STATS 305B: Methods for Applied Statistics I **Winter 2016/17**

Lecture 23: March 6

Lecturer: Jonathan Taylor *Scribes: Kenneth Tay*

23.1 Reproducible Kernel Smoothing ("Kernel Trick")

Say we have n observations (X_i, Y_i), with $X_i \in T$, where T is some space. We can set up the regression problem in the following way:

$$\underset{f \in \mathcal{F}}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^{n} \left[Y_i - f(X_i)\right]^2,$$

where $\mathcal{F}$ is some collection of "flexible" functions f.

One way to define $\mathcal{F}$ is through a "reproducing kernel".

Definition 23.1 *A* ***covariance function/positive semidefinite function*** *is a symmetric function $R : T \times T \to \mathbb{R}$ such that for all $(t_1, \ldots, t_k) \in T^k$ and $(a_1, \ldots, a_k) \in \mathbb{R}^k$,*

$$\sum_{i,j=1}^{n} a_i a_j R(t_i, t_j) \geq 0.$$

Example (Gaussian kernel): $T = \mathbb{R}$, $R(t, s) = \exp\left[-\frac{(t-s)^2}{2}\right]$.

Proposition 23.2 *Given a covariance function R, there exists a stochastic process $Z : \Omega \times T \to \mathbb{R}$ such that*

$$Cov(Z_t, Z_s) = R(t, s),$$

and

$$Var\left(\sum_{i=1}^{k} a_i Z_{t_i}\right) = \sum_{i,j=1}^{k} a_i a_j R(t_i, t_j).$$

(In fact, we may take Z to be Gaussian.)

Given a covariance function R, for each $t \in T$ we can define the function $R_t : T \to \mathbb{R}$ by $R_t(s) = R(t, s)$. In this setting, t is called a **knot**. We may also form linear combinations of these R_t's, giving rise to the reproducing kernel Hilbert space:

Definition 23.3 *Given a covariance function R, the* ***reproducing kernel Hilbert space*** *is*

$$\mathcal{H}_R = \left\{ h : h = \sum_i a_i R_{t_i} \text{ i.e. } h(s) = \sum_i a_i R(t_i, s),\ \|h\|_{\mathcal{H}}^2 < \infty \right\},$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm associated with the inner product

$$\left\langle \sum_i a_i R_{s_i}, \sum_j b_j R_{t_j} \right\rangle_{\mathcal{H}} = \sum_{i,j} a_i b_j R(s_i, t_j).$$

Note:

1. The inner product "reproduces" the kernel, in that $\langle R_t, R_s\rangle_{\mathcal{H}} = R(t,s)$.
2. (Evaluation property) For any $h \in \mathcal{H}$, $\langle h, R_t\rangle_{\mathcal{H}} = h(t)$.

With this set-up, we can reformulate our original regression problem as

$$\underset{f\in\mathcal{H}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n}[Y_i - f(X_i)]^2 + \lambda\|f\|_{\mathcal{H}}^2.$$

For a given T, there are many covariance functions, from smooth to rough. By choosing the covariance function appropriately, we could end up with a collection of functions with the desired level of smoothness.

Lemma 23.4 *Consider the problem*

$$\underset{f\in\mathcal{H}}{\textit{minimize}} \quad \sum_{i=1}^{n} L\left(Y_i, f(X_i)\right) + \lambda\|f\|_{\mathcal{H}}^2,$$

where L is some loss function. If there is a minimizer, then all minimizers are in the linear span $\mathcal{L} = span(R_{X_1}, \ldots, R_{X_n})$.

Proof: Let $\hat{f}$ be a minimizer. Then there exist weights $\hat{\omega} \in \mathbb{R}^n$ such that for $1 \le i \le n$,

$$\begin{aligned}\hat{f}(X_i) &= \sum_{j=1}^{n}\hat{\omega}_j R(X_i, X_j)\\ &= \left(\sum_{j=1}^{n}\hat{\omega}_j R_{X_j}\right)(X_i).\end{aligned}$$

If we let $\hat{g} = \sum_{j=1}^{n}\hat{\omega}_j R_{X_j}$, then $\hat{g} \in \mathcal{L}$ and $\hat{g}(X_i) = \hat{f}(X_i)$ for $1 \le i \le n$.

Let $\hat{\delta} = \hat{f} - \hat{g}$. Then $\hat{\delta}(X_i) = 0$ for all i. Note that

$$\begin{aligned}\langle \hat{g}, \hat{\delta}\rangle_{\mathcal{H}} &= \left\langle \sum_{j=1}^{n}\hat{\omega}_j R_{X_j}, \hat{\delta}\right\rangle_{\mathcal{H}}\\ &= \sum_{j=1}^{n}\hat{\omega}_j\hat{\delta}(X_j) && \text{(evaluation property)}\\ &= 0,\end{aligned}$$

and so

$$\begin{aligned}\|\hat{f}\|_{\mathcal{H}}^2 &= \|\hat{g} + \hat{\delta}\|_{\mathcal{H}}^2\\ &= \|\hat{g}\|_{\mathcal{H}}^2 + \|\hat{\delta}\|_{\mathcal{H}}^2\\ &\ge \|\hat{g}\|_{\mathcal{H}}^2.\end{aligned}$$

Since $\hat{f} = \hat{g}$ on the X_i's, it follows that $\hat{f}$ can only be a minimizer of the original objective function if $\|\hat{f}\|_{\mathcal{H}} \leq \|\hat{g}\|_{\mathcal{H}}$. Thus, we must have $\|\hat{\delta}\|_{\mathcal{H}} = 0$, i.e. $\hat{f} = \hat{g}$, which means that $f \in \mathcal{L}$. ■

This lemma allows us to reduce the regression problem to a finite dimensional problem!

Definition 23.5 *For covariance function R, define the* ***Gram matrix*** *G to be such that $G_{ij} = R(X_i, X_j)$.*

The regression problem can now be written as

$$\underset{\omega \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|Y - G\omega\|^2 + \frac{\lambda}{2}\omega^T G\omega.$$

This is like a generalized ridge regression problem. We will define

$$\hat{\omega}_\lambda := \underset{\omega \in \mathbb{R}^n}{\text{argmin}} \quad \frac{1}{2}\|Y - G\omega\|^2 + \frac{\lambda}{2}\omega^T G\omega,$$
$$\hat{f}_\lambda := \sum_j \hat{\omega}_{\lambda,j} R_{X_j}.$$

23.1.1 Reproducing Kernels & Gaussian Priors

First we define the **linear kernel**. For $t \in \mathbb{R}^p$, let the covariance of Z_t be $\gamma^T t$, where $\gamma \sim \mathcal{N}(0, \Sigma)$. The corresponding covariance function is

$$R_t(s) = R(t, s) = \text{Cov}(Z_t, Z_s) = s^T \Sigma t,$$

i.e. R_t maps $s \mapsto s^T(\Sigma t)$. As such, we have

$$\begin{aligned} \mathcal{H} &= \left\{ h = \sum_i a_i R_{t_i}, \|h\|_{\mathcal{H}}^2 < \infty \right\} \\ &= \left\{ h_a : h_a(x) = a^T x, \text{where } a \text{ is in the row space of } \Sigma \right\}. \end{aligned}$$

Here,

$$\langle h_a, h_b \rangle_{\mathcal{H}} = a^T \Sigma^{-1} b.$$