

# 1 Kernel Machines

Consider a linear model

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

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

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

Classical theory:

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

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

Where  $i$  is the Fisher information.

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

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

- If  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ , then  $\hat{\beta}^{\text{ols}}$  is the MLE of  $\beta^0$ . 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  $p \approx n$ ,  $\text{Var}(\hat{\beta}^{\text{ols}})$  explodes since  $X^T X$  is near singular.
- More generally, if  $p, n \rightarrow \infty$  then asymptotic normality can break down.

Recall the bias-variance decomposition:

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

We introduce bias to reduce the variance.

## 1.1 Ridge regression

Define

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

$\lambda$  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

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

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

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

*Proof.* We have

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

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

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

While we have variance

$$\begin{aligned} \operatorname{Var}(\hat{\beta}_\lambda^R) &= \mathbb{E} \|(X^T X + \lambda I)^{-1} X^T \varepsilon\|^2 \\ &= \sigma^2 [(X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}]. \end{aligned}$$

Then (\*) becomes

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

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

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

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

**Singular value decomposition**

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

$$X = UDV^T \quad (\text{“thin SVD”})$$

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

The fitted values in ridge regression are

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

where  $U_j$  denotes the  $j$ th column of  $U$ . For reference, in OLS regression

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

So ridge “projects” onto columns of  $U$ , but it shrinks  $j$ th component by a factor

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

Hence it shrinks small singular values to 0 rapidly.

**Note.** The matrix  $X(X^T X)^{-1} X^T Y$  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  $k$ th *principal component*  $U^{(k)}$  of  $X$  and *principal direction*  $v^{(k)}$  of  $X$  are defined recursively by

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

and

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

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

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

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

## 1.2 $v$ -fold cross-validation

We assume that  $(x_i, Y_i)$ ,  $i = 1, \dots, n$  is iid from some distribution (random design matrix). Let  $(x^*, Y^*)$  be another independent observation from this distribution. We may wish to pick  $\lambda$  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  $\lambda$ , using data splitting.

- Let  $(X^{(1)}, Y^{(1)}), \dots, (X^{(v)}, Y^{(v)})$  be groups of data points of roughly equal size. These are called *folds*.
- Let  $(X^{(-k)}, Y^{(-k)})$  denote all the folds except the  $k$ th.
- Let  $\kappa(i)$  be the fold to which sample  $i$  (i.e.  $(X_i, Y_i)$ ) belongs.

Our estimator of  $(\ddagger)$  is

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

Then define

$$\lambda_{\text{CV}} = \operatorname{argmin}_{\lambda \in \{\lambda_1, \dots, \lambda_m\}} \text{CV}(\lambda).$$

We use the estimator

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

How to choose  $v$ ?

**Note.**

- The expectation of each summand in  $\text{CV}(\lambda)$  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  $\text{CV}(\lambda)$  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  $\text{CV}(\lambda)$ , leading to high variance.
- A common choice is  $v = 5$  or  $v = 10$ .

### 1.3 Kernel trick

We have

$$\hat{Y}_\lambda^R = X(X^T X + \lambda I)^{-1} X^T Y.$$

Note that

$$\begin{aligned} X^T (X X^T + \lambda I) &= (X^T X + \lambda I) X^T \\ \implies (X^T X + \lambda I)^{-1} X^T &= X^T (X X^T + \lambda I)^{-1} \\ \implies X \underbrace{(X^T X + \lambda I)^{-1}}_{p \times p} X^T Y &= X X^T \underbrace{(X X^T + \lambda I)^{-1}}_{n \times n} Y. \end{aligned}$$

The computation cost of the LHS is  $\mathcal{O}(np^2 + p^3)$  while the RHS is  $\mathcal{O}(pn^2 + n^3)$ .

- When  $p \gg n$ , the 2nd expression is cheaper to compute;
- The fitted values in ridge regression only depend on  $X$  through the “Gram matrix”  $K = X X^T$ , with entries  $K_{ij} = \langle x_i, x_j \rangle$ .

Suppose we wish to fit a quadratic model:

$$Y_i = x_i^T \beta + \sum_{k,l} x_{ik} x_{il} \theta_{kl} + \varepsilon_i.$$

This can be done with a linear model where we replace the predictors  $x_i \in \mathbb{R}^p$  with a new “feature” vector:

$$\phi(x_i) = (x_{i1}, x_{i2}, \dots, x_{ip}, x_{i1}x_{i1}, x_{i1}x_{i2}, \dots, x_{ip}x_{ip}) \in \mathbb{R}^{p+p^2}.$$

We call  $\phi$  a “feature map”. Now we have  $\mathcal{O}(p^2)$  predictors. If  $p^2 \gg n$ , to compute ridge fitted values, we want to use the 2nd expression, with cost  $\mathcal{O}(p^2 n^2 + n^3)$ .

However, the part that scales as  $\mathcal{O}(p^2 n^2)$  is just the computation of the Gram matrix with entries  $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$ .

The *kernel trick* offers a shortcut for computing  $K$ .

**Idea:**

$$\begin{aligned} \left( \frac{1}{2} + x_i^T x_j \right)^2 - \frac{1}{4} &= \left( \frac{1}{2} + \sum_k x_{ik} x_{jk} \right)^2 - \frac{1}{4} \\ &= \sum_k x_{ik} x_{jk} + \sum_{k,l} x_{ik} x_{il} x_{jk} x_{jl} \\ &= \langle \phi(x_i), \phi(x_j) \rangle = K_{ij}. \end{aligned}$$

The LHS can be computed in  $\mathcal{O}(p)$  iterations, so we can obtain  $K$  in  $\mathcal{O}(n^2 p)$  iterations, and we can compute the fitted values in ridge regression in  $\mathcal{O}(n^2 p + n^3)$ , which is not worse than the linear model!

**Notes.**

- For many feature maps  $\phi$ , there are similar shortcuts.
- Instead of focusing on  $\phi$ , we can directly think of the function  $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$  as a measure of “similarity” between inputs  $x_i, x_j$ .

**Question:** for which similarities  $k$  is there a feature map  $\phi$  such that

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle?$$

**1.4 Kernels**

**Definition.** An *inner product space* is a real vector space  $\mathcal{H}$  endowed with a map  $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$  satisfying:

- (i) Symmetry: for all  $u, v \in \mathcal{H}$  we have  $\langle u, v \rangle = \langle v, u \rangle$ ;
- (ii) Bilinearity: for all  $a, b \in \mathbb{R}$  and all  $u, v, w \in \mathcal{H}$  we have

$$\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle.$$

- (iii) Positive-definiteness: we have  $\langle u, u \rangle \geq 0$  for all  $u \in \mathcal{H}$ , with equality if and only if  $u = 0$ .

Suppose that regression inputs  $x_1, \dots, x_n$  take values in an abstract set  $\mathcal{X}$  (so far we’ve had  $\mathcal{X} = \mathbb{R}^p$ , but the  $x_i$ ’s could be functions; images; graphs; etc.).

**Goal:** characterise similarity functions  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  for which there is an inner product space  $\mathcal{H}$  and a feature map  $\phi : \mathcal{X} \rightarrow \mathcal{H}$  such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}.$$

**Definition.** A (*positive-definite*) *kernel*  $k$  is a symmetric map  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  such that for all  $n \in \mathbb{N}$  and all  $x_1, \dots, x_n \in \mathcal{X}$ , the matrix  $K$  with entries  $K_{ij} = k(x_i, x_j)$  is positive semi-definite.

**Remark.** A kernel is not an inner product on  $\mathcal{X}$  in general. Indeed,  $\mathcal{X}$  does not even need to be a vector space, and  $k$  need not be bilinear. However, we do have a version of the Cauchy-Schwarz inequality for kernels.

**Proposition 1.3.** *Let  $k$  be a kernel on  $\mathcal{X}$ . Then*

$$k(x, x')^2 \leq k(x, x)k(x', x') \quad \forall x, x' \in \mathcal{X}.$$

*Proof.* Since  $k$  is a kernel,

$$\begin{pmatrix} k(x, x) & k(x, x') \\ k(x', x) & k(x', x') \end{pmatrix} \geq 0.$$

Hence this has non-negative determinant and  $k(x, x)k(x', x') - k(x, x')^2 \geq 0$ .  $\square$

**Proposition 1.4.** *Any similarity  $k$  defined by*

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}$$

*is a kernel.*

*Proof.* Symmetry of  $k$  is clear. Let  $x_1, \dots, x_n \in \mathcal{X}$  be arbitrary and take any vector  $\alpha \in \mathbb{R}^n$ . We need to show  $\alpha^T K \alpha \geq 0$ . Indeed

$$\begin{aligned} \alpha^T K \alpha &= \sum_{i,j} \alpha_i K_{ij} \alpha_j \\ &= \sum_{i,j} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle \alpha_j \\ &= \left\langle \sum_{i=1}^n \alpha_i \phi(x_i), \sum_{j=1}^n \alpha_j \phi(x_j) \right\rangle && \text{(linearity of } \langle \cdot, \cdot \rangle \text{)} \\ &\geq 0. && \text{(positive-definiteness of } \langle \cdot, \cdot \rangle \text{)} \end{aligned}$$

□