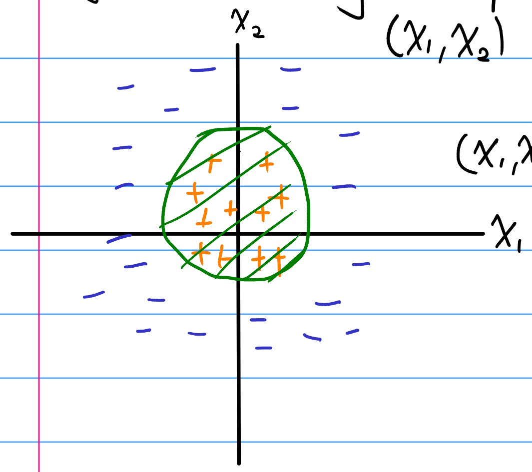
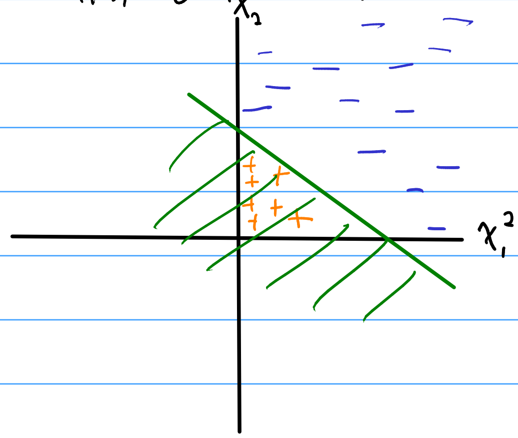
Kernels

Nonlinearly separable and nonlinear data

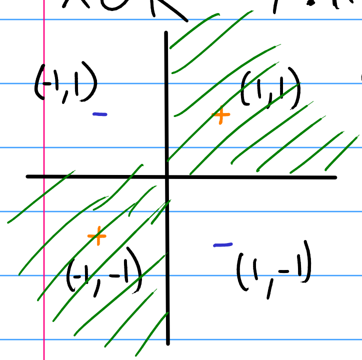


(x_1, x_2) $\quad x_1^2 + x_2^2 \leq r^2$

$(x_1, x_2) \rightarrow (x_1^2, x_2^2)$

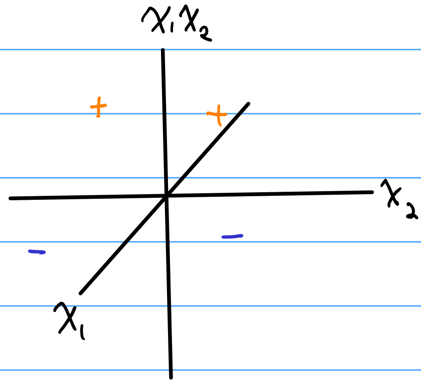


XOR $\quad \phi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$



$$\phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}$$

$\text{sign}(x_1 x_2)$



Feature Maps $\quad x \in \mathbb{R}^d, \; y \in \{\pm 1\}$

$$\phi(x) = \begin{bmatrix} x \\ 1 \end{bmatrix}, \quad w = \begin{bmatrix} p \\ b \end{bmatrix} \in \mathbb{R}^{d+1}$$

$\langle x, p \rangle > 0$

$\hookrightarrow \langle \phi(x), w \rangle = x^T p + b > 0$

$Q \in \mathbb{R}^{d \times d}, \; p \in \mathbb{R}^d, \; b \in \mathbb{R}$

$$\underbrace{x^T Q x}_{\sum (x_i x_j) Q_{ij}} + \sqrt{2}\, x^T p + b > 0$$

$$x x^T = \begin{bmatrix} x_1 x_1 & \cdots & x_1 x_d \\ x_2 x & & \\ \vdots & & \\ x_d x_1 & & x_d x_d \end{bmatrix} \xRightarrow{\text{Flatten}} \overrightarrow{x x^T} = \begin{bmatrix} x_1 x_1 \\ x_2 x_1 \\ x_3 x_1 \\ \vdots \\ x_d x_d \end{bmatrix} \quad \vec{Q} = \begin{bmatrix} Q_{11} \\ Q_{21} \\ Q_{31} \\ \\ Q_{dd} \end{bmatrix}$$

$$x^T Q x = \langle \overrightarrow{x x^T}, \vec{Q} \rangle$$

$$\phi(x) = \begin{bmatrix} \overrightarrow{x x^T} \\ \sqrt{2}\, x \\ 1 \end{bmatrix} \in \mathbb{R}^{d^2+d+1} \quad w = \begin{bmatrix} \vec{Q} \\ p \\ b \end{bmatrix}$$

$$\langle \phi(x), w \rangle \Leftrightarrow x^T Q x + \sqrt{2}\, x^T p + b$$

$$\phi : \mathbb{R}^d \to \mathbb{R}^{d^2+d+1}$$

- Computation $d \to O(d^2)$
- $\mathbb{R}^d \to \mathbb{R}^\infty$. $\quad (d=1) \; \phi(x) = [1, x, x^2, x^3, \ldots]^T$

x_i : ith point of training set

SVM (Dual)

$$\min_{C \geq \alpha \geq 0} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle - \sum \alpha_i$$

s.t. $\sum \alpha_i y_i = 0$.

$$\phi(x) = \begin{bmatrix} \overrightarrow{xx^T} \\ \sqrt{2}x \\ 1 \end{bmatrix}$$

$\langle \phi(x), \phi(x') \rangle = ?$ Naive : $O(d^2)$ time

$$\underbrace{\langle \overrightarrow{xx^T}, \overrightarrow{x'x'^T} \rangle}_{\sum_i \sum_j x_i x_j x'_i x'_j} + \underbrace{\langle \sqrt{2}x, \sqrt{2}x' \rangle}_{2\langle x, x' \rangle} + 1$$

$$= \sum x_i x'_i \left(\sum x_j x'_j \right) = \langle x, x' \rangle^2$$

$$\langle \phi(x), \phi(x') \rangle = \langle x, x' \rangle^2 + 2\langle x, x' \rangle + 1$$

$$= (\underbrace{\langle x, x' \rangle}_{O(d) \text{ computation}} + 1)^2$$

$$= k(x, x')$$

Kernels

$k: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel if $\exists$ feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$

$\langle \phi(x), \phi(x') \rangle = k(x, x')$

$\phi(x)$ may not be computable, but k still might be!

Examples:

Polynomial: $k(x,x') = (\langle x, x' \rangle + 1)^p$

Gaussian: $k(x,x') = \exp(-\|x - x'\|_2^2)$

What makes a valid kernel?

Take $x_1, \ldots, x_n$ (any dataset)

Let $K_{ij} = k(x_i, x_j) \in \mathbb{R}^{n \times n}$

$$\begin{array}{cc} & j \\ i & \begin{bmatrix} & & \\ & & k(x_i, x_j) \\ & & \end{bmatrix} \end{array}$$

Require:

- Symmetric: $K_{ij} = K_{ji}$
- Positive semidefinite: $v^T K v > 0 \quad \forall v \in \mathbb{R}^n$

$\exists \phi$ s.t. $\langle \phi(x_i), \phi(x_j) \rangle = K_{ij}$

Gram matrix. $\Phi = \begin{bmatrix} -\phi(x_1)- \\ \vdots \\ \phi(x_n) \end{bmatrix}$. $K = \Phi \Phi^T$

Using Kernels in SVMs

$$\min_{C \geq \alpha \geq 0} \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j \underbrace{K_{ij}}_{\langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j) = K_{ij}} - \sum \alpha_i \quad \text{s.t.} \sum \alpha_i y_i = 0$$

$$w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

$$\text{sign}(\langle w, \phi(x) \rangle) = \text{sign}\left(\sum \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle \right)$$
$$= \text{sign}\left(\sum \alpha_i y_i k(x_i, x) \right)$$

SVM (Linear Kernel): Train $O(nd)$, Test $O(d)$

General Kernel: Train $O(n^2 d)$, Test $O(nd)$