

1.2 Cross-validation

Suppose data $(Y_i, x_i)_{i=1}^n$ and new obs $(Y^*, x^*) \in \mathbb{R} \times \mathbb{R}^p$ are all i.i.d.
We'd like to pick λ to minimize

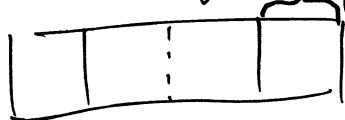
$$\mathbb{E} \{ (Y^* - x^{*T} \hat{\beta}_\lambda^R(x, Y))^2 \mid X, Y \}$$

An easier quantity to minimize is the expected prediction error

$$\mathbb{E} [\mathbb{E} \{ (Y^* - x^{*T} \hat{\beta}_\lambda^R(x', Y))^2 \mid X', Y \}]$$

\nwarrow dataset of size n

CV estimates this by splitting the data into v folds



$$(X^{(1)}, Y^{(1)}) \dots (X^{(v)}, Y^{(v)})$$

Let $(X^{(-k)}, Y^{(-k)})$ be all data except that in the k th fold. Write $K(i)$ for the fold to which the i th obs belongs. Define

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \{ Y_i - x_i^T \hat{\beta}_\lambda^R(X^{(-K(i))}, Y^{(-K(i))}) \}^2$$

often choose v as 5 or 10.

Writing λ_{cv} as the minimizer, estimate β^0 by $\hat{\beta}_{\lambda_{cv}}^R(X, Y)$.

Suppose we have computed $\hat{\beta}_\lambda^R$ on a grid of L values $\lambda_1 > \lambda_2 > \dots > \lambda_L$. Another approach is to minimize

$$\frac{1}{n} \sum_{i=1}^n \{ Y_i - \sum_{k=1}^L w_k x_i^T \hat{\beta}_{\lambda_k}^R(X^{(-K(i))}, Y^{(-K(i))}) \}^2$$

over $w \in \mathbb{R}^L$, subject to $w_k \geq 0$. This is known as stacking (Breiman, 1996) and often works better than CV.

1.3 The kernel trick

Alternative expression for fitted value of ridge

$$(X^T X + \lambda I) X^T = X^T (X X^T + \lambda I)$$

$$\overset{n \times n}{X X^T} (X X^T + \lambda I)^{-1} Y = X \overset{p \times p}{(X^T X + \lambda I)^{-1}} X^T Y = X \hat{\beta}_\lambda^R$$

Note

1. $X^T X$ is $p \times p$ $\mathcal{O}(np^2)$ to compute
 $X X^T$ is $n \times n$ $\mathcal{O}(n^2 p)$ to compute (if $p \gg n$, then more computationally efficient)
2. Key point: Fitted value only depends on inner products between obs

$$K = X X^T, \quad K_{ij} = x_i^T x_j$$

Suppose we want to fit a model with quadratic effects

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

Can still use ridge provided we expand our set of predictors

$$x_{i1}, \dots, x_{ip}, \quad x_{i1} x_{i1}, x_{i1} x_{i2}, \dots, x_{i1} x_{ip}$$

\vdots

$$x_{ip} x_{i1}, x_{ip} x_{i2}, \dots, x_{ip} x_{ip}$$

(interactions $x_{ik} x_{il}$ $k \neq l$ appear twice)

Given $\mathcal{O}(p^2)$ approach may be computationally infeasible for $p \gg n$ ($\mathcal{O}(p^2 n^2)$ to compute $X X^T$). new expanded X

Idea: compute K directly. Consider

$$\begin{aligned} (1 + x_i^T x_j)^2 &= \left(1 + \sum_k x_{ik} x_{jk}\right)^2 \\ &= 1 + 2 \sum_k x_{ik} x_{jk} + \sum_{k,l} x_{ik} x_{il} x_{jl} x_{jl} \end{aligned}$$

Equal to an inner product between vectors of the form

$$\begin{pmatrix} 1, \\ \sqrt{2} x_{i1}, \sqrt{2} x_{i2}, \dots, \sqrt{2} x_{ip}, \\ x_{i1} x_{i1}, x_{i1} x_{i2}, \dots, x_{i1} x_{ip}, \\ \vdots \\ x_{ip} x_{i1}, x_{ip} x_{i2}, \dots, x_{ip} x_{ip} \end{pmatrix}^T \quad (*)$$

If we set $K_{ij} = (1 + x_i^T x_j)^2$ and compute fitted value

$$K(K + \lambda I)^{-1} Y$$

it is exactly equivalent to fitted value from ridge fit to expanded set of predictors

given by (x) .

Computation of K is $\mathcal{O}(n^2 p)$.

Take home messages

1. Can apply this trick to any method that uses only inner products between obs.

2. Instead of fitting expanding our features using some feature map

$$\phi: x_i \rightarrow \phi(x_i) \text{ and fitting a model directly to } \begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_n) \end{pmatrix}$$

can directly compute $\langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$ to form

$$K_{ij} = k(x_i, x_j).$$