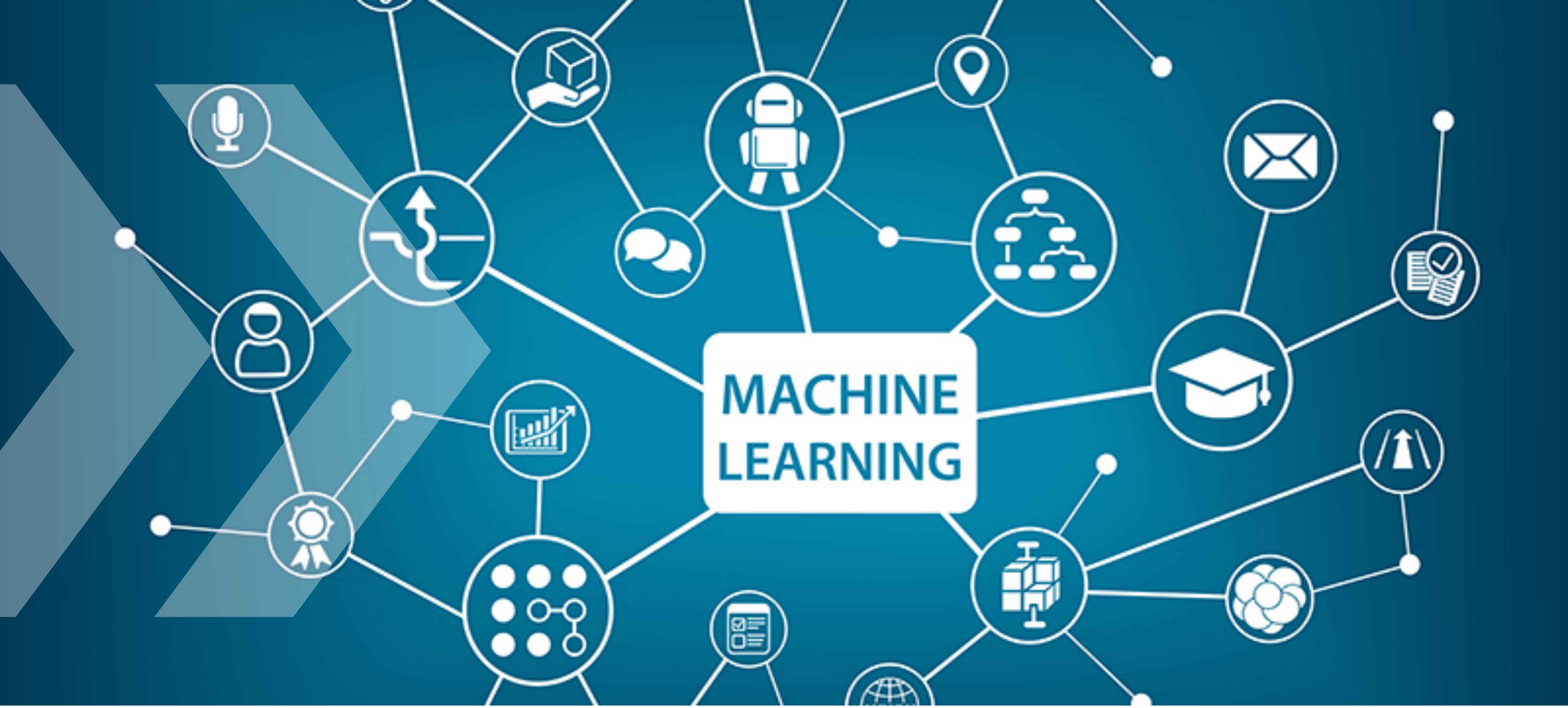


CS480/680: Intro to ML

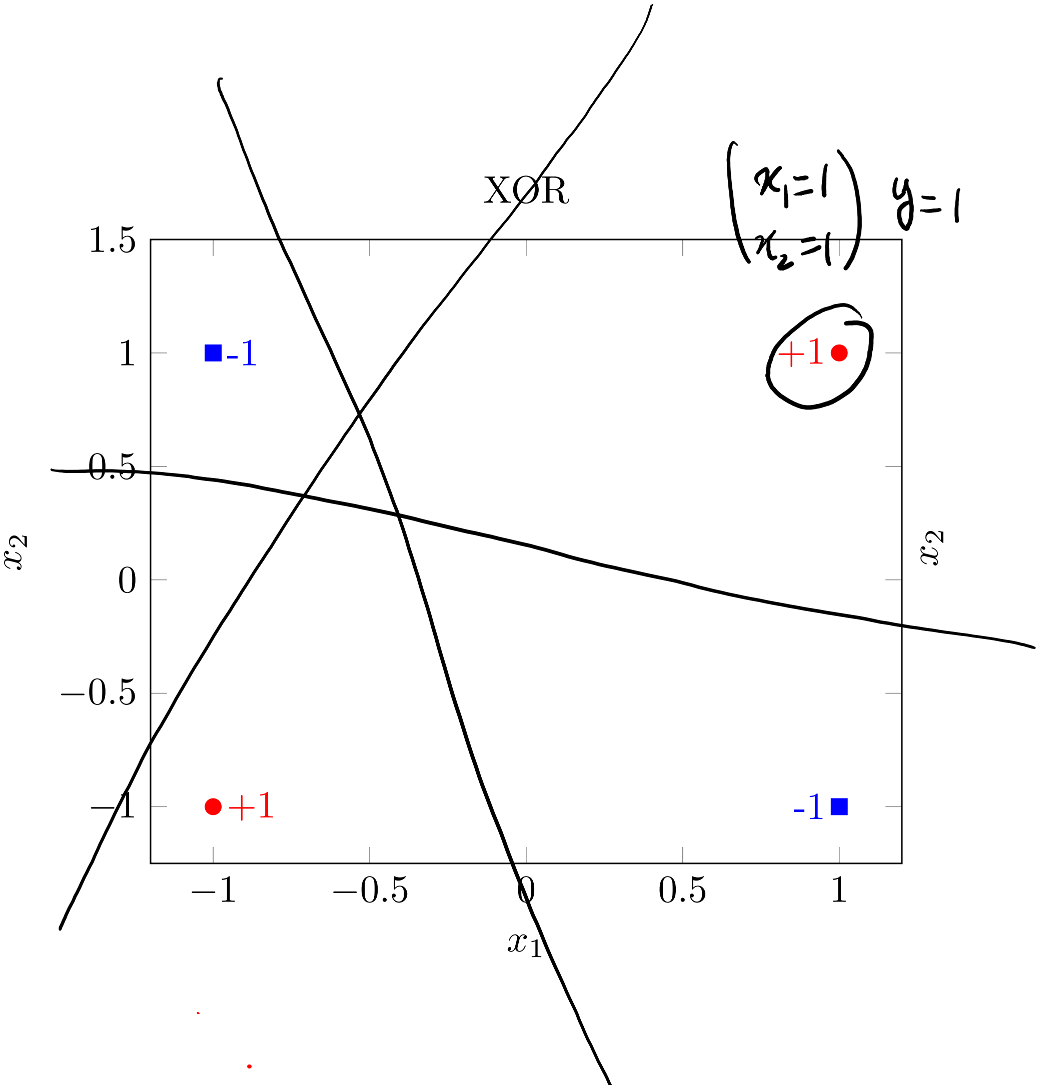
Lecture 09: Kernels

UNIVERSITY OF WATERLOO

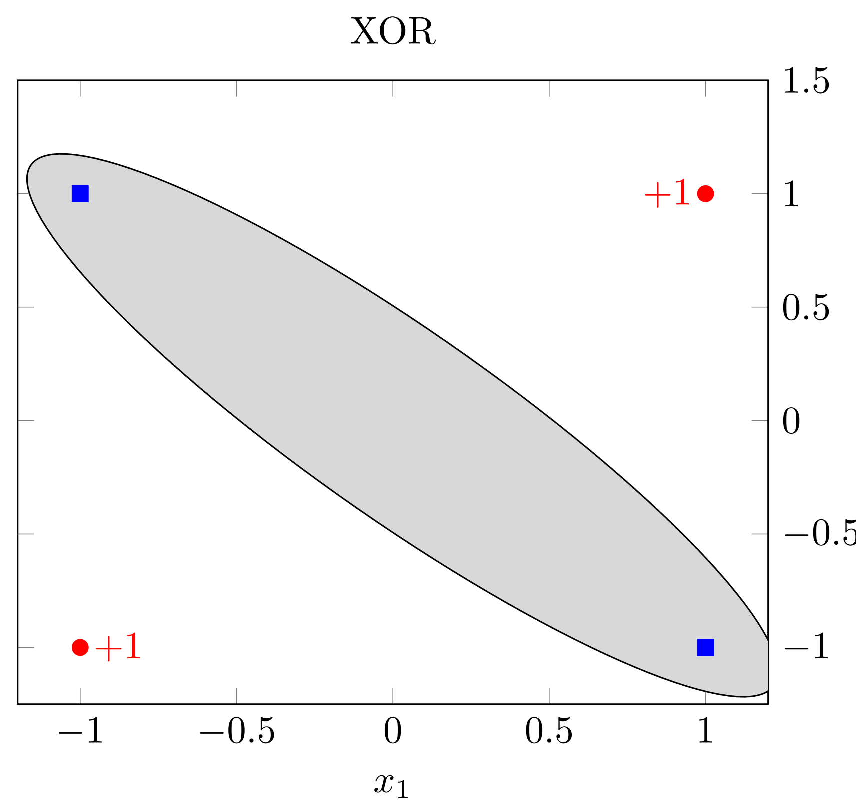
Outline

- Feature map
- Kernels
- The Kernel Trick

XOR



XOR
1.5
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
x_1
x_2
-1
+1
+1
-1
$x_1=1$
$x_2=1$
$y=1$



XOR
1.5
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
x_1
x_2
+1
+1

Quadratic classifier

$\sum_{ij} x_i Q_{ij} x_j$

$$\underline{\mathbf{x}^\top \boxed{Q} \mathbf{x} + \sqrt{2}\mathbf{x}^\top \boxed{\mathbf{p}} + \boxed{\gamma} \geq 0}$$

Weights
(to be learned)

$$\hat{y} = \mathrm{sign}(f(\mathbf{x}))$$

The power of lifting

$tr(AB) = tr(BA)$

$tr(x^\top Q x) = \mathbf{x}^\top Q \mathbf{x} + \sqrt{2}\mathbf{x}^\top \mathbf{p} + \gamma \geq 0$

$\|$

$tr(Q \cdot x x^\top)$

$R^d \Rightarrow R^{d*d+d+1}$

$\langle Q, x x^\top \rangle$

$$\| \quad \mathbf{w}^\top \phi(\mathbf{x}) \geq 0$$

$\phi(x)^\top w \geq 0$

Feature map

$$\phi(\mathbf{x}) = \begin{bmatrix} \overrightarrow{\mathbf{x}\mathbf{x}^\top} \\ \sqrt{2}\mathbf{x} \\ 1 \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} \vec{Q} \\ \mathbf{p} \\ \gamma \end{bmatrix}$$

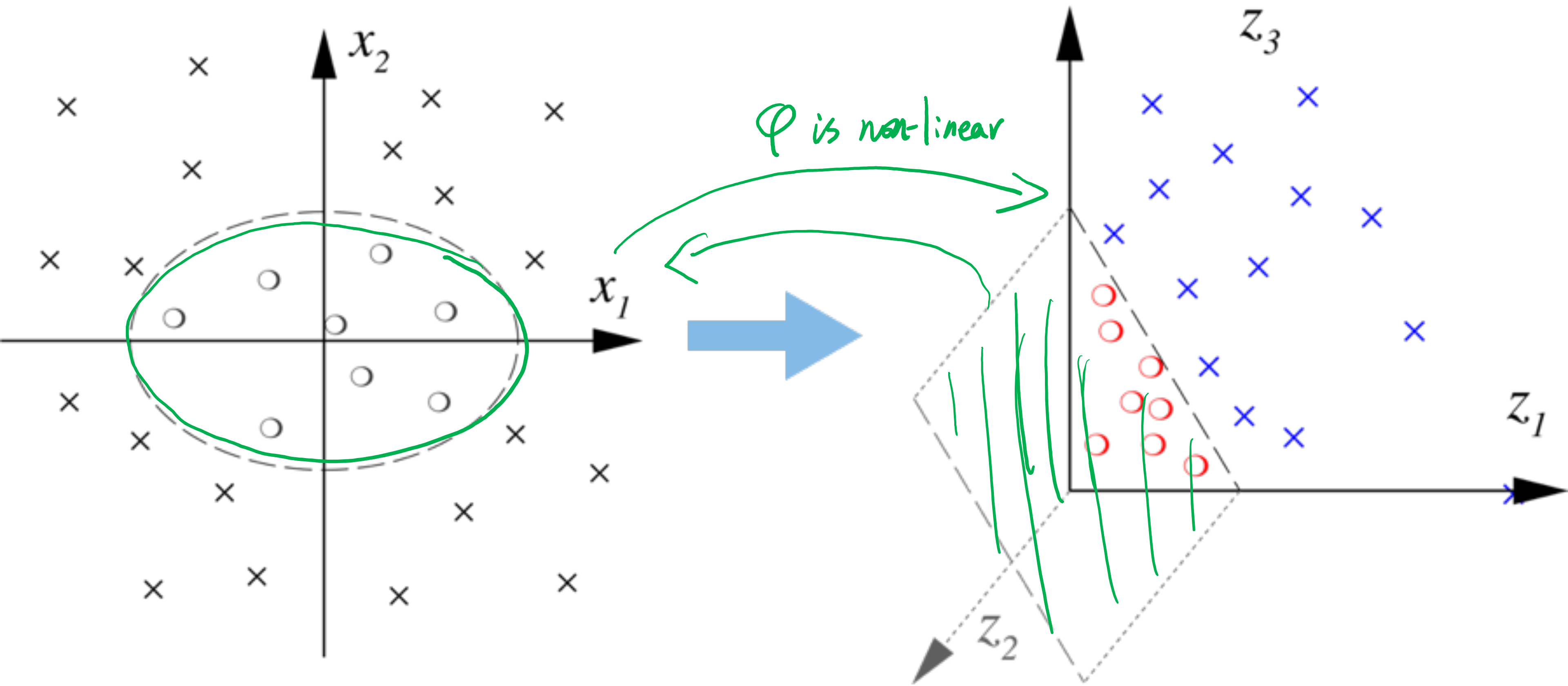
Example

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

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

Feature map may not be unique

Does it work?



Linear hyerplane in the feature space corresponds to nonlinear boundary in the original space!

And (almost) vice versa!!

UNIVERSITY OF WATERLOO

Curse of dimensionality?

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

computation in this space now

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

But, all we need is the dot product !!!

$$\underbrace{\phi(\mathbf{x})^\top \phi(\mathbf{x}')}_{\text{dot prod between } x \,\&\, x'} = (\mathbf{x}^\top \mathbf{x}')^2 + 2\mathbf{x}^\top \mathbf{x}' + 1$$

$$= (\mathbf{x}^\top \mathbf{x}' + 1)^2$$

- This is still computable in O(d)!

LHS: $O(d^2)$ RHS: $O(d)$

UNIVERSITY OF WATERLOO

Feature transform

$$\phi : \mathbf{R}^d \to \mathbf{R}^h$$ $h >> d$

- NN: learn ϕ simultaneously with **w**
- Here: choose a nonlinear ϕ so that for some *f* : R→R

$$\phi(\mathbf{x})^\top \phi(\mathbf{x}') = \boxed{f(\mathbf{x}^\top \mathbf{x}')}$$

save computation

UNIVERSITY OF WATERLOO

Outline

- Feature map
- Kernels
- The Kernel Trick

Reverse engineering

- Start with some function $k : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$, s.t. exists feature transform ϕ with

$$\phi(\mathbf{x})^\top \phi(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}')$$

- As long as *k* is efficiently computable, don't care the dim of ϕ (could be infinite!)

- Such *k* is called a (reproducing) kernel.

Examples

- Polynomial kernel

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}')^p$$

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + 1)^p$$

- Gaussian Kernel

sin θ

$\exp(t) = \sum_k \frac{t^k}{k!}$

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/\sigma)$$

$= \varphi(x)^\top \varphi(x')$

- Laplace Kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2/\sigma)$$

- Matérn Kernel

$$\frac{1}{2^{\nu-1}\Gamma(\nu)}\Big(\frac{2\sqrt{\nu}\|\mathbf{x} - \mathbf{x}'\|_2}{\theta}\Big)^\nu H_\nu\Big(\frac{2\sqrt{\nu}\|\mathbf{x} - \mathbf{x}'\|_2}{\theta}\Big)$$

Verifying a kernel

For any n, for any $x_1, x_2, \ldots, x_n$, the kernel matrix K with

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \qquad n \times n$$

is symmetric and positive semidefinite ($K \in \mathbb{S}^d_+$)

- Symmetric: $K_{ij} = K_{ji}$
- Positive semidefinite (PSD): for all $\boldsymbol{\alpha} \in \mathbf{R}^n$

$$\boldsymbol{\alpha}^\top K \boldsymbol{\alpha} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_{ij} \geq 0$$

Kernel calculus

- If k is a kernel, so is λk for any $\lambda \geq 0$

- If k_1 and k_2 are kernels, so is k_1+k_2
  - k_1 with φ_1, k_2 with φ_2 $\exists \varphi_i : k_i(x, x') = \varphi_i(x)^T \varphi_i(x')$, $i \in \{1, 2\}$
  - k_1+k_2 with ??

define $\varphi = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}$, then $\varphi(x)^T \varphi(x') = k_1(x, x') + k_2(x, x')$

- If k_1 and k_2 are kernels, so is $k_1 k_2$

Outline

- Feature map
- Kernels
- The Kernel Trick

Kernel SVM (dual)

$$\min_{C \geq \alpha \geq 0} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boxed{K_{ij}} - \sum_{i=1}^{n} \alpha_i$$

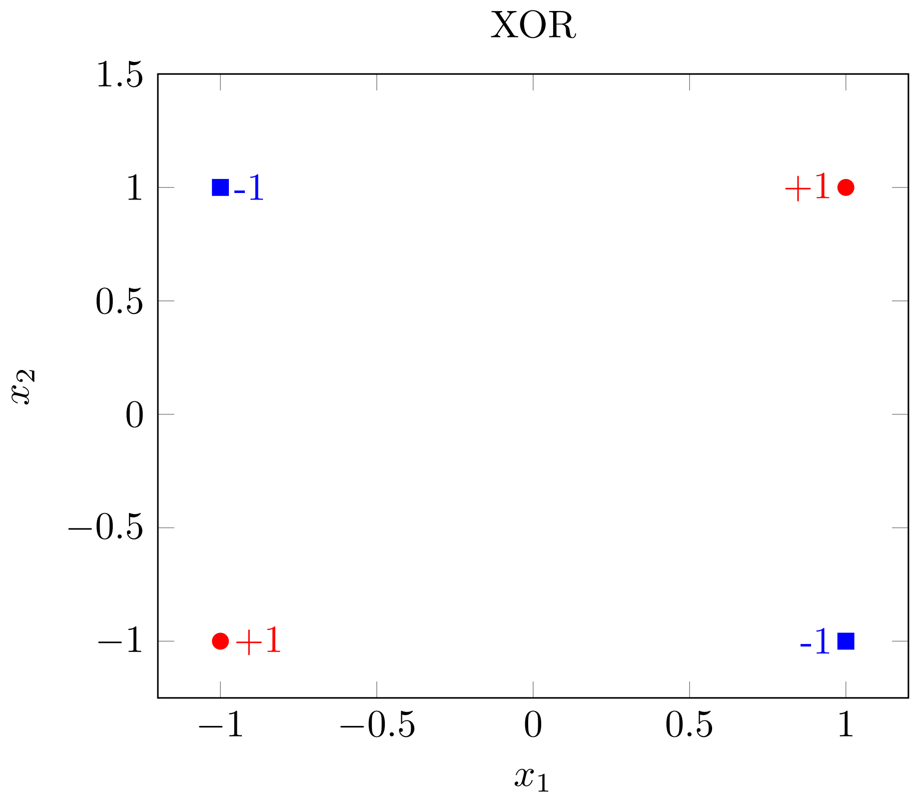
$$\text{s.t.} \sum_i \alpha_i y_i = 0$$

$x_i^\top x_j \to \varphi(x_i)^\top \varphi(x_j)$
$= \kappa(x_i, x_j)$
$= K_{ij}$

With **α**, $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \boxed{\phi(\mathbf{x}_i)}$ but **ϕ** is implicit…

UNIVERSITY OF WATERLOO

Does it work?



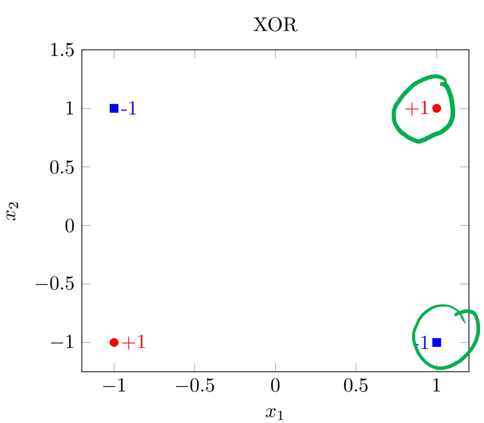
$$\phi(\mathbf{x}) = [x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1]$$

$$\boxed{k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)^2}$$

Does it work?

$$\min_{\cancel{C \ge \alpha \ge 0}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K_{ij} - \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \cancel{\sum_i \alpha_i y_i = 0}$$



$\begin{pmatrix}1\\1\end{pmatrix}$

$K_{11} = \left[\begin{pmatrix}1\\1\end{pmatrix}^T\begin{pmatrix}1\\1\end{pmatrix}+1\right]^2 = 9$

$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

$$\boxed{k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^2}$$

UNIVERSITY OF WATERLOO

Does it work?

$$9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 1$$

$$-\alpha_1 + 9\alpha_2 + \alpha_3 - \alpha_4 = 1$$

$$-\alpha_1 + \alpha_2 + 9\alpha_3 - \alpha_4 = 1$$

$$\alpha_1 - \alpha_2 - \alpha_3 + 9\alpha_4 = 1$$

$w = \sum_i \alpha_i y_i \phi(x_i)$

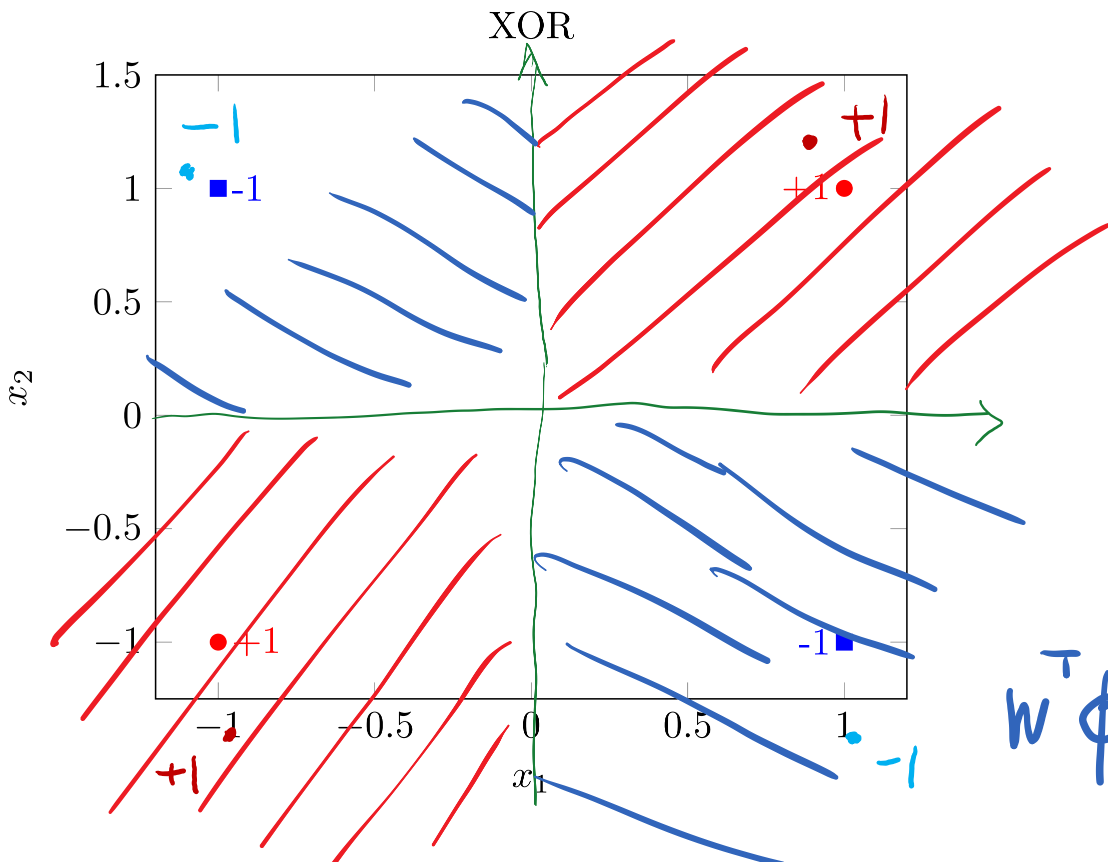
$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/8$$

$$\mathbf{w} = [0, \frac{1}{\sqrt{2}}, 0, 0, 0, 0]$$

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

UNIVERSITY OF WATERLOO

Does it work?



$w^T\phi(x) \geq 0$

$w^T\phi(x) \leq 0$

$$\mathbf{w}^\top \phi(\mathbf{x}) = x_1 x_2 \begin{matrix} \geq 0 \\ \leq 0 \end{matrix}$$

UNIVERSITY OF WATERLOO

Testing

- Given test sample **x**', how to perform testing?

$$\mathbf{w}^\top \phi(\mathbf{x}') = \sum_{i=1}^{n} \alpha_i y_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}')$$

w

$$= \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}')$$

No explicit access to ϕ, again!

dual variables

training set

kernel

Tradeoff

- Previously: training O(nd), test O(d)
- Kernel: training $O(n^2d)$, test O(nd)
- Nice to avoid explicit dependence on h (could be inf)
- But if n is also large… ~~(maybe later)~~

Learning the kernel (Lanckriet et al.'04)

$$\min_{C \geq \alpha \geq 0} \boxed{\max_{\zeta \geq 0}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left[\sum_{s=1}^{t} \zeta_s K_{ij}^{(s)} \right] - \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0$$

- Nonnegative combination of *t* pre-selected kernels, with coefficients ζ simultaneously learned

Logistic regression revisited

$$\min_{\mathbf{w} \in \mathbf{R}^d} \sum_i \log(1 + e^{-y_i \mathbf{w}^\top \mathbf{x}_i}) + \lambda \|\mathbf{w}\|_2^2$$

kernelize

$$\min_{\mathbf{w} \in \mathbf{R}^h} \sum_i \log(1 + e^{-y_i \mathbf{w}^\top \phi(\mathbf{x}_i)}) + \boxed{\lambda \|\mathbf{w}\|_2^2}$$

$\sum_j \alpha_j y_j \underbrace{\phi(x_j)^\top \phi(x_i)}_{K_{ji}}$

Representer Theorem (Wabha, Schölkopf, Herbrich, Smola, Dinuzzo, Yu…).

The optimal w has the following form:

$$\mathbf{w} = \sum_i \alpha_i y_i \phi(\mathbf{x}_i)$$

UNIVERSITY OF WATERLOO

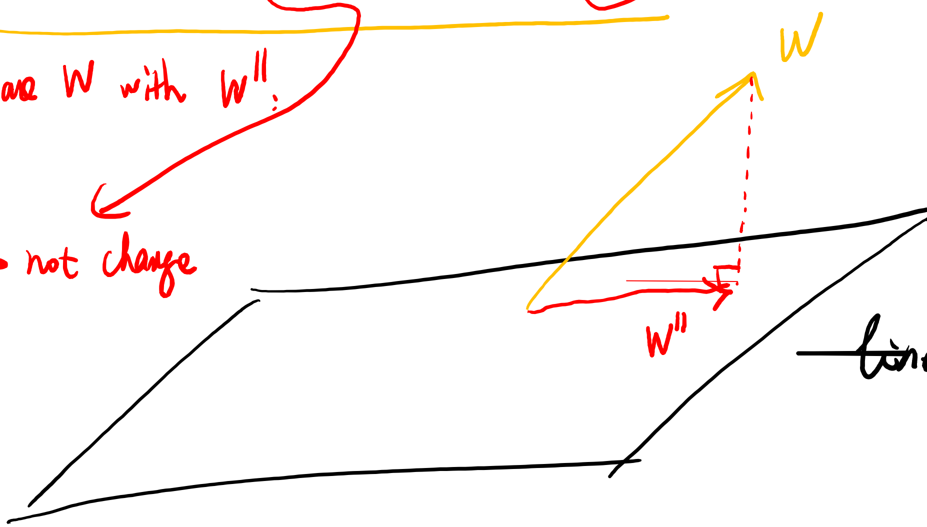
Orthogonal decomposition

$$\min_{w} \underbrace{\sum_i \ell(y_i w^T \phi(x_i))}_{} + \underbrace{\frac{\lambda}{2}\|w\|_2^2}_{}$$

decrease

replace w with w'':

~~does~~ not change



~~linear~~ span ~~of~~ $\{\phi(x_i)\}$

$$w = \sum_i \beta_i \phi(x_i) = \sum_i \underbrace{\alpha_i y_i}_{\beta_i} \phi(x_i)$$

Questions?

UNIVERSITY OF WATERLOO