CIS 5200: MACHINE LEARNING

KERNELS

Surbhi Goel

Content here draws from material by Jake/Shivani (UPenn), Christopher De Sa (Cornell)

Spring 2023

OUTLINE - TODAY

- Recap of SVMs
- Function Maps
- Kernel Functions
- Kernelization
- Demo!

SVM - SOFT-MARGIN

Primal:

$$\min_{w,b} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

such that $y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in [m]$

$\xi_i \geq 0, \forall i \in [m]$

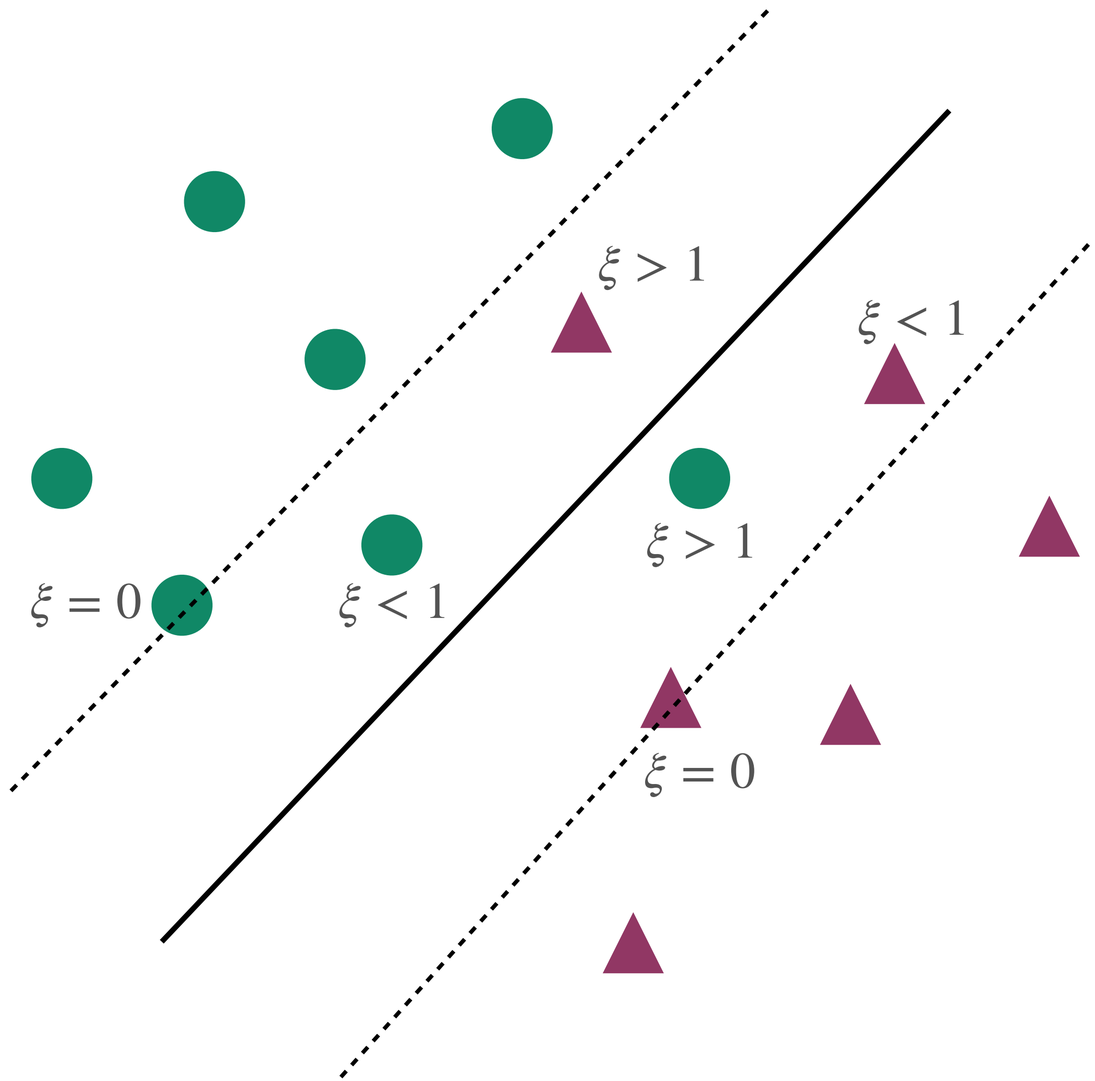
Slack

Dual:

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i^\top x_j) + \sum_{i=1}^{m}\alpha_i$$

such that $\sum_{i=1}^{m}\alpha_i y_i = 0$

$0 \leq \alpha_i \leq C, \forall i \in [m]$



SOFT-SVM - LOSS MINIMIZATION VIEW

$$\min_{w,b} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

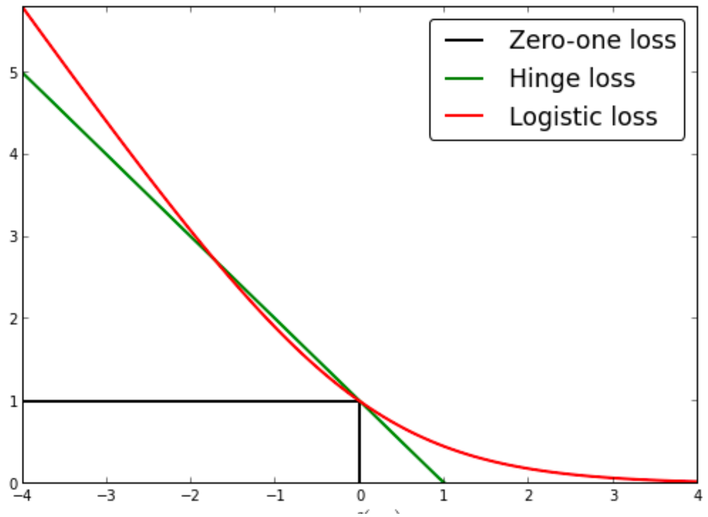
$$\text{such that} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in [m]$$

$$\xi_i \geq 0, \forall i \in [m]$$

Is equivalent to the following loss minimization problem for $C = \frac{1}{2\lambda m}$:

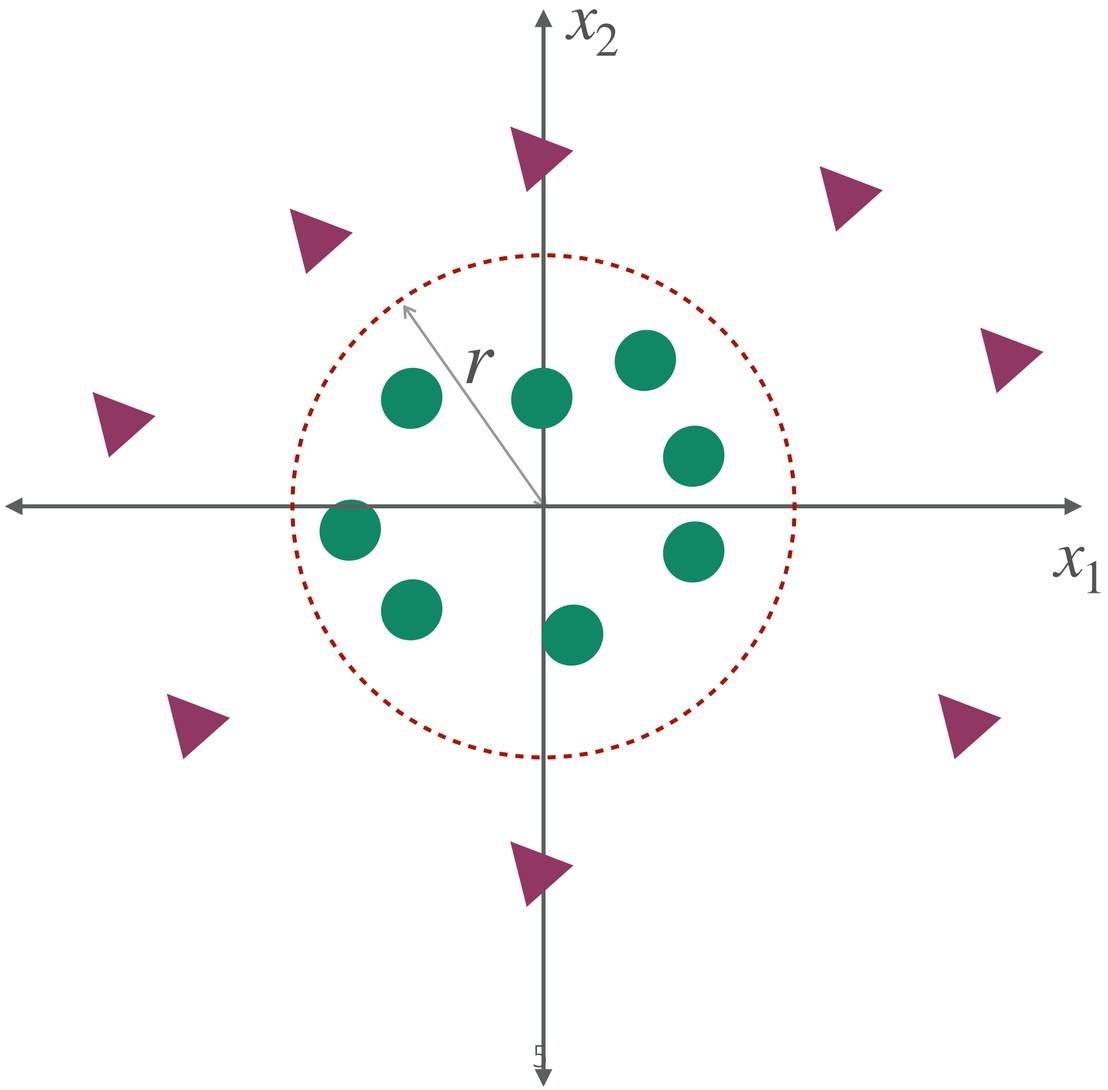
$$\min_{w,b} \frac{1}{m}\sum_{i=1}^{m}\max(0, 1 - y_i(w^\top x_i + b)) + \lambda\|w\|^2$$

ℓ_2-regularized hinge loss minimization



NON-SEPARABLE

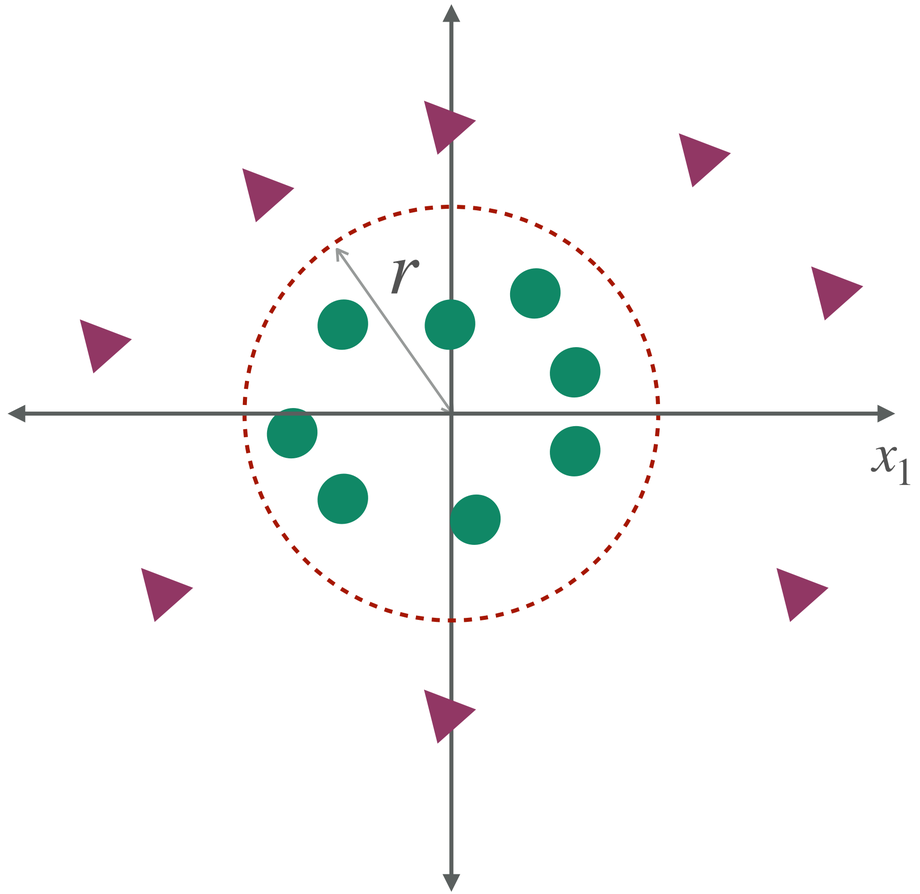
What can we do if data is like this?



FEATURE MAP - MAP TO HIGHER DIMENSIONS

Map data into to a higher dimensional space using feature map ϕ

$$x \mapsto \phi(x)$$

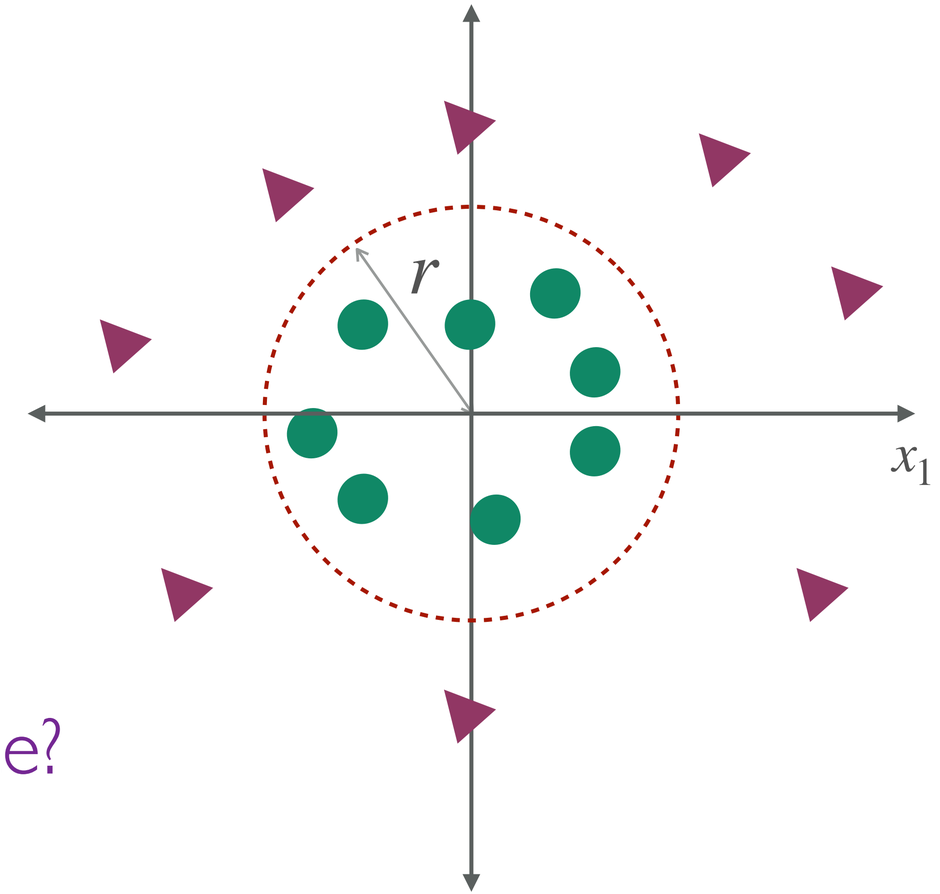


What features should we use?

FEATURE MAP - MAP TO HIGHER DIMENSIONS

Consider the following feature map:

$$\phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

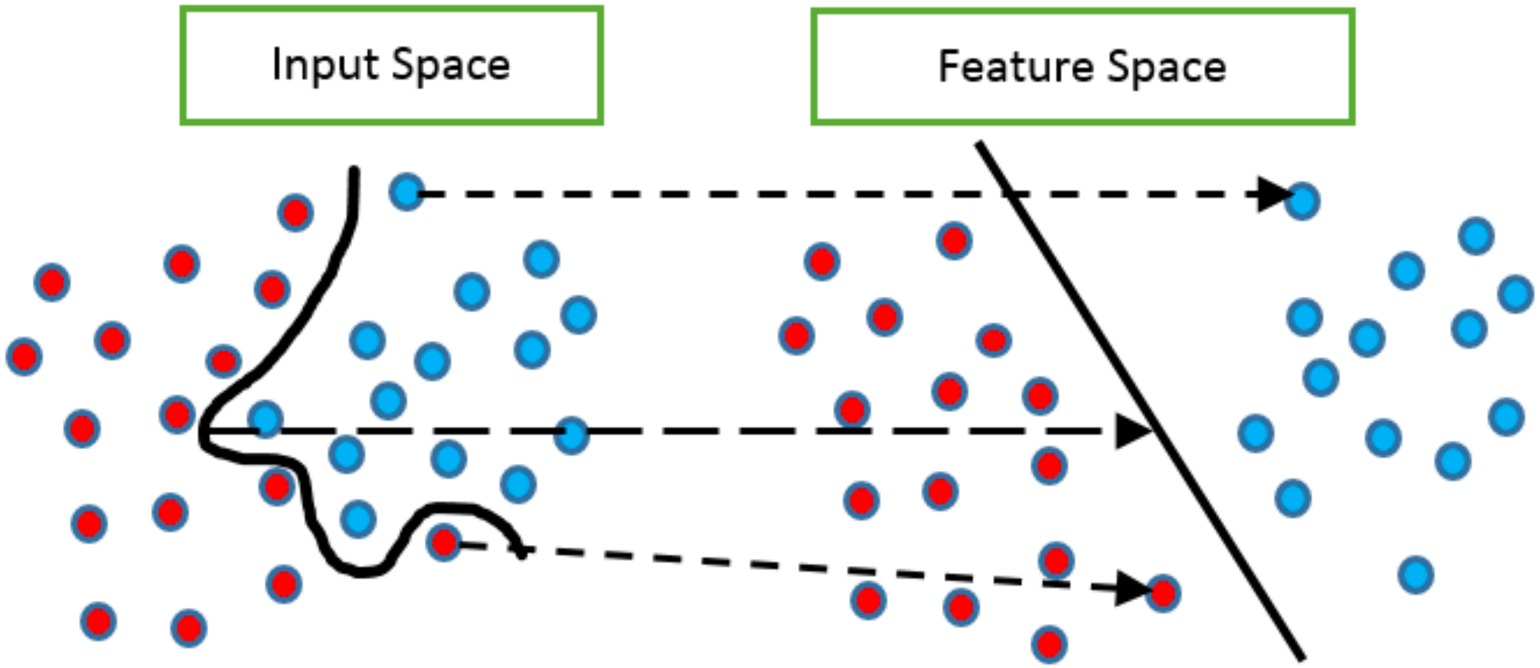


Is the data linearly separable in this feature space?

Let $w = [0,0,1,1]^\top$ and $b = r^2$, then we have

$$w^\top \phi(x) + b = x_1^2 + x_2^2 - r^2$$

FEATURE MAP - LINEAR TO NON-LINEAR



Training data: $\{(x_1, y_1), \ldots, (x_m, y_m)\} \mapsto \{(\phi(x_1), y_1), \ldots, (\phi(x_m), y_m)\}$

Predictor function: Linear functions $w^\top \phi(x) + b$

FEATURE MAP - CHALLENGE

Consider the following feature map:

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1^2 \\ x_1 x_2 \\ \vdots \\ x_d^2 \end{bmatrix}$$

Extension of the previous one to d dimensions

What is the dimension of this map?

$$D = (d+1)^2$$

What if we take all monomials to degree r?

$$D = (d+1)^r$$

This is huge!

RECALL - SOFT-SVM

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i^\top x_j) + \sum_{i=1}^{m}\alpha_i$$

such that

$$\sum_{i=1}^{m}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C, \forall i \in [m]$$

With feature map:

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j (\phi(x_i)^\top \phi(x_j)) + \sum_{i=1}^{m}\alpha_i$$

such that

$$\sum_{i=1}^{m}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C, \forall i \in [m]$$

We only need to compute inner products $\phi(x_i)^\top \phi(x_j)$

KERNEL TRICK - EXAMPLE

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1^2 \\ x_1 x_2 \\ \vdots \\ x_d^2 \end{bmatrix}$$

Let's compute inner product:

$$\phi(x)^\top \phi(x') = 1 + x^\top x' + (x^\top x')^2$$

What is the computational cost of this?

KERNEL FUNCTIONS

A kernel is a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies:

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

for some feature map ϕ that maps $\mathcal{X}$ to some inner product space $\mathcal{V}$.

For data $x_1, \ldots, x_m$ the kernel matrix $K \in \mathbb{R}^{m \times m}$

$$K_{ij} = k(x_i, x_j)$$

A kernel k is **valid** if for any $x_1, \ldots, x_m$: K is symmetric and positive semi-definite

$K = K^\top$

For any $x, x^\top K x \geq 0$

All eigenvalues are non-negative

KERNELS - EXAMPLES

- **Linear**:

$$k(x, x') = x^\top x'$$

- **Polynomial**: for degree r

$$k(x, x') = (1 + x^\top x')^r$$

- **Gaussian/Random Basis Function (RBF)**: for some parameter $\sigma > 0$

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

LETS KERNELIZE!

- Show that the solution to your problem lies in the span of the training points, $w = \sum_{i=1}^{m} \alpha_i x_i$
- Rewrite the algorithm and the predictor so that all training or test points are only accessed in inner-products ($x_i^\top x_j$) with other points
- Replace $x_i^\top x_j \rightarrow k(x_i, x_j)$ everywhere

EXAMPLE - SOFT-SVM

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i^\top x_j) + \sum_{i=1}^{m}\alpha_i$$

such that $\sum_{i=1}^{m}\alpha_i y_i = 0$

$$0 \le \alpha_i \le C, \forall i \in [m]$$

Optimal $w = \sum_{i=1}^{m}\alpha_i y_i x_i$ and prediction function is

$$\text{sign}(w^\top x + b) = \text{sign}\left(\sum_{i=1}^{m}\alpha_i y_i x_i^\top x + b\right)$$

EXAMPLE - KERNEL SVM

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{m}\alpha_i$$

such that $\sum_{i=1}^{m}\alpha_i y_i = 0$

$$0 \leq \alpha_i \leq C, \forall i \in [m]$$

Optimal $w = \sum_{i=1}^{m}\alpha_i y_i \phi(x_i)$ and prediction function is

$$\text{sign}(w^\top \phi(x) + b) = \text{sign}\left(\sum_{i=1}^{m}\alpha_i y_i k(x_i, x) + b\right)$$

EXAMPLE - PERCEPTRON

Algorithm 1: Perceptron

Initialize $w_1 = 0 \in \mathbb{R}^d$

for $t = 1, 2, \ldots$ **do**

  if $\exists i \in [m]$ *s.t.* $y_i \neq \text{sign}\left(w_t^\top x_i\right)$ **then** update $w_{t+1} = w_t + y_i x_i$

  else output w_t

end

Update is always in the feature space, so $w_* = \sum_{i=1}^{m} \alpha_i x_i$ for some $\alpha \in \mathbb{R}^m$

Can we write the algorithm in terms of α?

EXAMPLE - KERNEL PERCEPTRON

Algorithm 2: Perceptron - Dual

Initialize $\alpha_1 = 0 \in \mathbb{R}^d$

for $t = 1, 2, \ldots$ **do**

  if $\exists i \in \{1, \ldots, m\}$ *s.t.* $y_i \neq \text{sign}\left(\sum_{j=1}^{m} \alpha_{tj} x_j^\top x_i\right)$ **then** update $\alpha_{(t+1)i} = \alpha_{ti} + y_i$

  else output α_t

end

Now we can kernelize this since it only depends on inner products!

EXAMPLE - RIDGE REGRESSION

$$\min_{w} \quad \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda\|w\|_2^2 = \frac{1}{m}\|Y - Xw\|^2 + \lambda\|w\|_2^2$$

Can w be expressed as a linear combination of the input datapoints?

Proof by contradiction!

We have $w = \sum_{i=1}^{m} \alpha_i x_i = X^\top \alpha$ for some α

EXAMPLE - RIDGE REGRESSION

$$\min_{w} \quad \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda\|w\|_2^2 = \frac{1}{m}\|Y - XX^\top\alpha\|^2 + \lambda\alpha^\top XX^\top\alpha$$

Each element of $XX^\top$ is an inner product $x_i^\top x_j$ for some $i, j \in [m]$

Prediction is $w^\top x = \sum_{i=1}^{m} \alpha_i x_i^\top x = \alpha^\top Xx$

EXAMPLE - KERNEL RIDGE REGRESSION

$$\min_w \quad \frac{1}{m}\sum_{i=1}^{m}(y_i - w^\top \phi(x_i))^2 + \lambda\|w\|_2^2 = \frac{1}{m}\|Y - K\alpha\|^2 + \lambda\alpha^\top K\alpha$$

Here $K_{ij} = k(x_i, x_j)$ is the kernel/gram matrix

Prediction is $w^\top \phi(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) = \alpha^\top k_x$ where

$$k_x = [k(x, x_1) \ldots k(x, x_m)]^\top$$

DEMO

By folks at Cornell CS

POWER OF KERNELS

- Show that the solution to your problem lies in the span of the training points, $w = \sum_{i=1}^{m} \alpha_i x_i$

There is a general theorem called the Representer Theorem which tells us when this is true

- Rewrite the algorithm and the predictor so that all training or test points are only accessed in inner-products ($x_i^\top x_j$) with other points

- Replace $x_i^\top x_j \rightarrow k(x_i, x_j)$ everywhere for a valid kernel k

Super Powerful!

CHALLENGE

- How do we choose a good feature map ϕ?
- Feature map is the same for all inputs!

Can learn the feature map itself → deep learning!