1.7 Large-scale kernel machines

When n is large forming kernel matrix $K \in \mathbb{R}^{n \times n}$ and inverting $K + \lambda I$ (in the case of KRR) may be too computationally intensive. Furthermore, the fitted regression function is a sum over n terms

$$\hat{f}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i k(\cdot, x_i)$$

so evaluating $\hat{f}(x)$ at a new obs x would typically require $O(n)$ computations.

One approach to tackle these difficulties is to develop a random map $\hat{\phi} : \mathcal{X} \to \mathbb{R}^b$ (b small) s.t. $\mathbb{E}\, \hat{\phi}(x)^T \hat{\phi}(x') = k(x, x') \ \forall x, x' \in \mathcal{X}$.

To increase the quality of the approximation, we can use

$$x \longmapsto \frac{1}{\sqrt{L}} \underbrace{(\hat{\phi}_1(x), \ldots, \hat{\phi}_L(x))}_{\text{iid and } \hat{\phi}_1(x) \overset{d}{=} \hat{\phi}(x) \ \forall x \in \mathcal{X}}$$

Let Φ be the matrix with ith row $\frac{1}{\sqrt{L}}(\hat{\phi}_1(x_i), \ldots, \hat{\phi}_L(x_i))$. When approximate KRR with kernel k, can instead perform a 'regular' ridge regression on Φ.

$$\hat{\theta} = (\underbrace{\Phi^T \Phi + \lambda I}_{Lb \times Lb \text{ matrix}})^{-1} \Phi^T Y \qquad O(n(Lb)^2 + (Lb)^3)$$

To predict at a new x, compute

$$\frac{1}{\sqrt{L}} (\hat{\phi}_1(x), \ldots, \hat{\phi}_L(x)) \hat{\theta} \qquad O(Lb)$$

Rahimi & Recht (2007) propose a method for shift-invariant kernels, i.e. kernels $k(x, x') = h(x - x') \ \forall x, x' \in \mathcal{X} = \mathbb{R}^p$, some $h : \mathbb{R}^p \to \mathbb{R}$

<u>Theorem 8</u> (Bochner's thm)

Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant iff $\exists\, c > 0$, distribution F on $\mathbb{R}^p$, s.t. if $W \sim F$ then

$$k(x, x') = c\, \mathbb{E}\, e^{i(x-x')^T W} = c\, \mathbb{E} \cos((x-x')^T W)$$

Idea: Find $\hat{\phi}$ (depending on W) s.t.

$$\underset{(b=1)}{\hat{\phi}(x)\, \hat{\phi}(x')} = \cos((x-x')^T W)$$

Let $u \sim U[-\pi, \pi]$ and take $x, y \in \mathbb{R}$. Then

$$\mathbb{E} \cos(x+u)\cos(y+u) = \mathbb{E}(\cos x \cos u - \sin x \sin u)(\cos y \cos u - \sin y \sin u)$$

Now $u \overset{d}{=} -u$, so $\mathbb{E} \cos u \sin u = \mathbb{E} \cos(-u)\sin(-u) = -\mathbb{E}\cos(u)\sin(u) = 0$

$\cos^2 u + \sin^2 u = 1$, $\mathbb{E} \cos^2 u = \mathbb{E} \sin^2 u = \frac{1}{2}$

$$2\,\mathbb{E} \cos(x+u)\cos(y+u) = \cos x \cos y + \sin x \sin y = \cos(x-y)$$

Given a shift invariant k and associated distribution F, let $W \sim F$ and $u \sim U[-\pi, \pi]$ independently. Set $\hat{\phi}(x) = \sqrt{2c}\,\cos(W^T x + u)$

Then $\mathbb{E}\,\hat{\phi}(x)\hat{\phi}(x') = 2c\,\mathbb{E}\left[\mathbb{E}\{\cos(W^T x + u)\cos(W^T x' + u) \mid W\}\right]$

$$= 2c\,\mathbb{E}\cos(W^T(x-x')) = k(x, x')$$

For example, take $k(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$. Note if $W \sim N_p(0, \sigma^{-2} I)$, then $\mathbb{E}\, e^{it^T W} = e^{-\|t\|_2^2/(2\sigma^2)}$, so we can take $\hat{\phi}(x) = \sqrt{2}\cos(W^T x + u)$

Chapter 2 The Lasso and beyond

2.1 Model selection

Consider the linear model $Y = X\beta^0 + \varepsilon$, $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$.

Let $S = \{k : \beta^0_k \neq 0\}$ and $s = |S|$. In high-dim settings there is often reason to believe $s \ll p$.

First consider the low-dim setting where X has full col rank. The MSPE OLS:

$$\frac{1}{n}\mathbb{E}\|X\beta^0 - X\hat{\beta}^{OLS}\|_2^2 = \frac{1}{n}\mathbb{E}(\hat{\beta}^{OLS} - \beta^0)^T X^T X(\hat{\beta}^{OLS} - \beta^0)$$

$$= \frac{1}{n}\mathbb{E}\,\mathrm{tr}\{(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T X^T X\}$$

$$= \frac{1}{n}\mathrm{tr}(\sigma^2 (X^T X)^{-1} X^T X)$$

$$= \frac{\sigma^2 p}{n}$$

If we could find S and perform OLS just on X_S (submatrix of X formed from those cols of X indexed by S), we could achieve an MSPE of ~~σ^2/n~~ $\frac{\sigma^2 s}{n}$.

$\boxed{\text{EoSL } \begin{matrix} 57\text{-}61 \\ 61\text{-}73 \end{matrix}}$ print

Best subsets Consider regressing Y on X_M for every $M \subseteq \{1, \dots, p\}$ ($M \neq \emptyset$). Can pick the best model via CV. Infeasible for $p > 50$.

Forward selection 1. Start by fitting the intercept only model. 2. Add to the current model, the predictor that reduces the RSS by largest amount. 3. repeat 2 until model size exceeds some M.