Lecture 11

We recall and prove The Representer Theorem (Theorem 4) from last lecture:

Theorem 4 (The Representer Theorem). *Let $c : \mathcal{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a cost function, $\Omega : [0, \infty) \to \mathbb{R}$ a strictly monotonic increasing function, and $\mathcal{H}$ a RKHS. Then each minimizer of*

$$f^\star = \arg\inf_{f \in \mathcal{H}} \ \frac{1}{n} \sum_{i=1}^{n} c(x_i, y_i, f(x_i)) + \Omega(\|f\|_{\mathcal{H}})$$

has a representation of the form:

$$f^\star(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x). \tag{38}$$

Proof. Set $\|f\| = \|f\|_{\mathcal{H}}$ and define $\overline{\Omega}(\|f\|^2) = \Omega(\|f\|)$. Let U be the closed subspace of $\mathcal{H}$ defined as:

$$U = \left\{ g \in \mathcal{H} : g(x) = \sum_{i=1}^{n} \beta_i k(x_i, x) \right\}.$$

Since $\mathcal{H}$ is a Hilbert space, we can decompose $f \in \mathcal{H}$ as

$$f(x) = f_U(x) + f_\perp(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) + f_\perp(x),$$

where $f_U \in U$ and $\langle f_\perp, g \rangle = 0$ for all $g \in U$. Notice in particular that the last point implies:

$$\forall\, i = 1, \ldots, n, \quad \langle f_\perp, k(x_i, \cdot) \rangle = 0.$$

We now make two observations. The first is that:

$$\begin{aligned} \forall\, j = 1, \ldots, n, \quad f(x_j) &= \langle f, k(x_j, \cdot) \rangle, \\ &= \langle f_U + f_\perp, k(x_j, \cdot) \rangle, \\ &= \sum_{i=1}^{n} \alpha_i k(x_i, x_j) + \langle f_\perp, k(x_j, \cdot) \rangle, \\ &= \sum_{i=1}^{n} \alpha_i k(x_i, x_j). \end{aligned} \tag{39}$$

Thus $R_{\text{emp}}[f] = R_{\text{emp}}[f_U]$ depends only on f_U, and hence the values (39).

Secondly,

$$\begin{aligned}\Omega(\|f\|) &= \overline{\Omega}(\|f\|^2),\\ &= \overline{\Omega}(\|f_U\|^2 + \|f_\perp\|^2),\\ &\geq \overline{\Omega}(\|f_U\|^2),\\ &= \Omega\left(\left\|\sum_{i=1}^n \alpha_i k(x_i,\cdot)\right\|\right).\end{aligned}$$

Since Ω is strictly monotonically increasing, it is minimized only when $f_\perp = 0$. Since $R_{\text{emp}}[f] = R_{\text{emp}}[f_U]$, this implies that $f^\star \in U$, and hence can be represented as (38). □

Exercises

Exercise 21. Prove the Semiparametric Representer Theorem:

Theorem 5. *Let $c : \mathcal{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a cost function, $\Omega : [0,\infty) \to \mathbb{R}$ a strictly monotonic increasing function, and $\mathcal{H}$ a RKHS. Additionally, let $\{\psi_j : \mathcal{X} \to \mathbb{R}\}_{j=1}^m$ be a collection of m real valued functions with the property that the $n \times m$ matrix $(\psi_j(x_i))_{ij}$ has rank m. Finally, let $\mathcal{F}$ be the functional class:*

$$\mathcal{F} = \{\widetilde{f} = f + h : f \in \mathcal{H},\ \ h \in \text{span}\{\psi_j\}_{j=1}^m\}.$$

Then each minimizer:

$$\widetilde{f}^\star = \underset{\widetilde{f}=f+h\in\mathcal{F}}{\arg\inf}\ R_{\text{emp}}[\widetilde{f}] + \Omega(\|f\|_{\mathcal{H}}),$$

admits a representation of the form:

$$\widetilde{f}^\star(x) = \sum_{i=1}^n \alpha_i k(x_i, x) + \sum_{j=1}^m \beta_j \psi_j(x), \quad \alpha_i,\ \beta_j \in \mathbb{R}.$$

To understand better the relevance of the Semiparametric Representer Theorem, see [1, Chapter 4.8].

7.3 Kernel ridge regression

Let us return to the example at the end of Lecture 9. Now that we have introduced regularization, we can give the full story, which leads to the popular machine learning method, kernel ridge regression.

Recall that we are in the regression setting with a quadratic loss function $c(x, y, f(x)) = (y - f(x))^2$, and we have a feature map:

$$\Phi(x) = (\phi_1(x), \dots, \phi_m(x)) \in \mathbb{R}^m.$$

We considered the functional class $\mathcal{F} = \text{span}(\Phi)$, give by

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R} : f(x) = \sum_{j=1}^{m} w_j \phi_j(x), \;\; w_j \in \mathbb{R} \right\}.$$

Let us now try to minimize the regularized risk over training data $\{(x_i, y_i)\}_{i \le n}$, with regularization term $\|\mathbf{w}\|^2$:

$$\inf_{f \in \mathcal{F}} R_{\text{reg}}[f] = \inf_{\mathbf{w} \in \mathbb{R}^m} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \phi_j(x) \right)^2 + \lambda \|\mathbf{w}\|^2.$$

Similarly to before, we set:

$$F(\mathbf{w}) = F(w_1, \dots, w_m) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \phi_j(x) \right)^2 + \lambda \|\mathbf{w}\|^2.$$

Using the same notation

$$\Phi_{ji} = \phi_j(x_i) = \begin{pmatrix} \phi_1(x_1) & \phi_1(x_2) & \cdots & \phi_1(x_n) \\ \phi_2(x_1) & \phi_2(x_2) & \cdots & \phi_2(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_m(x_1) & \phi_m(x_2) & \cdots & \phi_m(x_n) \end{pmatrix}$$

Additionally, set:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} \in \mathbb{R}^m,$$

we obtain the following partial derivatives for F:

$$\frac{\partial F}{\partial w_\ell}(\mathbf{w}) = -2\Phi\mathbf{y}[\ell] + 2\Phi\Phi^T w[\ell] + 2\lambda\mathbf{w}[\ell].$$

Thus setting $\partial F/\partial w_\ell(\mathbf{w}) = 0$ for all $\ell = 1, \ldots, m$, we get:

$$\left(\Phi\Phi^T + \lambda I\right)\mathbf{w} = \Phi\mathbf{y} \Rightarrow \mathbf{w} = \left(\Phi\Phi^T + \lambda I\right)^{-1}\Phi\mathbf{y}.$$

We remark that if $\lambda > 0$, the matrix $\Phi\Phi^T + \lambda I$ is always invertible.

Let us now develop a kernelized version, similar to Exercise 19. To do so, note that we could have written $\mathbf{w}$ as:

$$\mathbf{w} = \lambda^{-1}\Phi\left(\mathbf{y} - \Phi^T\mathbf{w}\right).$$

Set:

$$\alpha = \lambda^{-1}\left(\mathbf{y} - \Phi^T\mathbf{w}\right). \tag{40}$$

Thus:

$$\mathbf{w} = \Phi\alpha \Rightarrow w_j = \sum_{i=1}^{n}\alpha_i\phi_j(x_i).$$

Note we saw this equation for w_j earlier in (37). But now with this specific regularization term, we can solve for α directly in terms of Φ and $\mathbf{y}$; indeed, using (40) we obtain:

$$\lambda\alpha = \mathbf{y} - \Phi^T\mathbf{w} = \mathbf{y} - \Phi^T\Phi\alpha,$$

which implies:

$$\left(\Phi^T\Phi + \lambda I\right)\alpha = \mathbf{y} \Rightarrow \alpha = \left(\Phi^T\Phi + \lambda I\right)^{-1}\mathbf{y}.$$

Similar to our earlier calculation we rewrite $f(x)$ as:

$$\begin{aligned} f(x) &= \sum_{j=1}^{m} w_j \phi_j(x), \\ &= \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \alpha_i \phi_j(x_i) \right) \phi_j(x), \\ &= \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \phi_j(x_i) \phi_j(x), \\ &= \sum_{i=1}^{n} \alpha_i \langle \Phi(x_i), \Phi(x) \rangle, \\ &= \sum_{i=1}^{n} \alpha_i k(x_i, x), \end{aligned}$$

where we have defined the kernel k as $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. Notice that Gram matrix of k, which is defined as $K_{ij} = k(x_i, x_j)$, is given by:

$$K = \Phi^T \Phi.$$

Thus we can write α as:

$$\alpha = (K + \lambda I)^{-1} \mathbf{y}.$$

To summarize, let $\mathcal{H}$ denote the RKHS of a kernel $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$, which is either defined by a feature map $\Phi : \mathcal{X} \to \ell^2$ or implicitly defines such a Φ through Mercer's Theorem. Then the optimizations

$$\inf_{f \in \overline{\text{span}(\Phi)}} R_{\text{reg}}[f] = \inf_{\mathbf{w} \in \ell^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{\infty} w_j \phi_j(x) \right)^2 + \lambda \|\mathbf{w}\|^2, \tag{41}$$

and

$$\inf_{f \in \mathcal{H}} R_{\text{reg}}[f] = \inf_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \tag{42}$$

are equivalent. Indeed, by the Representer Theorem, the minimizer of (42) has the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i),$$

which implies that (42) can be rewritten as:

$$\inf_{f\in\mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 = \inf_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \alpha_j k(x_i, x_j) \right)^2 + \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j). \tag{43}$$

The above calculations show that the two optimizations (41) and (43) result in the same minimizing function $f^\star$. Furthermore, they have closed form solutions for the weights given by:

$$\mathbf{w} = \left(\Phi\Phi^T + \lambda I\right)^{-1} \Phi\mathbf{y} \quad \text{and} \quad \alpha = (K + \lambda I)^{-1}\,\mathbf{y}.$$

The optimization (43) is called *kernel ridge regression*, which is popular machine learning algorithm.

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[5] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[6] J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[7] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.

[9] Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.