Properties of positive-definite kernels:

It is important to be able to construct kernels. We will discuss some basic kernels and also ways of constructing new kernels from old ones.

We have already seen polynomial kernels.

Let $X = \mathbb{R}$ and take our Hilbert space to be $\ell^2(\mathbb{N})$. Then we define $\Phi : \mathbb{R} \to \ell^2(\mathbb{N})$ by

$$\Phi(x) = e^{-x^2/2}\left(1, x, \frac{x^2}{\sqrt{2}}, \frac{x^3}{\sqrt{3!}}, \cdots \frac{x^n}{\sqrt{n!}}, \cdots\right)$$

Then $\langle \Phi(x), \Phi(x') \rangle =$

$$e^{-\frac{(x^2+x'^2)}{2}}\left(\sum_{n=0}^{\infty}\frac{x^n}{\sqrt{n!}}\right)\left(\sum_{m=0}^{\infty}\frac{x'^m}{\sqrt{m!}}\right)$$

$$= \exp\left[-\tfrac{1}{2}\left(x^2 + x'^2 - 2xx'\right)\right] = e^{-\frac{1}{2}(x-x')^2}$$

This is called the exponential kernel. We can easily do something similar for $\mathbb{R}^n$ & $\sigma > 0$ to define a kernel

$$K(\vec{x}, \vec{x}') = e^{-\frac{1}{2}\|\vec{x} - \vec{x}'\|^2}$$

This is commonly called the gaussian kernel. Note it is easy to compute this kernel even though the feature space is infinite dimensional.

Now for some operations on kernels

(1) <u>Normalization</u> Given a kernel $K : X \times X \to \mathbb{R}$ we define

$$\hat{K}(x,x') = \begin{cases} 0 \text{ if } & K(x,x) = 0 \text{ or } K(x',x') = 0 \\ \dfrac{K(x,x')}{\sqrt{K(x,x)\,K(x',x')}} & \text{otherwise} \end{cases}$$

__Prop__ If K is a psd kernel then so is $\hat{K}$.

__Proof__ Let $\{x_1, \dots x_n\} \subseteq X$ and let $\vec{v}$ be a vector in $\mathbb{R}^n$.
If $K(x_i, x_i) = 0$ then $K(x_i, x_j) = 0$ by Cauchy-Schwartz so $\hat{K}(x_i, x_j) = 0$ for all $j \in \{1, \dots n\}$. So assume $K(x_i, x_i) > 0$ for all $i \in \{1, \dots, n\}$. Then we have

$$\sum_{i,j=1}^{n} \frac{v_i v_j K(x_i, x_j)}{\sqrt{K(x_i, x_i) K(x_j, x_j)}} = \sum_{i,j=1}^{n} \frac{v_i v_j \langle \Phi(x_i), \Phi(x_j) \rangle}{\|\Phi(x_i)\| \; \|\Phi(x_j)\|}$$

where $\Phi : X \to H$ is the feature map into the RKHS H. We know such an RKHS exists.

But this is just

$$\left(\sum_{i=1}^{n} \frac{\Phi(v_i x_i)}{\|\Phi(x_i)\|} \right)^2 \geq 0.$$

Thus for any choice of $\vec{v}$ in $\mathbb{R}^n$ & any choice of x_i's in X $\vec{v}^T K(x_i, x_j) \vec{v}$ is positive semidefinite ∎

This is of course a very easy proof but it shows how one can leverage the fact that there is an underlying H and Φ.

Closure properties of kernels.

__Prop__ Kernels are closed under sum and product.

__Proof__ Suppose K, K' are two kernels and for some set of n points k, k' are the Gram matrices. So for any vector $\vec{v} \in \mathbb{R}^n$ we have $\vec{v} \cdot (k \vec{v}) \geq 0$ & $\vec{v} \cdot (k' \vec{v}) \geq 0$ so clearly $\vec{v} \cdot (k + k') \vec{v} \geq 0$. Thus $K + K'$ is psd.

The proof for ~~produ~~ products involves some matrix facts and can be read on pg 115 of Mohri et al.

There is another interesting operation on Kernels:

<u>Tensor product</u> We have two kernels K, K', we define $K \otimes K'$ by

$$\forall x_1, x_1'; x_2, x_2' \in X$$

$$(K \otimes K')(x_1, x_1', x_2, x_2') = K(x_1, x_2)\, K'(x_1', x_2').$$

Where does this come from?

Given 2 Hilbert spaces H_1 and H_2 we can define a new Hilbert space $H_1 \otimes H_2$. The underlying space is the closure of the span of all vectors of the form $h_1 \otimes h_2$ where $h_1 \in H_1$ & $H_2 \in h_2$. Note the <u>span</u>. For example if we are looking at $\mathbb{R}^2 \otimes \mathbb{R}^2$ and we define $e_0 \in \mathbb{R}_2$ to be $\binom{1}{0}$ and e_1 to be $\binom{0}{1}$ then $\mathbb{R}^2 \otimes \mathbb{R}^2$ contains

$$e_0 \otimes e_0 + e_1 \otimes e_1$$

There is no way to write this as ~~as~~ $u \otimes v$ where $u, v \in \mathbb{R}^2$ [Try it and see!] Thus the tensor product is far richer than just pairs of vectors. The inner product on $H_1 \otimes H_2$ is defined by

$$\langle a \otimes b, a' \otimes b' \rangle = \langle a \otimes a' \rangle_{H_1} \langle b \otimes b' \rangle_{H_2}$$

and extended linearly to the rest of the span.

Now suppose we have $\Phi_1 : X \to H_1$, $\Phi_2 : X \to H_2$ we can define $\Phi_1 \otimes \Phi_2 : X \times X \to H_1 \otimes H_2$ by

$$\Phi_1 \otimes \Phi_2 = \Phi(x_1) \otimes \Phi(x_2).$$

If you work out the algebra this will give exactly the formula for $K_1 \otimes K_2$.

The fact that it is psd is now immediate.

<u>Prop</u> If $\forall x, x' \in X, \lim_{n\to\infty} K_n(x,x')$ is defined for a family of kernels K_n then the function

$$K(x,x') := \lim_{n\to\infty} K_n(x,x') \text{ is a kernel.}$$

<u>Proof</u> Let k_n be the gramians for $\{x_1, \cdots x_m\} \subset X$.

Then $\forall n \quad \vec{v}\cdot(k_n\vec{v}) \geq 0 \Rightarrow \lim_{n\to\infty} \vec{v}\cdot(k_n\vec{v}) \geq 0$

$$= \vec{v}\cdot(k\vec{v}) \geq 0$$

where $k = \lim_{n\to\infty} k_n$ entrywise

— × —

SVM with kernels

Recall SVM in dual form

$$\max_{\vec{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j)$$

subject to constraints $0 \leq \alpha_i \leq C$ & $\sum_{i=1}^{n} \alpha_i y_i = 0$

Note the presence of the dot products. We can replace these dot products with kernels

$$\max_{\vec{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j)$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$

Representer Theorem

Let $K: X \times X \to \mathbb{R}$ be a kernel & $\Phi: X \to H$ the embedding into an RKHS. Then for any non-decreasing function $G: \mathbb{R} \to \mathbb{R}$ and any loss function $L: \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ the optimization problem for some fixed $x_1, \cdots x_n$

$$\operatorname*{argmin}_{h \in H} F(h) = \operatorname*{argmin}_{h \in H} G(\|h\|) + L(h(x_1), \cdots h(x_n))$$

has a solution of the form

$$h^* = \sum_{i=1}^{m} \alpha_i \Phi(x_i).$$

If G is increasing all solutions have this form.

Proof: Let $H_1 = \text{span}\{\Phi(x_i)\}$. Any h can be written as $h = h_1 + h_1^\perp$ where $h_1 \in H_1$ & $h_1^\perp \in H_1^\perp$ where we have $H = H_1 \oplus H_1^\perp$. Since G is non-decreasing

$$G(\|h_1\|) \leq G\left(\sqrt{\|h_1\|^2 + \|h_1^\perp\|^2}\right) = G(\|h\|).$$

Now
$$h(x_i) = \langle h, \Phi(x_i)\rangle = \langle h_1, \Phi(x_i)\rangle + \overset{0}{\langle h_1^\perp, \Phi(x_i)\rangle}$$
$$= \langle h_1, \Phi(x_i)\rangle = h_1(x_i)$$

Thus $L(h(x_1), \cdots, h(x_n)) = L(h_1(x_1), \cdots h_1(x_n))$

and $F(h_1) \leq F(h)$.

So for any solution there is an $h \in H_1$ with smaller F. If G is strictly increasing so $F(h_1) < F(h)$.

Note how the RKHS properties are used.

Remember $\forall x \in X \; \exists k_x \in H$ s.t. $\langle h, k_x\rangle = h(x)$ and $\Phi(x) = k_x$. This k_x is often written $K(x, \cdot)$.

Learning guarantees

We consider kernel based hypotheses coming from an RKHS. We bound the norm of the elements in H that we take as our hypotheses. Thus

$$\mathcal{H} = \{h \in H \mid \|h\| \leq \Lambda\} \text{ for some } \Lambda \geq 0.$$

$\forall h \in \mathcal{H}$ we have $h(x) = \langle h, K(x,\cdot)\rangle = \langle h, \Phi(x)\rangle$.

<u>Thm</u> Let $K: X \times X \to \mathbb{R}$ be a kernel and let $\Phi: X \to H$ be the associated feature map into the RKHS H constructed from K. Let $S \subseteq \{x \mid K(x,x) \leq r^2\}$ be a sample of size m i.e. $S = \{x_1, \cdots x_m\}$ where $\forall i \in \{1, \cdots m\}$ $K(x_i, x_i) \leq r^2$. Fix $\Lambda \geq 0$ and let

$$\mathcal{H} = \{x \mapsto \langle \omega, \Phi(x)\rangle = \omega(x) \mid \|\omega\| \leq \Lambda\}.$$

Then $\hat{\mathcal{R}}_S(\mathcal{H}) \leq \dfrac{\Lambda\sqrt{Tr[K]}}{m} \leq \sqrt{\dfrac{r^2\Lambda^2}{m}}$

<u>Proof</u> $\hat{\mathcal{R}}_S(\mathcal{H}) = \dfrac{1}{m} \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|\omega\| \leq \Lambda} \left\langle \omega, \sum_{i=1}^m \sigma_i \Phi(x_i)\right\rangle\right]$ Def of RC

$\leq \dfrac{\Lambda}{m} \mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^m \sigma_i \Phi(x_i)\right\|\right]$ Cauchy-Schwartz ~~equality~~ ~~case~~

$\leq \dfrac{\Lambda}{m}\left[\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^m \sigma_i \Phi(x_i)\right\|\right]^2\right]^{1/2}$ Jensen

$= \dfrac{\Lambda}{m}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sum_{i=1}^m \|\Phi(x_i)\|^2\right]\right]^{1/2}$ Look ma, no σ's!

$= \dfrac{\Lambda}{m}\left[\sum_{i=1}^m K(x_i, x_i)\right]^{1/2} = \dfrac{\Lambda}{m}\sqrt{Tr K}$

Now $K(x_i, x_i) \leq r^2$ so $Tr K \leq r^2 m$ so

$$\hat{\mathcal{R}}_S(\mathcal{H}) \leq \frac{\Lambda}{m}\sqrt{r^2 m} = \frac{\Lambda r}{\sqrt{m}}$$