1 Kernel Machines

Consider a linear model

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

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

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

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,*

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

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

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

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

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

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

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

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

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

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

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

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

A less ambitious goal is to minimise the MSPE

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

We can try to estimate this quantity for different values of λ, 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 kth.
- Let $\kappa(i)$ be the fold to which sample i (i.e (X_i, Y_i)) belongs.

Our estimator of ($\ddagger$) is

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

Then define

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

We use the estimator

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

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 $\hat{\beta}^R_\lambda(X^{(-k)}, Y^{(-k)})$ 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 $v = 10$.

1.3 Kernel trick

We have

$$\hat{Y}^R_\lambda = X(X^TX + \lambda I)^{-1}X^TY.$$

Note that

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

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" $K = XX^T$, 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 ϕ 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 $\mathcal{O}(p^2n^2 + n^3)$.

However, the part that scales as $\mathcal{O}(p^2n^2)$ 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:

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

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

Notes.

- For many feature maps ϕ, there are similar shortcuts.
- Instead of focusing on ϕ, 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 ϕ 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 $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ 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 $x_1, \ldots, x_n$ 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there is an inner product space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathcal{X}$, 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 $k(x, x)k(x', x') - k(x, x')^2 \geq 0$. □

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

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

□