EECS 598: Statistical Learning Theory, Winter 2014 **Topic 11**

Kernels

Lecturer: Clayton Scott *Scribe: Jun Guo, Soumik Chatterjee*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

These notes will introduce kernels which are the building blocks of modern kernel methods in machine learning.

2 Review of Hilbert Spaces

Definition 1. *A real* inner product space *(IPS) is a pair* $(V, \langle\cdot,\cdot\rangle)$*, where* V *is a real vector space and* $\langle\cdot,\cdot\rangle\colon\ V \times V \to \mathbb{R}$ *satisfies*

(a) $\langle u,u\rangle \geqslant 0, \forall u \in V$ *and* $\langle u,u\rangle = 0 \implies u = 0$

(b) $\langle u,v\rangle = \langle v,u\rangle, \forall u,v \in V$.

(c) $\forall a_1, a_2 \in \mathbb{R}, u_1, u_2, v \in V$, $\langle a_1u_1 + a_2u_2, v\rangle = a_1\langle u_1,v\rangle + a_2\langle u_2,v\rangle$.

Inner product spaces satisfy the Cauchy-Schwarz inequality:

Proposition 1. *Suppose* $(V, \langle\cdot,\cdot\rangle)$ *is an IPS. Then* $\forall u,v \in V$,

$$|\langle u,v\rangle| \leqslant \|u\|\,\|v\|\,.$$

where $\|u\| \triangleq \sqrt{\langle u,u\rangle}$.

Proof. Consider

$$G = \begin{bmatrix} \langle u,u\rangle & \langle u,v\rangle \\ \langle v,u\rangle & \langle v,v\rangle \end{bmatrix}.$$

Now

$$\begin{aligned} \text{G is a Gram matrix} &\Rightarrow \text{G is PSD} \\ &\Rightarrow \det(G) \geqslant 0 \\ &\Rightarrow \|u\|^2 \cdot \|v\|^2 - \langle u,v\rangle^2 \geqslant 0 \\ &\Rightarrow |\langle u,v\rangle| \leqslant \|u\|\,\|v\|\,. \end{aligned}$$

To show that G is a PSD matrix, observe that $\forall \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2$,

$$\begin{aligned} \begin{bmatrix} a & b \end{bmatrix} G \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} a\langle u,u\rangle + b\langle u,v\rangle \\ a\langle v,u\rangle + b\langle v,v\rangle \end{bmatrix} \\ &= \langle au+bv, au+bv\rangle \\ &\geqslant 0. \end{aligned}$$

□

Observation: The proof of the Cauchy-Schwarz inequality used all properties of inner products except definiteness: $\langle u,u\rangle = 0 \implies u = 0$. This will be important below.

Definition 2. *A metric space is a pair (M,d) where M is a set and $d : M \times M \to \mathbb{R}$ satisfies:*

(a) $d(x,y) \geqslant 0\ \forall x,y \in M$ *and* $d(x,y) = 0 \iff x = y$

(b) $\forall x,y \in M, d(x,y) = d(y,x)$

(c) $\forall x,y,z \in M, d(x,z) \leqslant d(x,y) + d(y,z)$.

An IPS is a metric space with induced metric $d(u,v) := \|u-v\| := \sqrt{\langle u-v, u-v\rangle}$. Verification of (a) and (b) is straightforward. To verify (c), it suffices to show

$$\forall u,v \in V, \quad \|u+v\| \leqslant \|u\| + \|v\|, \tag{1}$$

for then $d(x,z) = \|x-z\| = \|x-y+y-z\| \leqslant \|x-y\| + \|y-z\|$. To see (1), observe

$$\begin{aligned}(\|w\| - \|v\|)^2 &= \|w\|^2 - 2\|w\|\|v\| + \|v\|^2 \\ &\leqslant \|w\|^2 - 2\langle w,v\rangle + \|v\|^2 \\ &= \langle w-v, w-v\rangle \\ &= \|w-v\|^2 .\end{aligned}$$

So,

$$|\|u\| - \|v\|| \leqslant \|w-v\| .$$

Letting $w - v = u$, we have $\|u+v\| \leqslant \|u\| + \|v\|$.

Metric spaces allow us to study convergence sequences and continuity. We make the following definitions.

Definition 3. *Let (M,d) be a metric space.*

- *We say a sequence (x_n) converges iff $\exists x^* \in M$ such that $d(x_n,x^*) \to 0$ as $n \to \infty$.*
- *Suppose $(\tilde{M},\tilde{d})$ is another metric space, and $f : M \to \tilde{M}$. We say f* is continuous at *$x \in M$ iff $\forall$ sequences (x_n) in M converging to x, the sequence $(f(x_n))$ converges to $f(x)$ in $\tilde{M}$. We say f is* continuous *if it is continuous at all $x \in M$.*
- *A sequence (x_n) is a* Cauchy sequence *iff $\forall \epsilon > 0$, $\exists N \in \mathbb{N}\,(m,n \geqslant N \implies d(x_m,x_n) < \epsilon)$.*
- *A metric space (M,d) is said to be* complete *iff every Cauchy sequence of points in M converges to a point in M.*

Remark. Every convergent sequence is a Cauchy sequence, but not conversely. This is easy to check.

Inner products are continuous in one argument when the other argument is held fixed.

Lemma 1. *Let $(V,\langle\cdot\,,\cdot\rangle)$ be an IPS. Suppose $u_n \to u \in V$, i.e., $\lim_{n\to\infty}\|u_n - u\| = 0$, and let $v \in V$. Then,*

$$\lim_{n\to\infty}\langle u_n, v\rangle = \langle u,v\rangle,$$

i.e., the inner product is continuous in its first argument.

Proof. We need to show:

$$\lim_{n\to\infty}|\langle u_n,v\rangle - \langle u,v\rangle| = 0.$$

This follows from Cauchy-Schwarz: $|\langle u_n,v\rangle - \langle u,v\rangle| = |\langle u_n - u, v\rangle| \leqslant \|u_n - u\| \cdot \|v\| \to 0$, as $n \to \infty$, since $\|u_n - u\| \to 0$. ∎

Every metric space (M, d) has a completion (M^*, d^*), which is a complete metric space for which there exists $\varphi : M \to M^*$ satisfying:

(a) $\varphi(M)$ is dense in M^*.

(b) φ is an isometry, i.e., φ is injective and $\forall x, y \in M, d^*(\varphi(x), \varphi(y)) = d(x, y)$.

Furthermore, any two completions of (M, d) are isometric, i.e., there exists a bijective, distance preserving map between them. There is a standard construction of a completion. It is well worth understanding [1].

In the standard completion, it is not true that $M \subset M^*$, since M^* consists of equivalence classes of Cauchy sequences of points in M. However, in many cases there is a completion such that $M \subseteq M^*$.

Example. Consider $(\mathbb{Q}, |\cdot|)$, the rational numbers with the natural metric. This is not complete since one can take a sequence of decimal expansions of some irrational number, e.g., $3, 3.1, 3.14, 3.141, 3.1419, \ldots$. Such a sequence is Cauchy, but does not coverge to a point in $\mathbb{Q}$. $(\mathbb{R}, |\cdot|)$ is a completion of $(\mathbb{Q}, |\cdot|)$, and in this case φ is just the identity map.

Definition 4. *A* Hilbert space *is a complete inner product space, where completeness is with respect to the induced metric.*

Hilbert spaces satisfy nice properties like the projection theorem and the Riesz representation theorem. We'll use these later.

3 Kernels

Let $\mathcal{X}$ be a set.

Definition 5. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We say k is* symmetric *if and only if $k(x, x') = k(x', x), \forall x, x' \in \mathcal{X}$. We say k is* positive definite *(PD) if and only if $\forall n \in \mathbb{N}, \forall x_1, ..., x_n \in \mathcal{X}$, the $n \times n$ matrix $[k(x_i, x_j)]_{i,j=1}^n$ is positive semi-definite (PSD).*

Theorem 1. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The following are equivalent:*

(a) k is positive definite and symmetric.

(b) $\exists$ a Hilbert space $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ and a function $\Phi : \mathcal{X} \to \mathcal{F}$ s.t. $\forall x, x' \in \mathcal{X}, k(x, x') = \langle \Phi(x), \Phi(x') \rangle$

Definition 6. *If k satisfies the conditions of the previous theorem, we say k is a* kernel *on $\mathcal{X}$.*

Remarks and terminology. Kernels are some times called positive definite symmetric kernels or inner product kernels depending on which condition is being emphasized. Φ is called a feature map and $\mathcal{F}$ is called the feature space (we'll refer to $\mathcal{X}$ as the input space to avoid confusion). The feature map/space is not unique.

Proof of Theorem 1. ((b) $\Rightarrow$ (a)): Suppose (b) holds. Clearly k is symmetric because $\langle \cdot, \cdot \rangle$ is. Furthermore, for $x_1, ..., x_n \in \mathcal{X}$, $[k(x_i, x_j)]_{i,j=1}^n$=$[\langle \Phi(x_i), \Phi(x_j) \rangle]_{i,j=1}^n$, which is a Gram matrix, and therefore PSD. So k is PD.

((a) $\Rightarrow$ (b)): Assume k is positive definite and symmetric. Define

$$\mathcal{F}_0 = \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \;\middle|\; n \in \mathbb{N}, x_1, ..., x_n \in \mathcal{X}, \alpha_1, ..., \alpha_n \in \mathbb{R} \right\}.$$

For $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^m \beta_j k(\cdot, x'_j)$ in $\mathcal{F}_0$, define $\langle f, g \rangle = \sum_{(i,j)} \alpha_i \beta_j k(x_i, x'_j)$.

To check that this inner product is well-defined, i.e., independent of the particular representations of f and g, observe

$$\begin{aligned}\sum_{(i,j)} \alpha_i \beta_j k(x_i, x'_j) &= \sum_{j=1}^{m} \beta_j f(x'_j) \\ &= \sum_{i=1}^{n} \alpha_i g(x_i)\end{aligned}$$

So $\langle \cdot\,, \cdot \rangle$ only depends on the functions f and g, and not on their representations.

Next, we show that the properties of an inner product are satisfied.

- Symmetry: $\langle f, g\rangle = \sum_{i,j} \alpha_i \beta_j k(x_i, x'_j) = \sum_{j,i} \beta_j \alpha_i k(x'_j, x_i) = \langle g, f\rangle$, by symmetry of k.
- Linearity in first variable: Also an easy exercise.
- Nonnegativity: $\langle f, f\rangle = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \geqslant 0$, because k is PD.

Therefore the Cauchy-Schwarz inequality holds. To establish definiteness, suppose $\langle f, f\rangle = 0$. Then $\forall x \in \mathcal{X}$,

$$\begin{aligned}|f(x)| &= \left|\sum_{i=1}^{n} \alpha_i k(x_i, x)\right|, \\ &= |\langle f, k(\cdot\,, x)\rangle|\,, && \text{(since } k(\cdot\,, x) \in \mathcal{F}_0\text{)} \\ &\leqslant \sqrt{\langle f, f\rangle \cdot \langle k(\cdot\,, x), k(\cdot\,, x)\rangle} && \text{(Cauchy-Schwarz)} \\ &= 0.\end{aligned}$$

Thus $f(x) = 0, \forall x \in \mathcal{X}$, i.e., f is the zero function. The axioms of an inner product are established.

Let $(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mathcal{F}})$ be a completion of $\mathcal{F}_0$ with $\varphi : \mathcal{F}_0 \to \mathcal{F}$ the completion isometry. Define $\Phi : \mathcal{X} \to \mathcal{F}$ by $\Phi(x) = \varphi(k(\cdot\,, x))$. Then $\langle \Phi(x), \Phi(x')\rangle_{\mathcal{F}} = \langle k(\cdot\,, x), k(\cdot\,, x')\rangle_{\mathcal{F}_0} = k(x, x')$. This completes the proof. □

4 Examples and properties of kernels

Let $(\mathcal{F}, \langle\cdot,\cdot\rangle)$ be any Hilbert space. Then $k(x, x') = \langle x, x'\rangle$ is a kernel on $\mathcal{F}$, which can be seen by taking Φ to be the identity map.

Example. $k(x, x') = x^T x'$ (dot product) is a kernel in $\mathbb{R}^d$.

Lemma 2. *If k_1, k_2 are kernels on $\mathcal{X}$ and $\alpha \geq 0$, then αk and $k_1 + k_2$ are kernels.*

Proof. The proof is left as an exercise. □

Lemma 3. *If k_1 is a kernel on $\mathcal{X}_1$ and k_2 is a kernel on $\mathcal{X}_2$, then $k_1 \cdot k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$, where $(k_1 \cdot k_2)((x_1, x_2), (x'_1, x'_2)) := k_1(x_1, x_2)k_2(x'_1, x'_2)$*

Proof sketch. Denote $k = k_1 \cdot k_2$. Let $\mathcal{F}_i, \Phi_i$ be Hilbert spaces and the feature maps corresponding to $k_i, i = 1, 2$. Let $\mathcal{F}_1 \otimes \mathcal{F}_2$ be the tensor product of $\mathcal{F}_1$ and $\mathcal{F}_2$. This is an inner product space whose underlying set is $\mathcal{F}_1 \times \mathcal{F}_2$, but the vector space structure is different from the usual one given by the direct sum. The inner product on the tensor product is such that $\langle (f_1, f_2), (f'_1, f'_2)\rangle_{\mathcal{F}_1 \otimes \mathcal{F}_2} = \langle f_1, f'_1\rangle_{\mathcal{F}_1} \langle f_2, f'_2\rangle_{\mathcal{F}_2}$. Let $\mathcal{F}$ be a completion of $\mathcal{F}_1 \otimes \mathcal{F}_2$, and $\varphi : \mathcal{F}_1 \times \mathcal{F}_2 \to \mathcal{F}$ an associated completion embedding. Define

$\Phi : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{F}$ by $\Phi(x_1, x_2) = \varphi(\Phi_1(x_1), \Phi_2(x_2))$. Denoting $x = (x_1, x_2), x' = (x'_1, x'_2)$,

$$\begin{aligned} k(x, x') &= k_1(x_1, x'_1)k_2(x_2, x'_2) \\ &= \langle \Phi_1(x_1), \Phi_1(x'_1)\rangle_{\mathcal{F}_1} \langle \Phi_2(x_2), \Phi_2(x'_2)\rangle_{\mathcal{F}_2} \\ &= \langle (\Phi_1(x_1), \Phi_2(x_2)), (\Phi_1(x'_1), \Phi_2(x'_2))\rangle_{\mathcal{F}_1 \otimes \mathcal{F}_2} \\ &= \langle \varphi(\Phi_1(x_1), \Phi_2(x_2)), \varphi(\Phi_1(x'_1), \Phi_2(x'_2))\rangle_{\mathcal{F}} \\ &= \langle \Phi(x_1, x_2), \Phi(x'_1, x'_2)\rangle_{\mathcal{F}} \\ &= \langle \Phi(x), \Phi(x')\rangle_{\mathcal{F}} \end{aligned}$$

□

Lemma 4. *If $A : \tilde{\mathcal{X}} \to \mathcal{X}$ and k is a kernel on $\mathcal{X}$, then $\tilde{k}$ defined by $\tilde{k}(x, x') = k(A(x), A(x'))$ is a kernel on $\tilde{\mathcal{X}}$.*

Proof. The proof is left as an exercise. □

Corollary 1. *If k is a kernel on $\mathcal{X}$, then so is k^2 where $k^2(x, x') := k(x, x')^2$.*

Proof. By Lemma 2, $k \cdot k$ is a kernel on $\tilde{\mathcal{X}} \times \tilde{\mathcal{X}}$. Let $A : \mathcal{X} \to \mathcal{X} \times \mathcal{X}$ be defined by $A(x) = (x, x)$. By Lemma 3, $(k \cdot k)(A(x), A(x'))$ defines a kernel on $\mathcal{X}$ as

$$\begin{aligned} (k \cdot k)(A(x), A(x')) &= (k \cdot k)((x, x), (x', x')) \\ &= k(x, x')k(x, x') \\ &= k(x, x')^2 \end{aligned}$$

□

Example. Let $p(t) = \sum_{j=1}^{n} \alpha_j t^j,\ \alpha_j \geq 0$. Then $k(x, x') = p(\langle x, x'\rangle)$ is a kernel on $\mathbb{R}^d$. Taking $p(t) = (t+c)^m, c \geq 0$ gives the inhomogeneous polynomial kernel. An explicit feature map for $k(x, x') = (\langle x, x'\rangle + c)^m$ can be derived via

$$\begin{aligned} k(x, x') &= (\langle x, x'\rangle + c)^m \\ &= (x_1 x'_1 + x_2 x'_2 + \cdots + x_d x'_d + c)^m \\ &= \sum_{(j_1, \cdots, j_d) \in J(m)} \binom{m}{j_1, \cdots, j_d} (x_1 x'_1)^{j_1} \cdots (x_d x'_d)^{j_d} c^{m - \sum j_i} \\ &= \langle \Phi(x), \Phi(x')\rangle \end{aligned}$$

where $J(m) = \{(j_1, \cdots, j_d) \mid j_i \geq 0, \sum j_i \leq m\}$ and $\Phi(x)$ is the finite dimensional vector

$$\Phi(x) = \left(\sqrt{\binom{m}{j_1, \cdots, j_d} c^{m - \sum j_i}} x_1^{j_1} \cdots x_d^{j_d} c^{m - \sum j_i} x_1^{j_1} \cdots x_d^{j_d} \right)_{(j_1, \cdots, j_d) \in J(m)}.$$

Lemma 5. *Suppose the series $f(z) = \sum_{n=0}^{\infty} a_n z^n$ converges for $z \in (-r, r)$ where $r \in (0, \infty]$. If $a_n \geq 0\ \forall n$, then*

$$k(x, x') := \sum_{n=0}^{\infty} a_n \langle x, x'\rangle^n$$

is a kernel on $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\|_2 < \sqrt{r}\}$

Example. For an $\beta > 0$, $e^{\beta z} = \sum_{n=0}^{\infty} a_n z^n$ with $a_n = \frac{\beta^n}{n!}$, and this holds for all $z \in \mathbb{R}$ $(r = \infty)$. Therefore, $e^{\beta\langle x,x'\rangle}$ is a kernel on $\mathbb{R}^d$.

Proof. If $\|x\| < \sqrt{r}, \|x'\| < \sqrt{r}$ then $|\langle x, x'\rangle| < \|x\|\|x'\| < r$ and so $\sum_{n=0}^{\infty} a_n \langle x, x'\rangle^n$ converges, meaning k is well-defined. Since $f(z)$ converges on $(-r, r)$, it converges absolutely. Hence we can rearrange terms without affecting the limit.

$$\begin{aligned}
k(x, x') &= \sum_{n=0}^{\infty} a_n \langle x, x'\rangle^n \\
&= \sum_{n=0}^{\infty} a_n (x_1 x_1' + \cdots + x_d x_d')^n \\
&= \sum_{n=0}^{\infty} a_n \sum_{(j_1,\cdots,j_d):j_i\geq 0, \sum j_i = n} \binom{n}{j_1, \cdots, j_d} \prod_{i=1}^{d} (x_i x_i')^{j_i} \\
&= \sum_{(j_1,\cdots,j_d):j_i\geq 0} a_{j_1+\cdots+j_d} \frac{(\sum j_i)!}{\prod j_i!} \prod_{i=1}^{d} (x_i x_i')^{j_i} \\
&= \langle \Phi(x), \Phi(x')\rangle
\end{aligned}$$

where

$$\Phi(x) = \left(\sqrt{a_{j_1+\cdots+j_d} \frac{(\sum j_i)!}{\prod j_i!}} \prod x_i^{j_i} \right)_{(j_1,\cdots,j_d):j_i\geq 0}$$

and the Hilbert space is $\ell_2 := \{(c_1, c_2, \cdots) | \sum c_i^2 < \infty\}$. □

Lemma 6 (Normalized kernels)**.** *If k is a kernel on $\mathcal{X}$ such that $k(x, x) > 0$ for all x, then $\tilde{k}$ is also a kernel on $\mathcal{X}$, where $\tilde{k}(x, x') := \frac{k(x,x')}{\sqrt{k(x,x)k(x'.x')}}$.*

You are asked to remove the positivity condition in an exercise.

Proof. Let $\Phi : \mathcal{X} \to \mathcal{F}$ be a feature map for k. Define $\tilde{\Phi}(x) = \frac{\Phi(x)}{\|\Phi(x)\|}$. Then

$$\begin{aligned}
\tilde{k}(x, x') &= \frac{\langle \Phi(x), \Phi(x')\rangle}{\sqrt{\langle \Phi(x), \Phi(x)\rangle \langle \Phi(x'), \Phi(x')\rangle}} \\
&= \langle \tilde{\Phi}(x), \tilde{\Phi}(x')\rangle,
\end{aligned}$$

so $\tilde{k}$ is a kernel. □

Example. If $k(x, x') = e^{2\gamma\langle x,x'\rangle}, \gamma > 0$, then

$$\begin{aligned}
\tilde{k}(x, x') &= \frac{e^{2\gamma\langle x,x'\rangle}}{e^{\gamma\langle x,x\rangle} e^{\gamma\langle x',x'\rangle}} \\
&= e^{-\gamma(\|x\|^2 - 2\langle x,x'\rangle + \|x'\|^2)} \\
&= e^{-\gamma\|x-x'\|^2}
\end{aligned}$$

is the well known Gaussian kernel.

The above arguments emphasize the inner product characterization of kernels. For proofs of the above properties that leverage the positive definite symmetric characterization, see [2]. A good reference for additional theoretical properties of kernels is [4].

Exercises

1. Verify the linearity in the first argument axiom for the inner product defined in the proof of Theorem 1.

2. Properties of kernels

   (a) Prove Lemma 2

   (b) Prove Lemma 4

   (c) Suppose (k_n) is a sequence of kernels on $\mathcal{X}$ that converges pointwise to the function k, i.e., for all $x, x' \in \mathcal{X}$, $\lim_{n\to\infty} k_n(x, x') = k(x, x')$. Show that k is a kernel on $\mathcal{X}$.

   (d) Use the previous property to generalize Lemma 5 by replacing the dot product with an arbitrary kernel. Provide a revised theorem statement.

   (e) When we introduced normalized kernels above, we assumed $k(x, x') > 0 \, \forall x, x'$. Let's remove this restriction. Show that if k is a kernel, then so is

$$\tilde{k}(x, x') := \begin{cases} \frac{k(x,x')}{\sqrt{k(x,x)k(x',x')}}, & \text{if } k(x,x) > 0 \text{ and } k(x',x') > 0, \\ 0, & \text{otherwise.} \end{cases}$$

   (f) Give an alternate proof that the Gaussian kernel is a kernel without invoking normalized kernels.

3. A *radial kernel* on $\mathbb{R}^d$ is a kernel of the form $k(x, x') = g(\|x - x'\|)$ for some function g, where the norm is Euclidean norm.

   (a) Show that if $p \geq 0$, then

$$k(x, x') = \int_0^\infty e^{-u\|x-x'\|^2} p(u) du$$

   is a radial kernel. *Note:* It can be shown that if $g(\|x - x'\|)$ is a kernel for every dimension $d \geq 1$, then it must have the above form, where $p(u)du$ is generalized to $d\mu$ where μ is a finite measure [3].

   (b) (Laplacian kernel) Show that for $\alpha > 0$,

$$k(x, x') = e^{-\alpha\|x-x'\|}$$

   is a kernel. *Major hint:*

$$e^{-\alpha\sqrt{s}} = \int_0^\infty e^{-su} \frac{\alpha}{2\sqrt{\pi u^3}} e^{-\frac{\alpha^2}{4u}} du.$$

   (c) (Multivariate Student-type kernel) Show that for $\alpha, \beta > 0$,

$$k(x, x') = \left(1 + \frac{\|x - x'\|^2}{\beta}\right)^{-\alpha}$$

   is a kernel. *Hint:* Consider the MGF of a gamma distribution.

   (d) Show that

$$k(x, x') = e^{-\alpha\|x-x'\|_1}$$

   is a kernel, where $\|\cdot\|_1$ is the 1-norm. This kernel could be considered "radial" with respect to this alternative norm.

   The Laplacian and multivariate Student kernels are alternatives to the Gaussian kernel. They do not decay to zero as rapidly as the Gaussian kernel, and therefore are less likely to encounter numerical problems. Sometimes the entries of a Gaussian kernel matrix can be all zeros and ones, or the kernel matrix can be near-singular.

4. Learn about tensor products, including both the vector space and inner product space structure.

References

[1] http://www.math.columbia.edu/∼nironi/completion.pdf.

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, 2012.

[3] C. Scovel, D. Hush, I. Steinwart, and J. Theiler, "Radial kernels and their reproducing kernel Hilbert spaces," *J. Complexity*, vol. 26, pp. 641-660, 2010.

[4] I. Steinwart and A. Christman, *Support Vector Machines*, Springer, 2008.