Lecture 8

Exercises

Exercise 15. A *binary relation* R on $\mathcal{X}$ is a set of ordered pairs $(x, x') \in R \subset \mathcal{X} \times \mathcal{X}$. We often write $x \sim x'$ if $(x, x') \in R$. A binary relation is an *equivalence relation* if:

1. $x \sim x$.
2. $x \sim x' \iff x' \sim x$.
3. If $x \sim x'$ and $x' \sim x''$, then $x \sim x''$.

An example of an equivalence relation is the following. Let $\mathcal{X}$ be the set of all people in the world. Then for two people x and x', $x \sim x'$ if x and x' have the same birthday. Here is an example of a binary relation that is not an equivalence relation. Let $\mathcal{X} = \mathbb{R}$ and say $x \sim x'$ if $x \geq x'$. Then $2 \sim 1$ but $1 \nsim 2$.

The point of this exercise is to consider equivalence relations as kernels. To that end, consider a kernel $k : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ such that

$$\forall x \in \mathcal{X}, \quad k(x, x) = 1.$$

Prove that k is a positive semidefinite kernel if and only if it satisfies the following two properties:

$$\begin{aligned} \forall\, x, x' \in \mathcal{X}, &\quad k(x, x') = 1 \iff k(x', x) = 1, \\ \forall\, x, x', x'' \in \mathcal{X}, &\quad k(x, x') = k(x', x'') = 1 \implies k(x, x'') = 1. \end{aligned}$$

Exercise 16. Find examples of equivalence relations that lend themselves to an interpretation as similarity measures. Discuss whether there are other binary relations (that are not equivalence relations) that one might want to use as similarity measures.

5.3.3 Starting with a map Φ into an inner product space

In the previous section, we started with a positive semidefinite kernel k and showed that it induces a map Φ into an inner product space such that

$k(x,x') = \langle \Phi(x), \Phi(x') \rangle$. The other direction also holds! Indeed, suppose $\Phi : \mathcal{X} \to (V, \langle \cdot, \cdot \rangle)$ maps $\mathcal{X}$ into an inner product space. Then we can *define* a kernel k as $k(x,x') = \langle \Phi(x), \Phi(x') \rangle$, and this kernel is positive semidefinite since:

$$\begin{aligned}\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i,j=1}^{n} c_i c_j \langle \Phi(x_i), \Phi(x_j) \rangle, \\ &= \left\langle \sum_{i=1}^{n} c_i \Phi(x_i), \sum_{j=1}^{n} c_j \Phi(x_j) \right\rangle, \\ &= \left\| \sum_{i=1}^{n} c_i \Phi(x_i) \right\|^2 \geq 0.\end{aligned}$$

Thus we have proved:

Theorem 2. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semidefinite kernel. Then there exists an inner product space V and a map $\Phi : \mathcal{X} \to V$ such that*

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

Conversely, suppose $\Phi : \mathcal{X} \to V$ is a map from $\mathcal{X}$ into an inner product space V. Then the kernel $k(x,x') = \langle \Phi(x), \Phi(x') \rangle$ is positive semidefinite.

Theorem 2 is the basis for the *kernel trick*, which states that given an algorithm which is formulated in terms of an inner product $\langle \cdot, \cdot \rangle$, one can construct an alternative algorithm by replacing $\langle \cdot, \cdot \rangle$ with a positive definite kernel k. See [1, pages 34–35] for some interesting historical remarks regarding the kernel trick.

5.4 Reproducing kernel Hilbert spaces

Adapted from [1, Chapter 2.2.3]

In the previous section, we showed that positive semidefinite kernels k implicitly define a nonlinear map $\Phi : \mathcal{X} \to V$, where V is the inner product space of functions f taking the form (20), and the inner product $\langle \cdot, \cdot \rangle$ was defined by (21). The reproducing kernel Hilbert space (RKHS) associated to k is the completion of V with respect to the norm $\|f\| = \sqrt{\langle f, f \rangle}$:

$$\mathcal{H} = \overline{V}.$$

More precisely, let $\mathcal{X}$ be a nonempty set and let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{H}$ is a *reproducing kernel Hilbert space (RKHS)* if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:

1. k has the reproducing property:

$$\forall\, f \in \mathcal{H}, \quad \langle f, k(x, \cdot)\rangle = f(x). \tag{26}$$

2. k spans $\mathcal{H}$, i.e.,

$$\mathcal{H} = \overline{\operatorname{span}\{k(x, \cdot) : x \in \mathcal{X}\}}.$$

Note that the reproducing property (26) implies

$$k(x, x') = \langle k(x, \cdot), k(x', \cdot)\rangle,$$

from which it follows that k is a positive semidefinite kernel.

For those who have taken functional analysis, we remark that an equivalent definition of a RKHS is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that all evaluation functionals $T_x : \mathcal{H} \to \mathbb{R}$, $T_x f = f(x)$ are continous. In that case one can apply the Riesz representation theorem to infer the existence of k.

Exercises

Exercise 17. Prove that the kernel k of a RKHS is unique.

5.5 Mercer's Theorem

Adapted from [1, Chapter 2.2.4]

Previously, we showed that positive semidefinite kernels k give rise to an RKHS $\mathcal{H}$, and furthermore that the map $\Phi : \mathcal{X} \to \mathcal{H}$, defined as $\Phi(x) = k(\cdot, x)$, yield $k(x, x') = \langle \Phi(x), \Phi(x')\rangle$. In this section we define another Φ which maps $\mathcal{X}$ into ℓ^2.

Suppose now that $\mathcal{X} \subset \mathbb{R}^p$ is a compact set (that is, closed and bounded). Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric, bounded kernel, i.e., $k(x, x') = k(x', x)$ and $|k(x, x')| \leq C < \infty$ for all $x, x' \in \mathcal{X}$ (can be relaxed

to almost every $x \in \mathcal{X}$ if you know what this means). Define the *kernel integral operator* associated to k as:

$$T_k : \mathbf{L}^2(\mathcal{X}) \to \mathbf{L}^2(\mathcal{X}),$$
$$f(x) \mapsto T_k f(x) = \int_{\mathcal{X}} k(x, x') f(x')\, dx'.$$

We say that T_k is *positive semidefinite* if:

$$\forall\, f \in \mathbf{L}^2(\mathcal{X}), \quad \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') f(x) f(x')\, dx\, dx' \geq 0.$$

Using these notations we can then state the following theorem:

Theorem 3 (Mercer's Theorem). *Suppose $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous, symmetric kernel and $\mathcal{X} \subset \mathbb{R}^p$ is a compact set. If T_k is positive semidefinite, then T_k has a countable set of orthonormal eigenfunctions $\{\psi_j\}_{j=1}^N \subset \mathbf{L}^2(\mathcal{X})$ with associated eigenvalues $\{\lambda_j\}_{j=1}^N \subset (0, \infty)$, where $N \in \mathbb{N} \cup \{\infty\}$ and $\lambda_1 \geq \lambda_2 \geq \cdots$. Additionally:*

- $(\lambda_j)_{j=1}^N \in \ell^1$.
- *The kernel k can be written as:*

$$k(x, x') = \sum_{j=1}^{N} \lambda_j \psi_j(x) \psi_j(x'). \tag{27}$$

If $N = \infty$, the series converges absolutely and uniformly, i.e.,

$$\lim_{m \to \infty} \sup_{(x, x') \in \mathcal{X} \times \mathcal{X}} \left| k(x, x') - \sum_{j=1}^{m} \lambda_j \psi_j(x) \psi_j(x') \right| = 0. \tag{28}$$

Remark 1. Mercer's Theorem can be generalized so that k does not have to be continuous, and $\mathcal{X}$ can be any measure space $(\mathcal{X}, \mu)$ for which $\mu(\mathcal{X}) < \infty$. In this case, (27) holds for almost every $(x, x') \in \mathcal{X} \times \mathcal{X}$, and in (28) one must replace the sup with an ess sup.

We can define a feature map Φ as:

$$\Phi : \mathcal{X} \to \ell^2,$$
$$x \mapsto \Phi(x) = \left(\sqrt{\lambda_j} \psi_j(x) \right)_{j=1}^{N}. \tag{29}$$

Then by (27) we have

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

Note that if $N < \infty$, then $\Phi(x) \in \mathbb{R}^N$. On the other hand, if $N = \infty$, then by (28) we can approximate k to within any accuracy $\epsilon > 0$ with a finite number of terms m depending upon ϵ. That is, for each $\epsilon > 0$, there exists $m \in \mathbb{N}$ such that (for almost every (x, x') if k is not continuous),

$$\left| k(x, x') - \sum_{j=1}^{m} \lambda_j \psi_j(x) \psi_j(x') \right| < \epsilon.$$

This means, though, that we can define a finite dimensional feature map Φ^m as:

$$\begin{aligned} \Phi^m : \mathcal{X} &\to \mathbb{R}^m, \\ x &\mapsto \left(\sqrt{\lambda_j} \psi_j(x) \right)_{j=1}^{m}, \end{aligned}$$

such that (again, for almost every (x, x') if k is not continous),

$$|k(x, x') - \langle \Phi^m(x), \Phi^m(x') \rangle| < \epsilon.$$

In other words, if we compute Φ^m explicitly, we know that this embedding preserves the kernel similarity measure up to a small error; we'll come back to this idea later in manifold learning.

Kernels satisfying the conditions of Mercer's Theorem are called *Mercer kernels*. Mercer kernels are positive semidefinite by Theorem 2. Thus a Mercer kernel is also the kernel of a RKHS.

Let us see how the reproducing kernel map $x \mapsto k(\cdot, x)$ is related to the Mercer map (29). Let k be a Mercer kernel, and let us construct the RKHS $\mathcal{H}$ associated to k (which we know exists). By the definition of RKHS, it must contain all functions $f : \mathcal{X} \to \mathbb{R}$ of the form:

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

But by Mercer's Theorem,

$$f(x) = \sum_{i=1}^{\infty} \sum_{j=1}^{N} \lambda_j \psi_j(x) \psi_j(x_i).$$

Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be a candidate inner product for $\mathcal{H}$. We know that it must satisfy $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$. We claim that it is sufficient that

$$\langle \psi_j, \psi_\ell \rangle_{\mathcal{H}} = \frac{\delta(j-\ell)}{\lambda_j}.$$

Indeed, since $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ must satisfy the conditions of an inner product, we have:

$$\begin{aligned}
\langle f, k(\cdot, x) \rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{\infty} \alpha_i \sum_{j=1}^{N} \lambda_j \psi_j(\cdot) \psi_j(x_i), \sum_{\ell=1}^{N} \lambda_\ell \psi_\ell(\cdot) \psi_\ell(x) \right\rangle_{\mathcal{H}}, \\
&= \sum_{i=1}^{\infty} \alpha_i \sum_{j,\ell=1}^{N} \lambda_j \lambda_\ell \psi_j(x_i) \langle \psi_j, \psi_\ell \rangle_{\mathcal{H}} \psi_\ell(x), \\
&= \sum_{i=1}^{\infty} \sum_{j=1}^{N} \lambda_j \psi_j(x_i) \psi_j(x), \\
&= f(x).
\end{aligned}$$

For a list of common kernels see [1, Chapter 2.3].

Exercises

Exercise 18. Suppose k is a Mercer kernel. For $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$, let w be defined as:

$$w = \sum_{i=1}^{n} \alpha_i \Phi(x_i),$$

where Φ is either the RKHS feature map $\Phi(x) = k(\cdot, x)$ or the Mercer map (29). Show that the value of $\|w\|^2$ is the same regardless of which map you pick, where the norm is taken in the appropriate Hilbert space depending upon Φ.

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.