

Problem 1: Weighted Sum-of-Squares and Replicated Data

We consider the weighted sum-of-squares error function:

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

To find the optimal weights \mathbf{w}^* , we calculate the gradient with respect to \mathbf{w} and set it to zero.

$$\begin{aligned} \nabla_{\mathbf{w}} E_D(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}} \left[\frac{1}{2} \sum_{n=1}^N r_n \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 \right] \\ &= \frac{1}{2} \sum_{n=1}^N r_n \cdot 2 \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\} (-\phi(\mathbf{x}_n)) \\ &= - \sum_{n=1}^N r_n \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\} \phi(\mathbf{x}_n) \\ &= - \sum_{n=1}^N r_n t_n \phi(\mathbf{x}_n) + \sum_{n=1}^N r_n (\mathbf{w}^T \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n) \\ &= - \sum_{n=1}^N r_n t_n \phi(\mathbf{x}_n) + \left(\sum_{n=1}^N r_n \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right) \mathbf{w} \end{aligned}$$

Setting $\nabla_{\mathbf{w}} E_D(\mathbf{w}) = 0$:

$$\left(\sum_{n=1}^N r_n \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right) \mathbf{w} = \sum_{n=1}^N r_n t_n \phi(\mathbf{x}_n)$$

Let \mathbf{R} be a diagonal matrix with elements $R_{nn} = r_n$. In matrix notation, this becomes:

$$\begin{aligned} (\Phi^T \mathbf{R} \Phi) \mathbf{w} &= \Phi^T \mathbf{R} \mathbf{t} \\ \Rightarrow \mathbf{w}^* &= (\Phi^T \mathbf{R} \Phi)^{-1} \Phi^T \mathbf{R} \mathbf{t} \end{aligned}$$

Interpretation via Replicated Data: Consider a dataset where each original data point (\mathbf{x}_n, t_n) is replicated r_n times (assuming integer r_n). The total number of points is $N' = \sum r_n$. The standard sum-of-squares error on this expanded dataset is:

$$\begin{aligned} E'(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^N \underbrace{\sum_{k=1}^{r_n} (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2}_{r_n \text{ identical terms}} \\ &= \frac{1}{2} \sum_{n=1}^N r_n (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 \end{aligned}$$

This is exactly the weighted error function $E_D(\mathbf{w})$. Thus, minimizing the error on the replicated dataset yields the same solution \mathbf{w}^* .

Problem 2: Bayesian Linear Regression Posterior

We derive the posterior distribution $p(\mathbf{w} \mid \mathbf{t})$ given a Gaussian prior $p(\mathbf{w})$ and likelihood $p(\mathbf{t} \mid \mathbf{w})$.

$$\begin{aligned} \text{Likelihood: } p(\mathbf{t} \mid \mathbf{w}) &= \prod_{n=1}^N \mathcal{N}(t_n \mid \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \\ \text{Prior: } p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0) \end{aligned}$$

Log-Prior Expansion:

$$\begin{aligned} \ln p(\mathbf{w}) &= -\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) + \text{const} \\ &= -\frac{1}{2} \mathbf{w}^T \mathbf{S}_0^{-1} \mathbf{w} + \mathbf{w}^T \mathbf{S}_0^{-1} \mathbf{m}_0 - \frac{1}{2} \mathbf{m}_0^T \mathbf{S}_0^{-1} \mathbf{m}_0 + \text{const} \end{aligned}$$

Log-Likelihood Expansion:

$$\begin{aligned} \ln p(\mathbf{t} \mid \mathbf{w}) &= \sum_{n=1}^N \ln \mathcal{N}(t_n \mid \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \\ &= -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 + \text{const} \\ &= -\frac{\beta}{2} \sum_{n=1}^N (t_n^2 - 2t_n \mathbf{w}^T \phi(\mathbf{x}_n) + \mathbf{w}^T \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \mathbf{w}) + \text{const} \\ &= -\frac{1}{2} \mathbf{w}^T \left(\beta \sum_{n=1}^N \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right) \mathbf{w} + \mathbf{w}^T \left(\beta \sum_{n=1}^N t_n \phi(\mathbf{x}_n) \right) - \frac{\beta}{2} \sum_{n=1}^N t_n^2 \end{aligned}$$

Posterior Identification: The log-posterior is the sum of the log-prior and log-likelihood:

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2} \mathbf{w}^T (\mathbf{S}_0^{-1} + \beta \Phi^T \Phi) \mathbf{w} + \mathbf{w}^T (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t}) + \text{const}$$

We match this to the functional form of a Gaussian $\mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N)$:

$$\ln \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N) = -\frac{1}{2} \mathbf{w}^T \mathbf{S}_N^{-1} \mathbf{w} + \mathbf{w}^T \mathbf{S}_N^{-1} \mathbf{m}_N + \text{const}$$

Comparing the quadratic terms (in \mathbf{w}):

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \Phi^T \Phi$$

Comparing the linear terms (in \mathbf{w}):

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

Problem 3: Joint Posterior with Unknown Noise Precision

We consider the joint prior $p(\mathbf{w}, \beta) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta \mid a_0, b_0)$. We seek the posterior $p(\mathbf{w}, \beta \mid \mathbf{t}) \propto p(\mathbf{t} \mid \mathbf{w}, \beta) p(\mathbf{w} \mid \beta) p(\beta)$.

Expanding the Joint Log-Posterior:

$$\ln p(\mathbf{w}, \beta \mid \mathbf{t}) \propto \ln p(\mathbf{t} \mid \mathbf{w}, \beta) + \ln p(\mathbf{w} \mid \beta) + \ln p(\beta)$$

1. **Likelihood Term:**

$$\begin{aligned} \ln p(\mathbf{t} \mid \mathbf{w}, \beta) &= \frac{N}{2} \ln \beta - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 \\ &= \frac{N}{2} \ln \beta - \frac{\beta}{2} (\mathbf{w}^T \Phi^T \Phi \mathbf{w} - 2 \mathbf{w}^T \Phi^T \mathbf{t} + \mathbf{t}^T \mathbf{t}) \end{aligned}$$

2. **Prior Term (Gaussian on \mathbf{w}):** Note the precision is $\beta \mathbf{S}_0^{-1}$.

$$\begin{aligned} \ln \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) &= \frac{D}{2} \ln \beta - \frac{1}{2} \ln |\mathbf{S}_0| - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\ &= \frac{D}{2} \ln \beta - \frac{\beta}{2} (\mathbf{w}^T \mathbf{S}_0^{-1} \mathbf{w} - 2 \mathbf{w}^T \mathbf{S}_0^{-1} \mathbf{m}_0 + \mathbf{m}_0^T \mathbf{S}_0^{-1} \mathbf{m}_0) + \text{const} \end{aligned}$$

3. **Prior Term (Gamma on β):**

$$\ln \text{Gam}(\beta \mid a_0, b_0) = (a_0 - 1) \ln \beta - b_0 \beta + \text{const}$$

Combining and Matching Terms: We assume the posterior form $p(\mathbf{w}, \beta \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \text{Gam}(\beta \mid a_N, b_N)$.

Matching coefficients of $\beta \mathbf{w}^T (\dots) \mathbf{w}$ (Quadratic in \mathbf{w}):

$$\begin{aligned} -\frac{\beta}{2} \mathbf{w}^T (\Phi^T \Phi + \mathbf{S}_0^{-1}) \mathbf{w} &= -\frac{\beta}{2} \mathbf{w}^T \mathbf{S}_N^{-1} \mathbf{w} \\ \Rightarrow \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \Phi^T \Phi \end{aligned}$$

Matching coefficients of $\beta \mathbf{w}^T (\dots)$ (Linear in \mathbf{w}):

$$\begin{aligned} \beta \mathbf{w}^T (\Phi^T \mathbf{t} + \mathbf{S}_0^{-1} \mathbf{m}_0) &= \beta \mathbf{w}^T \mathbf{S}_N^{-1} \mathbf{m}_N \\ \Rightarrow \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \Phi^T \mathbf{t}) \end{aligned}$$

Matching coefficients of $\ln \beta$: The likelihood contributes $N/2$. The Gaussian prior contributes $D/2$ (which is absorbed into the normalization of the Gaussian posterior). The Gamma prior contributes $(a_0 - 1)$. The posterior Gamma term is $(a_N - 1) \ln \beta$. However, we must account for the normalization constant of the Gaussian likelihood regarding β . The likelihood contains $\beta^{N/2}$. The prior contains $\beta^{a_0 - 1}$.

$$\beta^{N/2} \cdot \beta^{a_0 - 1} = \beta^{a_0 + N/2 - 1} \implies a_N = a_0 + \frac{N}{2}$$

Matching coefficients of $-\beta$: Collecting all terms linear in β that are not dependent on \mathbf{w} :

$$-b_N\beta = -b_0\beta - \frac{\beta}{2}\mathbf{t}^T\mathbf{t} - \frac{\beta}{2}\mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 + \underbrace{\frac{\beta}{2}\mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N}_{\text{from completing square in } \mathbf{w}}$$

Thus:

$$b_N = b_0 + \frac{1}{2}(\mathbf{t}^T\mathbf{t} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N)$$