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

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

An easier quantity to minimize is the expected prediction error

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

((X',Y'): dataset of size n)

CV estimates this by splitting the data into v folds

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

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

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^n \left\{Y_i - x_i^T \hat{\beta}^R_\lambda(X^{(-\kappa(i))}, Y^{(-\kappa(i))})\right\}^2$$

Often choose v as 5 or 10.
Writing λ_{cv} as the minimizer, estimate β^* by $\hat{\beta}^R_{\lambda_{cv}}(X,Y)$.
Suppose we have computed $\hat{\beta}^R_\lambda$ on a grid of λ values $\lambda_1 > \lambda_2 > \ldots > \lambda_L$. Another approach is to minimize

$$\frac{1}{n}\sum_{i=1}^n \left\{Y_i - \sum_{\ell=1}^L w_\ell x_i^T \hat{\beta}^R_{\lambda_\ell}(X^{(-\kappa(i))}, Y^{(-\kappa(i))})\right\}^2$$

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

1.3 The kernel trick

Alternative expression for fitted vals of ridge

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

$$\underset{n\times n}{XX^T}(XX^T + \lambda I)^{-1}Y = X\underset{p\times p}{(X^TX + \lambda I)^{-1}}X^TY = X\hat{\beta}^R_\lambda$$

Note

1. $X^T X$ is $p \times p$ $\mathcal{O}(np^2)$ to compute

   XX^T is $n \times n$ $\mathcal{O}(n^2 p)$ to compute (if $p \gg n$, then more computationally efficient)

2. Key point: Fitted vals only depend on inner products between obs

   $K = XX^T$, $K_{ij} = x_i^T x_j$

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)

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

Idea: compute K directly. Consider

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

Equal to an inner product between vectors of the form

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

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

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

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

given by (*).
Computation of K is $O(n^2 p)$.

Take home messages

1. Can apply this trick to any method that uses only inner products between data.
2. Instead of ~~fitting~~ expanding our features using some feature map $\phi : x_i \to \phi(x_i)$ 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 $K_{ij} = k(x_i, x_j)$.