Modern Statistical Methods — Example Sheet 1

Lucas Riedstra

...

Question 1. *Consider minimising the following objective involving response $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n \times p}$ over $(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p$:*

$$\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta).$$

Here $J \colon \mathbb{R}^p \to \mathbb{R}$ is an arbitrary penalty function. Suppose $\bar{X}_k = 0$ for $k = 1, \ldots, p$. Assuming that a minimiser $(\hat{\mu}, \hat{\beta})$ exists, show that $\hat{\mu} = \bar{Y}$. Now take $J(\beta) = \lambda\|\beta\|_2^2$ so we have the ridge regression objective. Show that

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

From here onwards, whenever we refer to ridge regression, we will assume X has had its columns mean-centred.

Solution. If a minimiser $\hat{\mu}$ exists, then the derivative of the objective must be 0. We expand the objective function as

$$\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta) = \sum_{i=1}^n (Y_i - \mu - (X\beta)_i)^2 + J(\beta)$$

$$= \sum_{i=1}^n \left(Y_i - \mu - \sum_{j=1}^p X_{ij}\beta_j \right)^2.$$

Now, differentiating w.r.t. μ yields

$$\frac{\partial}{\partial \mu}\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta) = -2\sum_{i=1}^n \left(Y_i - \mu - \sum_{j=1}^p X_{ij}\beta_j \right)$$

$$= -2\left(\sum_{i=1}^n Y_i - n\mu - \sum_{i=1}^n \sum_{j=1}^p X_{ij}\beta_j \right)$$

$$= -2\left(\sum_{i=1}^n Y_i - n\mu - \sum_{j=1}^p \beta_j(\sum_{i=1}^p X_{ij}) \right)$$

$$= -2\left(\sum_{i=1}^n Y_i - n\mu \right),$$

since $\sum_{i=1}^p X_{ij} = 0$ for all i by assumption. Now, setting the derivative to 0 shows that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$ is the minimiser. Since the second derivative of the objective function is constantly $+2$, this $\hat{\mu}$ is indeed a minimum.

The computation of $\hat{\beta}$ is given in the lecture notes.

Question 2. *Consider performing ridge regression when $Y = X\beta^0 + \varepsilon$, where $X \in \mathbb{R}^{n \times p}$ has full column rank, and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Let the SVD of X be $UDV^\top$ and write $U^\top X\beta^0 = \gamma$. Show that*

$$\frac{1}{n}\mathbb{E}\left\|X\beta^0 - X\hat{\beta}_\lambda^R\right\|_2^2 = \frac{1}{n}\sum_{j=1}^p \left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2 \gamma_j^2 + \frac{\sigma^2}{n}\sum_{j=1}^p \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}.$$

Solution. We write $\hat{\beta} := \hat{\beta}_\lambda^R$. We write out

$$X\hat{\beta} = UDV^\top(V^\top D^2 V + \lambda I)^{-1}V^\top DU^\top Y = UD^2(D^2 + \lambda I)^{-1}U^\top Y = \sum_j \frac{D_{jj}^2}{\lambda + D_{jj}^2}U_j U_j^\top Y.$$

Since $X\beta^0 = UDV^\top\beta^0$ lies in the range of U, we have $X\beta^0 = \sum_j U_j U_j^\top X\beta^0$.

$$X\beta^0 - X\hat{\beta} = \sum_j U_j U_j^\top X\beta^0 - \sum_j \frac{D_{jj}^2}{\lambda + D_{jj}^2}U_j U_j^\top Y$$

$$= \sum_j \left(U_j^\top X\beta^0 - \frac{D_{jj}^2 U_j^\top Y}{\lambda + D_{jj}^2}\right)U_j$$

$$= \sum_j \left(\frac{D_{jj}^2 U_j^\top(X\beta^0 - Y) + \lambda U_j^\top X\beta^0}{\lambda + D_{jj}^2}\right)$$

and therefore

$$\frac{1}{n}\mathbb{E}\left\|X\beta^0 - X\hat{\beta}\right\|_2^2 = \frac{1}{n}\sum_j \mathbb{E}\left(\frac{D_{jj}^2 U_j^\top(X\beta^0 - Y) + \lambda U_j^\top X\beta^0}{\lambda + D_{jj}^2}\right)^2$$

$$\overset{*}{=} \frac{1}{n}\sum_j \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}\mathbb{E}[(U_j^\top(X\beta^0 - Y))^2] + \frac{1}{n}\sum_j \left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2(U_j^\top X\beta^0)^2$$

$$\overset{**}{=} \frac{\sigma^2}{n}\sum_j \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2} + \frac{1}{n}\sum_j \left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2\gamma_j^2.$$

Note that the cross-terms in $(*)$ disappear since they are linear combinations of $\mathbb{E}(X\beta^0 - Y) = 0$, while the equality $**$ holds since

$$\mathbb{E}[(U_j^\top(X\beta^0 - Y))^2] = \mathrm{Var}(U_j^\top(X\beta^0 - Y)) = U_j^\top\mathrm{Var}(X\beta^0 - Y)U_j = \sigma^2 U_j^\top U_j = \sigma^2.$$

Question 3. *Show that the ridge regression estimates can be otained by ordinary least squares regression on an augmented data set with $\sqrt{\lambda}I$ added to the botten of X, and p zeroes added to the end of the response Y.*

Solution. Let $\tilde{X}, \tilde{Y}$ be the new data set, then our ordinary least squares objective function is

$$L(\beta) := \left\|\tilde{Y} - \tilde{X}\beta\right\|^2$$

$$= \|Y - X\beta\|^2 + \left\|\sqrt{\lambda}\beta\right\|^2$$

$$= \|Y - X\beta\|^2 + \lambda\|\beta\|^2,$$

which is exactly our ridge regression objective function.

Question 4. *In the following, assume that forming AB where $A \in \mathbb{R}^{a \times b}, B \in \mathbb{R}^{b \times c}$ requires $O(abc)$ computational operations, and that if $M \in \mathbb{R}^{d \times d}$ is invertible, then forming M^{-1} requires $O(d^3)$ operations.*

(a) *Suppose we wish to apply ridge regression to data $(Y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$ with $n \gg p$. A complication is that the data is split into m separate datasets of size $n/m \in \mathbb{N}$,*

$$Y = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{pmatrix}, \quad X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(m)} \end{pmatrix},$$

with each dataset located on a different server. Moving large amounts of data between servers is expensive. Explain how one can produce ridge estimates $\hat{\beta}_\lambda$ by communicating only $O(p^2)$ numbers from each server to some central server. What is the total order of the computation time required at each server, and at the central server for your approach?

(b) *Now suppose instead that $p \gg n$ and it is instead the variable sthat are split across m servers, so each server has only a subset of $p/m \in \mathbb{N}$ variables for each observation, and some central server stores Y. Explain how one can obtain the fitted values $X\hat{\beta}_\lambda$ communicating only $O(n^2)$ numbers from each server to the central server. What is the total order of the computation time required at each server, and at the central server for your approach?*

Solution. (a) Note that

$$X^\top Y = \left(X^{(1)\top} \dots X^{(m)\top} \right) \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{pmatrix} = \sum_{i=1}^m X^{(i)\top} Y^{(i)}.$$

Each of the $X^{(i)\top} Y^{(i)}$ can be computed at the data server (requiring $O(np)$ computation) and produces p numbers per data server.

Analogously, we have

$$X^\top X = \sum_{i=1}^m X^{(i)\top} X^{(i)},$$

which can also be computed at the data server (requiring $O(np^2)$ computation) and produces p^2 numbers per data server. In the end, each server will therefore have $O(p^2)$ numbers to send, having done $O(np^2)$ computations.

When each server sends their numbers to the central data server, the central data server can then compute $(X^\top X + \lambda I)^{-1}$ using $O(p^3)$ computations and then multiply that with $X^\top Y$ which will require $O(np)$ computations.

(b) In this case, we have
$$X = \left(X^{(1)} \dots X^{(m)} \right).$$

We use the alternative form of $\hat{\beta}_\lambda$, namely

$$X\hat{\beta}_\lambda^R = XX^\top (XX^\top + \lambda I)^{-1} Y.$$

Now, we have $XX^\top = \sum_{i=1}^m X^{(i)} X^{(i)\top}$, and each $X^{(i)} X^{(i)\top}$ can be computed at the data server, requiring $O(n^2 p)$ computations and producing $O(n^2)$ numbers to send.

After sending these numbers over to the central server, the inverse $(XX^\top + \lambda I)^{-1}$ can be computed in $O(n^3)$ time, then multiplied from the left with $XX^\top$ in $O(n^3)$ time, and multiplied with Y in $O(n^2)$ time. The total amount of computation at the central server will therefore be $O(n^3)$ as well.

Question 5. *Prove Proposition 4 in the notes.* Hint: for part (ii) it may help to consider the eigen-decompositions of positive semi-definite matrices $K^{(1)}$ and $K^{(2)}$ derived from kernels k_1 and k_2 in the form $K^{(1)} = PDP^\top = \sum_{i=1}^n P_i P_i^\top D_{ii}$ for example.

Proof. Proposition 4 is the following:

Suppose $k_1, k_2, \ldots$ are kernels.

   (i) If $\alpha_1, \alpha_2 \geq 0$ then $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel. If $\lim_{k\to\infty} k(x, x') =: k(x, x')$ exists for all $x, x' \in \mathcal{X}$, then k is a kernel.

   (ii) The pointwise product $k = k_1 k_2$ is a kernel.

Let $x_1, \ldots, x_n \in \mathcal{X}$ and define the positive definite matrix $K^{(m)}$ by $K_{i,j}^{(m)} = k_m(x_i, x_j)$.

(i) Since $(\alpha_1 k_1 + \alpha_2 k_2)(x_i, x_j) = \alpha_1 k_1(x_i, x_j) + \alpha_2 k_2(x_i, x_j)$, the kernel matrix $K_{ij} := (\alpha_1 k_1 + \alpha_2 k_2)(x_i, x_j)$ equals $\alpha_1 K_1 + \alpha_2 K_2$, and is therefore clearly symmetric positive semi-definite. This shows that $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel.

For the second part (the pointwise limit of kernels is a kernel), it suffices to show that the entrywise limit of positive semi-definite matrices is again positive semidefinite. But this is clear: if $K^{(m)} \to K$ entrywise, then for any $\mathbf{x} \in \mathbb{R}^n$ we have

$$\mathbf{x}^\top K \mathbf{x} = \mathbf{x}^\top \left(\lim_{m\to\infty} K^{(m)} \right) \mathbf{x} = \lim_{m\to\infty} \mathbf{x}^\top K^{(m)} \mathbf{x} \geq 0,$$

so K is positive semi-definite.

(ii) Let $K^{(1)}$ and $K^{(2)}$ be positive semi-definite, then we must show that the pointwise product $K^{(1)} \circ K^{(2)}$ is positive semi-definite as well. Write $K^{(1)} = \sum_{\ell=1}^n \lambda_\ell P_\ell P_\ell^\top$ and $K^{(2)} = \sum_{\ell=1}^n \mu_\ell Q_\ell Q_\ell^\top$, where $\lambda_i, \mu_i \geq 0$ and $\{P_\ell\}, \{Q_\ell\}$ are orthonormal bases of $\mathbb{R}^n$.

Then we find

$$(K^{(1)} \circ K^{(2)})_{ij} = \left(\sum_m \lambda_m P_{im} P_{jm} \right) \left(\sum_\ell \mu_\ell Q_{i\ell} Q_{j\ell} \right) = \sum_{m,\ell} \lambda_m \mu_\ell (P_{im} P_{jm} Q_{i\ell} Q_{j\ell})$$

$$= \sum_{m,\ell} \lambda_m \mu_\ell \big[(P_m \circ Q_\ell)(P_m \circ Q_\ell)^\top \big]_{ij},$$

and therefore $K^{(1)} \circ K^{(2)} = \sum_{m,\ell} \lambda_m \mu_\ell \big[(P_m \circ Q_\ell)(P_m \circ Q_\ell)^\top \big]$, which is a nonnegative combination of positive semi-definite matrices and therefore positive semi-definite.

$\square$

Question 6. *Let $\mathcal{X} := \{ x \in \mathbb{R}^d : \|x\| < 1 \}$. Show that $k(x, x') := (1 - x^\top x')^{-\alpha}$ defined on $\mathcal{X} \times \mathcal{X}$, where $\alpha > 0$, is a kernel.*

Proof. Note by Cauchy-Schwarz we have for $x, x' \in \mathcal{X}$ that $|x^\top x'| \leq \|x\| \|x'\| < 1$, so k is well-defined. Symmetry of k is clear. Define the function

$$\varphi \colon (-\infty, 1) \to \mathbb{R} \colon x \mapsto (1 - x)^{-\alpha}.$$

Then φ is smooth, and we can compute its Taylor series around 0. It is computed that

$$\varphi^{(k)}(x) = (1-x)^{-\alpha-k} \prod_{j=0}^{k-1} (\alpha + j) \implies \varphi^{(k)}(0) = \prod_{j=0}^{k-1} (\alpha + j).$$

4

Therefore, the Taylor series of φ is given by

$$\sum_{k=0}^{\infty}\left(\prod_{j=0}^{k-1}(\alpha+j)\right)\frac{x^k}{k!},$$

and its radius of convergence is given by

$$\lim_{k\to\infty}\frac{\frac{1}{k!}\prod_{j=0}^{k-1}(\alpha+j)}{\frac{1}{(k+1)!}\prod_{j=0}^{k}(\alpha+j)}=\lim_{k\to\infty}\frac{k}{\alpha+k}=1,$$

so φ agrees with its Taylor series on $(-1,1)$. By Cauchy-Schwarz, for $x,x'\in\mathcal{X}$ we have $x_i^\top x_j\leq\|x_i\|\|x_j\|<1$, so we can write for $x,x'\in\mathcal{X}$

$$k(x,x')=\varphi(x^\top x')=\sum_{k=0}^{\infty}\left(\prod_{j=0}^{k-1}(\alpha+j)\right)\frac{(x^\top x')^k}{k!}.$$

By proposition 4, every term in the sum is a kernel, and therefore the limit is a kernel as well. $\qquad\square$

Question 7. *Suppose we have a matrix of predictors $X\in\mathbb{R}^{n\times p}$ where $p\gg n$. Explain how to obtain the fitted values of the following ridge regression using the kernel trick:*

Minimise over $\beta\in\mathbb{R}^p$, $\vartheta\in\mathbb{R}^{p(p-1)/2}$, $\gamma\in\mathbb{R}^p$,

$$\sum_{i=1}^{n}\left(Y_i-\sum_{k=1}^{p}X_{ik}\beta_k-\sum_{k=1}^{p}\sum_{j=1}^{k-1}X_{ik}X_{ij}\vartheta_{jk}-\sum_{k=1}^{p}X_{ik}^2\gamma_k\right)^2+\lambda_1\|\beta\|^2+\lambda_2\|\vartheta\|^2+\lambda_3\|\gamma\|^2.$$

Note that we have indexed ϑ with two numbers for convenience.

Proof. Our linear model has predictors

$$\{X_{ik}\mid 1\leq k\leq p\}\cup\{X_{ij}X_{ik}\mid 1\leq j<k\leq p\}\cup\left\{X_{ik}^2\mid 1\leq k\leq p\right\}$$

for $i=1,\ldots,k$. We want to assume that $\lambda_1=\lambda_2=\lambda_3$: to this end, define $\xi=\frac{\lambda_1}{\lambda_2}$ and $\eta=\frac{\lambda_3}{\lambda_1}$, and replace ϑ by $\xi\vartheta$ and γ by $\eta\gamma$. This means we will have to replace our predictors by

$$Z_i:=(X_{ik}\mid 1\leq k\leq p)^\top\cup(\xi X_{ij}X_{ik}\mid 1\leq j<k\leq p)^\top\cup(\gamma X_{ik}^2\mid 1\leq k\leq p)^\top$$

We use the kernel trick: note that

$$\langle Z_i,Z_j\rangle=\sum_k X_{ik}X_{jk}+\xi^2\sum_{k<\ell}X_{ik}X_{i\ell}X_{jk}X_{j\ell}+\gamma^2\sum_k X_{ik}^2 X_{jk}^2$$

$$=\left(\frac{1}{\sqrt{2}\xi}+\frac{\xi}{\sqrt{2}}\sum_k X_{ik}X_{jk}\right)^2+(\gamma^2-\frac{\xi^2}{2})\sum_k X_{ik}^2 X_{jk}^2-\frac{1}{2\xi^2}.$$

We have therefore rewritten $\langle Z_i,Z_j\rangle$ into something that can be computed in $O(p)$, which is exactly the aim of the kernel trick. $\qquad\square$

Question 8. *Let $\hat{\alpha}$ be a minimiser of $\|Y-K\alpha\|^2+\lambda\alpha^\top K\alpha$, with K being a kernel matrix as usual. Show that $K\hat{\alpha}=K(K+\lambda I)^{-1}Y$.*

5

Solution. Suppose $\hat{\alpha}$ minimises $Q(\alpha) := \|Y - K\alpha\|^2 + \lambda \alpha^\top K \alpha$. Then we have

$$0 = \nabla_\alpha Q(\hat{\alpha}) = -2K(Y - K\hat{\alpha}) + 2\lambda K\hat{\alpha}$$
$$\lambda K\hat{\alpha} = KY - K^2 \hat{\alpha}$$
$$(K + \lambda)K\hat{\alpha} = KY$$
$$K\hat{\alpha} = (K + \lambda)^{-1}KY.$$

Since K and $(K + \lambda)^{-1}$ are simultaneously diagonalisable, they commute, and therefore $K\hat{\alpha} = K(K + \lambda)^{-1}Y$.

Question 9. *Consider minimising*

$$c(Y, X, f(x_1) + \mu, \ldots, f(x_n) + \mu) + J(\|f\|_\mathcal{H}^2),$$

over $f \in \mathcal{H}$ and $\mu \in \mathbb{R}$ where $\mathcal{H}$ is an RKHS. Here c is an arbirary loss function and J is strictly increasing. Let k be the reproducing kernel of $\mathcal{H}$. Show that any minimiser $\hat{g}(\cdot) = \hat{f}(\cdot) + \hat{\mu}$ may be written as

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

where $\hat{\alpha}_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

Proof. As in the proof of the representer theorem, we project f onto $U = \text{Span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$, and find that the problem is equivalent to finding

$$\underset{u \in U, \mu \in \mathbb{R}}{\arg\min} c(Y, X, u(x_1) + \mu, \ldots, u(x_n) + \mu) + J(\|u\|^2),$$

and writing $u = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$, we find that any minimiser $\hat{g}$ satisfies

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

as required. □

Question 10. *This question proves a result needed for Theorem 7 in the notes. Let $\mathcal{H}$ be a RKHS of functions on $\mathcal{X}$ with reproducing kernel k and suppose $f^0 \in \mathcal{H}$. Let $x_1, \ldots, x_n \in \mathcal{X}$ and let K be the kernel matrix $K_{ij} = k(x_i, x_j)$. Show that*

$$\left(f^0(x_1), \ldots, f^0(x_n)\right)^\top = K\alpha,$$

for some $\alpha \in \mathbb{R}^n$ and moreover that $\|f^0\|_\mathcal{H}^2 \geq \alpha^\top K\alpha$.

Proof. Let $U = \text{Span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$. Then we have $f^0 = u + v$ for some $u \in U, v \in U^\perp$, and writing $u = \sum_{m=1}^n \alpha_m k(\cdot, x_m)$ we see that

$$f^0(x_i) = \langle k(\cdot, x_i), u + v \rangle = \langle k(\cdot, x_i), u \rangle = u(x_i) = \sum_{m=1}^n \alpha_m k(x_i, x_m) = (K\alpha)_i,$$

and from this it follows immediately that

$$\|f^0\|^2 = \|u\|^2 + \|v\|^2 \geq \|u\|^2 = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K\alpha.$$

□

Question 11. *Show from first principles that the Sobolev kernel is indeed a (positive definite) kernel.*

Solution. Let $\mathcal{X} = [0, 1]$ and $k(x, x') = \min(x, x')$ the Sobolev kernel. We must then show that for any $x_1, \ldots, x_n \in \mathcal{X}$, the matrix $K_{ij} = \min(x_i, x_j)$ is positive semi-definite (symmetry of k is clear).

To this end, assume $x_1 \leq \cdots \leq x_n$ (this may require relabeling of the x_i which does not affect positive semi-definiteness of K). Now let $\mathbf{f}_i$ denote the vector $f_{i1}, \ldots, f_{i(i-1)} = 0$, $f_{ii}, \ldots, f_{in} = 1$.

$$
K = \begin{bmatrix}
x_1 & x_1 & x_1 & \cdots & x_1 \\
x_1 & x_2 & x_2 & \cdots & x_2 \\
x_1 & x_2 & x_3 & \cdots & x_3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_1 & x_2 & x_3 & \cdots & x_n
\end{bmatrix}
= x_1 \mathbf{f}_1 \mathbf{f}_1^\top + (x_2 - x_1) \mathbf{f}_2 \mathbf{f}_2^\top + \cdots + (x_n - x_{n-1}) \mathbf{f}_n \mathbf{f}_n^\top.
$$

Since this is a nonnegative combination of positive semidefinite matrices, we conclude that K is positive semidefinite.

Question 12. *Let $\mathcal{H}$ be a RKHS with reproducing kernel k. Show that if $h_x \in \mathcal{H}$ has the property that $\langle h_x, f \rangle = f(x)$ for all $f \in \mathcal{H}$, then $h_x(\cdot) = k(\cdot, x)$.*

Proof. For any $y \in \mathcal{X}$ we have

$$
h_x(y) = \langle k(\cdot, y), h_x \rangle = \langle h_x, k(\cdot, y) \rangle = k(x, y) = k(y, x),
$$

from which it follows that $h_x(\cdot) = k(\cdot, x)$. $\qquad\square$

Question 13. *Prove that if k is a reproducing kernel for RKHS's $\mathcal{H}_1$ and $\mathcal{H}_2$, then $\mathcal{H}_1 = \mathcal{H}_2$, so the RKHS is uniquely determined by k.* Hint: First argue that it is enough to show the result for $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Next consider decomposing each $f \in \mathcal{H}_2$ as $f = u + v$ with $u \in \mathcal{H}_1, v \in \mathcal{H}_1^\perp$ and argue that $v = 0$.

Proof. Without loss of generality we can assume $\mathcal{H}_1 = \mathcal{H}$, where $\mathcal{H}$ is the RKHS constructed in the proof of Theorem 5, so $\mathcal{H}$ is the completion of $\text{Span}\{k(\cdot, x) \mid x \in \mathcal{X}\}$. Now we can embed $\mathcal{H}$ into $\mathcal{H}_2$ by mapping $k(\cdot, x)$ to k_x. Clearly this mapping respects the inner product by construction, so this is a linear embedding. Therefore, we may assume $\mathcal{H}_1 \subseteq \mathcal{H}_2$.

Now suppose $\mathcal{H}_1 \subseteq \mathcal{H}_2$ and decompose each $f \in \mathcal{H}_2$ as $f = u + v$ with $u \in \mathcal{H}_1$, $v \in \mathcal{H}_1^\perp$. By the previous question we know that for every x there exists a unique k_x with $\langle k_x, f \rangle = f(x)$ for all f, and clearly we have $k_x \in \mathcal{H}_1$.

Then for each $f \in \mathcal{H}_2, x \in \mathcal{X}$ we have (since $k_x \in \mathcal{H}_1$) that

$$
f(x) = \langle k_x, f \rangle = \langle k_x, u \rangle + \langle k_x, v \rangle = \langle k_x, u \rangle = u(x),
$$

which implies $v = 0$. Since this is the case for every $f \in \mathcal{H}_2$, we conclude that $\mathcal{H}_1^\perp = 0$ and therefore $\mathcal{H}_1 = \mathcal{H}_2$ (since Hilbert spaces are closed). $\qquad\square$