Statistical Learning II

Lecture 7 - Ridge regression

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
06.11.2024

Marvels and pitfalls of OLS

Recall that:

$$\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star + \frac{1}{n}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}$$

$$= \boldsymbol{\theta}_\star + \sum_{j=1}^{d} \frac{1}{\sigma_j}\langle \boldsymbol{u}_j, \boldsymbol{\varepsilon} \rangle \boldsymbol{v}_j$$

Hence:

- signal is stronger in directions with larger s.v.
- noise dominates directions with smaller s.v.

OLS has larger variance for data with small "effective dimension".

What to do?

Classical strategies to mitigate variance:

- Dimensionality reduction: PCA, random projections (sketching), etc.

- Variable subset selection: Stepwise selection, best Subset Selection, etc.

- Regularisation: ridge, LASSO, etc.

Ridge regression

Ridge regression

Note the averaged norm of the OLS is given by:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[||\hat{\boldsymbol{\theta}}_{OLS}||_2^2\right] = ||\boldsymbol{\theta}_\star||_2^2 + \sigma^2 \sum_{j=1}^{d} \frac{1}{\sigma_j^2}$$

Therefore, small s.v.s lead to larger expected norm.

Ridge regression

Note the averaged norm of the OLS is given by:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[||\hat{\boldsymbol{\theta}}_{OLS}||_2^2\right] = ||\boldsymbol{\theta}_\star||_2^2 + \sigma^2 \sum_{j=1}^{d} \frac{1}{\sigma_j^2}$$

Therefore, small s.v.s lead to larger expected norm.

💡 Key idea: penalise the norm.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{2n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Ridge regression

Note the averaged norm of the OLS is given by:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[||\hat{\boldsymbol{\theta}}_{OLS}||_2^2\right] = ||\boldsymbol{\theta}_\star||_2^2 + \sigma^2 \sum_{j=1}^{d} \frac{1}{\sigma_j^2}$$

Therefore, small s.v.s lead to larger expected norm.

<u>Key idea</u>: penalise the norm.

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

Least squares empirical risk

Regularisation or "ridge" penalty

Ridge regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Ridge regression

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) := \frac{1}{2n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Ridge regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) := \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

Remarks:

- The regularised empirical risk is a strongly convex function of $\boldsymbol{\theta} \in \mathbb{R}^d$

$$\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) = -\frac{1}{n}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta}$$

$$\nabla_{\boldsymbol{\theta}}^2 \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_d \succ 0$$

$$(= \hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_n)$$

Ridge regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) := \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

Remarks:

- The regularised empirical risk is a strongly convex function of $\boldsymbol{\theta} \in \mathbb{R}^d$

$$\nabla_{\boldsymbol{\theta}}\hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) = -\frac{1}{n}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta}$$

$$\nabla_{\boldsymbol{\theta}}^2\hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_d \succ 0$$

$$(= \hat{\boldsymbol{\Sigma}}_n + \lambda\boldsymbol{I}_n)$$

In other words, minimiser always exist and is unique.

Ridge regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathscr{R}}_n^\lambda(\boldsymbol{\theta}) := \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

$$\nabla_{\boldsymbol{\theta}}\hat{\mathscr{R}}_n^\lambda(\boldsymbol{\theta}) = -\frac{1}{n}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta} \stackrel{!}{=} \mathbf{0}$$

Ridge regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) := \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

$$\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) = -\frac{1}{n}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta} \overset{!}{=} \boldsymbol{0}$$

$$\Updownarrow$$

$$\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta} + \lambda\boldsymbol{I}_d\right)\boldsymbol{\theta} \overset{!}{=} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{y}$$

Ridge regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) := \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

$$\nabla_{\boldsymbol{\theta}}\hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) = -\frac{1}{n}\boldsymbol{X}^\top\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right) + \lambda\boldsymbol{\theta} \stackrel{!}{=} \mathbf{0}$$

$$\Updownarrow$$

$$\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta} + \lambda\boldsymbol{I}_d\right)\boldsymbol{\theta} \stackrel{!}{=} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{y}$$

$$\Updownarrow$$

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X},\boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Ridge regression

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) := \frac{1}{2n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

The unique solution is given by:

$$\hat{\boldsymbol{\theta}}_{\lambda}(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n} \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Ridge regression

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\mathscr{R}}_n^{\lambda}(\boldsymbol{\theta}) := \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

The unique solution is given by:

$$\hat{\boldsymbol{\theta}}_{\lambda}(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$$

For $\lambda \to 0^+$, $\hat{\boldsymbol{\theta}}_{\lambda} \to \hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$

Ridge regression

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

<u>Remarks:</u>

- As before, consider s.v.d. of $\boldsymbol{X} = \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$

Ridge regression

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

<u>Remarks:</u>

- As before, consider s.v.d. of $\boldsymbol{X} = \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \frac{\sigma_j}{\sigma_j^2 + n\lambda} \langle \boldsymbol{u}_j, \boldsymbol{y} \rangle \boldsymbol{v}_j$$

Ridge regression

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

<u>Remarks:</u> • As before, consider s.v.d. of $\boldsymbol{X} = \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \frac{\sigma_j}{\sigma_j^2 + n\lambda} \langle \boldsymbol{u}_j, \boldsymbol{y} \rangle \boldsymbol{v}_j$$

Ridge performs shrinkage:
small s.v.s are suppressed!