4. non-linear classification

4.1 polynomial function

↳ transforms your input into a new set of features by including all possible combinations up to a certain degree d

↳ they include

1. original features (x_1, x_2)
2. power of features (x_1^2, x_2^3)
3. interaction terms $(x_1x_2, x_1^2x_2)$

↳ say you original data is $x = [x_1, x_2] \in \mathbb{R}^2$ & you perform a polynomial transformation of degree $d=2$

the new features are:

1. degree 0 : a constant term (ex. 1)
2. degree 1 : original features (x_1, x_2)
3. degree 2 : squares (x_1^2, x_2^2) & interactions terms (x_1, x_2)

the new feature vector $\phi(x) \Rightarrow$ includes all combinations of powers of x_1 & x_2 where total power $\leq d$

$$\phi(x) = [1, x_1, x_2, x_1^2, x_2^2, x_1x_2]$$

$1 \leq 2$ $1 \leq 2$ $2 \leq 2$ $2 \leq 2$ $2 \leq 2$

now $x \in \mathbb{R}^2$ is transformed into $\phi(x) \in \mathbb{R}^6$

to apply factorial normalization for each term:

j_1 & j_2 represent the features x_1 & x_2 j_3 is there to hold the degree when not used

degree 0 (constant term)

- term = 1

- degree $\Rightarrow d = 0$, $j_1 = 0 = j_2$, $j_3 = 2$

$$\sqrt{\frac{d!}{j_1! \cdot j_2! \cdot j_3!}} \cdot \sqrt{\frac{2!}{0! \cdot 0! \cdot 2!}} = \sqrt{1} = 1$$

So here, the degree is 0 but the ~~transformation is~~ of degree 2, so we put the 2 in j_3 to ensure x_1 & x_2 are of degree 0. Same with the rest

degree 1 (linear terms)

- term = x_1, x_2

- degree $\Rightarrow d = 1$

for $x_1 \Rightarrow j_1 = 1, j_2 = 0, j_3 = 1$

for $x_2 \Rightarrow j_1 = 0, j_2 = 1, j_3 = 1$

$$x_1 \Rightarrow \sqrt{\frac{2!}{1! \cdot 0! \cdot 1!}} = \sqrt{2}, \quad x_2 \Rightarrow \sqrt{\frac{2!}{0! \cdot 1! \cdot 1!}} = \sqrt{2}$$

degree 2 (quadratic terms)

- term = x_1^2, x_2^2, $x_1 x_2$

- degree $\Rightarrow d = 2$

for $x_1^2 \Rightarrow j_1 = 2, j_2 = 0, j_3 = 0$

for $x_2^2 \Rightarrow j_1 = 0, j_2 = 2, j_3 = 0$

for $x_1 x_2 \Rightarrow j_1 = 1, j_2 = 1, j_3 = 0$

$$x_1^2 \Rightarrow \sqrt{\frac{2!}{2! \cdot 0! \cdot 0!}} = 1 \qquad x_2^2 \Rightarrow \sqrt{\frac{2!}{2! \cdot 0! \cdot 0!}} = 1$$

$$x_1 x_2 = \sqrt{\frac{2!}{1! \cdot 1! \cdot 0!}} = \sqrt{2}$$

the classifier now has the following form

$$h(\phi(x); \theta, \theta_0) = sign(\theta^T \phi(x) + \theta_0)$$

$$= sign(\theta_1 x_1 + \theta_2 x_2 + \theta_3 (x_1^2) \ldots + \theta_0)$$

the size of the vector $\phi(x) =$

$$\binom{n+d}{d} = \frac{(n+d)!}{d!\,n!}$$

4.2 the Kernel trick

↳ allows for dot product calculation $\phi(x)^T \phi(x')$ in higher dimensional feature space without explicitly computing $\phi(x)$

$$K(x, x') = \phi(x)^T \phi(x')$$

↳ Kernel functions

1. additivity: sum of 2 valid Kernels is a valid Kernel
2. Scaler multiplication: valid Kernel · +ve scaler = valid Kernel
3. product of Kernels: product of 2 valid Kernels is a valid Kernel
4. exponentiation

4.3 Kernel least square

$$\min_{\beta} \|y - \phi(x)\beta\|_2^2$$

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

- limitations
  1. complex patterns
  2. overfitting
  3. curse of dimensionality
  4. Kernel & parameter tuning