1 Kernel Machines

Consider a linear model

$$Y_i = x_i^T \beta^0 + \varepsilon_i,\ i = 1, \ldots, n,\ x_i \in \mathbb{R}^p \text{ fixed}$$

where $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$. We have

$$\begin{aligned} \hat{\beta}^{\text{ols}} &= \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - x_i^T \beta)^2 \\ &= \text{argmin}_{\beta \in \mathbb{R}^p} ||Y - X\beta||^2 \\ &= (X^T X)^{-1} X^T Y. \end{aligned}$$

Classical theory:

- $\hat{\beta}^{\text{ols}}$ unbiased,

  $$\mathrm{Var}(\hat{\beta}^{\text{ols}}) = \sigma^2 (X^T X)^{-1} = i^{-1}(\beta^0)$$

  Where i is the Fisher information.

- Cramér-Rao lower bound: if an estimator $\tilde{\beta}$ is unbiased then

  $$\mathrm{Var}(\tilde{\beta}) - i^{-1}(\beta^0) \quad \underbrace{\geq}_{\text{positive semi-definite}} \quad 0.$$

- If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $\hat{\beta}^{\text{ols}}$ is the MLE of β^0. Furthermore $\sqrt{n}(\hat{\beta}^{\text{ols}} - \beta^0) \sim \mathcal{N}(0, n\sigma^2 (X^T X)^{-1})$. From this we can derive confidence intervals, hypothesis test, etc.

In a general model with parameter $\theta \in \mathbb{R}^p$, n independent observations, under regularity, we have asymptotic normality, i.e $\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta^0))$ (with p fixed).

Question: what happens when p is large relative to n?

- If $p > n$, $\hat{\beta}^{\text{ols}}$ is not even defined.
- If $p \approx n$, $\mathrm{Var}(\hat{\beta}^{\text{ols}})$ explodes since $X^T X$ is near singular.
- More generally, if $p, n \to \infty$ then asymptotic normality can break down.

Recall the bias-variance decomposition:

$$\begin{aligned} \mathrm{mse}(\tilde{\beta}) &= \mathbb{E}_{\beta^0, \sigma^2} \left[(\tilde{\beta} - \beta^0)(\tilde{\beta} - \beta^0) \right] \\ &= \mathbb{E}_{\beta^0, \sigma^2} \left\| \tilde{\beta} - \mathbb{E}\tilde{\beta} + \mathbb{E}\tilde{\beta} - \beta^0 \right\| \\ &= \mathrm{Var}(\tilde{\beta}) + \left\| \mathbb{E}(\tilde{\beta}) - \beta^0 \right\|^2. \end{aligned}$$

We introduce bias to reduce the variance.

1.1 Ridge regression

Define

$$(\hat{\mu}^R_\lambda, \hat{\beta}^R_\lambda) = \text{argmin}_{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}\left[||Y-\mu\mathbf{1}-X\beta||^2 + \underbrace{\lambda||\beta||^2}_{\text{penalty for large }\beta}\right].$$

λ is called a *regularisation* or *tuning* parameter. We shall assume the columns of X have been standardised (mean 0, variance 1).

After standardisation, we can show that

$$\hat{\mu}^R_\lambda = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}.$$

Hence, if we replace Y with $Y - \mathbf{1}\bar{Y}$ we can write

$$\begin{aligned}\hat{\beta}^R_\lambda &= \text{argmin}_{\beta\in\mathbb{R}^p}\left[||Y-X\beta\|^2+\lambda||\beta||^2\right]\\ &= \underbrace{(X^TX+\lambda I_p)^{-1}}_{\text{always invertible}}X^TY.\end{aligned}$$

Theorem 1.1. *For $\lambda > 0$ sufficiently small,*

$$\mathbb{E}\|\hat{\beta}^{ols}-\beta^0\|^2 - \mathbb{E}\|\hat{\beta}^R_\lambda-\beta^0\|^2 > 0. \tag{$*$}$$

Proof. We have

$$Y = X\beta^0+\varepsilon.$$

The bias of $\hat{\beta}^R_\lambda$ is

$$\begin{aligned}\mathbb{E}(\hat{\beta}^R_\lambda-\beta^0) &= (X^TX+\lambda I)^{-1}X^TX\beta^0-\beta^0\\ &= (X^TX+\lambda I)^{-1}(X^TX+\lambda I-\lambda I)\beta^0-\beta^0\\ &= -\lambda(X^TX+\lambda I)^{-1}\beta^0.\end{aligned}$$

While we have variance

$$\begin{aligned}\text{Var}(\hat{\beta}^R_\lambda) &= \mathbb{E}\left\|(X^TX+\lambda I)^{-1}X^T\varepsilon\right\|^2\\ &= \sigma^2\left[(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}\right].\end{aligned}$$

Then $(*)$ becomes

$$\begin{aligned}&\mathbb{E}\|\hat{\beta}^{\text{ols}}-\beta^0\|^2-\mathbb{E}\|\hat{\beta}^R_\lambda-\beta^0\|^2\\ &=\sigma^2(X^TX)^{-1}-\sigma^2(X^TX+\lambda I)X^TX(X^TX+\lambda I)^{-1}\\ &\quad -\lambda^2(X^TX+\lambda I)^{-1}\beta^0(\beta^0)^T(X^TX+\lambda I)^{-1}\\ &= \ \vdots \qquad\qquad\qquad\qquad (\text{use SVD } X=UDU^T)\\ &=\lambda(X^TX+\lambda I)^{-1}\left[\sigma^2\left\{2I_p+\lambda(X^TX)^{-1}\right\}-\lambda\beta^0(\beta^0)^T\right](X^TX+\lambda I)^{-1}.\end{aligned}$$

We want to show this is positive definite. This is equivalent to

$$\begin{aligned} &\sigma^2\left[2I + \lambda(X^TX)^{-1}\right] - \lambda\beta^0(\beta^0)^T > 0 \\ \iff\ & 2\sigma^2 I - \lambda\beta^0(\beta^0)^T > 0 \\ \iff\ & 2\sigma^2||z||^2 - \lambda(z^T\beta^0)^2 > 0 \quad \forall z \in \mathbb{R}^p. \qquad (\dagger) \end{aligned}$$

We also have $(z^T\beta^0)^2 \leq ||z||^2||\beta^0||^2$ by Cauchy-Schwartz. Hence (†) holds for all $\lambda < \frac{2\sigma^2}{||\beta^0||^2}$. □