Machine Learning

Maximum Likelihood of the Normal Distribution with Restrictions

Irving Gómez Méndez

UNIVERSIDAD IBEROAMERICANA
CIUDAD DE MÉXICO ®

Let $y|X \sim \mathcal{N}(X^T\beta, \sigma^2)$, $\beta \in \mathbb{R}^p$ and assume that we want to maximize the log-likelihood subject to some linear conditions, that is, we want to maximize:

$$\ell(\beta, \sigma^2) = -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) - \frac{n}{2}\log(\sigma^2)$$
$$\text{subject to } \mathbf{K}\beta = \mathbf{m}$$

where $\mathbf{K}$ is a $q \times p$ matrix of range q, $q \leq p$.

The Lagrange function is given by

$$L(\sigma^2, \beta, \lambda) = -\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)-\frac{n}{2}\log(\sigma^2)-\lambda^T(\mathbf{K}\beta-\mathbf{m})$$

Let be $\hat{\beta}_R$ and $\hat{\sigma}_R^2$ the values that maximize the log-likelihood and satisfy the restrictions.These values must satisfy

$$\left.\frac{\partial L}{\partial \sigma^2}\right|_{(\hat{\sigma}_R^2, \hat{\beta}_R, \hat{\lambda})} = 0, \quad \left.\frac{\partial L}{\partial \beta}\right|_{(\hat{\sigma}_R^2, \hat{\beta}_R, \hat{\lambda})} = 0, \quad \text{and} \quad \left.\frac{\partial L}{\partial \lambda}\right|_{(\hat{\sigma}_R^2, \hat{\beta}_R, \hat{\lambda})} = 0$$

Thus,

$$\begin{aligned} \left.\frac{\partial L}{\partial \sigma^2}\right|_{(\hat{\sigma}_R^2, \hat{\beta}_R, \hat{\lambda})} &= 0 \\ \Leftrightarrow \frac{1}{2\hat{\sigma}_R^4}(\mathbf{Y} - \mathbf{X}\hat{\beta}_R)^T(\mathbf{Y} - \mathbf{X}\hat{\beta}_R) - \frac{n}{2\hat{\sigma}_R^2} &= 0 \\ \Leftrightarrow \hat{\sigma}_R^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\beta}_R)^T(\mathbf{Y} - \mathbf{X}\hat{\beta}_R) &\equiv \frac{1}{n}SSR(\hat{\beta}_R) \end{aligned}$$

$$
\begin{aligned}
\left.\frac{\partial L}{\partial \beta}\right|_{\left(\hat{\sigma}_R^2, \hat{\beta}_R, \hat{\lambda}\right)} &= 0 \\
\Leftrightarrow -\frac{1}{2\hat{\sigma}_R^2}\left(2\mathbf{X}^T\mathbf{X}\hat{\beta}_R - 2\mathbf{X}^T\mathbf{Y}\right) - \mathbf{K}^T\hat{\lambda} &= 0 \\
\Leftrightarrow \mathbf{X}^T\mathbf{X}\hat{\beta}_R + \hat{\sigma}_R^2\mathbf{K}^T\hat{\lambda} &= \mathbf{X}^T\mathbf{Y} \\
\Leftrightarrow \hat{\beta}_R + \hat{\sigma}_R^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{K}^T\hat{\lambda} &= \hat{\beta} \\
\Leftrightarrow \mathbf{K}\hat{\beta}_R + \hat{\sigma}_R^2\mathbf{K}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{K}^T\hat{\lambda} &= \mathbf{K}\hat{\beta}
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\left.\frac{\partial L}{\partial \lambda}\right|_{\left(\hat{\sigma}_R^2, \hat{\beta}_R, \hat{\lambda}\right)} &= 0 \\
\Leftrightarrow \mathbf{K}\hat{\beta}_R &= \mathbf{m}
\end{aligned}
$$

From the last expression, we have

$$\mathbf{m} + \hat{\sigma}_R^2 \mathbf{K} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{K}^T \hat{\lambda} = \mathbf{K}\hat{\beta}$$

$$\Leftrightarrow \hat{\lambda} = \frac{1}{\hat{\sigma}_R^2} \left[\mathbf{K} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{K}^T\right]^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)$$

Substituting for $\hat{\beta}_R$,

$$\hat{\beta}_R + \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{K}^T \left[\mathbf{K} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{K}^T\right]^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) = \hat{\beta}$$

$$\Leftrightarrow \hat{\beta}_R = \hat{\beta} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{K}^T \left[\mathbf{K} \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{K}^T\right]^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)$$