ELE 535: Machine Learning and Pattern Recognition
Homework 9

Zachary Hervieux-Moore

Friday 21$^{\text{st}}$ December, 2018

Exercise 1: Kernels:

a) Let $\mathcal{A}$ be a finite set and for each subset $\mathcal{U} \subseteq \mathcal{A}$ let $|\mathcal{U}|$ denote the number of element in $\mathcal{U}$. For $\mathcal{U}, \mathcal{V} \subset \mathcal{A}$, let $k(\mathcal{U}, \mathcal{V}) = |\mathcal{U} \cap \mathcal{V}|$. By finding a suitable feature map, show that $k(\cdot, \cdot)$ is a kernel on the power set $\mathcal{P}(\mathcal{A})$ of all subsets of $\mathcal{A}$.

b) Show that $k(x, z) = \sum_{i=1}^{n} \cos^2(x_i - z_i)$ is a kernel on $\mathbb{R}^n$.

c) Let $P \in \mathbb{R}^{n \times n}$ be symmetric PSD. Show that $k(x, z) = e^{-\frac{1}{2}(x-z)^T P(x-z)}$ is a kernel on $\mathbb{R}^n$.

d) $k(x, y) = h_t(Ax)^T h_t(Ay)$ where $A \in \mathbb{R}^{n \times n}$ and h_t is a thresholding function that maps $z = [z_i]$ to $h_t(z) = [\tilde{z}_i]$ with $\tilde{z}_i = z_i$ if $|z_i| > t$ and 0 otherwise.

Answer:

a) Let us use the indicator $I_{\mathcal{U}}$ as our feature map $\phi(\mathcal{U})$ in the L^2 Hilbert space which has the norm

$$\int_{\mathcal{A}} f(a) g(a) d\mu(a)$$

Thus, our kernel becomes

$$\begin{aligned} k(\mathcal{U}, \mathcal{V}) &= \int_{\mathcal{A}} \phi(\mathcal{U}) \phi(\mathcal{V}) d\mu(a) \\ &= \int_{\mathcal{A}} I_{\mathcal{U}}(a) I_{\mathcal{V}}(a) d\mu(a) \\ &= |\mathcal{U} \cap \mathcal{V}| \end{aligned}$$

We conclude that the kernel is defined on the powerset $\mathcal{P}(\mathcal{A})$.

b) We make use of the trig identity,

$$\cos^2(x_i - z_i) = \cos^2(x_i)\cos^2(z_i) + \sin^2(x_i)\sin^2(z_i) + 2\cos(x_i)\sin(x_i)\cos(z_i)\sin(z_i)$$
$$= \begin{bmatrix}\cos^2(x_i) & \sin^2(x_i) & \sqrt{2}\cos(x_i)\sin(x_i)\end{bmatrix}\begin{bmatrix}\cos^2(z_i) \\ \sin^2(z_i) \\ \sqrt{2}\cos(z_i)\sin(z_i)\end{bmatrix}$$

Thus we use the feature map

$$\phi(x) = \begin{bmatrix}\cos^2(x_1) \\ \sin^2(x_1) \\ \sqrt{2}\cos(x_1)\sin(x_1) \\ \vdots \\ \cos^2(x_n) \\ \sin^2(x_n) \\ \sqrt{2}\cos(x_n)\sin(x_n)\end{bmatrix}$$

Which is in $\mathbb{R}^{3n}$ and hence a Hilbert space. This feature map coupled with the standard inner product yields our kernel.

c) We know that the Gaussian kernel is a valid kernel:

$$k_G(x, z) = e^{-\gamma\|x-z\|_2^2}$$

Now, using the fact that P is PSD and hence diagonalizable, say $P = U^T \Sigma U$, we can write the kernel in question as:

$$\begin{aligned}
k(x, z) &= e^{-\frac{1}{2}(x-z)^T P(x-z)} \\
&= e^{-\frac{1}{2}(x'-z')^T U^T P U(x'-z')} \\
&= e^{-\frac{1}{2}(x'-z')^T \Sigma (x'-z')} \\
&= e^{-\frac{\sigma_1}{2}\|x'-z'\|_2^2} \cdot \ldots \cdot e^{-\frac{\sigma_n}{2}\|x'-z'\|_2^2} \\
&= K_{G_1}(x', z') \cdot \ldots \cdot K_{G_n}(x', z')
\end{aligned}$$

We use the transformation $x' = Ux$. A product of kernels is a kernel and so we conclude that the original is a kernel.

d) We have that $h_t(Ax)$ is a composition of feature maps $\phi(x) = Ax$ and $\varphi(x) = h_t(x)$ where they both map from $\mathbb{R}^n$ to $\mathbb{R}^n$ and so this is the standard Euclidean Hilbert space. Thus our kernel is $k_{\phi\circ\varphi}(x, y) = h_t(Ax)^T h_t(Ay)$.

Exercise 2: Let k_j be a kernel on $\mathcal{X}$ with feature map $\phi_j : \mathcal{X} \to \mathbb{R}^q$, $j = 1, 2$. In each part below, find a simple feature map for the kernel k in terms of feature maps for the kernels k_j. By this means, give an interpretation for the new kernel k.

a) $k(x, z) = k_1(x, z) + k_2(x, z)$

b) $k(x, z) = k_1(x, z)k_2(x, z)$

c) $k(x, z) = k_1(x, z)/\sqrt{k_1(x, x)k_1(z, z)}$

Answer:

a) We have that

$$\begin{aligned} k(x, z) &= k_1(x, z) + k_2(x, z) \\ &= \langle \phi_1(x), \phi_1(z) \rangle + \langle \phi_2(x), \phi_2(z) \rangle \\ &= \left\langle \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \end{bmatrix}, \begin{bmatrix} \phi_1(z) \\ \phi_2(z) \end{bmatrix} \right\rangle \end{aligned}$$

So our new feature map is $\phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \end{bmatrix}$ thus adding kernels is equivalent to stacking in the feature space.

b) A product in kernel space yields,

$$\begin{aligned} k(x, z) &= k_1(x, z)k_2(x, z) \\ &= \langle \phi_1(x), \phi_1(z) \rangle \langle \phi_2(x), \phi_2(z) \rangle \\ &= \langle \phi_1(x) \otimes \phi_2(x), \phi_1(z) \otimes \phi_2(z) \rangle \end{aligned}$$

Where the last line is the trace operator on matrix multiplication; the standard inner product for matrices. So our new feature map is $\phi(x) = \phi_1(x) \otimes \phi_2(x)$ thus multiplying kernels is equivalent to tensor product in the feature space.

c) Simplifying the expression yields,

$$\begin{aligned} k(x, z) &= k_1(x, z)/\sqrt{k_1(x, x)k_1(z, z)} \\ &= \frac{\langle \phi_1(x), \phi_1(z) \rangle}{\|\phi_1(x)\| \|\phi_1(z)\|} \\ &= \left\langle \frac{\phi_1(x)}{\|\phi_1(x)\|}, \frac{\phi_1(z)}{\|\phi_1(z)\|} \right\rangle \end{aligned}$$

So our new feature map is $\phi(x) = \frac{\phi_1(x)}{\|\phi_1(x)\|}$ thus this transformation in kernel space is normalization in feature space.

Exercise 3: A binary labelled set of data in $\mathbb{R}^2$ is used to learn a SVM using the homogeneous quadratic kernel. By writing the equation for the decision boundary in terms of a quadratic form, reason about the types of decision boundaries that are possible in $\mathbb{R}^2$. In each case, give a neat sketch.

Answer: The kernel SVM problem is as follows,

$$\begin{gathered}
\max_{\alpha \in \mathbb{R}^m} \mathbf{1}^T \alpha - \frac{1}{2} \alpha^T K \alpha \\
y^T \alpha = \mathbf{0} \\
\alpha \leq C\mathbf{1} \\
\alpha \geq \mathbf{0}
\end{gathered}$$

Where K is the kernel matrix of $y_i \langle \phi(x_i), \phi(x_j) \rangle y_j$. In this case, our kernel is the quadratic kernel of $k(x, z) = (x^T z)^2$. After solving the above optimization problem, the decision boundary becomes

$$\begin{gathered}
\sum_{i \in A} \alpha_i^* y_i (x_i^T x)^2 + b^* \geq 0 \\
\sum_{i \in A} \alpha_i^* y_i x^T x_i x_i^T x + b^* \geq 0 \\
x^T \left(\sum_{i \in A} \alpha_i^* y_i x_i x_i^T \right) x + b \geq 0 \\
x^T Q x + b^* \geq 0
\end{gathered}$$

Thus, the decision boundaries if, $Q \neq 0$, are simply ellipses in $\mathbb{R}^2$. If $Q = 0$, then this is the degenerate case of assigning all entries to the sign of b^*.

Exercise 4: Consider the definite integral

$$\int_0^\infty e^{-(ax^2+\frac{b}{x^2})}dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}e^{-2\sqrt{ab}}$$

Show that it can be rewritten as

$$\frac{\alpha}{\sqrt{\pi}}\int_0^\infty e^{-\frac{s}{t^2}}e^{-(\frac{\alpha t}{2})^2}dt = e^{-\alpha\sqrt{s}}, \quad \alpha > 0$$

Now use the above results to show that $k(x,z) = e^{-\alpha\|x-z\|_2}$ is a kernel.

Answer: by letting $\alpha = 2\sqrt{a}$ and $s = b$, the change is immediate. Then, using this result, we have that

$$\begin{aligned}
k(x,z) &= e^{-\alpha\|x-z\|_2} \\
&= e^{-\alpha\sqrt{\|x-z\|_2^2}} \\
&= \frac{\alpha}{\sqrt{\pi}}\int_0^\infty e^{-\frac{\|x-z\|_2^2}{t^2}}e^{-(\frac{\alpha t}{2})^2}dt
\end{aligned}$$

Now, using Thereom 16.4.1 about smoothed kernels, we have that $p(t) = \frac{\alpha}{\sqrt{\pi}}e^{-(\frac{\alpha t}{2})^2}$ and $k(t,x,z) = e^{-\frac{\|x-z\|_2^2}{t^2}}$ which is the Gaussian kernel parameterized by $1/t^2$. As $p(t)$ is integrable on $[0,\infty)$, Theorem 16.4.1 applies. Thus, we conclude that $e^{-\alpha\|x-z\|_2}$ is a kernel.

Exercise 5: Let $a > 0$ and $L_2[0, a]$ denote the set of real valued square integrable functions on the interval $[0, a]$. $L_2[0, a]$ is a Hilbert space under the inner product $\langle g, h\rangle = \int_0^a g(s)h(s)ds$. Show that for each $a > 0$,

a) $k(x, z) = \min(x, z)$ is a kernel on $[0, a]$.

b) $k(x, z) = a - \max(x, z)$ is a kernel on $[0, a]$.

c) $k(x, z) = e^{-(\max(x,z)-\min(x,z))}$ is a kernel on $[0, a]$. Plot the function $\max(x, z) - \min(x, z)$ and use this to simplify the result further.

d) for each $a > 0$, and $\gamma \geq 0$, $k(x, z) = e^{-\gamma|x-z|}$ is a kernel on $[-a, a]$.

e) for each $\gamma \geq 0$, $k(x, z) = e^{-\gamma|x-z|}$ is a kernel on $\mathbb{R}$.

Answer:

a) Let $\phi(\cdot) = g_x(t) = \mathbf{1}_{\{t\leq x\}}$ be our feature map, then we get that the kernel is

$$\begin{aligned}
k(x, z) &= \langle g_x, g_z\rangle \\
&= \int_0^a \phi_x(t)\phi_z(t)dt \\
&= \int_0^a \mathbf{1}_{\{t\leq x\}}\mathbf{1}_{\{t\leq z\}}dt \\
&= \int_0^a \mathbf{1}_{\{t\leq x\cap t\leq z\}}dt \\
&= \int_0^a \mathbf{1}_{\{t\leq \min(x,z)\}}dt \\
&= \int_0^{\min(x,z)} dt \\
&= \min(x, z)
\end{aligned}$$

b) By changing the sign of the inequality in the indicator above, we get that the feature map is $\phi(\cdot) = g_x(t) = \mathbf{1}_{\{x \leq t\}}$ and the kernel becomes

$$\begin{aligned}
k(x,z) &= \langle g_x, g_z \rangle \\
&= \int_0^a \phi_x(t)\phi_z(t)dt \\
&= \int_0^a \mathbf{1}_{\{x \leq t\}}\mathbf{1}_{\{z \leq t\}}dt \\
&= \int_0^a \mathbf{1}_{\{x \leq t \cap z \leq t\}}dt \\
&= \int_0^a \mathbf{1}_{\{\max(x,z) \leq t\}}dt \\
&= \int_{\max(x,z)}^a dt \\
&= a - \max(x,z)
\end{aligned}$$

c) Simplifying the kernel, we get that

$$\begin{aligned}
k(x,z) &= e^{-(\max(x,z)-\min(x,z))} \\
&= e^{\min(x,z)} \cdot e^{-a} \cdot e^{a-\max(x,z)}
\end{aligned}$$

From parts a) and b), we get that $e^{\min(x,z)}$ and $e^{a-\max(x,z)}$ are kernels by 16.2.1. We also get that the above is the product of two kernels with a scalar. By the properties of kernels in 16.2.1, we have that the entire result is a kernel. We also have that $\max(x,z) - \min(x,z) = |x - z|$ (the plot is a v-shaped trough in the first quadrant) and so we conclude that $k(x,z) = e^{-|x-z|}$ is a kernel on $[0,a]$.

d) From the previous part, $k(x,z) = -|x-z| = \min(x,z)-a+a-\max(x,z)$ which is the sum of kernels and hence a kernel. By non-negative scaling, we also have that $k(x,z) = -\gamma|x-z|$ is a kernel. Again, by properties of 16.2.1, we have that $k(x,z) = e^{-\gamma|x-z|}$ is a kernel on $[0,a]$. We can extend this to $[-a,a]$ by repeating the steps in part a) to c) but with the inner product $\langle g, h \rangle = \int_{-a}^a g(s)h(s)ds$ and using $k(x,z) = \min(x,z) + a$ for part a).

e) From part d), we have kernels $\{K_a\}_{a>0}$, taking the limit $\lim_{a\to\infty} K_a$ yields a kernel $k(x,z) = e^{-\gamma|x-z|}$ for $\gamma \geq 0$ and valid on $\mathbb{R}$.