Lecture 6

Some examples of Banach and Hilbert spaces are:

- $\mathbb{R}^d$ is a Hilbert space of dimension d.
- Let $E \subseteq \mathbb{R}^d$ and $p \geq 1$. The set $\mathbf{L}^p(E)$, defined as:

$$\mathbf{L}^p(E) = \left\{ f : E \to \mathbb{R} : \int_E |f(x)|^p \, dx < \infty \right\},$$

  is an infinite dimensional vector space (so long as E has nonzero Lebesgue measure). We can define a norm as follows:

$$\|f\|_p = \left(\int_E |f(x)|^p \, dx \right)^{1/p},$$

  which makes it a Banach space (since it is complete). When $p = 2$, we can also define an inner product:

$$\langle f, g \rangle = \int_E f(x) g(x) \, dx,$$

  which makes $\mathbf{L}^2(E)$ a Hilbert space.

- Let $C[0,1]$ be the space of real valued continuous functions on the interval $[0,1]$. This is a vector space, and we can make it an inner product space via:

$$\langle f, g \rangle = \int_0^1 f(x) g(x) \, dx.$$

  However, it is *not* a Hilbert space because it is not complete! Here is an example showing why. Define a sequence of functions $\{f_n\}_{n \geq 1} \subset C[0,1]$:

$$f_n(x) = \begin{cases} nx & \text{if } x \in [0, 1/n], \\ 1 & \text{if } x \in (1/n, 0]. \end{cases}$$

  You can show that f_n is a Cauchy sequence in the norm:

$$\|f\|^2 = \int_0^1 f(x)^2 \, dx.$$

But the limit as $n \to \infty$ is:

$$\lim_{n\to\infty} f_n(x) = f(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x \in (0,1] \end{cases} \notin C[0,1].$$

A subset U of a normed space V is *dense* if every point $v \in V$ either belongs to U or is a limit point of U. For any normed space V, one can construct a complete normed space $\overline{V}$, which contains V as a dense subspace. The space $\overline{V}$ is called the *completion* of V. As an example, the completion of $C[0,1]$ (as defined above) is $\mathbf{L}^2[0,1]$. Since every inner product space can be completed to be a Hilbert space, inner products spaces are sometimes referred to as *pre-Hilbert spaces*.

Some properties of finite dimensional inner product spaces (which you studied in linear algebra), carry over to general (infinite dimensional) Hilbert spaces. One such property is the notion of a *projection*. Let $\mathcal{H}$ be a Hilbert space and $U \subset \mathcal{H}$ a closed subspace. Then every $v \in \mathcal{H}$ can be written uniquely as:

$$v = z + z^{\perp}, \quad z \in U, \ \langle z^{\perp}, u\rangle = 0, \forall\, u \in U.$$

The vector z is the unique element of U such that:

$$\|v - z\| = \inf_{u\in U} \|v - u\|.$$

The map $v \mapsto z$ is called the projection of x onto U, and is denoted:

$$P_U v = Pv = z.$$

The projection operator P is a linear map.

Let A be an index set (e.g., $\mathbb{N}$), which is possibly uncountably infinite (e.g., $\mathbb{R}$). A collection of vectors $\{v_i\}_{\alpha\in A} \subset \mathcal{H}$ is an orthonormal system if

$$\langle v_\alpha, v_\beta\rangle = \delta_{\alpha,\beta} = \begin{cases} 1 & \text{if } \alpha = \beta, \\ 0 & \text{if } \alpha \neq \beta. \end{cases}$$

A set $\mathcal{B} = \{e_\alpha\}_{\alpha\in A} \subset \mathcal{H}$ is an *orthonormal basis (ONB)* if it is an orthonormal system and if

$$\forall\, \alpha \in A, \ \langle v, e_\alpha\rangle = 0 \Longrightarrow v = 0.$$

A Hilbert space $\mathcal{H}$ is *separable* if it has a countable ONB; this means we can take the index set A to be $\mathbb{N}$.

Every vector $v \in \mathcal{H}$ (separable or not) can be represented in an ONB $\mathcal{B} = \{e_\alpha\}_{\alpha \in A}$ as:

$$v = \sum_{\alpha \in A} \langle v, e_\alpha \rangle e_\alpha. \tag{14}$$

Parseval's Theorem proves that:

$$\|v\|^2 = \sum_{\alpha \in A} |\langle v, e_\alpha \rangle|^2.$$

Note that the right hand side of (14) may have an infinite number of nonzero terms, unlike an algebraic basis for an infinite dimensional vector space.

Another example of a Banach space is ℓ^p, where $1 \leq p \leq \infty$. For $1 \leq p < \infty$, define ℓ^p as:

$$\ell^p = \left\{ x = (x[i])_{i \in \mathbb{N}} : x[i] \in \mathbb{R}, \ \ \|x\|_p = \left(\sum_{i \in \mathbb{N}} |x[i]|^p \right)^{1/p} < \infty \right\}.$$

Define ℓ^∞ as:

$$\ell^\infty = \left\{ x = (x[i])_{i \in \mathbb{N}} : x[i] \in \mathbb{R}, \ \ \|x\|_\infty = \sup_{i \in \mathbb{N}} |x[i]| < \infty \right\}.$$

We remark that the series $(1/n)_{n \in \mathbb{N}}$ is in ℓ^p for $p > 1$.

The space ℓ^2 (so $p = 2$) is a Hilbert space with inner product:

$$\langle x, y \rangle = \sum_{i \in \mathbb{N}} x[i] y[i].$$

It is a separable Hilbert space, since the following countable set $\mathcal{B}$ is an ONB for ℓ^2:

$$\mathcal{B} = \{ e_k : k \in \mathbb{N}, \ \ e_k[i] = \delta(k - i) \}.$$

5 Kernels

5.1 Introduction

Adapted from [1, Chapter 2.1]

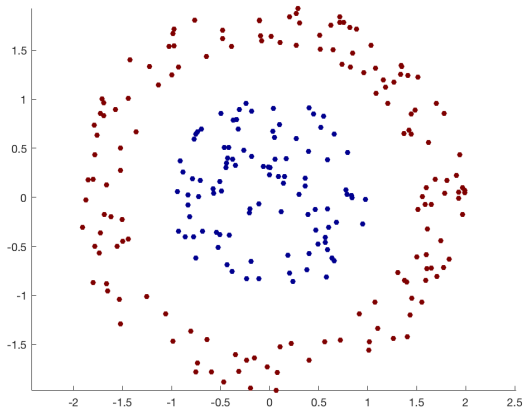
Kernels $k(x, x')$ allow us to build nonlinearity into our machine learning algorithms via the similarity measure between $x, x' \in \mathcal{X}$. For unsupervised learning, this will allow us to untangle highly nonlinear data, that linear methods like PCA cannot. For supervised learning, many binary classification kernel algorithms have the form:

$$y = f(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i k(x, x_i) + b\right), \tag{15}$$

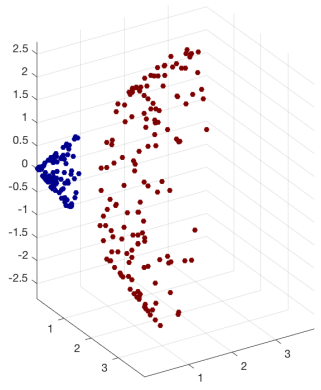
while in kernel regressions we have:

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) + b.$$

Let's look at a couple of examples to motivate the use of kernels. Consider the following labeled training data in Figure 7(a). The decision boundary we need to learn is a circle. Linear learning algorithms, like the one described in Section 2, can only learn linear decision boundaries, which in this case are lines. Thus we need a nonlinear algorithm.



(a) Two class training data in which the decision boundary is a circle.



(b) Data from Figure 7(a) after nonlinear mapping implicitly defined by the kernel $k(x, x') = \langle x, x' \rangle^2$.

Figure 7: Illustration of the implicit nonlinear mapping of the nonlinear polynomial kernel.

Let $x, x' \in \mathcal{X} \subset \mathbb{R}^2$ and consider the following nonlinear quadratic polynomial kernel:

$$k(x, x') = \langle x, x' \rangle^2.$$

If we expand this kernel, we get:

$$\begin{aligned} k(x,x') &= \langle x, x' \rangle^2, \\ &= \left(x[1]x'[1] + x[2]x'[2]\right)^2, \\ &= x[1]^2x'[1]^2 + x[2]^2x'[2]^2 + 2x[1]x[2]x'[1]x'[2], \\ &= \langle \Phi(x), \Phi(x') \rangle, \end{aligned}$$

where

$$\begin{aligned} \Phi : \mathbb{R}^2 &\to \mathbb{R}^3, \\ x = (x[1], x[2]) &\mapsto \Phi(x) = (x[1]^2, x[2]^2, \sqrt{2}x[1]x[2]). \end{aligned}$$

Thus $k(x,x')$ is a linear kernel between x and x', *after* the nonlinear mapping $x \mapsto \Phi(x)$. In other words, it is linear in the "feature space" defined by the image of Φ. Figure 7(b) shows that this nonlinearity is enough to transform the nonlinear decision boundary in Figure 7(a) into a linear decision boundary (which is a 2D hyperplane since we are now in $\mathbb{R}^3$). Pretty amazing! To drive this point home, if we examine this kernel in the classifier (15) we get:

$$\begin{aligned} f(x) &= \operatorname{sgn}\left(\sum_{i=1}^n \alpha_i k(x, x_i) + b\right), \\ &= \operatorname{sgn}\left(\sum_{i=1}^n \alpha_i \left(x[1]^2x_i[1]^2 + x[2]^2x_i[2]^2 + 2x[1]x[2]x_i[1]x_i[2]\right) + b\right), \\ &= \operatorname{sgn}\Bigg(\underbrace{\left(\sum_{i=1}^n \alpha_i x_i[1]^2\right)}_{w_1} x[1]^2 + \underbrace{\left(\sum_{i=1}^n \alpha_i x_i[2]^2\right)}_{w_2} x[2]^2 + \cdots \\ &\quad \cdots \underbrace{\left(\sum_{i=1}^n \alpha_i \sqrt{2} x_i[1]x_i[2]\right)}_{w_3} \sqrt{2}x[1]x[2] + b\Bigg), \\ &= \operatorname{sgn}\left(w_1x[1]^2 + w_2x[2]^2 + w_3\sqrt{2}x[1]x[2] + b\right). \end{aligned}$$

So the kernel is learning weights in the nonlinear feature space $\Phi(x)$, and linearly combining the weighted features to classify x.

Exercises

Exercise 12. Sketch a "proof" that the kernel $k(x, x') = \langle x, x' \rangle^2$ can learn any ellipsoidal decision boundary in $\mathbb{R}^2$ centered at the origin, given sufficient (even say infinite) training data. More specifically, let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{-1, +1\}$, with class -1 lying inside the ellipse, and class $+1$ lying outside the ellipse (see Figure 8). Explain convincingly that the following binary classifier:

$$y = f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i k(x, x_i) + b\right).$$

can learn this ellipsoidal boundary, if it has enough training data. This does not have to be a rigorous proof, but you should use equations and mathematical ideas to make your points.

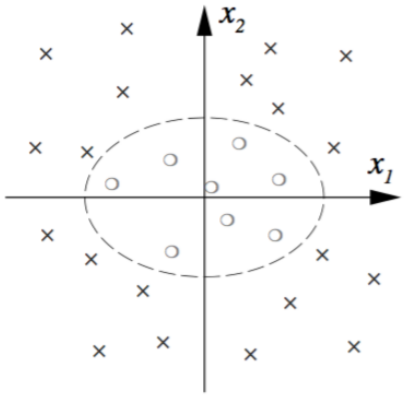
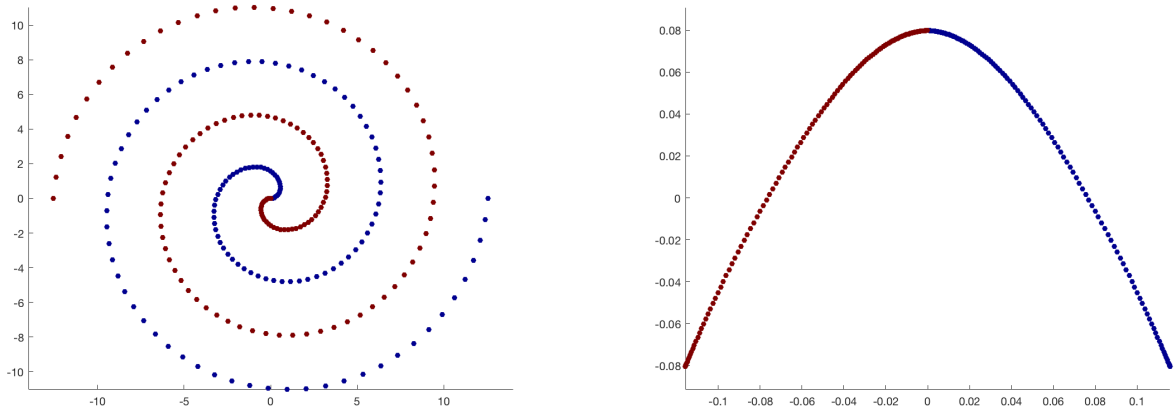


Figure 8: Ellipsoidal decision boundary

Figure 9(a) illustrates another example: two intertwined spirals, each with a different class label. The class boundary here is highly nontrivial, since the two spirals are intertwined. On the other hand, the data lies on a 1D curve, so if we can "untangle" this curve the problem will become much simpler. This is exactly what nonlinear dimension reducing methods do. In this case, we used something called "diffusion maps," which uses heat diffusion to learn that this data lies on a 1D curve, and then maps this data into a much simpler curve given in Figure 9(b). The algorithm is built upon the Gaussian kernel, which is defined as:

$$k(x, x') = \exp(-\|x - x'\|^2 / 2\sigma^2),$$

where σ gives the width of the kernel. We'll come back to "diffusion maps" and other such algorithms later in the course. But it is clear from Figure 9(b) that the decision boundary for the classification problem can now be taken as a line!



(a) Two intertwined spirals, each with a different class label.

(b) Data from Figure 9(a) after nonlinear mapping learned, unsupervised, via diffusion maps.

Figure 9: Illustration of nonlinear manifold learning to unwind highly curved data.

Exercises

Exercise 13. Let $x, x' \in \mathbb{R}^p$. Prove that the kernel

$$k(x, x') = (\langle x, x' \rangle + c)^2, \quad c \in \mathbb{R}$$

induces a feature map $\Phi : \mathbb{R}^p \to \mathcal{H}$ into all monomials up to degree 2. What is the dimension of $\mathcal{H}$? Discuss the role of c.

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[5] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[6] J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[7] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.