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

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

Classical theory:

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

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

  Where i is the Fisher information.

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

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

- If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $\hat{\beta}^{\mathrm{ols}}$ is the MLE of β^0. Furthermore $\sqrt{n}(\hat{\beta}^{\mathrm{ols}} - \beta^0) \sim \mathcal{N}(0, n\sigma^2 (X^T X)^{-1})$. 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 $\sqrt{n}(\hat{\theta}^{\mathrm{MLE}} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta^0))$ (with p fixed).

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

- If $p > n$, $\hat{\beta}^{\mathrm{ols}}$ is not even defined.

- If $p \approx n$, $\mathrm{Var}(\hat{\beta}^{\mathrm{ols}})$ 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

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

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

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

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

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

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

$$\text{mse}(\hat{\beta}^{ols}) - \text{mse}(\hat{\beta}_\lambda^R) = \mathbb{E}||\hat{\beta}^{ols} - \beta^0||^2 - \mathbb{E}||\hat{\beta}_\lambda^R - \beta^0||^2 > 0. \qquad (*)$$

Proof. We have
$$Y = X\beta^0 + \varepsilon.$$

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

$$\mathbb{E}(\hat{\beta}_\lambda^R - \beta^0) = (X^TX + \lambda I)^{-1}X^TX\beta^0 - \beta^0$$
$$= (X^TX + \lambda I)^{-1}(X^TX + \lambda I - \lambda I)\beta^0 - \beta^0$$
$$= -\lambda(X^TX + \lambda I)^{-1}\beta^0.$$

While we have variance

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

Then $(*)$ becomes

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

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

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

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

Singular value decomposition

Suppose $n \geq p$, so we can always write $X \in \mathbb{R}^{n \times p}$ as

$$X = UDV^T \qquad \text{(``thin SVD'')}$$

where $U \in \mathbb{R}^{n \times p}, V \in \mathbb{P}^{p \times p}$, with orthonormal columns, $D \in \mathbb{R}^{p \times p}$ diagonal with $D_{11} \geq D_{22} \geq \ldots \geq D_{pp} \geq 0$.

The fitted values in ridge regression are

$$\begin{aligned}
\hat{Y}_\lambda^R = X\hat{\beta}_\lambda^R &= X(X^TX + \lambda I)^{-1}X^TY \\
&= UDV^T(VD^2V^T + \lambda I)^{-1}VDU^TY \quad \text{(using } VV^T = V^TV = I\text{)} \\
&= UD(D^2 + \lambda I)^{-1}DU^TY \\
&= \sum_{j=1}^{p} U_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda}U_j^TY
\end{aligned}$$

where U_j denotes the jth column of U. For reference, in OLS regression

$$\hat{Y}^{ols} = X\hat{\beta}^{\text{ols}} = X(X^TX)^{-1}X^TY = \sum_{j=1}^{p} U_j U_j^T Y.$$

So ridge "projects" onto columns of U, but it shrinks jth component by a factor

$$\frac{D_{jj}^2}{D_{jj}^2 + \lambda}.$$

Hence it shrinks small singular values to 0 rapidly.

Note. The matrix $X(X^TX)^{-1}X^TY$ is known as the "hat matrix" and it represents an orthogonal projection onto the column space of X.

The SVD of X is related to principal component analysis.

Definition. The kth *principal component* $U^{(k)}$ of X and *principal direction* $v^{(k)}$ of X are defined recursively by

$$v^{(k)} = \operatorname{argmax}_{v \in \mathbb{R}^p} ||Xv||^2 \text{ subject to } ||v|| = 1, \ (v^{(j)})^T X^T X v = 0 \ \forall j < k$$

and
$$u^{(k)} = Xv^{(k)}.$$

Lemma 1.2. *If $D_{jj} > 0$ for all $j \in \{1, \ldots, p\}$ then $v^{(k)} = V_k$, $u^{(k)} = D_{kk}U_k$.*

Message: ridge is good when the signal (β^0) is large for the top principal components of X.

Computation: we can compute $\hat{Y}_\lambda^R$ for any value of λ quickly after doing an SVD, which has cost $\mathcal{O}(np^2)$.

1.2 v-fold cross-validation

We assume that (x_i, Y_i), $i = 1, \ldots, n$ is iid from some distribution (random design matrix). Let (x^*, Y^*) be another independent observation from this distribution. We may wish to pick λ minimising the mean-squared prediction error (MSPE) conditional on (X, Y):

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X,Y))^2 | (X,Y)\}.$$

A less ambitious goal is to minimise the MSPE

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X,Y))^2\} = \mathbb{E}\left[\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X,Y))^2 | (X,Y)\}\right]. \quad (\ddagger)$$

We can try to estimate this quantity for different values of λ, using data splitting.

- Let $(X^{(1)}, Y^{(1)}), \ldots, (X^{(v)}, Y^{(v)})$ be groups of data points of roughly equal size. These are called *folds*.

- Let $(X^{(-k)}, Y^{(-k)})$ denote all the folds except the kth.

- Let $\kappa(i)$ be the fold to which sample i (i.e (X_i, Y_i)) belongs.

Our estimator of $(\ddagger)$ is

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i - x_i^T \underbrace{\hat{\beta}_\lambda^R(X^{(-\kappa(i))}, Y^{(-\kappa(i))})}_{\text{using all folds except the ones containing } (x_i, Y_i)} \right\}^2.$$

Then define

$$\lambda_{\mathrm{CV}} = \mathrm{argmin}_{\lambda \in \{l_1, \ldots, l_m\}} \mathrm{CV}(\lambda).$$

We use the estimator

$$\hat{\beta}_{\lambda_{\mathrm{CV}}}^R(X, Y).$$

How to choose v?

Note.

- The expectation of each summand in $\mathrm{CV}(\lambda)$ is almost the same as $\ddagger$, which is what we want to estimate. The only difference is the size of the training set. Hence the bias of $\mathrm{CV}(\lambda)$ is small when v is large [the extreme of this is $v = n$, called "leave one out" cross-validation].

- When v is large, the estimator $\hat{\beta}_\lambda^R(X^{(-k)}, Y^{(-k)})$ is similar for different values of k, which leads to positively correlated summands in $\mathrm{CV}(\lambda)$, leading to high variance.

- A common choice is $v = 5$ or $v = 10$.