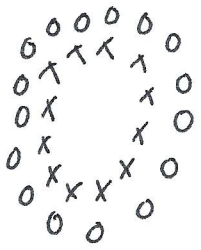
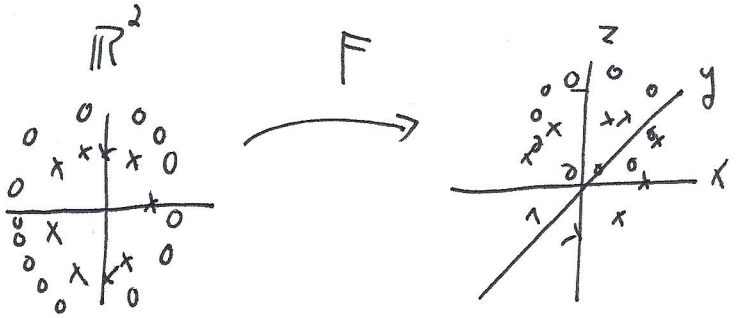
Lecture #17: F of SML



- SVM are notable to handle data that cannot be separated linearly, e.g. where $\times = -1$, $\circ = +1$.
- Kernel methods allow to handle such classification problems, in which non-linear separation is required. We will focus on their applications to supervised learning, but they are pervasive in unsupervised learning as well, e.g. graph clustering.
- The intuition is to map data into a high dimensional space, where separation is cleaner. In the case of data drawn from distinct concentric circles $\{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 = r_i^2\}$ for $r_1 \neq r_2$, we can embed this data in $\mathbb{R}^2$ into $\mathbb{R}^3$ via the map $F: \mathbb{R}^2 \to \mathbb{R}^3$, $(x,y) \mapsto (x, y, x^2+y^2)$. Then the data is linearly separable in the third coordinate:



Badly drawn, but the level sets in the z-coordinate don't intersect! In fact, they are linearly separable.

- The example of concentric circles suggests a simple map F. Can we find a "universal" F? Sort of, but it needs to be ∞-dimensional. In this case, things get weird. The cute insight underlying much of kernel method literature, is that F is actualy not needed.
- Indeed, what I really need is $\|F(x) - F(y)\|_2^2$

$$= \langle F(x), F(x) \rangle + \langle F(y), F(y) \rangle - 2\langle F(x), F(y) \rangle.$$

- So, if $\langle F(x), F(y) \rangle$ can be quickly computed $\forall x, y$, we are in good shape.
- The kernel trick is this observation, coupled with fast ways of getting these high-dimensional inner products.

Defn: A function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel over $\mathcal{X}$.

ex: $\mathcal{X} = \mathbb{R}^D$, $K(x, x') = \exp(-\|x - x'\|_2^2)$

- We will aim to write $K(x,x') = \langle \Phi(x), \Phi(x') \rangle$ for some embedding function Φ that is left implicit. If $\Phi: \mathbb{R}^D \to \mathcal{F}$, we may think of $\mathcal{F}$ as a feature space and K as a similarity measure in the feature space.
- There is clearly a method to construct K given Φ; simply define $K(x,x') = \langle \Phi(x), \Phi(x') \rangle$.
Is it possible to get Φ given only K? Yes, under mild conditions:

Theorem (Mercer): Let $\mathcal{X} \subset \mathbb{R}^N$ be compact and let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be

(a.) continuous

(b.) symmetric, i.e. $K(x,x') = K(x',x)$ $\forall x \neq x'$.

Then $\exists$ constants $\{a_n\}_{n=0}^{\infty}$, $a_n > 0$, functions $\phi_n: \mathcal{X} \to \mathbb{R}$ s.t. $K(x,x') = \sum_{n=0}^{\infty} a_n \phi_n(x) \phi_n(x')$ (uniform convergence)

$\Updownarrow$

For any $c \in L^2(\mathcal{X})$, $\iint_{\mathcal{X} \times \mathcal{X}} c(x) c(x') K(x,x') \, dx \, dx' \geq 0$. (★)

- The condition (★) is equivalent to (→ we will define this) that the kernel K is positive semidefinite, and mirrors the analogous condition for matrices $A \in \mathbb{R}^{N \times N}$ that $x^T A x \geq 0$ $\forall x \in \mathbb{R}^{N \times 1}$.

·Indeed, Mercer's condition is sort of like an "infinite dimensional" positive-semidefiniteness condition.

Defn: A Kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite symmetric (PDS) if $\forall \{x_1, \dots, x_m\} \subseteq \mathcal{X}$, the induced kernel matrix $K = \left(K(x_i, x_j)\right)_{i,j=1}^{m}$ is a symmetric, positive semidefinite matrix. (SPSD).

·Recall that $K \in \mathbb{R}^{m \times m}$ is SPSD if the following hold:

(a.) $K_{ij} = K_{ji} \quad \forall i,j$

(b.) All the eigenvalues of K are non-negative.

·Note that (b.) is equivalent to $c^T K c \geq 0 \quad \forall c \in \mathbb{R}^{m \times 1}$.

·Given a sample $S = \{x_i\}_{i=1}^{m} \subset \mathbb{R}^D$, we call the associated $K \in \mathbb{R}^{m \times m}$ the Gram matrix associated to K and S.

ex: For any constant $c > 0$, a polynomial kernel of degree $d \in \mathbb{Z}_+$ is the kernel $K: \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ defined as $K(x, x') = (\langle x, x' \rangle + c)^d$.

ex: For any constant $\sigma > 0$, the Gaussian kernel with scaling parameter σ is defined over $\mathbb{R}^D$ as $K(x, x') = \exp(-\|x - x'\|_2^2 / \sigma^2)$. It is not hard to see that by Taylor expanding $K'(x, x') = \exp(\langle x, x' \rangle / \sigma^2)$ that

$$K'(x, x') = \sum_{j=0}^{\infty} \frac{(\langle x, x' \rangle)^j}{\sigma^{2j} j!},$$

so that K itself is an (infinite) linear combination of polynomial kernels.

ex: For any $a, b \in \mathbb{R}_{\geq 0}$, a sigmoid kernel is defined over $\mathbb{R}^D$ as

$$K(x,x') = \tanh(a\langle x,x'\rangle + b)$$

$$= \frac{\exp(a\langle x,x'\rangle + b) - \exp(-a\langle x,x'\rangle - b)}{\exp(a\langle x,x'\rangle + b) + \exp(-a\langle x,x'\rangle - b)}.$$

- We will show PDS kernels induce an inner product in a Hilbert space.

Lemma (C.S. for PDS kernels): Let K be a PDS kernel. Then $\forall x, x' \in \mathcal{X}$,

$$K(x,x')^2 \leq K(x,x)\,K(x',x').$$

Proof: This is immediate by noting that the determinant of $\begin{pmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{pmatrix}$

is non-negative, by PDS. ∎

Theorem (Reproducing Kernel Hilbert Spaces): Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel. Then there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi: \mathcal{X} \to \mathcal{H}$ s.t.

(a.) $\forall x, x' \in \mathcal{H}$, $K(x,x') = \langle \Phi(x), \Phi(x') \rangle$

(b.) $\forall h \in \mathcal{H}$, $\forall x \in \mathcal{X}$, $h(x) = \langle h, K(x,\cdot) \rangle$.

Proof: Define $\Phi(x): \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ as $\Phi(x)(x') = K(x,x')$. Define

$$\mathcal{H}_0 = \left\{ \sum_{i \in I} a_i \Phi(x_i) \;\middle|\; a_i \in \mathbb{R},\ x_i \in \mathcal{X},\ |I| < \infty \right\}.$$

Now, introduce the operation $\langle \cdot, \cdot \rangle$ on $\mathcal{H}_0 \times \mathcal{H}_0$ defined for $f = \sum_{i \in I} a_i \Phi(x_i)$, $g = \sum_{j \in J} b_j \Phi(x_j')$ as $\langle f, g \rangle = \sum_{(i,j) \in I \times J} a_i b_j K(x_i, x_j')$.

Clearly $\langle \cdot, \cdot \rangle$ is symmetric, and note that $\langle f, g \rangle$ is representation independent, since

$$\langle f, g \rangle = \sum_{j \in J} b_j f(x_j') = \sum_{i \in I} a_i g(x_i).$$

Moreover, $\langle \cdot, \cdot \rangle$ is bilinear. Finally, $\langle f, f \rangle \geq 0$ clearly, so that $\langle \cdot, \cdot \rangle$ is a positive semidefinite bilinear form. More generally, for any $f_1, \dots, f_n$ and any $c_1, \dots, c_m \in \mathbb{R}$,

$$\sum_{i,j=1}^{m} c_i c_j f_i f_j = \left\langle \sum_{i=1}^{m} c_i f_i, \sum_{j=1}^{n} c_j f_j \right\rangle \geq 0 \Rightarrow \langle \cdot, \cdot \rangle \text{ is a PDS kernel}$$

on $\mathcal{H}_0$. By our C.S. lemma,

$$\langle f, \Phi(x) \rangle^2 \leq \langle f, f \rangle \cdot \langle \Phi(x), \Phi(x) \rangle.$$

By definition of $\langle \cdot, \cdot \rangle$, for any $f = \sum_{i \in I} a_i \Phi(x_i) \in \mathcal{H}_0$,

$$f(x) = \sum_{i \in I} a_i K(x_i, x) = \langle f, \Phi(x) \rangle.$$

Thus, $|f(x)|^2 \leq \langle f, f \rangle K(x, x)$ $\forall x \in X$, so that $\langle \cdot, \cdot \rangle$ is definite. We have thus shown $\langle \cdot, \cdot \rangle$ defines an inner product and that (a.), (b.) hold on $\mathcal{H}_0$. Completing $\mathcal{H}_0$ to $\mathcal{H}$ by closure under the norm induced by $\langle \cdot, \cdot \rangle$ yields a Hilbert space $\mathcal{H}$ in which (a.), (b.) still hold, since $\mathcal{H}_0$ is by construction dense in $\mathcal{H}$. ∎

- We call $\mathcal{H}$ the RKHS associated to K.
- PDS kernels implicitly (through Φ) define a feature embedding; this Φ is important to practical performance.