Computational Statistics & Machine Learning

Lecture 8

Hilbert spaces

Mark Girolami
mag92@cam.ac.uk

Department of Engineering
University of Cambridge

November 3, 2021

Overview

- Subspace of Banach space - Inner Product
- Pre-Hilbert Space
- Hilbert Space
- l^2 and L^2 Spaces
- Reproducing Kernel Hilbert Space

Subspace of Banach space - Inner Product

- In finite dimensions normed spaces are always complete - Banach spaces
- In infinite dimensions normed spaces may not be complete
- The l^p and L^p spaces are complete have you been able to prove this yet ?
- A subspace of Banach space which is important is one which has an inner product defined
- Why is an inner product space important for Machine Learning ?
- Only for $p = 2$ will l^p and L^p spaces have an inner product - source of norm is inner-product

Pre-Hilbert Space

- Inner product spaces that are incomplete are called pre-Hilbert spaces
- Consider the sequence of functions $\{f_n\}$ defined on the unit line as

$$f_n(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq \frac{1}{2} \\ 1 - 2n\left(x - \frac{1}{2}\right) & \text{for } \frac{1}{2} \leq x \leq \frac{1}{2}\left(1 + \frac{1}{n}\right) \\ 0 & \text{for} \frac{1}{2}\left(1 + \frac{1}{n}\right) \leq x \leq 1 \end{cases}$$

- It is clear $\{f_n\}$ converges to a discrete step function with a jump at $\frac{1}{2}$

Pre-Hilbert Space

- The distance between each f_n induced by the L^2 norm is

$$\begin{aligned} \|f_n - f_m\| &= \sqrt{\int_0^1 (f_n(x) - f_m(x))^2 dx} \\ &= \left(1 - \frac{n}{m}\right)\sqrt{\frac{1}{6n}} \to 0 \text{ as } m, n \to \infty \end{aligned}$$

- $\{f_n\}$ is therefore a Cauchy sequence that converges to

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq \frac{1}{2} \\ 0 & \text{if } \frac{1}{2} \leq x \leq 1 \end{cases}$$

- This is not a continuous function and hence not in the original space of functions
- The space is incomplete and a pre-Hilbert space

David Hilbert

- David Hilbert 1862 - 1943

Hilbert Space

- A *Hilbert Space* is a complete normed space endowed with an inner product
- Hilbert space hugely important in Machine Learning applications of function approximation (almost all)
- Functions can be considered as points in an infinite dimensional space defined by an infinite dimensional orthonormal basis

Hilbert Space

- Let $\{e_i\}$ be an infinite set of orthonormal vectors in Hilbert space V
- For any function f (a point in V) denote via the inner product the coefficients $c_i = \langle e_i, f \rangle$
- The finite sum from the first n terms in the infinite sequence $f_n = \sum_{i=1}^{n} c_i e_i$ and $\|f_n\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle e_i, e_j \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \delta_{i,j} = \sum_{i=1}^{n} c_i^2$
- $|\langle f, f_n \rangle|^2 \leq \|f\|^2 \cdot \|f_n\|^2 = \|f\|^2 \cdot \sum_{i=1}^{n} c_i^2$ by Cauchy-Schwarz
- The inner product $\langle f, f_n \rangle = \sum_i c_i \langle f, e_i \rangle = \sum_{i=1}^{n} c_i^2$
- Putting both together we obtain

$$\sum_{i=1}^{n} c_i^2 \leq \|f\|^2$$

Hilbert Space

- This holds for $n = \infty$ - Bessel's Inequality

$$\sum_{i=1}^{\infty} c_i^2 \leq \|f\|^2$$

- Therefore

$$\lim_{n\to\infty} f_n = \lim_{n\to\infty} \sum_{i=1}^{n} c_i e_i = \sum_{i=1}^{\infty} c_i e_i$$

- This has a finite norm (from Bessel) so it is indeed a convergent Cauchy sequence of functions
- The question is whether $f \in V$ thus V being a complete Hilbert space

Hilbert Space

- In an infinite dimensional space a set of orthonormal vectors can be a basis only when it is complete
- This can be checked using Parseval's Identity
- For any f in a Hilbert space V with an infinite set of orthonormal vectors $\{e_i\}$ then

$$\{e_i\} \text{ is complete } \iff \|f\|^2 = \sum_{i=1}^{\infty} c_i^2 \text{ with } c_i = \langle e_i, f \rangle$$

- Parseval's identity provides a criterion for the availability of an infinite set of vectors $\{e_i\}$ as a basis
- Both l^2 and L^2 spaces are complete Hilbert spaces could you show (prove) this?

L^2 Space

- Convergence of sequences of functions in L^2 space is important
- For any small $\epsilon > 0$ then

$$n > N \implies \|f_n - f\| < \epsilon \text{ with } f_n, f \in L^2$$

- The infinite sequence converges in the norm of the L^2 space what does this mean?
- Remember that we are using Lebesgue integral in defning the norm - hence the L in L^p spaces
- Therefore

$$\|f_n - f\| = \sqrt{\int_a^b |f(x) - f_n(x)|^2 dx}$$

which tests convergence in the mean
- The space L^2 can be considered as the *completion* of the space of continuous functions

L^2 Space

- This convergence in the mean indicates that while the L^2 norm indicates convergence of f_n to f there may be points in the domain where $f_n(x) \neq f(x)$
- This is described as *convergence almost everywhere* - you will see a.e. in texts and papers
- This is not feasible with either pointwise or uniform notions of convergence
- Remember that these are Lebesgue integrals defining the norm as such there may be sets in the domain which have measure zero whose function evaluations do not contribute to the integral
- This suggests equivalence everywhere *expect for sets of measure zero*
- This also means that the space L^2 is a space of equivalence classes of functions

Equivalence of l^2 and L^2 Spaces

- Let us represent a function by its Fourier series

$$f(x) = \sum_{k=-\infty}^{\infty} c_k \exp(ikx)$$

- This can be generalised as follows - consider a set of square integrable orthonormal functions $\{\phi_k\}$ in the norm of L^2
- The numbers $c_k = (f, \phi_k)$ are called the Fourier coefficients of $f \in L^2$
- The series

$$\sum_{k=1}^{\infty} c_k \phi_k$$

is called the Fourier series of f with respect to $\{\phi_k\}$

Equivalence of l^2 and L^2 Spaces

- As c_k is an inner product $\langle f, \phi_k \rangle$ it yields Bessel's inequality
$$\sum_{k=1}^{\infty} c_k^2 \leq \|f\|$$
- As $f \in L^2$ the norm remains finite and Bessel ensures the convergence of the infinite series
- Therefore the sequence of Fourier coefficients $\{c_k\}$ is an element of the space l^2 whichever orthonormal functions are chosen
- The elements $f \in L^2$ and $c = (c_1, c_2, c_3, \cdots) \in l^2$ are connected via $c_k = \langle f, \phi_k \rangle$ the Fourier coefficient
- The spaces l^2 and L^2 are *isomorphic* there is a one to one correspondence between elements of l^2 and L^2 that preserves algebraic structures - l^2 is a coordinate system for L^2

Function Approximation

- The development of function approximation in Hilbert space relies on definition of orthonormal basis functions with appropriate convergence properties
- A subspace of Hilbert space that has important properties is worth considering for function approximation - the RKHS
- Enormously influential in Machine Learning for function learning over last 20 years very many researchers, papers, and books available vast subject area
- Will briefly introduce the RKHS and how functions can be approximated from data

Function Approximation

- RKHS is a Hilbert space of functions where point evaluation is continuous
- We have already seen that in L^2 two functions f and g can be close in L^2 norm $\|f - g\|_{L^2}$ but not so in pointwise manner $|f(x) - g(x)|$
- The smaller Hilbert space, the RKHS, is such that the evaluation of functions is continuous that is if functions are close in norm they are also close pointwise
- The RKHS is associated with a *reproducing kernel* function when the inner product is taken the function evaluation is reproduced $f(x) = \langle f, K_x \rangle$ where K_x is the reproducing kernel defined as a function $K(x, y)$ at the evaluation points x and y

Function Approximation

- An RKHS defines a reproducing kernel function (which is symmetric and positive definite)
- A reproducing kernel function uniquely defines an RKHS (Moore-Aronszajn)
- It is shown that for a data set $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N) \in \mathcal{X} \times \mathbb{R}$ representing $f(x_i) = y_i$
- The regularised empirical error

$$\frac{1}{N}\sum_{n=1}^{N}|f(x_n) - y_n|^2 + \lambda\|f\|^2$$

is minimised by functions of the form

$$\hat{f}(\cdot) = \sum_{n=1}^{N}\alpha_n K(\cdot, x_n)$$

- This essentially reduces an infinite dimensional problem to one that operates in an N-dimensional subspace

Function Approximation

- Show optimal values for $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)$ satisfying

$$\underset{\alpha \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} |f(x_n) - y_n|^2 + \lambda \|f\|^2$$

  are given by the solution of the linear systems

$$(K + N\lambda I)\alpha = y$$

  where λ is a positive scalar, K is the $N \times N$ matrix with elements $K(x_i, x_j)$, I is the $N \times N$ identity matrix, and y is the $N \times 1$ vector with elements y_i
- This then enables function approximation via

$$\hat{f}(\cdot) = \sum_{n=1}^{N} \alpha_n K(\cdot, x_n)$$

  where each α_n is given above
- How do you interpret the norm based penalisation term if the L^2 norm is employed ? what is being penalised?