CMSE 820 HW 8

Hao Lin

November 10, 2018

1 Question 1

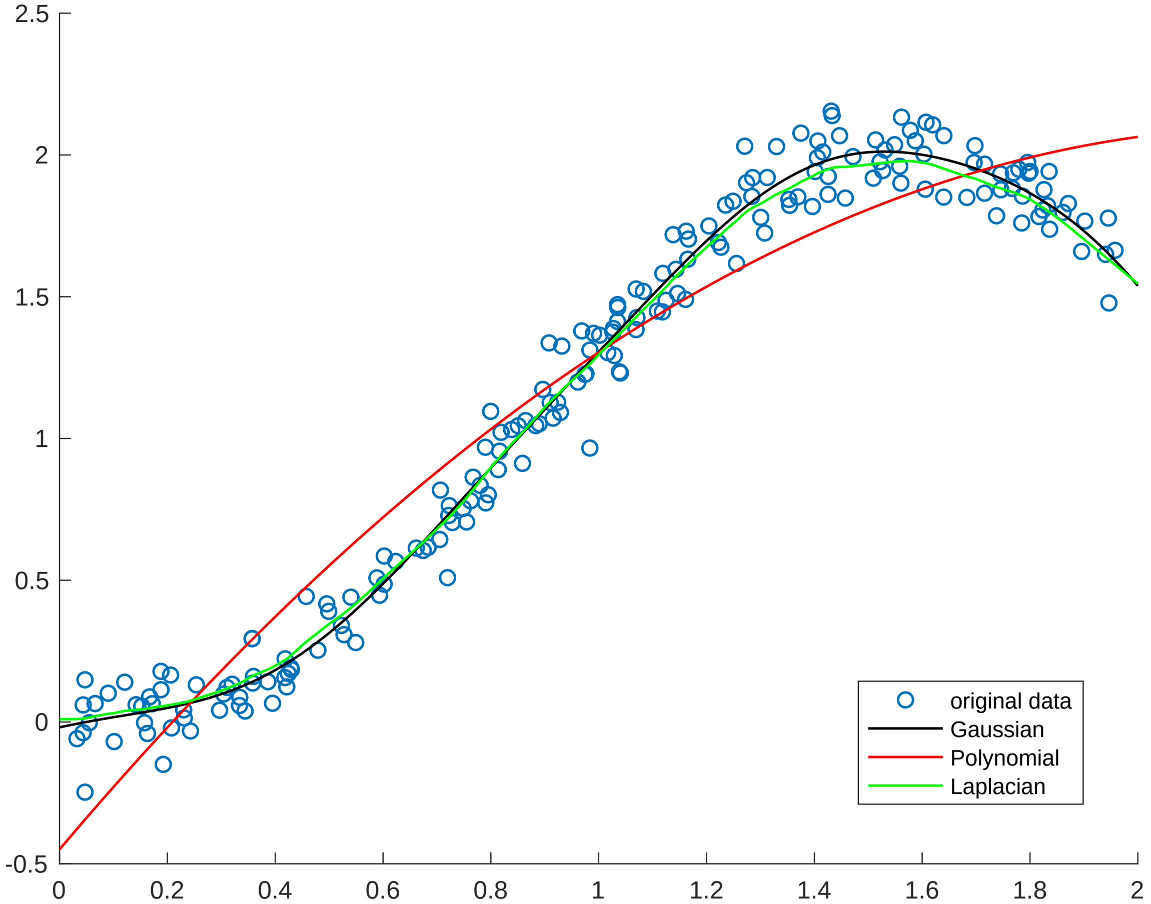

Figure 1: Plot of kernel regression functions $f(x)$ vs x and original data.

The regression kernels and tuning parameter λ used are
(1) Gaussian kernel: $k(x, x') = \exp(-\|x - x'\|^2/0.25)$ with $\lambda = 0.01$;
(2) Polynomial kernel: $k(x, x') = (\langle x, x' \rangle + 1)^2$ with $\lambda = 0.1$;
(3) Laplacian kernel: $k(x, x') = \exp(-\|x - x'\|)$ with $\lambda = 1$.

By comparing the RSS for the regression fit, we find that the quality of Gaussian kernel regression and Laplacian kernel regression is much better than that of Polynomial kernel regression. We can explain this by noting that the feature maps of Gaussian kernel and Laplacian kernel map $\mathcal{X}$ to an inifinite dimensional space, while the feature map of the polynomial kernel is of a finite dimension.

2 Question 2

In this project, we perform binary classifications for pairs of $1, 2, 3, 4$, i.e., $\{1, 2\}$,$\{1, 3\}$,...,$\{3, 4\}$. The data set for each pair is taken from the MNIST data set and split into two parts: training data set (60%) and test data set (40%). RKHS ridge regression with the quadratic polynomial $k(x, x') = (\langle x, x' \rangle + 1)^2$ is performed. The tuning parameter λ is fixed at $\lambda = 10$ for simplicity. (One can vary λ for better performance.)

The accuracy for the test set of different pairs:

	2	3	4
1	99.70%	99.68%	99.81%
2		99.19%	99.73%
3			99.95%

The performance is remarkably good.

3 Question 3

Prove the Semiparametric Representation Theorem: Let $c : \mathcal{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be a cost function, $\Omega : [0, \infty) \to \mathbb{R}$ be a strictly monotonic increasing function, and $\mathcal{H}_k$ be an 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 $\mathcal{F} = \{\tilde{f} = f + h : f \in \mathcal{H}_k,\ h \in \text{span}\{\psi_j\}_{j=1}^m\}$ be a functional class. Then each minimizer

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

admits a representation of the form

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

Proof: Let $\mathcal{B} = \{f_\alpha \in \mathcal{H}_k : f_\alpha = \sum_{i=1}^n \alpha_i k(x_i, \cdot), \alpha_i \in \mathbb{R}\}$ be a subspace of $\mathcal{H}_k$. Since B is of finite dimension, $\mathcal{B}$ is closed, and therefore $\mathcal{H}_k = \mathcal{B} \oplus \mathcal{B}^\perp$. That is, $\forall\, f \in \mathcal{H}_k$, there exists uniquely $f_\alpha \in \mathcal{B}$ and $f' \in \mathcal{B}^\perp$ such that $f = f_\alpha + f'$. Moreover, for any $i = 1, ..., n$,

$$f(x_i) = \langle f_\alpha + f', k(x_i, \cdot)\rangle = \langle f_\alpha, k(x_i, \cdot)\rangle = f_\alpha(x_i).$$

Note also that, for any $f \in \mathcal{H}_k$, $c[x_i, y_i, f(x_i)] = c[x_i, y_i, f_\alpha(x_i)]$, and

$$\Omega(\|f\|_{\mathcal{H}_k}) = \Omega(\|f_\alpha + f'\|_{\mathcal{H}_k}) = \Omega\Big(\sqrt{\|f_\alpha\|^2_{\mathcal{H}_k} + \|f'\|^2_{\mathcal{H}_k}}\Big) \geq \Omega(\|f_\alpha\|_{\mathcal{H}_k}).$$

So the minimizer $\tilde{f}^\star \in \mathcal{B} \oplus \text{span}\{\psi_j\}_{j=1}^m \subseteq \mathcal{F}$. That is, there exists $\alpha_i \in \mathbb{R}$ for all integers $1 \leq i \leq n$ and $\beta_j \in \mathbb{R}$ for all integers $1 \leq j \leq n$ such that

$$\tilde{f}^\star = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) + \sum_{j=1}^{m} \beta_j \psi_j.$$

Once α is determined, the coefficients β can be found from

$$\Psi\beta = \hat{Y} - \mathbf{K}\alpha,$$

where $\hat{Y} = (y_1, y_2, .., y_n) \in \mathbb{R}^n$, $K_{ij} = k(x_i, x_j)$ and $\Psi_{ij} = \psi_j(x_i)$. Since Ψ is of full rank, it is invertible and β can be uniquely determined by

$$\beta = \Psi^{-1} P_\Psi(\hat{Y} - \mathbf{K}\alpha) = (\Psi^T\Psi)^{-1}\Psi^T(\hat{Y} - \mathbf{K}\alpha).$$

4 Question 4

Given a p.s.d. kernel k, prove that

$$\tilde{k} = \frac{k(x, x')}{\sqrt{k(x, x)k(x', x')}}$$

is a p.s.d. kernel.
Proof: It is obvious that $\tilde{k}$ is symmetric by construction. ($\tilde{k}$ is well-defined when $k(x, x) > 0$ for all $x \in \mathcal{X}$.) For any positive integer $N \in \mathbb{N}^+$, for any real number $c_i \in \mathbb{R}$, for any $x_i \in \mathcal{X}$ where $i = 1, ..., N$, we have

$$\sum_{ij} c_i c_j \frac{k(x_i, x_j)}{\sqrt{k(x_i, x_i)k(x_j, x_j)}} = \sum_{ij} \left(\frac{c_i}{\sqrt{k(x_i, x_i)}}\right)\left(\frac{c_j}{\sqrt{k(x_j, x_j)}}\right) k(x_i, x_j) \geq 0.$$

The last inequality follows from the positive semi-definiteness of k. By the statement above, $\tilde{k}$ is also p.s.d.

5 Question 5

Let $k(\cdot, \cdot)$ be a p.s.d. kernel satisfying the conditions in Mercer's theorem. Define

$$\mathcal{H} = \{f = \sum_{j=1}^{N} w_j \sqrt{\lambda_j} \psi_j : \{w_j\} \in l_2\}.$$

1. Solve

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|w\|^2, \quad (*)$$

where $w = (w_1, w_2, ..., w_N)^T$ and $\|w\|^2 = w^T w$.

Let Ψ be an $n \times N$ matrix such that $\Psi_{ij} = \sqrt{\lambda_j}\psi_j(x_i)$ for $i = 1, ..n$ and $j = 1, ...N$. Define $y = (y_1, y_2, ..., y_n)^T \in \mathbb{R}^n$, $\hat{y} = (f(x_1), f(x_2), ..., f(x_n))^T = \Psi w \in \mathbb{R}^n$. We an rewrite the original problem as

$$\hat{w} = \arg\min_{w \in \mathbb{R}^N} \|Y - \Psi w\|^2 + \lambda \|w\|^2.$$

The corresponding Lagrangian is $\mathcal{L}(w) = \|Y - \Psi w\|^2 + \lambda \|w\|^2$. Setting its partial derivative to 0 yields

$$\frac{\partial \mathcal{L}}{\partial w} = -2\Psi^T (Y - \Psi w) + 2\lambda w = 0 \implies \hat{w} = (\Psi^T \Psi + \lambda \mathbf{I})^{-1} \Psi^T Y.$$

So $\hat{f} = \sum_{j=1}^{N} \hat{w}_j \sqrt{\lambda_j} \psi_j$.

2. Show that this formulation is equivalent to RKHS Ridge regression.

Recall the RKHS Ridge regression problem

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \iff \hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \|Y - \mathbf{K}\alpha\|^2 + \lambda \alpha^T \mathbf{K} \alpha,$$

where $\mathbf{K}$ is an $n \times n$ matrix such that $\mathbf{K}_{ij} = k(x_i, x_j)$. The solution to this problem is given by

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

Note also that by the Mercer's theorem

$$\mathbf{K} = \Psi \Psi^T.$$

We will show that the RKHS Ridge regression solution is also a solution to the problem (*). Firstly, We need to show that the residual sums of squares are the same

$$\begin{aligned}
& \|Y - \hat{Y}_{\hat{\alpha}}\|^2 = \|Y - \mathbf{K}\hat{\alpha}\|^2 = \|Y - \Psi\hat{w}\|^2 = \|Y - \hat{Y}_{\hat{w}}\|^2 \\
\Longleftarrow \quad & \mathbf{K}\hat{\alpha} = \Psi\hat{w} \\
\Longleftrightarrow \quad & \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}Y = \Psi(\Psi^T\Psi + \lambda\mathbf{I})^{-1}\Psi^T Y \\
\Longleftrightarrow \quad & \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1} - \Psi(\Psi^T\Psi + \lambda\mathbf{I})^{-1}\Psi^T = 0 \\
\Longleftrightarrow \quad & \mathbf{K} - \Psi(\Psi^T\Psi + \lambda\mathbf{I})^{-1}\Psi^T(\mathbf{K} + \lambda\mathbf{I}) = 0 \\
\Longleftrightarrow \quad & \mathbf{K} - \Psi(\Psi^T\Psi + \lambda\mathbf{I})^{-1}\Psi^T(\Psi\Psi^T + \lambda\mathbf{I}) = 0 \\
\Longleftrightarrow \quad & \mathbf{K} - \Psi(\Psi^T\Psi + \lambda\mathbf{I})^{-1}(\Psi^T\Psi + \lambda\mathbf{I})\Psi^T = 0 \\
\Longleftrightarrow \quad & \mathbf{K} - \Psi\Psi^T = 0.
\end{aligned}$$

Secondly, we need to show that the penalty terms are equal as well. Let $w_{\hat{\alpha}}$ denote the corresponding w of the model $\hat{\alpha}$ that satisfies $f_{\hat{\alpha}}(\cdot) = \sum_j w_{\hat{\alpha}}\sqrt{\lambda_j}\psi_j(\cdot) = \sum_i \alpha_i k(\cdot, x_i)$. In particular, considering the data points x_i only, we have $\Psi w_{\hat{\alpha}} = \mathbf{K}\alpha$. Since $\mathbf{K}\hat{\alpha} = \Psi\hat{w}$,

$$\Psi w_{\hat{\alpha}} = \Psi\hat{w} \implies w_{\hat{\alpha}} = \hat{w} \implies \|w_{\hat{\alpha}}\| = \|\hat{w}\|.$$

Thus, the model $f_{\hat{\alpha}}$ given by $\hat{\alpha}$ also solves the problem (*). Since both the problem (*) and the RKHS Ridge regression problem are strictly convex optimization problems, so the solution is unique, implying that $f_{\hat{\alpha}} = f_{\hat{w}}$ solves both of the problems. Hence, the two problems are equivalent.