Ridge Regression and PCA

Archer Yang

McGill University

September 25, 2024

- **Reading assignment:** Chapter 3.2-3.4 The Elements of Statistical Learning

Efficient Computation of Ridge Regression

Ridge regression

- The ridge penalized least squares estimator of $\hat{\boldsymbol{\beta}}^{(\lambda)}$ is defined by

$$\hat{\boldsymbol{\beta}}^{(\lambda)} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{1}$$

  where $\lambda \geq 0$

- Solve $\partial f(\boldsymbol{\beta}^{(\lambda)}) = 0$ for $\boldsymbol{\beta}^{(\lambda)}$, which is

$$\begin{aligned} -2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}^{(\lambda)} + 2\lambda\boldsymbol{\beta}^{(\lambda)} &= 0 \\ (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)\boldsymbol{\beta}^{(\lambda)} &= \boldsymbol{X}^\top\boldsymbol{y} \qquad (2) \\ \boldsymbol{\beta}^{(\lambda)} &= (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \qquad (3) \end{aligned}$$

Efficient computation

- (2) can be solved with a linear system solver
- Or use the less efficient closed-form solution in (3)
- Both methods have to be recomputed for a different value of λ, computationally expensive when p is large and inefficient if we wish to compute $\boldsymbol{\beta}^{(\lambda)}$ for multiple values of λ.

Efficient computation

- Using the full SVD of $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{U} = \mathbb{R}^{n\times p}$, $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{D} = \mathbb{R}^{p\times p}$

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{V}^\top$$

- So write (2) as

$$(\boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{V}^\top + \lambda\boldsymbol{I}_p)\boldsymbol{\beta}^{(\lambda)} = \boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{U}^\top\boldsymbol{y}$$

  and replacing $\boldsymbol{I}_p$ with $\boldsymbol{V}\boldsymbol{V}^\top$ gives

$$\boldsymbol{V}(\boldsymbol{D}^\top\boldsymbol{D} + \lambda\boldsymbol{I}_p)\boldsymbol{V}^\top\boldsymbol{\beta}^{(\lambda)} = \boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{U}^\top\boldsymbol{y}$$

- $\boldsymbol{V}(\boldsymbol{D}^\top\boldsymbol{D} + \lambda\boldsymbol{I}_p)\boldsymbol{V}^\top$ is the eigen-decomposition of $\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p$, which is positive semi-definite.

Efficient computation

- Assuming that $\lambda > 0$,

$$\begin{aligned}
\boldsymbol{\beta}^{(\lambda)} &= \boldsymbol{V}(\boldsymbol{D}^\top \boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{V}^\top \boldsymbol{V} \boldsymbol{D}^\top \boldsymbol{U}^\top \boldsymbol{y} \\
&= \boldsymbol{V}(\boldsymbol{D}^\top \boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{D}^\top \boldsymbol{U}^\top \boldsymbol{y} \\
&= \boldsymbol{V} \boldsymbol{M} \boldsymbol{U}^\top \boldsymbol{y}
\end{aligned}$$

where $\boldsymbol{M} = (\boldsymbol{D}^\top \boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{D}^\top = \text{diag}(m_1, \cdots, m_p) \in \mathbb{R}^{p \times p}$ is diagonal where

$$m_j = \begin{cases} d_j / (d_j^2 + \lambda) & j = 1, \cdots, r \\ 0 & j = r, \cdots, p \end{cases}$$

Efficient computation

- We can use the reduced SVD on $\boldsymbol{X}$, for rank($\boldsymbol{X}$) = $r \leq \min(n, p)$,

$$\boldsymbol{X} = \boldsymbol{U}_r \boldsymbol{D}_r \boldsymbol{V}_r^\top$$

where $\boldsymbol{U}_r = \boldsymbol{U}_{[,1,\cdots,r]}$ and $\boldsymbol{V}_r = \boldsymbol{V}_{[,1,\cdots,r]}$, therefore

$$\boldsymbol{\beta}^{(\lambda)} \quad = \quad \boldsymbol{V}_r \boldsymbol{M}_r \boldsymbol{U}_r^\top \boldsymbol{y}$$

where $\boldsymbol{M}_r = \text{diag}(m_1, \cdots, m_r) \in \mathbb{R}^{r \times r}$ with diagonal values $m_j = d_j / (d_j^2 + \lambda)$.

Connection between Ridge Regression and PCA

OLS solution

Using the SVD we can write the OLS fitted vector as

$$\begin{aligned}\hat{\boldsymbol{y}}^{\mathrm{OLS}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{y}.\end{aligned}$$

- $\boldsymbol{U}^\top\boldsymbol{y}$ are the coordinates of $\boldsymbol{y}$ with respect to the basis $\boldsymbol{U}$.
- We see that $\boldsymbol{X}\boldsymbol{V} = \boldsymbol{X}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r) = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r)\boldsymbol{D} = \boldsymbol{U}\boldsymbol{D}$.
- $\boldsymbol{u}_j$ can be viewed as scaled version of the coordinates $\mathbf{z}_j$, because

$$\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{v}_j = \boldsymbol{u}_j d_j \qquad \mathrm{Var}(\boldsymbol{z}_j) = \mathrm{Var}(\boldsymbol{X}\boldsymbol{v}_j) = \frac{d_1^2}{N}$$

- Can view $\hat{\boldsymbol{y}}^{\mathrm{OLS}}$ is a reconstruced version of $\boldsymbol{y}$

Ridge solutions

On the other hand, the ridge solutions are

$$\begin{aligned}
\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{ridge}} &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} \\
&= \sum_{j=1}^{p}\boldsymbol{u}_j\frac{d_j^2}{d_j^2+\lambda}\boldsymbol{u}_j{}^\top\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{H}\boldsymbol{U}^\top\boldsymbol{y}
\end{aligned}$$

where

$$\boldsymbol{H} = \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{D}$$

is diagonal with

$$h_j = d_j^2/(d_j^2+\lambda) \qquad j = 1, \cdots, p$$

Ridge solutions

- Like $\hat{\boldsymbol{\beta}}^{\text{OLS}}$, $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ projects $\mathbf{y}$ to $\mathbf{U}^T\mathbf{y}$ with respect to the orthonormal basis $\mathbf{U}$.
- It then shrinks these projected $\mathbf{y}$ by the factors $d_j^2/(d_j^2 + \lambda)$.
- And then reconstructs $\mathbf{y}$ using $\mathbf{U}\mathbf{H}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2+\lambda} \boldsymbol{u}_j{}^\top \boldsymbol{y}$. This means that a greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.