AI2611 Machine Learning (Spring 2023) Assignment 1

Xiangyuan Xue (521030910387)

Ridge regression is a generalized linear regression which introduces an L-2 regularization term. The object function can be specified as

$$L(\boldsymbol{\beta}) = (\boldsymbol{Z\beta} - \boldsymbol{y})^T (\boldsymbol{Z\beta} - \boldsymbol{y}) + \lambda \cdot \boldsymbol{\beta}^T \boldsymbol{\beta}$$

where $\boldsymbol{Z} \in \mathbb{R}^{m \times n}$, $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$. Let the derivative of be zero, we have

$$\nabla_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = 2 \left(\boldsymbol{Z}^T \boldsymbol{Z\beta} - \boldsymbol{Z}^T \boldsymbol{y} + \lambda \boldsymbol{\beta} \right) = 0$$

which yields

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \left(\boldsymbol{Z}^T \boldsymbol{Z} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{Z}^T \boldsymbol{y}$$

namely the closed-form solution of ridge regression.

However, numerical computation of $\left(\boldsymbol{Z}^T \boldsymbol{Z} + \lambda \boldsymbol{I} \right)^{-1}$ is expensive. Instead, we could use singular value decomposition (SVD) to lower the computation cost. Suppose

$$\boldsymbol{Z} = \boldsymbol{UDV}^T$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p)$ is an $n \times p$ orthogonal matrix, $\boldsymbol{D} = \operatorname{diag}(d_1, d_2, \ldots, d_p)$ is a $p \times p$ diagonal matrix consisting singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ is a $p \times p$ orthogonal matrix. Note that $\boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{V}^T \boldsymbol{V} = \boldsymbol{I}_p$ and $\boldsymbol{D}^T \boldsymbol{D} = \boldsymbol{D}^2$, thus it holds that

$$\begin{aligned}
\boldsymbol{Z}^T \boldsymbol{Z} &= \left(\boldsymbol{UDV}^T \right)^T \left(\boldsymbol{UDV}^T \right) \\
&= \boldsymbol{VD}^T \boldsymbol{U}^T \boldsymbol{UDV}^T \\
&= \boldsymbol{VD}^T \boldsymbol{DV}^T \\
&= \boldsymbol{VD}^2 \boldsymbol{V}^T
\end{aligned}$$

Then the closed-form solution can be rewritten as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} &= \left(\boldsymbol{VD}^2 \boldsymbol{V}^T + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{VD}^T \boldsymbol{U}^T \boldsymbol{y} \\
&= \left[\boldsymbol{V} \left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p \right) \boldsymbol{V}^T \right]^{-1} \boldsymbol{VDU}^T \boldsymbol{y} \\
&= \left(\boldsymbol{V}^T \right)^{-1} \left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{V}^{-1} \boldsymbol{VDU}^T \boldsymbol{y} \\
&= \boldsymbol{V} \left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{DU}^T \boldsymbol{y}
\end{aligned}$$

Since $\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p = \operatorname{diag}_j \left(d_j^2 + \lambda\right)$, we have $\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1} = \operatorname{diag}_j \left(\frac{1}{d_j^2+\lambda}\right)$, then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} &= \boldsymbol{V} \operatorname{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) \boldsymbol{D} \boldsymbol{U}^T \boldsymbol{y} \\
&= \boldsymbol{V} \operatorname{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) \operatorname{diag}_j \left(d_j\right) \boldsymbol{U}^T \boldsymbol{y} \\
&= \boldsymbol{V} \operatorname{diag}_j \left(\frac{d_j}{d_j^2 + \lambda}\right) \boldsymbol{U}^T \boldsymbol{y}
\end{aligned}$$

which is just the required form. Since SVD can be computed efficiently, the computation of ridge regression can be largely accelerated.