3 Regression

3.1 Ridge regression models

i.

$$\text{Let } \mathbf{X} := \begin{bmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ & \vdots & \\ x_{n1} & \dots & x_{nm} \end{bmatrix}, \quad \mathbf{w} := \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \text{ and } \mathbf{y} := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

The ordinary least squares regression cost function can be written in matrix notation is given by,

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y}).$$

Adding the quadratic penalty term, we get

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}$$

ii. $J(\mathbf{w})$ can be simplified as:

$$\begin{aligned} J(\mathbf{w}) &= \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \\ &= \frac{1}{2}\left[(\mathbf{w}^\top\mathbf{X}^\top - \mathbf{y}^\top)(\mathbf{X}\mathbf{w} - \mathbf{y})\right] + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \\ &= \frac{1}{2}\left[\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{y}^\top\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y})\right] + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \\ &= \frac{1}{2}\left[\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{y}^\top\mathbf{X}\mathbf{w} + \mathbf{y}^\top\mathbf{y}\right] + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \end{aligned}$$

Using results from task 2.1 [Supplement-2], the gradient of $J(\mathbf{w})$:

$$\begin{aligned} \nabla_{\mathbf{w}} J(\mathbf{w}) &= \frac{1}{2}\left[\left(\mathbf{X}^\top\mathbf{X} + (\mathbf{X}^\top\mathbf{X})^\top\right)\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}\right] + \frac{\lambda}{2}[2\mathbf{w}] \\ &= \frac{1}{2}\left[2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{y}\right] + \frac{\lambda}{2}[2\mathbf{w}] \\ &= \mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{y} + \lambda\mathbf{w} \end{aligned}$$

Setting gradient $\nabla_{\mathbf{w}} J(\mathbf{w})$ to zero, we obtain:

$$\begin{aligned} 0 &= \mathbf{X}^\top\mathbf{X}\mathbf{w}^* - \mathbf{X}^\top\mathbf{y} + \lambda\mathbf{w}^* \\ \mathbf{w}^* &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \end{aligned}$$