means that if our likelihood is as follows:

$$p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1})$$

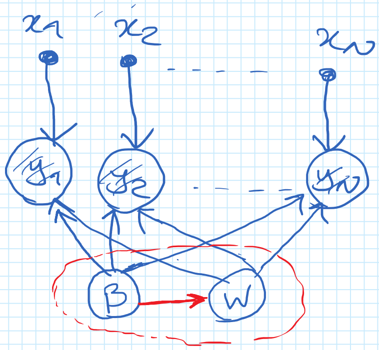
Then the conjugate prior for both $\boldsymbol{w}$ and $\beta$ is

$$p(\boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_0, \beta^{-1}\boldsymbol{S}_0)\mathrm{Gamma}(\beta \mid a_0, b_0)$$

Show that the posterior distribution takes the same form as the prior, i.e.

$$p(\boldsymbol{w}, \beta \mid \mathcal{D}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \beta^{-1}\boldsymbol{S}_N)\mathrm{Gamma}(\beta \mid a_N, b_N)$$

Also be sure to give the expressions for $\boldsymbol{m}_N$, $\boldsymbol{S}_N$, $a_N$, and $b_N$.




Normal Distribution

$$\mathcal{N}(x \mid \mu, \Sigma) = (2\pi)^{-\frac{M}{2}} |\Sigma|^{-1} \exp\{(x-\mu)^T \Sigma^{-1} (x-\mu)\}$$

$$\mathcal{N}_{can}(x \mid \eta, \Lambda) \propto \exp\{\eta^T x - \frac{1}{2} x^T \Lambda x\}$$

for example $\Rightarrow P(w) \propto \exp\{\underbrace{[1.0, 0.5, 0.3]}_{\eta} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \frac{1}{2} x^T \underbrace{I}_{\Lambda} x\}$

How to convert canonical form to standard form

$\mu = \Lambda^{-1}\eta \qquad \Sigma = \Lambda^{-1}$

$$P(w \mid D, \beta) = \frac{P(w, \beta \mid D)}{\cancel{P(\beta \mid D)}} \propto P(w, \beta \mid D)$$

$P(w \mid \beta) P(\beta) = P(w, \beta)$

$$\propto P(D \mid w, \beta)\, P(w, \beta)$$

$y^T y \rightarrow$ neglect $\qquad m_0^T S_0^{-1} m_0 \rightarrow$ neglect

$$\propto \exp\{-\frac{\beta}{2}(y - \Phi w)^T (y - \Phi w)\} \ldots$$

$y$ → neglect    $m_0^T S_0^{-1} m_0$ → neglect

$$\propto \exp\left\{-\frac{\beta}{2}(y-\Phi w)^T(y-\Phi w)\right\} \exp\left\{-\frac{\beta}{2}(w-m_0)^T S_0^{-1}(w-m_0)\right\}$$

$$\propto \exp\left\{-\frac{\beta}{2}\left[w^T\Phi^T\Phi w - 2y^T\Phi w\right]\right\} \exp\left\{-\frac{\beta}{2}\left[w^T S_0^{-1} w - 2m_0^T S_0^{-1} w\right]\right\}$$

$$\propto \exp\left\{\underbrace{\left[\beta\Phi^T y + \beta S_0^{-1} m_0\right]^T}_{\eta} w - \frac{1}{2} w^T \underbrace{\left[\beta\Phi\Phi^T + \beta S_0^{-1}\right]}_{\Lambda} w\right\}$$

$$\mathcal{N}_{can}\left(w \mid \beta\Phi^T y + \beta S_0^{-1} m_0,\ \beta\Phi^T\Phi + \beta S_0^{-1}\right)$$

conven → $\mathcal{N}\left(w \mid (\Phi^T\Phi + S_0^{-1})^{-1}(\Phi^T y + S_0^{-1} m_0),\ \frac{1}{\beta}(\Phi^T\Phi + S_0^{-1})^{-1}\right)$

$$\Rightarrow\ S_N = \Phi^T\Phi + S_0^{-1},\quad m_N = S_N(\Phi^T y + S_0^{-1} m_0)$$

---

$$P(\beta \mid D) = \frac{P(w,\beta \mid D)}{P(w \mid \beta, D)} = \frac{P(w,\beta)\,P(D \mid w,\beta)}{P(w \mid \beta, D)}$$

$$= \frac{\mathcal{N}(w \mid m_0, \beta^{-1}S_0)\ Gama(\beta \mid a_0, b_0)\ \prod_n \mathcal{N}(y_n \mid x_n, w)}{\mathcal{N}(w \mid m_N, \beta^{-1}S_N)}$$

$$= \frac{|\beta^{-1}S_0|^{-\frac{1}{2}} \exp\left\{-\frac{\beta}{2}(w-m_0)^T S_0^{-1}(w-m_0)\right\} e^{-b_0\beta}\, \beta^{a_0-1} \times \beta^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}(y-\Phi w)^T(y-\Phi w)\right\}}{|\beta^{-1}S_N|^{-\frac{1}{2}} \exp\left\{-\frac{\beta}{2}(w-m_N)^T S_N^{-1}(w-m_N)\right\}}$$

$\beta^{\frac{M}{2}} \times \beta^{a_0-1} \times \beta^{\frac{N}{2}}$ ...

$$= \frac{\cancel{\beta^{\frac{M}{2}}} \times \beta^{a_0 - 1} \times \beta^{\frac{N}{2}}}{\cancel{\beta^{\frac{M}{2}}}} \exp\left\{-\beta\left[\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0) + b_0 + \frac{1}{2}(Y-\Phi w)^T(Y-\Phi w) - \frac{1}{2}(w-m_N)^T S_N^{-1}(w-m_N)\right]\right\}$$

- $p(\beta|D)$ is not a function of $w$, but we see some terms contain $w$.

According to the values for $m_N, S_N$ (computed in previous part), the above expression becomes independent from $w$.

Because:

The terms containing $w$ = $w^T\underbrace{\left(S_0^{-1} + \Phi^T\Phi - S_N^{-1}\right)}_{0}w - 2\underbrace{\left(m_0^T S_0^{-1} + Y^T\Phi - m_N^T S_N^{-1}\right)}_{0}w$

$$= \beta^{a_0 + \frac{N}{2} - 1} \exp\left\{-\beta\left[\frac{1}{2} m_0^T S_0^{-1} m_0 - \frac{1}{2} m_N^T S_N^{-1} m_N + b_0 + \frac{1}{2} Y^T Y\right]\right\}$$

$$= \text{Gamma}\left(\beta \,\middle|\, a_0 + \frac{N}{2},\ b_0 + \frac{1}{2}\left(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + Y^T Y\right)\right)$$