Lecture 6 — February 9

Lecturer: Martin Wainwright *Scribe: Anna Rafferty*

Note: These lecture notes are still rough, and have only have been mildly proofread.

6.1 Recap

In the last lecture, we saw that there is a link between reproducing kernel Hilbert spaces (RKHS) and kernels: for any RKHS, there exists a positive semi-definite kernel function, and conversely, given any positive semi-definite kernel function, we can construct an RKHS such that $R_x(\cdot) = K(\cdot, x)$ is the representer of evaluation. We also developed the representer theorem, which introduces the idea that solutions for minimizing the arguments of suitably regularized loss functions over $\mathcal{H}$ take the form $f(\cdot) = \sum_i^n \alpha_i \mathbb{K}(\cdot, x^{(i)})$. More formally:

Theorem 6.1. *Representer theorem: Let* Ω: $[0, +\infty) \to \mathbb{R}$ *be strictly increasing and let* ℓ: $(X \times Y \times \mathbb{R})^n \to \mathbb{R} \cup \{+\infty\}$ *by a loss function. Consider:*

$$min_{f \in \mathcal{H}} \ell(x^{(i)}, y^{(i)}, f(x^{(i)})) + \lambda_n \Omega(\|f\|^2_{\mathcal{H}}) \tag{6.1}$$

where $\mathcal{H}$ *is an RKHS with kernel* $\mathbb{K}$*. Then any optimal solution has the following form:*

$$f(\cdot) = \sum_i^n \alpha_i \mathbb{K}(\cdot, x^{(i)}) \tag{6.2}$$

In the above form, the α_i are the data dependent weights and $\mathbb{K}(\cdot, x^i)$ is the kernel function centered at x^i.

6.2 Kernel Ridge Regression

In the previous lecture, we developed a linear form of kernel ridge regression. We now seek to generalize this formulation.
We assume a model of form $y = f(x) + w$, where we are trying to estimate f (the regression function). In this formula, w is some additive noise, y is the response variable which we assume is an element of $\mathbb{R}$, and the x are the $\mathbb{R}^d$ covariates or predictors. Pairs $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times \mathbb{R}$ are observed, and we seek f^* such that $y^i = f^*(x^i) + w^{(i)}$ for $i = 1, \ldots, n$. Given an RKHS $\mathcal{H}$ with kernel $\mathbb{K}$, we can estimate f^* by solving an optimization problem over the RKHS:

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^n (y^{(i)} - f(x^{(i)}))^2 + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}} \tag{6.3}$$

The first term in this minimization problem is the data term, and the second is a regularization term, penalizing functions f whose RKHS norms are too large. (The parameter $\lambda > 0$ is a the regularization constant; for now, think about it as fixed, but later we will discuss methods to choose it.)

By Theorem 6.1, we conclude that any solution $\hat{f}$ takes the form

$$\hat{f}(\cdot) = \sum_{j=1}^{n} \alpha_j \mathbb{K}(\cdot, x^{(j)}) \tag{6.4}$$

We define $y = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{pmatrix} \in \mathbb{R}^n$ and $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $K_{ij} = \mathbb{K}(x^{(i)}, x^{(j)})$. Then, substituting the representation (6.4) into our original problem and simplifying, we obtain the following equivalent problem:

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{2}\|y - \mathbf{K}\alpha\|_2^2 + \frac{\lambda}{2}\alpha^T \mathbf{K}\alpha \tag{6.5}$$

The final term can be shown to be the equivalent of $\frac{\lambda}{2}\|\sum_{j=1}^{n} \alpha_j \mathbb{K}(\cdot, x^{(j)})\|_{\mathcal{H}}^2$:

$$\begin{aligned} \frac{\lambda}{2}\|\sum_{j=1}^{n} \alpha_j \mathbb{K}(\cdot, x^{(j)})\|_{\mathcal{H}}^2 &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle \mathbb{K}(\cdot, x^{(i)}), \mathbb{K}(\cdot, x^{(j)})\rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \mathbb{K}(x^{(i)}, x^{(j)}) = \sum_{i,j=1}^{n} \alpha_i \alpha_j K_{ij} \\ &= \alpha^T \mathbf{K}\alpha \end{aligned} \tag{6.6}$$

We now take the gradient of 6.5, using the fact that $\mathbf{K}$ is symmetric:

$$\nabla C(\alpha) = -\mathbf{K}y + \mathbf{K}^2 y\alpha + \lambda \mathbf{K}\alpha \tag{6.7}$$

Any solution should satisfy $\nabla C(\alpha) = 0$. Reorganizing, we get that $\mathbf{K}(\mathbf{K} + \lambda \mathbf{I})\alpha = \mathbf{K}y$. One solution to this is $\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} y$. This turns out to be the only solution we care about due to the form of $\hat{f}(\cdot)$; any other solution won't affect the final form of $\hat{f}$. Hence our estimate is:

$$\hat{f}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i \mathbb{K}(\cdot, x^{(i)}) \tag{6.8}$$

This is the generic method of kernel ridge regression. We now turn to several examples to gain some intuition.

6.3 Examples of Kernel Ridge Regression

6.3.1 Linear Kernels

For a linear kernel, we have $\mathbb{K}(x, z) = \sum_{j=1}^{d} x_j z_j$. Thus, $\mathbf{K} = \mathbf{X}\mathbf{X}^T$, where

$$\mathbf{X} = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \\ (x^{(n)})^T \end{bmatrix}$$

is a matrix in $\mathbb{R}^{n \times d}$.

We compute the solution

$$\hat{\alpha} = (\mathbf{X}\mathbf{X} + \lambda \mathbf{I})^{-1} y,\ \hat{y} = \hat{f}(z) = \hat{\alpha}^T \mathbf{X} z,\ z \in \mathbb{R}^d (\text{new sample}) \tag{6.9}$$

We can also think of this as $\hat{f}(z) = \hat{\theta}^T z$, making $\hat{\theta} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I})^{-1} y$. We can compute this inner product using only the kernel Gram matrix $\mathbf{K}$.

6.3.2 Polynomial Kernels

We consider polynomial kernels of degree $m \geq 2$. Then:

$$\mathbb{K}(x, z) = (1 + \sum_{j=1}^{d} x_j z_j)^m$$

$$\mathcal{H} = \text{span}\{\mathbb{K}(\cdot, z)\} = \text{span}\{\prod_{j=1}^{d} (\cdot)_j^{\alpha_j} | 0 \leq \alpha_j \leq m,\ \sum_{j=1}^{d} \alpha_j \leq m\} = \sum_{k=0}^{m} \binom{m}{k} (\sum_{j=1}^{d} x_j z_j)^k$$

Let $m = 2$, $d = 2$. Then:

$$(1 + x_1 z_1 + x_2 z_2)^2 = 1 + 2z_1 x_1 + 2z_2 x_2 + z_1^2 x_1^2 + z_2 x^2 + 2z_1 z_2 x_1 x_2 \tag{6.10}$$

This polynomial scales with the size of d. By performing ridge regression in the kernel space, we pay this penalty only in terms of computing the kernels, not exponentially. Our algorithm for polynomial ridge regression is as follows:
1. Compute $K_{ij} = (1 + \langle x^{(i)}, x^{(j)} \rangle_{\mathbb{R}^d})^m, \forall i, j = 1, \ldots, n$.
2. Then $\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} y$.
3. Finally, form $f(\cdot) = \sum_{i=1}^{n} \alpha_i (1 + \langle \cdot, x^{(i)} \rangle_{\mathbb{R}^\alpha})^m$.

6.3.3 Sobolev Spaces

Recall our definition of a first-order Sobolev space:

$$\mathcal{H} = \left\{ f : [0, 1] \to \mathbb{R} \;\middle|\; \begin{array}{l} f(0) = 0 \\ f \text{ is differentiable almost everywhere} \\ \|f\|_{\mathcal{H}}^2 = \int_0^1 (f'(t))^2 dt < +\infty \end{array} \right. \tag{6.11}$$

This is an RKHS with $\langle f, g\rangle_{\mathcal{H}} = \int_0^1 f'(t)g'(t)dt$ and kernel $\mathbb{K}(x,z) = \min(x,z)$. Then, $\mathcal{H} = \text{span}\{min(\cdot, z)\}$.
Let us have samples $(x^{(i)}, y^{(i)}) \in \mathbb{R} \times \mathbb{R}$. Then:

$$K_{ij} = \min(x^{(i)}, x^{(j)}),\ \hat{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}y,\ \hat{f}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i \min(\cdot, x^{(i)}) \tag{6.12}$$

This fits something piecewise linear, approximately like drawing line segments between each successive data point; the degree to which the resulting function approximates line segments between each data point depends on the regularization parameter λ. Our original problem was $\min_{f\in\mathcal{H}} \left\{\frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - f(x^{(i)}))^2 + \lambda\|f\|_{\mathcal{H}}^2\right\}$. Thus, as $\lambda \to \infty$, $\hat{f}$ goes to the all zeros function. (Exercise: play with λ in Matlab to get a sense of how the function changes with different regularization parameters.)

6.3.4 Practical Kernel Issues

In our examples, we've left out several practical issues. First, how do we choose λ? Often, we may meed to choose a sequence of λ_n that depend on the sample size. We also have not considered model selection: how do we decide what kernel to choose? This might be choosing a kernel family, or choosing a specific kernel from a family (e.g., choosing the order of a polynomial kernel).

6.4 Mercer's Theorem

We now turn to developing Mercer's theorem and the concept of eigenfunctions; we'll return to this topic more fully next class.

Theorem 6.2. *Let $X \subseteq \mathbb{R}^d$ that is bounded and closed. Given a kernel $\mathbb{K} : X \times X \to \mathbb{R}$, assume that $\int_X \int_X \mathbb{K}^2(x,y)dxdy < +\infty$. In this case, we can define a mapping $T_{\mathbb{K}} : L^2(X) \to L^2(X)$, where $L^2(X) = \{f : X \to \mathbb{R} | \int_X f^2(x)dx < +\infty$. Then $T_{\mathbb{K}}(f) = \int_y \mathbb{K}(x,y)f(y)dy$.*

$T_{\mathbb{K}}(f)$ is called a linear operator. We call ϕ an *eigenfunction* of $T_{\mathbb{K}}$ with eigenvalue λ if $\int_X \mathbb{K}(x,y)\phi(y)dx = \lambda\phi(x)$; that is, $T_{\mathbb{K}}(\phi) = \lambda\phi$.