Kernel functions

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 \quad \rightarrow \text{data point w/ 2 features}$$

Kernel: $$\phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix} \in \mathbb{R}^3$$

original: $h_\theta(x) = \text{sign}(\theta w) = \begin{cases} +1 & \text{if } \theta w > 0 \\ 0 & \text{if } \theta w = 0 \\ -1 & \text{if } \theta w < 0 \end{cases}$

new:

$$h_\theta(\phi(x)) = \theta \phi(x)$$

$$= \text{sign}\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 (x_1^2 + x_2^2)\right)$$

SUMs and the kernel trick

to make predictions w SUMs:

$$h_w(x) = \text{sign}(wx + b)$$

with transformations:

$$h_w(\phi(x)) = \text{sign}(w\phi(x) + b)$$

from SUM dual formulation: $w = \sum_{i=1}^{n} \alpha_i y_i x_i$

$$\Rightarrow h_w(x) = \text{sign}\left(\underbrace{\left(\sum_{i=1}^{n} \alpha_i y_i \, x_i \right) \cdot x}_{\sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x)} + b \right)$$

$$\Rightarrow h_w(\phi(x)) = \text{sign}\left(\underbrace{\left(\sum_{i=1}^{n} \alpha_i y_i \, \phi(x_i) \right) \cdot \phi(x)}_{\sum_{i=1}^{n} \alpha_i y_i (\phi(x_i) \cdot \phi(x))} + b \right)$$

this is computationally expensive, so:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

↙ kenel function, depending on approach

these are equivalent. another perspective from the optimization problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

↙ inner product of 2 points

using $\phi(x)$:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\phi(x_i) \cdot \phi(x_j))$$

↙ inner product of 2 transformed points

Polynomial Kernel

let's say we have $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, goal: transform $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^{K>2}$ (higher dimension)

Polynomial Kernel:

$$\phi(x) = \begin{bmatrix} \dots \frac{\sqrt{d!}}{\sqrt{j_1!\, j_2! \cdots j_{n+1}!}} \cdots x_1^{j_1} x_2^{j_2} \cdots x_n^{j_n} \cdot 1^{j_{n+1}} \dots \end{bmatrix}$$

- $d!$: highest degree in polynomial
- $j_1! \, j_2! \cdots j_{n+1}!$: original # of features
- $\frac{\sqrt{d!}}{\sqrt{j_1! \, j_2! \cdots j_{n+1}!}}$: normalization constant for each new feature
- $x_1^{j_1} x_2^{j_2} \cdots x_n^{j_n}$: feature
- $1^{j_{n+1}}$: to account for constant term
- $\dots$: equation for each new feature | polynomial term s.t. each term has a different combination of $\sum j_i = d$

$$\sum_{i=1}^{n+1} j_i = d$$

calculate for all possible j_i combinations given this constraint

$$\text{size of } \phi(x) = \binom{n+d}{d} = \frac{(n+d)!}{d!\, n!}$$

$\phi(x) \in \mathbb{R}^{\binom{n+d}{d}}$

the new terms in $\phi(x)$ will look like:

- constant term $1 \rightarrow j_1 = 0, \; j_2 = 0, \; \dots, \; j_{n+1} = d$
- original terms $(x_1) \rightarrow j_1 = 1, \; j_2 = 0, \; \dots, \; j_{n+1} = d-1$
- interaction terms $(x_1 x_2) \rightarrow j_1 = 1, \; j_2 = 1, \; \dots, \; j_{n+1} = d-2$
- power of features $(x_1^d) \rightarrow j_1 = d, \; j_2 = 0, \; \dots, \; j_{n+1} = 0$

example : given $n = 2$ and $d = 2$:

$$\text{size of } \phi(x) = \frac{(2+2)!}{2!\,2!} = 6$$

term #1 $\quad j_1 = 0 \quad j_2 = 0 \quad j_3 = 2 \quad \rightarrow \sum j_i = d = 2$

$$\frac{\sqrt{2!}}{\sqrt{0!\,0!\,2!}} (x_1)^0 (x_2)^0 (1)^2 = (1)(1) = 1$$

term #2 $\quad j_1 = 1 \quad j_2 = 0 \quad j_3 = 1$

$$\frac{\sqrt{2!}}{\sqrt{1!\,0!\,1!}} (x_1)^1 (x_2)^0 (1)^1 = (\sqrt{2})(x_1) = \sqrt{2}\, x_1$$

term #3 $\quad j_1 = 0 \quad j_2 = 1 \quad j_3 = 1$

$$\frac{\sqrt{2!}}{\sqrt{0!\,1!\,1!}} (x_1)^0 (x_2)^1 (1)^1 = (\sqrt{2})(x_2) = \sqrt{2}\, x_2$$

term #4 $\quad j_1 = 1 \quad j_2 = 1 \quad j_3 = 0$

$$\frac{\sqrt{2!}}{\sqrt{1!\,1!\,0!}} (x_1)^1 (x_2)^1 (1)^0 = (\sqrt{2})(x_1 x_2) = \sqrt{2}\, x_1 x_2$$

term #5 $j_1 = 2 \quad j_2 = 0 \quad j_3 = 0$

$$\frac{\sqrt{2!}}{\sqrt{2!\ 0!\ 0!}} (x_1)^2 (x_2)^0 (1)^0 = (1)(x_1^2) = x_1^2$$

term #6 $j_1 = 0 \quad j_2 = 2 \quad j_3 = 0$

$$\frac{\sqrt{2!}}{\sqrt{0!\ 2!\ 0!}} (x_1)^0 (x_2)^2 (1)^0 = (1)(x_2)^2 = x_2^2$$

finally: $\phi(x) = \begin{bmatrix} 1 \\ \sqrt{2}\, x_1 \\ \sqrt{2}\, x_2 \\ \sqrt{2}\, x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix} \in \mathbb{R}^6$

Kernel trick

the dot product of $K(A \cdot B) = K(A_1 B_1 + A_2 B_2 + \cdots + A_n B_n)$

binomial theorem $(A + B)^n = \sum_{k=0}^{n} \binom{n}{k} A^{n-k} B^k$

taking the above $\phi(x)$ example:

$\phi(x) \cdot \phi(z) = 1 + 2x_1 z_1 + 2x_2 z_2 + 2x_1 x_2 z_1 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2$

using $K(x, z) = (x \cdot z + 1)^2$ and the binomial theorem:

$$(x \cdot z + 1)^2 = (x_1 z_1 + x_2 z_2 + 1)^2$$

$$= 1 + 2x_1 z_1 + 2x_2 z_2 + 2x_1 x_2 z_1 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2$$

$$K(x, z) = (x \cdot z + 1)^d = \phi(x) \cdot \phi(z)$$

↳ equivalent ↲

RBF Kernel

transformation is polynomial with $d = \infty$

$$K(x, z) = \exp\left(-\gamma \|x - z\|^2\right)$$

Proving that $K(x, z) = \phi(x) \cdot \phi(z)$ w RBF is too hectic, but lets use $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$:

- $\|x - z\|^2 = (x_1 - z_1)^2 + (x_2 - z_2)^2$

  all interaction & power terms ↑

  $= x_1^2 + z_1^2 - 2x_1 z_1 + x_2^2 + z_2^2 - 2x_2 z_2$

- $\exp(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!}$

- $\exp\left(-\gamma \|x - z\|^2\right) = \sum_{k=0}^{\infty} \frac{1}{k!} \left(-\gamma \|x - z\|^2\right)^k$

this is somehow equivalent to $\langle \phi(x), \phi(z) \rangle$, doesnt matter

Kernel least squares

Simply using $\phi(x)$ in $\min_{\beta} \| y - w\beta \|^2$:

$$\min_{\beta} \| y - \phi(x)\beta \| = (y - \phi\beta)^T (y - \phi\beta)$$

$$= y^T y - 2\beta^T \phi^T y + \beta^T \phi^T \phi \beta$$

$$\frac{dJ}{d\beta} = -2\phi^T y + 2\phi^T \phi \beta = 0$$

$$\phi^T \phi \beta = \phi^T y$$

$$\boxed{\beta = (\phi^T \phi)^{-1} \phi^T y}$$

with regularization:

$$\min_{\beta} \|y - \phi\beta\| + \lambda \|\beta\|$$

$$\Rightarrow \quad \boxed{\beta = (\phi^T \phi + \lambda I)^{-1} \phi^T y}$$