Hilbert space

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

A Hilbert space is an inner product space that is complete: every Cauchy sequence of its elements has a limit that again belongs to the space. Three examples are used throughout machine learning (ML): the Euclidean space of a fixed dimension, the space of random variables (RVs) with finite variance on a common probability space, and a reproducing kernel Hilbert space (RKHS). Linear regression uses a Euclidean space as its feature space and to parameterize its hypothesis space. Optimal estimation amounts to a projection in a Hilbert space consisting of RVs. A kernel method uses an RKHS as its transformed feature space and as its hypothesis space.

Definition

Consider some machine learning (ML) method with a hypothesis space that is equipped with a metric, i.e., a metric space. An iterative training method improves a hypothesis step by step, producing a sequence $h^{(1)}, h^{(2)}, \ldots$ of hypotheses. To ensure that this sequence converges to some optimal hypothesis $h^{\star}$, it is necessary that the hypotheses are increasingly close to each other, i.e., that they form a Cauchy sequence. Being a Cauchy sequence is not sufficient on its own. What turns it into convergence for every such sequence is a property of the underlying metric space, namely that it is complete. A prime example of a complete metric space is a Euclidean space $\mathbb{R}^d$ of finite dimension d.

A Hilbert space is a generalization of a Euclidean space to possibly infinite dimensions. In particular, a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is an inner product space in which every Cauchy sequence of its elements has a limit that again belongs to $\mathcal{H}$ (Bauschke and Combettes, 2011, Sect. 1.12) (see Fig. 1).

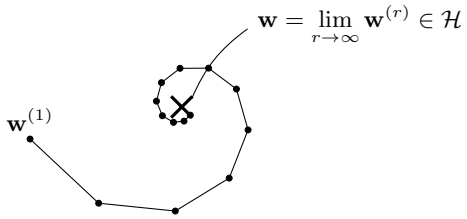

Figure 1: Completeness. The elements $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ of a Cauchy sequence (dots) lie ever closer to each other, and their limit $\mathbf{w} = \lim_{r\to\infty} \mathbf{w}^{(r)}$ (cross) again belongs to the Hilbert space. Data generated by `pythondemos/hilbertspace.py`

The elements of $\mathcal{H}$ are called vectors, whether they are arrays of numbers, random variables (RVs), or functions: what makes them vectors is that they can be added and scaled (see vector space). The inner product induces a norm $\|\mathbf{u}\|_{\mathcal{H}} := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ and, in turn, a metric $d(\mathbf{u}, \mathbf{v}) := \|\mathbf{u} - \mathbf{v}\|_{\mathcal{H}}$ (see inner product). A sequence $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots$ of vectors of $\mathcal{H}$ is a Cauchy sequence if its vectors eventually lie within any prescribed distance of each other: for every distance $\epsilon > 0$, however small, there is an index N such that $\left\|\mathbf{u}^{(r)} - \mathbf{u}^{(r')}\right\|_{\mathcal{H}} < \epsilon$ for all $r, r' \geq N$ (Rudin, 1976, Definition 3.8).

Three examples of Hilbert spaces are used throughout ML: the Euclidean space $\mathbb{R}^d$ of feature vectors, the space of RVs with finite variance on a common probability space, and a reproducing kernel Hilbert space (RKHS) of functions. The first of them carries the standard inner product $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$, and its completeness follows from that of the real numbers (Rudin, 1976, Thm. 3.11). Linear regression uses the Euclidean space $\mathbb{R}^d$ in two roles. The feature vectors $\mathbf{x} \in \mathbb{R}^d$ of the data points it is applied to are vectors of this space, and so are the weights $\mathbf{w} \in \mathbb{R}^d$ that parameterize its hypothesis space $\left\{h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle \;:\; \mathbf{w} \in \mathbb{R}^d\right\}$. Gradient descent (GD) searching for those weights produces a sequence $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ of model parameters. These are vectors in the second role, so the question the opening paragraph asks has a reassuring answer here: $\mathbb{R}^d$ is complete, and a Cauchy sequence of model parameters converges to model parameters.

The second example is the set of all RVs x with finite variance, defined on a common probability space (Gray and Davisson, 2004, Sect. 5.8.1). Here, two RVs are identified whenever the expectation of their squared difference is zero, $\mathbb{E}\{(x - x')^2\} = 0$, i.e., whenever $x = x'$ with probability one. With this identification, the expectation $\langle x, x' \rangle := \mathbb{E}\{xx'\}$ is an inner product, and the resulting space is complete (Gray and Davisson, 2004, Lem. 5.1).

In the Hilbert space of finite-variance RVs on a common probability space, optimal linear estimation is an orthogonal projection (see Fig. 2). A linear estimator $\widehat{y}$ of an RV y from an observed RV x is an element of the subspace $\{ax \;:\; a \in \mathbb{R}\}$ spanned by x. The corresponding estimation error $y - \widehat{y}$ is

measured by the induced norm $\sqrt{\langle y-\widehat{y}, y-\widehat{y}\rangle}$. The smallest error is obtained by the linear estimator whose error is orthogonal to x (Gray and Davisson, 2004, Thm. 4.9),

$$\begin{aligned} 0 &= \langle y-\widehat{y}, x\rangle \\ &= \mathbb{E}\big\{\big(y-\widehat{y}\big)x\big\}. \end{aligned}$$

When the RVs y, x are zero-mean, this orthogonality means that the error $y-\widehat{y}$ is uncorrelated with x.

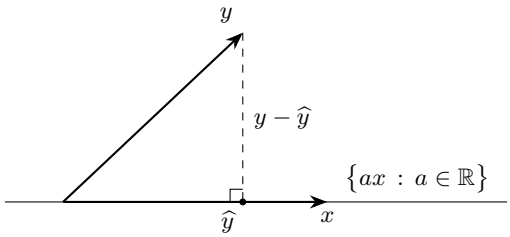

Figure 2: Optimal linear estimation as an orthogonal projection. The RV y, the observed RV x, and the estimator $\widehat{y}$ are vectors of the Hilbert space of finite-variance RVs. The estimators that are linear in x form the subspace $\{ax : a \in \mathbb{R}\}$ spanned by x (the straight line through it), and the estimation error $y-\widehat{y}$ is smallest for the $\widehat{y}$ whose error is orthogonal to x, the right angle marked by the small square at $\widehat{y}$

Optimal nonlinear estimation can also be represented as an orthogonal projection, but on a larger subspace. Projecting y onto the RVs that are functions of x, rather than onto the multiples ax alone, gives the conditional expectation $\mathbb{E}\{y \mid x\}$: it has the smallest squared error among all estimators computed from x (Gray and Davisson, 2004, Thm. 4.5).

The third example is an RKHS. It is a Hilbert space $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$ whose inner product reproduces point evaluations. In particular, each RKHS is associated with a kernel $k\big(\cdot,\cdot\big)$ such that $k\big(\mathbf{x},\cdot\big) \in \mathcal{H}$ for every $\mathbf{x} \in \mathcal{X}$ and

$$h(\mathbf{x}) = \langle h, k\big(\mathbf{x},\cdot\big)\rangle \quad \text{for every } h \in \mathcal{H}.$$

Thus, each vector $h \in \mathcal{H}$ is itself a hypothesis map $\mathcal{X} \to \mathbb{R}$, evaluated by taking an inner product with $k\big(\mathbf{x},\cdot\big)$. This construction of a hypothesis map generalizes the linear model, which uses the standard inner product of the Euclidean space. The map $\mathbf{x} \mapsto k\big(\mathbf{x},\cdot\big)$ is a feature transformation, carrying the feature vector of a data point into $\mathcal{H}$. A kernel method uses an RKHS in two roles. It is the transformed feature space, in which a data point is represented by $k\big(\mathbf{x},\cdot\big)$, and it is the hypothesis space from which a hypothesis is learned (Hastie et al., 2009, Sect. 12.3.3; Schölkopf and Smola, 2002). These are the two roles of $\mathbb{R}^d$ in linear regression, with one difference: there a vector parameterizes a hypothesis, here a vector is one.

Cross References

inner product, vector space, norm, Cauchy sequence, Euclidean space, RV, variance, expectation, conditional expectation, RKHS, kernel method, orthogonal projection

References

1. Bauschke and Combettes (2011). Convex Analysis and Monotone Operator Theory in Hilbert Spaces. 1st ed., Springer Science+Business Media, New York, NY, USA.

2. Gray and Davisson (2004). An Introduction to Statistical Signal Processing. Cambridge Univ. Press, Cambridge, U.K.

3. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

4. Rudin (1976). Principles of Mathematical Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.

5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.