Homework 4

Atlas42

October 2021

1 Posterior transformation

Bayes theorem:

$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$
$\Leftrightarrow \quad posterior = \frac{likelihood \times prior}{evidence}$
$\Rightarrow \quad p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \frac{p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{x}, \mathbf{t}, \alpha, \beta)}$

$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$ is a posterior. While likelihood is given the parameter how the parameter fit the data, posterior is given the data, what is the probability of parameter. In the posterior, we also include our belief.
We expect to maximize the posterior to find $\mathbf{w}$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Because $p(\mathbf{x}, \mathbf{t}, \alpha, \beta)$ is dependent on $\mathbf{w}$

Suppose $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) = (\frac{\alpha}{2\pi})^{\frac{M+1}{2}} e(-\frac{\alpha}{2}\mathbf{w^T w})$

So, now we maximize the posterior to find $\mathbf{w}$

$$\begin{aligned} p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) &= \textstyle\sum_{n=1}^{N} \log(\mathcal{N}(t_n|y(X_n, w), \beta^{-1})) \times \log(\mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) \\ &= -\tfrac{\beta}{2} \textstyle\sum_{n=1}^{N} (y(x_n, \mathbf{w} - t_n))^2 - \tfrac{\alpha}{2}\mathbf{w}^T\mathbf{w} \end{aligned}$$

$\Rightarrow$ maximum of posterior is given by minimizing:

$$\tfrac{\beta}{2} \textstyle\sum_{n=1}^{N} (y(x_n, \mathbf{w} - t_n))^2 + \tfrac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Thus, we have:

$$\mathrm{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \lambda = \begin{bmatrix} 1 & \lambda_1 \\ 1 & \lambda_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & \lambda_n \end{bmatrix}, \mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \mathrm{t} = \begin{bmatrix} t_1 \\ t_2 \\ \cdot \\ \cdot \\ t_n \end{bmatrix}$$

$$\Rightarrow \text{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} w_0 + x_1 W_1 \\ w_0 + x_2 W_1 \\ \cdot \\ \cdot \\ w_0 + x_n W_1 \end{bmatrix} = \text{xw and } \begin{bmatrix} w_0 + \lambda_1 W_1 \\ w_0 + \lambda_2 W_1 \\ \cdot \\ \cdot \\ w_0 + \lambda n W_1 \end{bmatrix} = \lambda \mathbf{w^T w}$$

Our loss function is: $\left\|Xw - t\right\|_2^2 + \lambda \mathbf{w^T w} = \begin{bmatrix} t_1 - y_1 \\ t_2 - y_2 \\ \cdot \\ \cdot \\ t_n - y_n \end{bmatrix} + \begin{bmatrix} \lambda_1 W_1 \\ \lambda_2 W_1 \\ \cdot \\ \cdot \\ \lambda_n W_1 \end{bmatrix}$

Now, to minimize our loss function with $\frac{\partial l}{\partial w} = 0$

$$\Rightarrow \frac{\partial L}{\partial w} = \begin{bmatrix} \frac{t_1 - y_1}{\partial w} \\ \frac{t_2 - y_2}{\partial w} \\ \cdot \\ \cdot \\ \frac{t_n - y_n}{\partial w} \end{bmatrix} + \begin{bmatrix} \frac{\lambda_1 W_1}{\partial w} \\ \frac{\lambda_2 W_2}{\partial w} \\ \cdot \\ \cdot \\ \frac{\lambda_n W_n}{\partial w} \end{bmatrix} = \begin{bmatrix} 2X_1 \frac{t_1 - x_1 W_1}{\partial w} \\ 2X_2 \frac{t_2 - x_2 W_2}{\partial w} \\ \cdot \\ \cdot \\ 2X_n \frac{t_1 - x_1 W_1}{\partial w} \end{bmatrix} + \begin{bmatrix} 2\frac{\lambda_1 W_1}{\partial w} \\ 2\frac{\lambda_2 W_2}{\partial w} \\ \cdot \\ \cdot \\ 2\frac{\lambda_n W_n}{\partial w} \end{bmatrix} = 0$$

$$\Leftrightarrow 2X^T(t - XW) + 2\lambda W = 0$$

$$\Leftrightarrow X^T X - X^T X W + \lambda W = 0$$

$$\Leftrightarrow X^T X W + \lambda W = X^T t$$

$$\Leftrightarrow W(X^T X + \lambda \mathcal{I}) = X^T t$$

$$\Leftrightarrow W = (X^T X + \lambda \mathcal{I})^{-1} X^T t$$