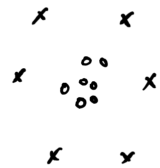
Recap: Nonlinear Features

Kernels

Kernel Regression

Example: $\underline{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} \in \mathbb{R}^2$

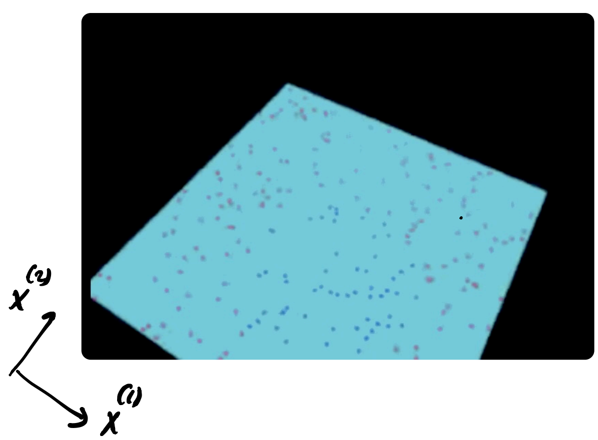


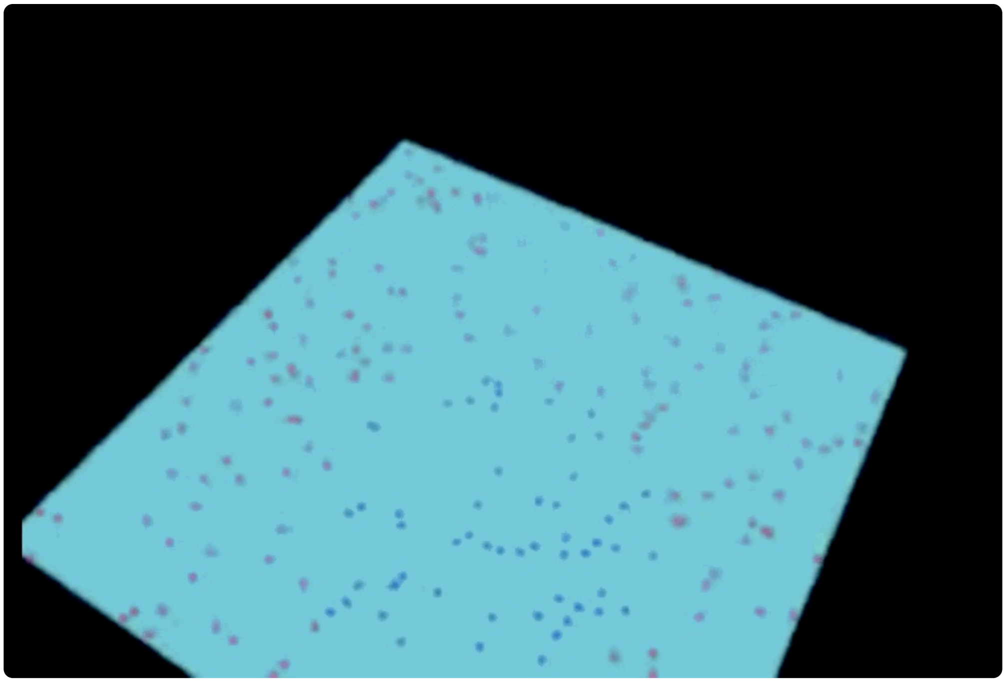
$$\Phi(x) = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(1)} \cdot x^{(2)} \\ (x^{(1)})^2 \\ (x^{(2)})^2 \end{bmatrix}$$

$$y \to \text{sign}\left\{ (x^{(1)} - c_1)^2 + (x^{(2)} - c_2)^2 - r^2 \right\}$$

$$(x^{(1)})^2 + c_1^2 - 2c_1 x^{(1)} + (x^{(2)})^2 + c_2^2 - 2c_1 x^{(2)} - r^2$$

$$\begin{bmatrix} x^{(1)} \\ x^{(2)} \\ (x^{(1)})^2 + (x^{(2)})^2 \end{bmatrix}$$





Inner Product Kernel

In the prev. approach issue is n is large
we can't compute $\Phi(x)$ explicitly

→ Many ML algs depend on $\Phi(x)$ via inner products

$$\langle \Phi(\underline{x}), \Phi(\underline{x}') \rangle$$

→ For many Φ $K(\underline{x},\underline{x}') = \langle \Phi(x), \Phi(x') \rangle$

we can compute efficiently even if m is large or infinite!

Example: $d=2$ $K(x,y) = (x^T y)^2 = \langle x, y \rangle^2$

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

$$= x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2$$

$$\Phi(x) = \begin{bmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{bmatrix} \Rightarrow \qquad = \langle \Phi(x), \Phi(y) \rangle$$

In k Nearest Neighbors: $\|x_i - x\|_2^2 = \langle x_i - x, x_i - x \rangle$

$$= \langle x_i, x_i \rangle - 2\langle x, x_i \rangle + \langle x, x \rangle$$

<u>Example:</u> d arbitrary

$$\begin{aligned} k(x,y) &= (x^T y)^2 \\ &= \Big(\sum_i x_i y_i\Big)^2 \\ &= \Big(\sum_i x_i y_i\Big) \cdot \Big(\sum_j x_j y_j\Big) \\ &= \sum_{ij} x_i y_i x_j y_j \quad = \langle \Phi(x), \Phi(y) \rangle \end{aligned}$$

$$\Phi(x) = \begin{bmatrix} x_1^2 \\ \vdots \\ x_d^2 \\ x_1 x_2 \\ \vdots \\ x_1 x_d \\ \vdots \\ x_{d-1} x_d \end{bmatrix} \in \mathbb{R}^m \qquad m = d + \binom{d}{2}$$

Exercise $\quad k(x,y) = (x^T y)^3 \Rightarrow \Phi(x) = ?$

Generalizing the above ex.

$$k(x,y) = (x^T y)^p$$

multinomial coefficient

$$\frac{p!}{j_1! \cdots j_d!}$$

$$= \sum_{\substack{j_1, \dots, j_d \\ \sum_i j_i = p}} \binom{p}{j_1 \cdots j_d} x_1^{j_1} \cdots x_d^{j_d} \cdot y_1^{j_1} \cdots y_d^{j_d}$$

$$\Rightarrow \Phi(x) = \left[\cdots, \sqrt{\binom{p}{j_1 \cdots j_d}} x_1^{j_1} \cdots x_d^{j_d}, \cdots \right]^T$$

all monomials of degree p.

A real inner product space is a vector space V on which we can def. a func. $\langle x, y \rangle$ (inner-product) s.t

(a) $\langle \alpha_1 x_1 + \alpha_2 x_2, y \rangle = \alpha_1 \langle x_1, y \rangle + \alpha_2 \langle x_2, y \rangle$

b) $\langle x, y \rangle = \langle y, x \rangle$

c) $\langle x, x \rangle \geq 0 \quad \forall x \in V$ with eq. iff $x = 0$

When can we find $k(x,y) = \langle \Phi(x), \Phi(y) \rangle$ for some inner product space V and $\Phi : \mathbb{R}^d \to V$

Defn: $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is an inner product kernel if $\exists$ an inner prod. space V and feature map $\Phi : \mathbb{R}^d \to V$ s.t $k(x,y) = \langle \Phi(x), \Phi(y) \rangle \quad \forall x, y \in \mathbb{R}^d$.

Φ and V may not be unique

Positive Definite Kernels

Defn: k is symmetric if $k(x,y) = k(y,x) \quad \forall x, y$

k is positive s. definite if it satisfies

SPD. Sym. and pos. def

(matrix K with entries K_{ij})

$K_{ij} = k(x_i, x_j)$ ↳ is PSD for all $x_1, \ldots, x_n$.

Theorem: (Mercer's). k is a SPD. kernel $\Longleftrightarrow$ k is an inner product kernel

A is a P.D. matrix $\Longleftrightarrow x^T A x > 0 \quad \forall x.$

Examples:

1. Homogeneous polynomial kernel
$$k(x,y) = (x^T y)^p$$

2. Inhomogeneous polynomial kernel
$$k(x,y) = (x^T y + 1)^p$$

3. Gaussian Kernel
$$k(x,y) = e^{-\frac{1}{2\sigma^2} \cdot \|x-y\|_2^2} \qquad (\sigma > 0)$$

V is an infinite dim space

$$\Phi(x) = e^{-x^2/2\sigma^2} \cdot \left[1, \sqrt{\frac{1}{1!\sigma^2}}x, \sqrt{\frac{1}{2!\cdot\sigma^2}}x^2, \sqrt{\frac{1}{3!\sigma^3}}\cdot x^3, \ldots\right]$$

Kernel trick:

1. Select an inner product kernel k
2. Formulate a linear learning method that only depends on inner products $\langle x, x' \rangle$
3. Replace $\langle x, x' \rangle$ with $k(x, x')$

$$XX^T \longrightarrow K$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$$XX^T = \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \cdots & x_1^T x_n \\ \vdots & & & \\ x_n^T x_1 & \cdots & & \end{bmatrix}$$

Ridge regression $\quad \hat{w} = (\tilde{X}^T\tilde{X} + \lambda I)^{-1}\tilde{X}^T\tilde{y} \qquad \hat{b} = \bar{y} - \hat{w}^T\bar{x}$

$$\tilde{x}_i = x_i - \bar{x} \qquad \bar{y} = \frac{1}{n}\sum y_i$$

$$\tilde{y} = y - \bar{y} \qquad \bar{x} = \frac{1}{n}\sum x_i$$

Given a test sample x : $\quad \hat{f}(x) = \hat{b} + \hat{w}^T x$

$$= \bar{y} + \hat{w}^T \cdot (x - \bar{x})$$

$\hat{f}(x)$ depends on $x_1, \ldots, x_n$ only in terms of inner products $\langle x_i, x_j \rangle$ and $\langle x_i, x \rangle$.

Matrix Inversion Lemma: (Woodbury formula) to $(\tilde{X}^T\tilde{X} + \lambda I)^{-1}$

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

(to make it look like $\tilde{X}\tilde{X}^T = \begin{bmatrix} \langle \tilde{x}_1, \tilde{x}_1 \rangle & \cdots & \langle \tilde{x}_1, \tilde{x}_2 \rangle \\ \vdots & & \vdots \end{bmatrix}$)

$A = \lambda I \quad U = \tilde{X}^T \quad V = \tilde{X} \quad C = I$

$$(\lambda I + \tilde{X}^T\tilde{X})^{-1} = \frac{1}{\lambda}\cdot I - \frac{1}{\lambda}\cdot I \cdot \tilde{X}^T \cdot \left(I + \tilde{X}\frac{1}{\lambda}\cdot I \cdot \tilde{X}^T \right)^{-1} \cdot \tilde{X} \cdot \frac{1}{\lambda}\cdot I$$

$$= \frac{1}{\lambda} I - \frac{1}{\lambda}\tilde{X}^T \left(I + \frac{\tilde{X}\tilde{X}^T}{\lambda} \right)^{-1} \tilde{X}$$

$$= \frac{1}{\lambda} \cdot \left[I - \tilde{X}^T \cdot (\lambda I + \tilde{X}\tilde{X}^T)^{-1} \tilde{X} \right]$$