Answers to TP Theoretical Questions

Introduction Supervised Learning IMA205

Abdennour K.

13 mars 2025

Ordinary Least Square (OLS)

Let C be an unbiased linear estimator. Let D such that $C = H + D$ and $\tilde{\beta} = C\mathbf{y} = \hat{\beta} + D\mathbf{y}$, then :

$$\mathbb{E}(\tilde{\beta}) = \beta + \mathbb{E}(D\mathbf{y}) = \mathbb{E}(D(\mathbf{x}\beta + \varepsilon)) = \beta + D\mathbf{x}\beta$$

C is unbiased so $\mathbb{E}(C\mathbf{y}) = \beta$ and then $D\mathbf{x} = 0$. It follows :

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}(C\mathbf{y}) \\
&= C\,\mathrm{Var}(\mathbf{y})C^t \\
&= C\,\mathrm{Var}(\mathbf{x}\beta + \varepsilon)C^t \\
&= C\sigma^2 I_d C^t = \sigma^2 CC^t
\end{aligned}$$

Since $C = H + D$:

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \sigma^2(HH^t + HD^t + DH^t + DD^t) \\
&= \sigma^2((\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{x}(\mathbf{x}^t\mathbf{x})^{-1} + (\mathbf{x}^t\mathbf{x})^{-1}(D\mathbf{x})^t + D\mathbf{x}(\mathbf{x}^t\mathbf{x})^{-1} + DD^t) \\
&= \sigma^2((\mathbf{x}^t\mathbf{x})^{-1} + DD^t) = \mathrm{Var}(\hat{\beta}) + \sigma^2 DD^t > \mathrm{Var}(\hat{\beta})
\end{aligned}$$

We used $\mathbb{E}(\varepsilon\varepsilon^t) = \sigma^2 I_d$, that the error vector is centered ($\mathbb{E}(\varepsilon) = 0$) and the samples $\mathbf{x}$ are non-stochastic.

Ridge estimator

- Let's compute the gradient of the objective function to minimize :

  $$\nabla f(\beta) = -2\mathbf{x}_c^t(\mathbf{y}_c - \mathbf{x}_c\beta) + 2\lambda\beta$$

  The minimum is reached at :

  $$\beta^*_{ridge} = (\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\mathbf{x}_c^t\mathbf{y}_c$$

  Then the bias is given by :

  $$B = \mathbb{E}(\beta^*_{ridge}) - \beta^*_{ridge} = (\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\mathbf{x}_c^t\mathbf{x}_c\beta - \beta = [(\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\mathbf{x}_c^t\mathbf{x}_c - I_d]\beta \begin{cases} \neq 0 & \textbf{if } \lambda > 0 \\ = 0 & \textbf{if } \lambda = 0 \end{cases}$$

  The estimator is unbiased iff $\lambda = 0$.

- Note $\mathbf{x}_c = UDV^t$.

$$\begin{aligned}\beta^*_{ridge} &= (VD^TU^tUDV^t + \lambda I_d)^{-1}VDU^t\mathbf{y}_c \\ &= V(D^2 + \lambda I_d)^{-1}V^tVDU^t\mathbf{y}_c \\ &= V(D^2 + \lambda I_d)^{-1}U^t\mathbf{y}_c\end{aligned}$$

  It is far more easy to invert $D^2 + \lambda I_d$ that is a diagonal matrix than $\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d$. By noting d_k the k-th coefficient of D we have :

$$[D^2 + \lambda I_d]_{k,k} = \frac{1}{d_k^2 + \lambda}$$

  It is very useful especially when $\mathbf{x}_c$ is high dimension and so the inversion present a great computational cost.

- Let's compute $\mathrm{Var}(\beta^*_{ridge})$:

$$\begin{aligned}\mathrm{Var}(\beta^*_{ridge}) &= \mathrm{Var}((\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\mathbf{x}_c^t\mathbf{y}) \\ &= (\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\mathbf{x}_c^t \,\mathrm{Var}(y)\mathbf{x}_c(\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1} \\ &= \sigma^2(\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\mathbf{x}_c^t\mathbf{x}_c(\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\end{aligned}$$

  For $\lambda > 0$, $\mathbf{x}_c^t\mathbf{x}_c < \mathbf{x}_c\mathbf{x}_c^t + \lambda I_d$. Then :

$$\begin{aligned}\mathrm{Var}(\beta^*_{ridge}) &= \sigma^2(\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1}\mathbf{x}_c^t\mathbf{x}_c(\mathbf{x}_c^t\mathbf{x}_c + \lambda I_d)^{-1} \\ &< \sigma^2(\mathbf{x}_c^t\mathbf{x}_c^t)^{-1}\mathbf{x}_c^t\mathbf{x}_c^t(\mathbf{x}_c^t\mathbf{x}_c^t)^{-1} = \sigma^2(\mathbf{x}_c^t\mathbf{x}_c^t)^{-1} = \mathrm{Var}(\hat{\beta}_{OLS})\end{aligned}$$

- The higher λ is the higher the bias is and lower $\mathrm{Var}(\beta^*_{ridge})$ is, and vice-versa.

- If $\mathbf{x}_c^t\mathbf{x}_c = I_d$:

$$\beta^*_{ridge} = (I_d + \lambda I_d)^{-1}\mathbf{x}_c^t\mathbf{y}_c = (\lambda + 1)^{-1}\mathbf{x}_c^t\mathbf{y}_c$$

  Since $\mathbf{x}_c^t\mathbf{x}_c = I_d$, $\hat{\beta}_{OLS} = (\mathbf{x}_c^t\mathbf{x}_c)^{-1}\mathbf{x}_c^t\mathbf{y}_c = \mathbf{x}_c^t\mathbf{y}_c$ and then :

$$\beta^*_{ridge} = \frac{\hat{\beta}_{OLS}}{\lambda + 1}$$

Elastic Net

The objective function to minimize is convex but non-differentiable. The Fermat rule gives that a minimum β_{ElNet} verifies :

$$0 \in \partial f(\beta_{ElNet})$$

With $\partial f(\cdot)$ the sub-gradient of f. Let's compute it :

$$\partial f(\beta) = \begin{cases} \{2\mathbf{x}_c^t(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda_2 2\beta + \lambda_1\} & \text{si } \beta > 0 \\ \{2\mathbf{x}_c^t(\mathbf{x}_c - \mathbf{x}_c\beta) + \lambda_2 2\beta - \lambda_1\} & \text{si } \beta < 0 \\ \left[2\mathbf{x}_c^t\mathbf{y}_c - \lambda_1, 2\mathbf{x}_c^t\mathbf{y}_c + \lambda_1\right] & \text{si } \beta = 0 \end{cases}$$

Then

$$\begin{aligned}0 \in \partial f(\beta_{ElNet}) &\iff 2\lambda_2\mathbf{x}_c^t(\mathbf{y}_c - \mathbf{x}_c\beta_{ElNet}) + \lambda_2 2\beta_{ElNet} \pm \lambda_1 = 0 \\ &\iff 2\beta_{ElNet}(\lambda_2 I_d + \mathbf{x}_c^t\mathbf{x}_c) = 2\mathbf{x}_c^t\mathbf{y}_c \pm \lambda_1 \\ &\iff \beta_{ElNet} = \frac{\hat{\beta}_{OLS} \pm \frac{\lambda_2}{2}}{\lambda_2 + 1}\end{aligned}$$