Non-linear classification (continued)

- As discussed earlier the Adaline can readily be extended to training data that is not linearly seperable.
- This was done by introducing a non-linear function

$$\underline{\Phi}: \mathbb{R}^n \longrightarrow \mathbb{R}^m$$

from n-dim feature space to an m-dim space for sufficiently large m so that the training data becomes linearly seperable again.

- This can also be done with the SVM which in dual formulation reads

max $\tilde{L}(\alpha)$ for

$$\tilde{L}(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle$$

subject to $0 \leq \alpha_i \leq \mu$, $i = 1 \ldots M$, $\sum_{j=1}^{M} \alpha_j y^{(j)} = 0$

- The scalar product $\langle x^{(i)}, x^{(j)} \rangle$ must only be replaced by

$$S(x^{(i)}, y^{(j)}) := \langle \underline{\Phi}(x^{(i)}), \underline{\Phi}(x^{(j)}) \rangle_{\mathcal{H}}$$

where $\underline{\Phi}$ must be chosen such that

$$\underline{\Phi}: \mathbb{R}^n \longrightarrow \mathcal{H} = \text{Hilbert space with } \dim \mathcal{H} \gg n$$

- However, this is computationally very costly as computation of $S(x,y)$ $\sim \mathcal{O}(\dim \mathcal{H})$.

- <u>Idea:</u> replace the scalar product $S(x,y)$ by a function $K(x,y)$ that mimics the properties of a scalar product.

This lead to the following definition:

<u>DEF (pos. def. sym. Kernels)</u> A map $K: \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ is called positive definite symmetric kernel if

$\forall \{x_1, \dots, x_m\} \subset \mathbb{R}^n$: the matrix $\left(K(x_i, x_j)\right)_{\substack{1 \le i \le m \\ 1 \le j \le m}}$

is symmetric and positive semi-definite

<u>Recall:</u> A matrix M is sym. & pos. semi-def. if it is symmetric and $\forall v \in \mathbb{R}^m: \langle v, M v\rangle \geq 0$

The idea is that for such kernel K there is a Φ such that $K(x,y) = \langle \Phi(x), \Phi(y)\rangle$.

<u>Example:</u> Polynomial Kernel

We define $\forall x, y \in \mathbb{R}^n$:

$$K(x,y) = (x \cdot y + c)^d$$

for tuning parameters $c \in \mathbb{R}$, $d \in \mathbb{N}$

For $n=2$, $d=2$ we find

$$\Phi: \mathbb{R}^2 \to \mathbb{R}^6, \quad \Phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2c}\, x_1 \\ \sqrt{2c}\, x_2 \\ c \end{pmatrix}$$

because

$$(x \cdot y + c)^2 = (x_1 y_1 + x_2 y_2 + c)^2$$

$$= (x_1 y_1 + x_2 y_2)^2 + 2(x_1 y_1 + x_2 y_2)c + c^2$$

$$= x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2 + 2(x_1 y_1 + x_2 y_2) c + c^2$$

LEM (Cauchy-Schwarz): k pos. def. sym., then $\forall x,y \in \mathbb{R}^n$,

$$k(x,y)^2 \leq k(x,x)\,k(y,y)$$

proof: Def. matrix

$$M := \begin{pmatrix} k(x,x) & k(x,y) \\ k(y,x) & k(y,y) \end{pmatrix}$$

Since k pos. def. sym. M is symmetric and pos. semi-def.

$$\Rightarrow \underbrace{\det M}_{} \geq 0$$

$$= k(x,x)\,k(y,y) - k(x,y)^2 \quad \square$$

THM: Let $k: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a pos. def. sym. kernel, then

i) $\exists$ a Hilbert space $\mathcal{H}$ of real-valued functions and a map $\Phi: \mathbb{R}^n \to \mathcal{H}$ s.t.

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$$

ii) $\forall f \in \mathcal{H}, x \in \mathbb{R}^n$

$$f(x) = \langle f, k(x,\cdot) \rangle_{\mathcal{H}}$$

Proof: For $x,y \in \mathbb{R}^n$ define

$$\Phi(x)(y) := k(x,y)$$

Def. $\mathcal{H}^0 := \left\{ \sum_{i \in I} a_i \Phi_i(x_i) \;\middle|\; a_i \in \mathbb{R}, x_i \in \mathbb{R}^n, |I| < \infty \right\}$

and an inner product

$\langle \cdot, \cdot \rangle: \mathcal{H}^0 \times \mathcal{H}^0 \to \mathbb{R}$ s.t.

such that for $f, g \in \mathcal{H}^0$, i.e.,

$$f = \sum_{i \in I} a_i \Phi(x_i)$$

$$g = \sum_{j \in J} b_j \Phi(x_j)$$

$$\langle f, g \rangle := \sum_{\substack{i \in I \\ j \in J}} a_i b_j k(x_i, y_j)$$

$$= \sum_{i \in I} a_i g(x_i)$$

$$= \sum_{j \in J} b_j f(x_j)$$

Hence, i) $\langle f, g \rangle = \langle g, f \rangle$

ii) representation independent

iii) $\langle f, f \rangle = \sum_{\substack{i \in I \\ j \in J}} a_i b_j k(x_i, x_j) \geq 0$

i.e. pos. semi-def.

Hence, $\langle \cdot, \cdot \rangle$ is a pos. def sym. kernel on $\mathcal{H}^0$.

But $\forall f \in \mathcal{H}^0$

$$\langle f, \Phi(x) \rangle^2 \overset{(*)}{\leq} \langle f, f \rangle \langle \Phi(x), \Phi(x) \rangle$$

and for $f = \sum_{i \in I} a_i \Phi(x_i)$

$$\left.\begin{aligned} f(x) &= \sum_{i \in I} a_i \Phi(x_i)(x) \\ &= \sum_{i \in I} a_i k(x_i, x) \end{aligned}\right\} (**)$$

$$= \langle f, \Phi(x) \rangle$$

Hence,

$$|f(x)|^2 = \langle f, \Phi(x) \rangle^2 \leq \langle f, f \rangle k(x, x)$$

$$\Rightarrow \quad f = 0 \Leftrightarrow \langle f, f \rangle = 0$$

Thus $\langle \cdot, \cdot \rangle$ is an inner product on $\mathcal{H}^0$.

$\mathcal{H} = \overline{\mathcal{H}^0}^{\langle\cdot,\cdot\rangle}$, i.e., completion w.r.t. $\langle\cdot,\cdot\rangle$,

is again a Hilbert space with inner product

$\langle\cdot,\cdot\rangle$.

As $f \mapsto \langle f, \Phi(x)\rangle$ is Lipschitz

continuous due to (*) and by construction

$\mathcal{H}^0$ is dense (**) holds on all of $\mathcal{H}$. $\square$

<u>Conclusion:</u> SVM for non-linear problems

1) choose pos. def. sym. kernel K

2) max $\tilde{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y^{(i)} y^{(j)} \underbrace{K(x^{(i)}, y^{(j)})}_{= \Phi(x^{(i)})\cdot\Phi(y^{(j)})}$

subject to $\mu = \alpha_i + \beta_i$, $\alpha_i \geq 0$, $\beta_i \geq 0$

$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$ for $i = 1 \ldots m$

3) activation function

KKT(I)

i) $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$

ii) $\underline{w} = \sum_{i=1}^{m} \alpha_i y^{(i)} \Phi(\underline{x}^{(i)})$

iii) $\mu = \alpha_i + \beta_i$ for $i = 1 \ldots m$

KKT(II)

iv) $\alpha_i \left(1 - y^{(i)}(w_0 + \underline{w}\cdot\Phi(\underline{x}^{(i)})) - \xi_i\right) = 0$

for $\alpha_i \geq 0$ $i = 1 \ldots m$

v) $\beta_j \xi_j = 0$

vi) for $\beta_j \geq 0$ $j = 1 \ldots m$

compute

$\underline{w} = \sum_{i=1}^{m} \alpha_i y^{(i)} \underline{x}^{(i)}$

for vectors fulfilling $\alpha_i \neq 0$

- these are $(y^{(i)}, x^{(i)})$ fulfilling

$1 - y^{(i)} \left(w_0 + \underline{w}\cdot\Phi(\underline{x}^{(i)})\right) - \xi_i = 0$

- lont must fulfill

$$\beta_i \xi_i = 0$$

$$\Rightarrow \begin{cases} \xi_i = 0 & \text{support vector } y^{(i)}(w_0 + w \cdot \Phi(\underline{x}^{(i)})) = 1 \\ \xi_i \neq 0 & \text{outlier } y^{(i)} w \cdot x^{(i)} = 1 - \xi_i \end{cases}$$

can be detected by

$$\beta_i = 0 \Rightarrow \alpha_i = \mu$$

So choose support vector $(x^{(i)}, y^{(i)})$ to compute

$$y^{(i)}(w_0 + \underline{w} \cdot \Phi(\underline{x}^{(i)})) = 1 \Rightarrow w_0 = y^{(i)} - \underline{w} \cdot \Phi(\underline{x}^{(i)})$$

This gives $w = (w_0, \underline{w})$ and defines the activation function:

$$h(x) = \sigma(w_0 + \underline{w} \cdot \Phi(\underline{x}))$$

$$= \sigma\left(y^{(i)} - \underline{w} \cdot \Phi(\underline{x}^{(i)}) + \underline{w} \cdot \Phi(\underline{x})\right)$$

$$= \sigma\Big(y^{(i)} - \sum_{j=1}^{m} \alpha_j y^{(j)} k(\underline{x}^{(j)}, \underline{x}^{(i)}) + \sum_{j=1}^{m} \alpha_j y^{(j)} k(\underline{x}^{(j)}, \underline{x})\Big)$$

LEM: Pos. def. sym. kernel are closed under sum, product, tensor product, point-wise limit and composition with power series $\sum_{n=0}^{\infty} a_n x^n$ for $a_n \geq 0$.

Proof: sum) clear

product) give two pos. def. sym. kernels

$$k(x,y),\ k'(x,y)$$

def. kernel matrix for $k(x,y)$

$$k_{ij} := k(x_i, x_j),\ 1 \leq i,j \leq m$$

$\exists\, R$ s.t. $K = R^* R$ since k pos. semidef

Then $k(x,y) k'(x,y)$ has kernel matrix

$$k_{ij} k'_{ij}$$

$$\forall v \in \mathbb{R}^n: \sum_{i=1}^{m} \sum_{j=1}^{m} v_i k_{ij} k'_{ij} v_j$$

$$= \sum_{i,j=1}^{m} \sum_{k=1}^{m} v_i R^*_{ik} k'_{ij} R_{kj} v_j = \sum_{i=1}^{m} w$$

$$= \sum_{k=1}^{n} \underbrace{\langle w_k, K' w_k \rangle}_{\geq 0 \text{ because } K' \text{ pos. def. sym.}} \quad \text{for } (w_k)_j = R_{kj} v_j$$

furthermore sym.

tensor product) $k, \hat{k}$ pos. def. sym kernels

Say $K : (x_1, x_1', x_2, x_2') \to k(x_1, x_2)$

$\hat{K} : (x_1, x_1', x_2, x_2') \to \hat{k}(x_1', x_2')$

tensor product

$$k \otimes k'(x_1, x_1', x_2, x_2') = k(x_1, x_1', x_1, x_2')$$

then argument as with product.

point-wise limit) Say $(k_n)_{n \in \mathbb{N}}$ pos. def. sym. kernels and $k_n(x,y) \to k(x,y)$ then for all $v \in \mathbb{R}^n$

$$\langle v, k_n v \rangle \geq 0 \Rightarrow \lim_{n \to \infty} \langle v, k_n v \rangle \geq 0$$

and symmetry likewise

power series) Say $\sum_{n=0}^{\infty} a_n x^n$ converges for conv. radius $\delta > 0$ and suppose $|k(x,y)| < \delta \ \forall x, y \in \mathbb{R}^n$, then $\sum_{n=0}^{N} a_n k^n$ pos. def. sym. by product) & sum) and $a_n \geq 0$ and $\sum_{n}^{\infty} a_n k^n$ by limit). $\square$

Normalization of Kernels

for any kernel k we def the normalized kernel

$$\hat{k}(x,y) := \begin{cases} 0 & \text{if } k(x,x) = 0 \\ & \text{or } k(y,y) = 0 \\ \dfrac{k(x,y)}{\sqrt{k(x,x)\, k(y,y)}} & \end{cases}$$

LEM: Let k be pos. def. sym.

$\Rightarrow$ the corresponding $\hat{k}$ is pos. def. sym.

Proof: Let $\{x_1, \dots, x_M\} \subset \mathbb{R}^n$, $c \in \mathbb{R}^M$.

If $k(x_i, x_i) = 0 \Rightarrow k(x_i, x_j) = 0$
$\Rightarrow \hat{k}(x_i, x_j) = 0$.

Hence, we only need to treat $k(x_i, x_i) > 0$.

Then

$$\sum_{i,j=1}^{M} c_i \frac{k(x_i, x_j)}{\sqrt{k(x_i, x_i)\, k(x_j, x_j)}} c_j$$

$$= \sum_{i,j=1}^{M} c_i \frac{\langle \Phi(x_i), \Phi(x_j) \rangle}{\|\Phi(x_i)\| \, \|\Phi(x_j)\|} c_j$$

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

Symmetry follows from the definition. □

Example (Gaussian kernel)

$$\hat{k}(x, y) = \exp\left(- \frac{|x-y|^2}{2\sigma}\right), \quad \sigma \geq 0$$

Let us first consider

$$k(x, y) = \exp\left(+ \frac{x \cdot y}{\sigma}\right)$$

$$= \lim_{N \to \infty} \sum_{n=0}^{N} \frac{|x-y|^{2n}}{n! \, \sigma^n}$$

is a pos. def. sym. kernel since $x \cdot y$ is and LEM product, sum, power series).

$$\hat{k}(x, y) = \frac{k(x, y)}{\sqrt{k(x, x)\; k(y, y)}} = \frac{e^{\frac{x \cdot y}{\sigma^2}}}{e^{-\frac{x^2}{2\sigma^2}} \; e^{-\frac{y^2}{2\sigma^2}}}$$

$$= \exp\left(- \frac{|x-y|^2}{2\sigma^2}\right)$$