Lecture 7

5.2 Positive semidefinite kernels

Adapted from [1, Chapter 2.2.1]

Starting with this section we will try to answer the following question: Which types of kernels $k(x, x')$ induce a nonlinear feature map $\Phi : \mathcal{X} \to \mathcal{H}$, from a set $\mathcal{X}$ into a Hilbert space $\mathcal{H}$, so that $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$?

Given a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and sampled data $\mathcal{X}_n = \{x_1, \dots, x_n\} \subset \mathcal{X}$, the $n \times n$ matrix

$$K_{ij} = k(x_i, x_j)$$

is the *Gram matrix* of k with respect to $\mathcal{X}_n$.

A real valued Gram matrix K satisfying

$$\sum_{i,j=1}^{n} c_i c_j K_{ij} \geq 0$$

for all $c_i \in \mathbb{R}$ is *positive semidefinite*. A symmetric matrix is positive semidefinite if and only if all of its eigenvalues are nonnegative.

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is symmetric, and which for all $n \in \mathbb{N}$ and all $x_1, \dots, x_n \in \mathcal{X}$ gives rise to a positive semidefinite Gram matrix, is a *positive semidefinite kernel*. Positive semidefinite kernels are nonnegative on the diagonal (check this!):

$$\forall\, x \in \mathcal{X}, \quad k(x, x) \geq 0. \tag{16}$$

Kernels can be regarded as generalized inner products. However, they are not linear! (so linearity in the arguments does not hold) They do satisfy a type of Cauchy Schwarz inequality though:

Proposition 1. *If k is a positive semidefinite kernel, then*

$$\forall\, x, x' \in \mathcal{X}, \quad k(x, x')^2 \leq k(x, x) k(x', x').$$

Proof. Take $x_1 = x$ and $x_2 = x'$. Then for all $c_1, c_2 \in \mathbb{R}$,

$$c_1^2 k(x_1, x_1) + c_2^2 k(x_2, x_2) + 2c_1 c_2 k(x_1, x_2) \geq 0. \tag{17}$$

Take:

$$c_1 = k(x_1, x_2) \qquad c_2 = -k(x_1, x_1). \tag{18}$$

Plugging (18) into (17):

$$\begin{aligned} k(x_1, x_2)^2 k(x_1, x_1) + k(x_1, x_1)^2 k(x_2, x_2) - 2k(x_1, x_2)^2 k(x_1, x_1) &\geq 0, \\ k(x_1, x_1)^2 k(x_2, x_2) - k(x_1, x_1) k(x_1, x_2)^2 &\geq 0, \\ k(x_1, x_1) \left[k(x_1, x_1) k(x_2, x_2) - k(x_1, x_2)^2 \right] &\geq 0. \end{aligned}$$

But $k(x_1, x_1) \geq 0$, so this implies that

$$k(x_1, x_1) k(x_2, x_2) - k(x_1, x_2)^2 \geq 0,$$

which completes the proof. □

Exercises

Exercise 14. Prove (16).

5.3 The reproducing kernel map

Adapted from [1, Chapter 2.2.2]

Just a reminder, k is a real valued positive semi-definite kernel; also let $\mathcal{X}$ be nonempty. Let

$$\mathbb{R}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathbb{R}\} = \text{ set of all functions mapping } \mathcal{X} \text{ to } \mathbb{R},$$

and define:

$$\begin{aligned} \Phi :& \mathcal{X} \to \mathbb{R}^{\mathcal{X}}, \\ & x \mapsto \Phi(x) = k(\cdot, x). \end{aligned}$$

So $\Phi(x) \in \mathbb{R}^{\mathcal{X}}$, and we have

$$\forall\, x' \in \mathcal{X}, \quad \Phi(x)(x') = k(x', x) = k(x, x').$$

Thus this map Φ represents $x \in \mathcal{X}$ be measuring its similarity to all other points in $\mathcal{X}$. See Figure 10 for an illustration of the map Φ.

We are going to systematically:

1. Turn the image of Φ into a vector space.
2. Define an inner product on this vector space.
3. Show this inner product satisfies $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$.

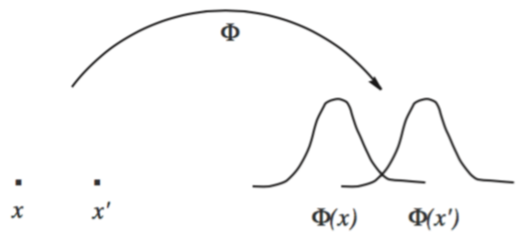

Figure 10: Visualization of the feature map Φ, which represents each $x \in \mathcal{X}$ by a kernel shaped function sitting on x. In this sense, each data point is represented by its similarity to all other points in $\mathcal{X}$. In the picture, the kernel is assumed to be bell shaped, e.g., a Gaussian $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$.

5.3.1 Making the image of Φ a vector space

Let $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in \mathcal{X}$ all be arbitrary. Linear combinations of $\Phi(x_1), \ldots, \Phi(x_n)$ take the form:

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i). \tag{19}$$

As you can verify, the collection of all such f (19) defines a vector space V. Note that two different collections of points $\{x_i\}_{i \leq n}$ and coefficients $\{\alpha_i\}_{i \leq n}$ may give the same f! In other words, there may exist $m \in \mathbb{N}$, $\beta_1, \ldots, \beta_m \in \mathbb{R}$, and $x'_1, \ldots, x'_m \in \mathcal{X}$ such that:

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) = \sum_{j=1}^{m} \beta_j k(\cdot, x'_j).$$

5.3.2 Defining an inner product

Let f be as in (19) and let g be:

$$g(\cdot) = \sum_{j=1}^{m} \beta_j k(\cdot, x_j').$$

Define the inner product between f and g as:

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x_j'). \tag{20}$$

Before checking the properties of an inner product, we first need to make sure it is "well defined." Indeed, it depends upon the points $\{x_i\}_{i \leq n}$ and $\{x_j'\}_{j \leq m}$, and the coefficients $\{\alpha_i\}_{i \leq n}$ and $\{\beta_j\}_{j \leq m}$, used to represent f and g, respectively. To check this, first observe:

$$\begin{aligned}\langle f, g \rangle &= \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \beta_j k(x_i, x_j'), \\ &= \sum_{i=1}^{n} \alpha_i g(x_i). \end{aligned} \tag{21}$$

Thus $\langle f, g \rangle$ does not depend on the representation of g. Similarly,

$$\langle f, g \rangle = \sum_{j=1}^{m} \beta_j f(x_j'),$$

and so it does not depend on the representation of f either.

Let us now show that $\langle \cdot, \cdot \rangle$ satisfies the properties of an inner product; we begin with additivity. By the previous calculation:

$$\begin{aligned}\langle f + h, g \rangle &= \sum_{j=1}^{m} \beta_j (f(x_j') + h(x_j')), \\ &= \sum_{j=1}^{m} \beta_j f(x_j') + \sum_{j=1}^{m} \beta_j h(x_j'), \\ &= \langle f, g \rangle + \langle h, g \rangle. \end{aligned}$$

It is also homogeneous since for $a \in \mathbb{R}$:

$$\langle af, g\rangle = \sum_{j=1}^{m} \beta_j (af(x_j')) = a \sum_{j=1}^{m} \beta_j f(x_j') = a\langle f, g\rangle.$$

Additionally it is symmetric since k is symmetric:

$$\langle f, g\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x_j') = \sum_{j=1}^{m}\sum_{i=1}^{n} \beta_j \alpha_i k(x_j', x_i) = \langle g, f\rangle.$$

The function $\langle \cdot, \cdot \rangle$ is nonnegative because k is a positive semi-definite kernel:

$$\langle f, f\rangle = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

This property, along with additivity and homogeneity, implies that the kernel $\rho(f, g) = \langle f, g\rangle$, defined on the image of Φ, is a positive semidefinite kernel. Indeed, for any $\gamma_1, \ldots, \gamma_n \in \mathbb{R}$ and $f_1, \ldots, f_n \in \text{image}(\Phi)$,

$$\sum_{i,j=1}^{n} \gamma_i \gamma_j \rho(f_i, f_j) = \sum_{i,j=1}^{n} \gamma_i \gamma_j \langle f_i, f_j\rangle = \left\langle \sum_{i=1}^{n} \gamma_i f_i, \sum_{j=1}^{n} \gamma_j f_j \right\rangle \geq 0.$$

To show that $\langle \cdot, \cdot \rangle$ is strictly positive, we observe that using (21) gives:

$$\langle k(\cdot, x), f\rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x). \tag{22}$$

This is a remarkable property! It is why these positive semidefinite kernels are also called *reproducing kernels*. Notice it implies:

$$\langle k(\cdot, x), k(\cdot, x')\rangle = k(x, x'). \tag{23}$$

Using (22) and Proposition 1 (kernel version of Cauchy-Schwarz) applied to the kernel $h(f, g) = \langle f, g\rangle$, we get:

$$\begin{aligned} |f(x)|^2 &= |\langle k(\cdot, x), f\rangle|^2, \\ &\leq \langle k(\cdot, x), k(\cdot, x)\rangle \langle f, f\rangle, \\ &= k(x, x)\langle f, f\rangle. \end{aligned}$$

Thus $\langle f, f \rangle = 0$ clearly implies $f(x) = 0$ for all $x \in \mathcal{X}$, and so at last we have proven that $\langle \cdot, \cdot \rangle$ is an inner product!

Since we defined $\Phi(x) = k(\cdot, x)$, in light of (23) we have:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle. \tag{24}$$

Therefore, the inner product space $(\text{image}(\Phi), \langle \cdot, \cdot \rangle)$ defines a "feature space" for the kernel k, in which evaluation of $k(x, x')$ corresponds to computing an inner product between $\Phi(x)$ and $\Phi(x')$.

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.