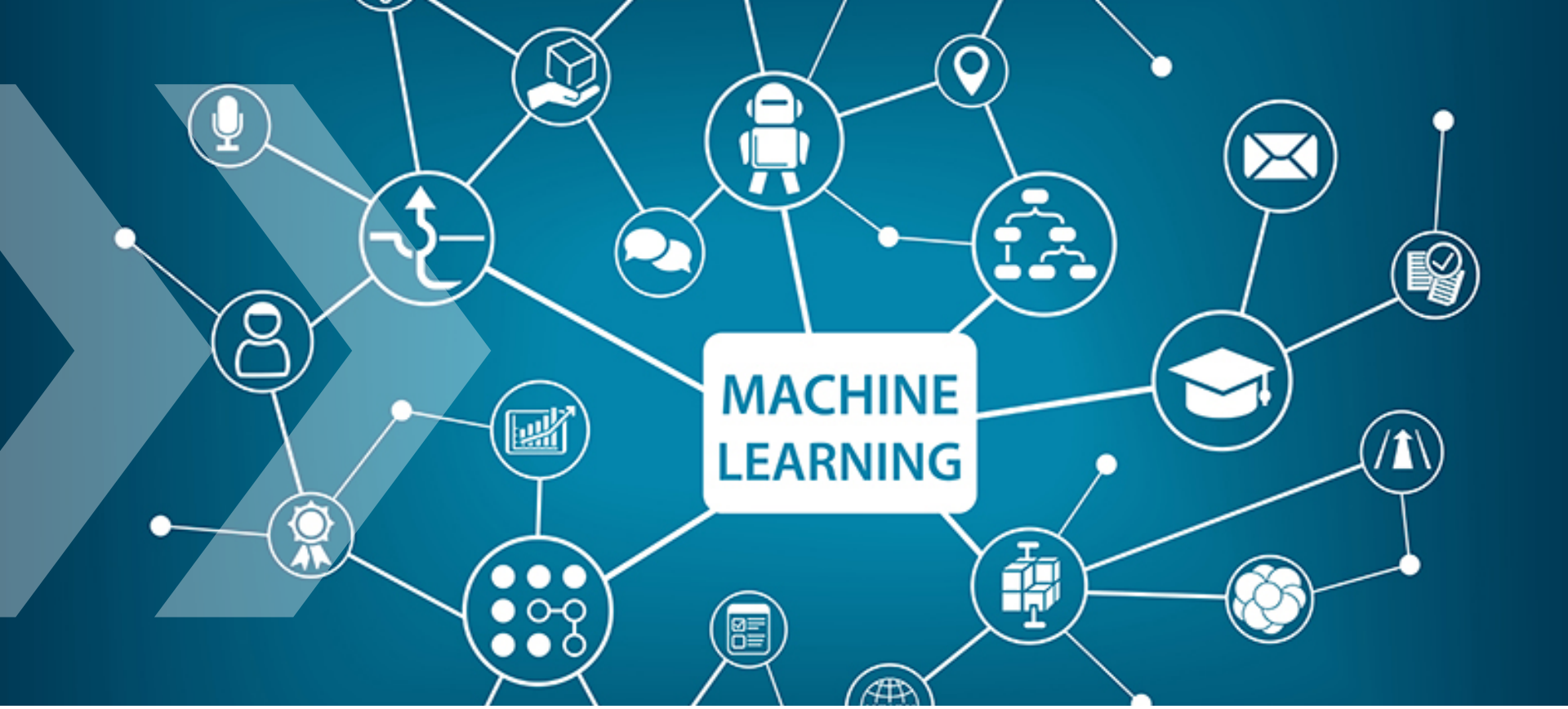


CS489/698: Intro to ML

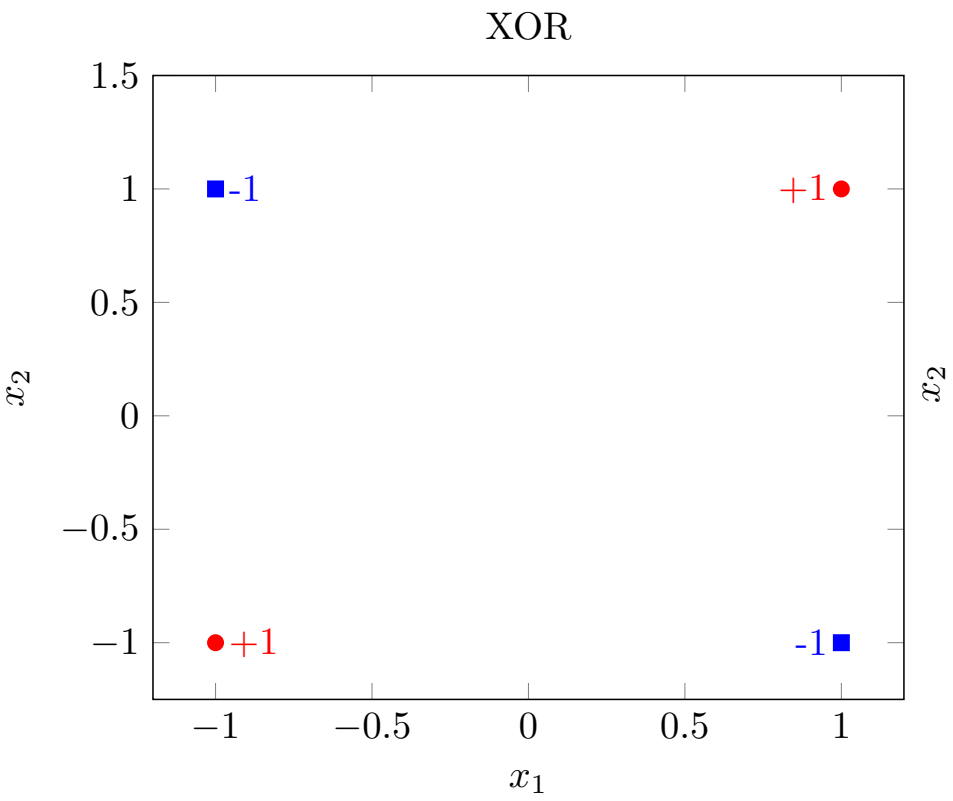
Lecture 08: Kernels

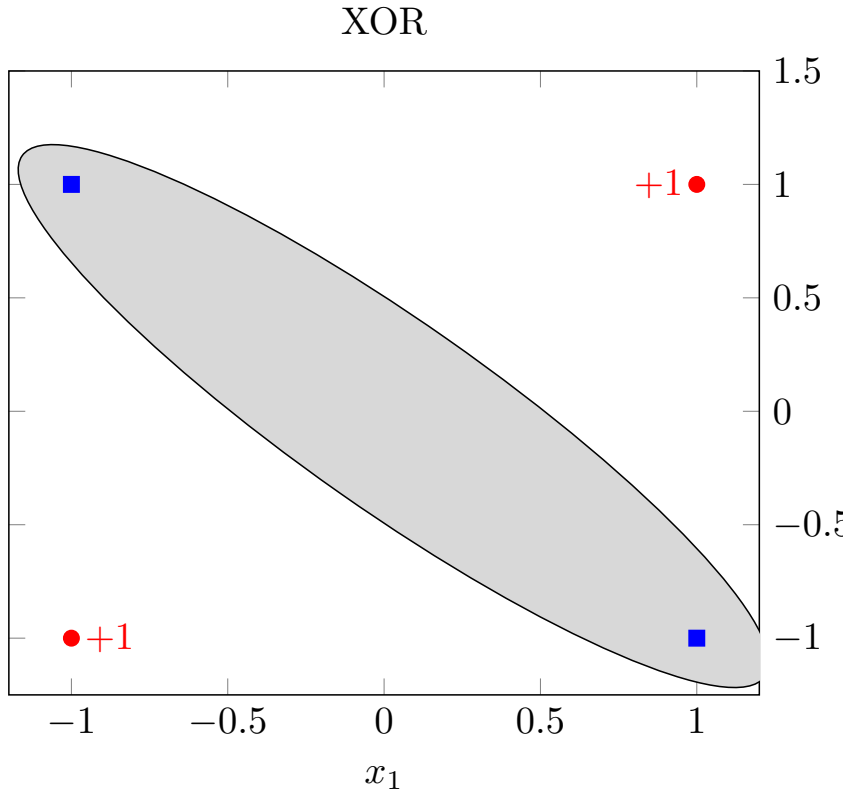
UNIVERSITY OF WATERLOO

Outline

- Feature map
- Kernels
- The Kernel Trick
- Advanced

XOR revisited





Quadratic classifier

Weights
(to be learned)

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

$$\hat{y} = f(\mathbf{x}) = 1$$

The power of lifting

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

$$\Updownarrow$$

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

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

Feature map

$$\phi(\mathbf{x}) = \begin{bmatrix} \overrightarrow{\mathbf{x}\mathbf{x}^\top} \\ \sqrt{2}\mathbf{x} \\ 1 \end{bmatrix} \begin{bmatrix} x_1\vec{x} \\ x_2\vec{x} \\ \vdots \\ x_d\vec{x} \end{bmatrix}$$

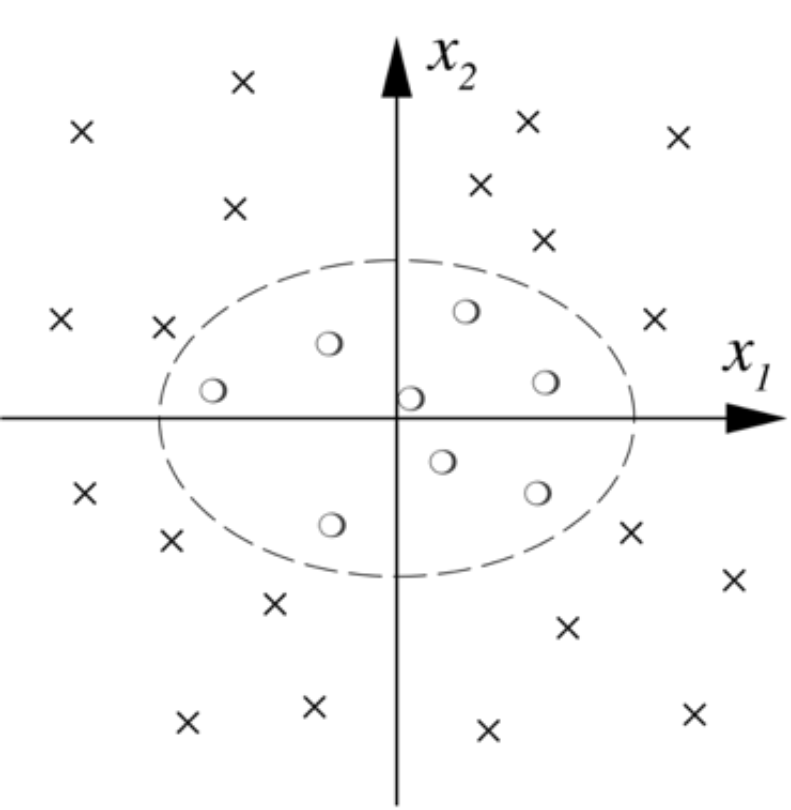
$$\mathbf{w} = \begin{bmatrix} \overrightarrow{Q} \\ \mathbf{p} \\ \gamma \end{bmatrix}$$

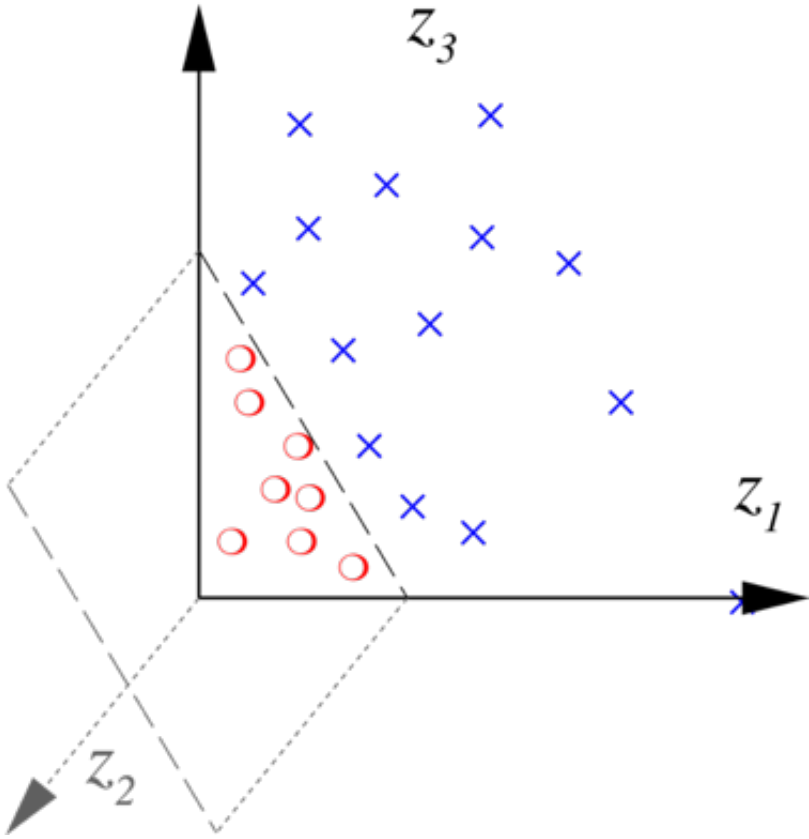
Example

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

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

Does it work?





Curse of dimensionality?

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

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 !!!

$$\begin{aligned} \phi(\mathbf{x})^\top \phi(\mathbf{x}') &= (\mathbf{x}^\top \mathbf{x}')^2 + 2\mathbf{x}^\top \mathbf{x}' + 1 \\ &= (\mathbf{x}^\top \mathbf{x}' + 1)^2 \end{aligned}$$

- This is still computable in O(d)!

Feature transform

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

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

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

save computation

Outline

- Feature map
- Kernels
- The Kernel Trick
- Advanced

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

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

- Laplace Kernel

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

- Matérn Kernel

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

Verifying a kernel

$\kappa: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\exists \varphi$ s.t. $\forall x, x'$, $\kappa(x,x') = \varphi(x)^\top \varphi(x')$

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

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

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

necessity: (1) $K_{ij} = \varphi(x_i)^\top \varphi(x_j) = \varphi(x_j)^\top \varphi(x_i) = K_{ji}$

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

(2) $\sum_i \sum_j \alpha_i \alpha_j \varphi(x_i)^\top \varphi(x_j) = \left\| \sum_i \alpha_i \varphi(x_i) \right\|^2 \geq 0$

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

Kernel calculus

- If k is a kernel, so is λk for any λ ≥ 0
- If k_1 and k_2 are kernels, so is k_1+k_2
- If k_1 and k_2 are kernels, so is k_1k_2

$\sum_i \gamma_i K_i$ is a kernel if K_i is and $\gamma_i \geq 0$

Outline

- Feature map
- Kernels
- The Kernel Trick
- Advanced

Kernel SVM (dual)

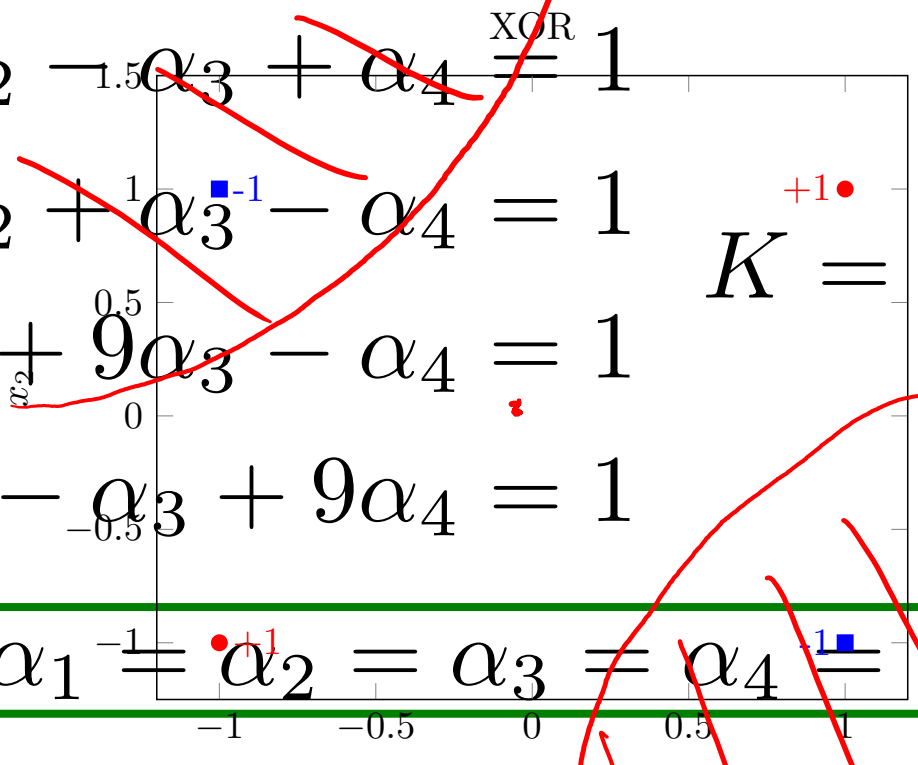
$$\min_{C \geq \alpha \geq 0} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K_{ij} - \sum_{i=1}^{n} \alpha_i$$

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

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

Does it work?

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



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

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

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

$$\mathbf{w}^\top \phi(\mathbf{x}) = -x_1 x_2$$

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

Testing

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

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

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

No explicit access to φ, again!

dual variables

training set

kernel

Tradeoff

- Previously: training O(nd), test O(d)
- Kernel: training $O(n^2)$, test O(n)
- 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)

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

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

- 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

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

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

The optimal w has the following form:

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

Kernel Logistic Regression (primal)

$$\min_{\boldsymbol{\alpha}\in\mathbf{R}^n} \sum_i \log(1+e^{-y_i \boldsymbol{\alpha}^\top K_{:i}}) + \lambda \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

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

$$\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}_t - \eta_t [\nabla^2 f(\boldsymbol{\alpha}_t)]^{-1} \cdot \nabla f(\boldsymbol{\alpha}_t)$$

$$\nabla^2 f(\boldsymbol{\alpha}_t) = \sum_i p_i(1-p_i) K_{:i} K_{:i}^\top + 2\lambda K$$

$$\nabla f(\boldsymbol{\alpha}_t) = K^\top (\mathbf{p} - \frac{\mathbf{y}+1}{2}) + 2\lambda K \boldsymbol{\alpha}_t$$

$\mathbf{y} \in \{-1,1\}^n$

$$p_i = \frac{1}{1+e^{-\boldsymbol{\alpha}_t^\top K_{:i}}}$$

uncertain predictions get bigger weight

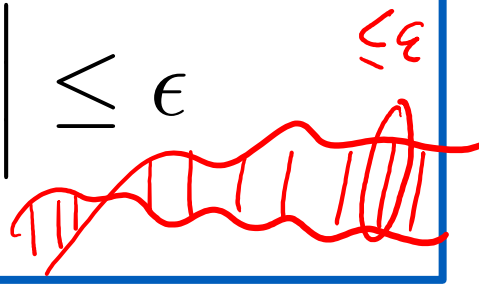
Outline

- Feature map
- Kernels
- The Kernel Trick
- Advanced

Universal approximation (Micchelli, Xu, Zhang'06)

Universal kernel. For any compact set Z, for any continuous function $f : Z \to \mathbf{R}$, for any $\varepsilon > 0$, there exist $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in Z and $\alpha_1, \alpha_2, \ldots, \alpha_n$ in **R** such that

$$\max_{\mathbf{x} \in Z} \left| f(\mathbf{x}) - \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i) \right| \leq \epsilon$$



decision boundary

kernel methods

Example. The Gaussian kernel. $\exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / \sigma)$

Kernel mean embedding (Smola, Song, Gretton, Schölkopf, ...)

$$\mathbf{P} \mapsto \mu_{\mathbf{P}} := \mathbf{E}(\phi(X)), \quad \text{where } X \sim \mathbf{P}$$

feature map of some kernel

- Characteristic kernel: the above mapping is 1-1
- Completely preserve the information in the distribution P
- Lots of applications

Questions?