Machine learning : Sheet 2

Author : Djordje Zivanovic

1. On the lectures we got the following equation:

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{2}\log(2\pi\sigma^2)$$

   By setting derivative over σ to 0, we get the following:

$$\begin{aligned}\frac{\partial \mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma)}{\partial \sigma} &= \frac{-2\sigma^{-3}}{2}(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{2}\frac{2\pi 2\sigma}{2\pi\sigma^2} \\ &= -\sigma^{-3}(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{\sigma} = 0\end{aligned}$$

   Multiplying the last expression with σ^3, because variance cannot be 0, and dividing by N we get

$$\sigma^2 = \frac{(\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} - \mathbf{y})}{N}$$

   Since optimal $\mathbf{w}$ is calculated by setting gradient to zero, we insert $\mathbf{w}_{\mathbf{ML}}$ to the last equation and get the necessary σ.

2. By using vector multiplication rules we can simplify the expression:

$$\begin{aligned}\mathcal{L}_{ridge}(\mathbf{w}, b) &= (\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y})^{\mathrm{T}}(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y}) + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}b\mathbf{1} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + b\mathbf{1}^{\mathrm{T}}\mathbf{X}\mathbf{w} + b^2N - b\mathbf{1}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} - by^{\mathrm{T}}\mathbf{1} + \mathbf{y}^{\mathrm{T}}\mathbf{y} + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}\end{aligned}$$

   Deriving $\mathcal{L}_{ridge}(\mathbf{w}, b)$ per b and equaling to zero we get:

$$\begin{aligned}\frac{\partial \mathcal{L}_{ridge}(\mathbf{w}, b)}{\partial b} &= 0 + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{1} - 0 + \mathbf{1}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 2bN - \mathbf{1}^{\mathrm{T}}\mathbf{y} - 0 - y^{\mathrm{T}}\mathbf{1} + 0 + 0 \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{1} + \mathbf{1}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 2bN - 2\mathbf{y}^{\mathrm{T}}\mathbf{1} = 0\end{aligned}$$

   Using the property that product of $\mathbf{A}^{\mathrm{T}}B = B^{\mathrm{T}}A$ when **A,B** are column vectors we got the simplification of the last equation. Further, $\mathbf{X}^{\mathrm{T}}$ when multiplied by $\mathbf{1}$ from column matrix whose cells are the sum of the same features. From the beginning condition that sum for every possible set of features is 0, the first two addends will give 0 as a result (associativity of matrix multiplication). Thus, we get $b = \mathbf{y}^{\mathrm{T}}$ which is equivalent to $\hat{b} = \frac{1}{N}\sum_{i=1}^{N} y_i$.
   Deriving $\mathcal{L}_{ridge}(\mathbf{w}, b)$ per $\mathbf{w}$ and equaling to zero we get:

$$\begin{aligned}\frac{\partial \mathcal{L}_{ridge}(\mathbf{w}, b)}{\partial \mathbf{w}} &= 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} + 0 - \mathbf{X}^{\mathrm{T}}\mathbf{y} + 0 + 0 - 0 - \mathbf{X}^{\mathrm{T}}\mathbf{y} \;\; - 0 + 0 + 2\lambda\mathbf{w} \\ &= 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\lambda\mathbf{w} = 0\end{aligned}$$

   In the previous derivation we simplified some expression (set them to zero) using the same observation $\mathbf{X}^{\mathrm{T}}\mathbf{1} = 0$. Using distributivity of matrix multiplication

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_D)\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

   Multiplying the both sides by $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_D)$ we get the $\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
   Consequence of centering $\mathbf{y}$ gives us that bias is unnecessary in this case (we would need neither to calculate nor to use it).