CIS3990-002: Representer Theorem Fall 2023

Lecture: Linear Algebra

Date: October 4th, 2023 *Author: Eric Wong*

1 Functional Analysis

We will now embark into the land of function spaces. This will be based in part on these notes: https://www.stat.cmu.edu/~larry/=sml/functionspaces.pdf

- Function spaces are simply vector spaces where the elements of your vector space are functions.
- You can think of $\mathbb{R}^d$ as equivalent (homomorphic) to a function space, where the space consists of linear functions of d variables (i.e. for all $w \in \mathbb{R}^d$, we can create a bijection to the function $f(x) = x^\top w$)
- Typicaly, we want to find the best function in a function space that fits the data. However, we don't want to simply interpolate all the data—we want a function that behaves "nicely".
- All the characteristics of a vector space carry over to the function space
- Function space have a basis, i.e. $f = \sum_i \alpha_i b_i$ where b_i are basis functions.
- We can define an inner product between functions, such as $\langle f, g \rangle = \int_0^1 f(x)g(x)dx$
- The inner product then induces a norm, $\|f\|^2 = \langle f, f \rangle$.
- Functions are orthogonal if $\langle f, g \rangle = 0$.
- An orthonormal basis for a function space satisfies norm 1 and orthogonality.
- We can consider subspaces of a function space, its orthogonal complement, and projections onto a subspace.

2 Hilbert Space

- To define a Hilbert space, we need to define the notion of completeness.
- Intuitively, completeness means that as 2 points get closer and closer together, they converge to some point.
- A sequence $x_1, x_2, \dots$ is a Cauchy sequence if $\|x_m - x_n\| \to 0$ as $m, n \to \infty$.
- Cauchy sequences represent the notion that 2 "points" get closer and closer together.
- A space is complete if every Cauchy sequence converges to a limit.

- Example of an incomplete space: The space of continuous functions $C[0,1]$ with the norm $\|f\|_2^2 = \int_0^1 |f(x)|^2 dx$. Then, consider the sequence

$$f_n(x) = \begin{cases} -1 \text{ if } x \in [0, 1/2 - 2^{-n}] \\ (x - \frac{1}{2}) \cdot 2^n \\ 1 \text{ if } x \in [1/2 + 2^{-n}, 1] \end{cases}$$

  The limit of this sequence is the discontinuous function $f(x) = \begin{cases} -1 & \text{if } x \in [0, 1/2] \\ 1 & \text{if } x \in (1/2, 1] \end{cases}$

- This limit is what the 2 "points" converge to
- A complete inner product space is a Hilbert space
- A complete vector space with a norm is called a Banach space.
- Every inner product space defines an induced norm, therefore every Hilbert space is a Banach space
- However, not every Banach space is a Hilbert space. For example, the supremum norm $\|f\| = \sup_x f(x)$ can not be given by an inner product.
- Example: $\mathbb{R}^d$ with the standard inner product $\langle u, v \rangle = \sum_i v_i w_i$ is a Hilbert space.
- Example: the set of square integrable functions $f \in L^2(a,b) = \{f : \|f\|_2 < \infty\}$, i.e. functions such that $\int_a^b f(x)^2 dx < \infty$ and inner product $\langle f, g \rangle = \int_a^b f(x)g(x)dx$, is a Hilbert space.
- One can generalize the $L^2[a,b]$ space of functions to arbitrary p-norm where $\|f\|_p^p = \int_a^b |f(x)|^p dx$ and say $L_p(a,b) = \{f : \|f\|_p < \infty\}$
- For $L^2(a,b)$, we can get a countable orthonormal basis $\phi_1, \phi_2, \ldots$ such that $\|\phi_j\| = 1$ for all j and $int_a^b \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$. Then, every square integrable function can be written as the sum of basis functions $f = \sum_i \alpha_i \phi_i$.
- Example: Fourier basis on $[0,1]$ is $\phi_1(x) = 1$, and

$$\phi_{2j}(x) = \frac{1}{\sqrt{2}} \cos(2j\pi x), \ \phi_{2j+1}(x) = \frac{1}{\sqrt{2}} \sin(2j\pi x)$$

- Example: Cosine basis on $[0,1]$ is $\phi_0(x) = 1$ and

$$\phi_j(x) = \sqrt{2} \cos(2\pi j x)$$

3 Kernels

We can define a class of smooth functions using a construct called a kernel.

- A Mercer kernel is a continuous function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

1. $K(x, y) = K(y, x)$
2. K is positive semidefinite, i.e.

$$\sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) c_i c_j \geq 0$$

for all finite set of points $x_1, \ldots, x_N$ and real numbers $c_1, \ldots, c_N$. This can be written as $c^\top K(X) c$ for all $c \in \mathbb{R}^N$ and kernel matrices $K(X)$ where $K(X)_{ij} = K(x_i, x_j)$. In other words, the kernel matrix is a positive semidefinite matrix.

- Example: Gaussian kernel, $K(x, y) = \exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right)$

- An aside for Eigenfunctions: Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric and $K(x, y) < \infty$. Consider the linear operator $T_K : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ where $[T_K f](x) = \int_{\mathcal{X}} K(x, y) f(y)$. You can think of this as smoothing $f(x)$ around x where the weight of $f(y)$ for $f(x)$ is given by $K(x, y)$.

- Suppose T_K is positive semidefinite, i.e. $\int_{\mathcal{X}} \int_{\mathcal{X}} f(x) K(x, y) f(y) dx dy \geq 0$ for any $f \in L^2(\mathcal{X})$. Let λ_i, Ψ_i be eigenfunctions and eigenvectors of T_k, i.e.

$$T_K \Psi_i = \lambda_i \Psi_i \Leftrightarrow \int_{\mathcal{X}} K(x, y) \Psi_i(y) dy = \lambda_i \Psi_i(x)$$

Then, $\sum_i \lambda_i < \infty$, $\sup_x \Psi_i(x) < \infty$, and

$$K(x, y) = \sum_i \lambda_i \Psi_i(x) \Psi_i(y)$$

This representation is known as Mercer's Theorem.

- If K_1, K_2 are Mercer kernels, then so are $K(x, y) =$

1. $K_1(x, y) + K_2(x, y)$
2. $cK_1(x, y)$ for $c \geq 0$
3. $K_1(x, y) + c$ for $c \geq 0$
4. $K_1(x, y) K_2(x, y)$
5. $f(x) f(y)$ for $f : \mathcal{X} \to \mathbb{R}$
6. $K_1(x, y)^d$
7. $\exp(K_1(x, y))$

4 Reproducing Kernel Hilbert Space

- Let $K(x, y)$ be a kernel, and let $K_x(\cdot) = K(x, \cdot)$ be the kernel with the first argument fixed.

- Note $K_x(y) = K(x, y)$

- Consider the set of all possible linear combinations of the kernel:

$$\mathcal{H}_0 = \{f : f = \sum_j \alpha_j K_{x_j}\}$$

- For this set of functions, let $f = \sum_i \alpha_i K_{x_i}$ and $g = \sum_j \beta_j K_{y_j}$. Then we can define an inner product as
$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j K(x_i, y_j)$$
with the usual induced norm $\|f\| = \sqrt{\langle f, f \rangle}$

- The completion of $\mathcal{H}_0$ with respect to this norm is a Hilbert space called the RKHS generated by K, or $\mathcal{H}_K$.

- An RKHS is named after the reproducing property

- That is, let $\mathcal{H}_K$ be an RKHS of functions from a domain $\mathcal{X}$ to $\mathbb{R}$. Then, for every $x \in \mathcal{X}$, there exists a function δ_x such that for all $f \in \mathcal{H}_K$,
$$f(x) = \sum_i \alpha_i K_{x_i}(x) = \sum_i \alpha_i K(x_i, x) = \langle f, K_x \rangle$$
where the inner product comes from taking $g = K_x$

- In other words, the inner product of a function with K_x evaluates that function at x.

- This also implies that $\langle K_x, K_y \rangle = K_x(y) = K(x, y)$. K is called the reproducing kernel. K_x is called the representer.

- You can check that this is a well-defined Hilbert space, i.e.
  - $\langle f, g \rangle = \langle g, f \rangle$
  - $\langle cf + dg, h \rangle = c\langle f, h \rangle + c\langle g, h \rangle$
  - $\langle f, f \rangle = 0$ iff $f = 0$

- To verify the last one, suppose $\langle f, f \rangle = 0$. Pick any x. Then, using Cauchy-Schwarz,
$$0 \le f(x)^2 = \langle f, K_x \rangle^2 = \langle f, K_x \rangle \langle f, K_x \rangle \le \|f\|^2 \|K_x\|^2 = \langle f, f \rangle \|K_x\|^2 = 0$$
Therefore $0 \le f(x)^2 \le 0 \Rightarrow f(x) = 0$

- Evaluation functional: δ_x assigns a real number to each function, defined as $\delta_x f = f(x)$

- In an RKHS, the evaluation functional is $\delta_x f = \langle f, K_x \rangle = f(x)$ from the reproducing property

- Theorem: A Hilbert space is an RKHS if and only if the evaluation functionals are continuous

- Continuous means that if $f_n \to f$, then $\delta_x f_n \to \delta_x f$.

- This is not always true: Let $f(x) = 0$ and $f_n(x) = \sqrt{n} 1[x < 1/n^2]$. Then, $\|f_n - f\| = \|f_n\| = \sqrt{\int f_n(x)^2 dx} = \sqrt{\int_0^{1/n^2} n} = \frac{1}{\sqrt{n}} \to 0$. However, $\delta_0 f_n = \sqrt{n}$ which does not converge to $\delta_0 f = 0$. This is because a Hilbert space in general can contain very unsmooth functions.

- Every RKHS has a unique reproducing kernel. Moore-Aronszajn states that every PD function $K(\cdot, \cdot)$ defines a unique RKHS with K as its reproducing kernel.

- We have no assumption on the domain $\mathcal{X}$.

5 Representer Theorem

We will now prove a representer theorem. There are many representer theorems—we will prove a general version from https://people.eecs.berkeley.edu/~bartlett/courses/281b-sp08/8.pdf.

Theorem 1. *Fix a kernel k and let H be the corresponding RKHS. Let $\Omega : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function and let the SVM optimization problem be expressed as*

$$J(f^*) = \min_{f \in \mathcal{H}} J(f) = \sum_i \ell(f(x_i), y_i) + \Omega(\|f\|^2_{\mathcal{H}}$$

Then, the solution can be expressed as

$$f^* = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot)$$

Furthermore, if Ω is strictly increasing, then all solutions have this form.

We will do in in the following steps:

1. First, we will use orthogonality to show that Ω is a fuction of the sum of norms in the span and the complement of the span of kernels.
2. Second, we will use the reproducing property to rewrite $f(x_i)$ as an inner product in the Hilbert space, and in particular the span of the kernels.
3. Therefore, any minimizer will necessarily eliminate the orthogonal component, resulting in the global solution laying in the span of the kernels.

For step one, consider the subspace

$$U = \text{span}\{k(x_i, \cdot) : i \in (1, \dots, N)\}$$

Let f be any function. Then we can project f onto this subspace and its orthogonal complement:

$$f = f_s + f_\perp$$

Since these spaces are orthogonal, we have

$$\|f\|^2 = \|f_s\|^2 + \|f_\perp\|^2$$

To see this, let $b_1, \dots, b_k$ be a basis for S and $c_1, \dots c_k$ be a basis for the complement of S. Then,

$$\|f\|^2 = \langle f_s + f_\perp, f_s + f_\perp \rangle = \|f_s\|^2 + \|f_\perp\|^2 + 2\langle f_s, f_\perp \rangle$$

and note that the last term is zero since

$$\langle f_s, f_\perp \rangle = \langle \sum_i \alpha_i b_i, \sum_j \beta_j c_j \rangle = \sum_{ij} \langle b_i c_j \rangle = 0$$

Therefore, since Ω is non-decreasing,

$$\Omega(\|f\|_{\mathcal{H}}\|^2) \geq \Omega(\|f_s\|_{\mathcal{H}}^2)$$

which means for any f, Ω can be made smaller when f lands in the subspace f_s.

For step two, use the reproducing property to conclude that

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle = \langle f_S, k(x_i, \cdot)\rangle + \langle f_\perp, k(x_i, \cdot)\rangle = \langle f_s, k(x_i, \cdot)\rangle = f_s(x_i)$$

Therefore, $\sum_i \ell(f(x_i), y_i) = \sum_i \ell(f_s(x_i), y_i)$ only depends on f_s.

For the third step, note that the loss only depends on f_s (i.e. it is independent of the orthogonal subspace), and that the regularizer is minimized if f lies within S. Therefore, $J(f)$ is minimized if f lies within S and we can express $f^*(x) = \sum_i \alpha_i k(x_i, x)$ as a sum of the basis vectors of $\mathcal{H}$.