LECTURE 19

THE KERNEL METHOD

- The proof of Grothendieck's Inequality in Lecture 18 is based on the existence of maps $\phi, \psi : \mathbb{R}^d \to \mathbb{R}^N$ s.t.

$$\langle \phi(u), \psi(v) \rangle = \sin\left(\frac{\pi}{2}\langle u, v\rangle\right) \quad \forall u, v \in \mathbb{R}^d \text{ unit}$$

realizes nonlinearity.

- This idea leads to the kernel trick in Machine Learning (ML)

Our first example of a ML task:

BINARY CLASSIFICATION

- Want: automatic cancer diagnostics

cell $\in \mathbb{R}^d$ → [computer] → diagnosis: 1 = cancer, −1 = no cancer

Each cell = d parameters, e.g. features of a cell (radius, perimeter, texture, symmetry, ...) $\in \mathbb{R}^d$

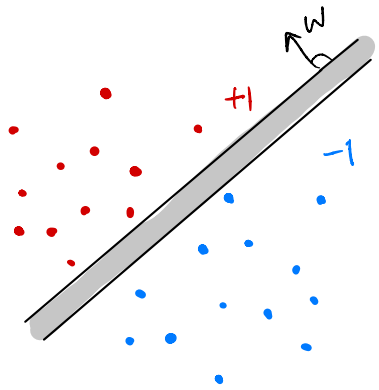
$\Rightarrow$ Want $f : \mathbb{R}^d \to \{\pm 1\}$.

$f(x) = y$

(x: test, y: prediction)

- Training data: (x_i, y_i), $i = 1, \dots, n$

= n cells that are already classified by doctors



- First, assume linear separation:

$$\exists w \in \mathbb{R}^d : \begin{cases} \langle w, x_i \rangle > 1 & \text{if } y_i = 1 \\ \langle w, x_i \rangle < -1 & \text{if } y_i = -1 \end{cases}$$

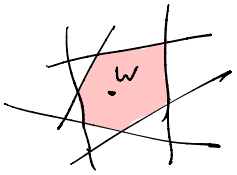
$$\Leftrightarrow \boxed{\langle w, x_i \rangle y_i > 1 \quad \forall i} \qquad (*)$$

- Learning: find $w \in \mathbb{R}^d$ that satisfies $(*)$

$(*)$ = Intersection of half-spaces

$\Rightarrow$ Linear (feasibility) program.

Tractable.



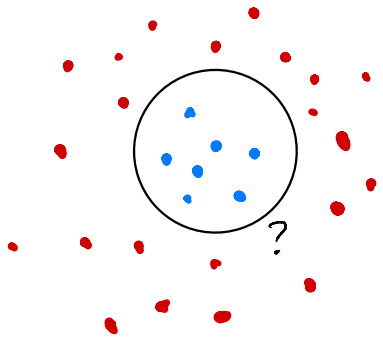
REMARK: generally, we include offset: $\langle w, x \rangle + b = 0$ but can include it into w: $\langle w \oplus b, x \oplus 1 \rangle = 0$

- Prediction: $\forall$ new cell $x \in \mathbb{R}^d$, diagnose it:

$$f(x) = \begin{cases} 1 \text{ (cancer)} & \text{if } \langle w, x \rangle > 1 \\ -1 \text{ (no cancer)} & \text{if } \langle w, x \rangle < -1 \end{cases} = \boxed{\operatorname{sign} \langle w, x \rangle}$$

- "Support Vector Machine" (SVM), "Perceptron"

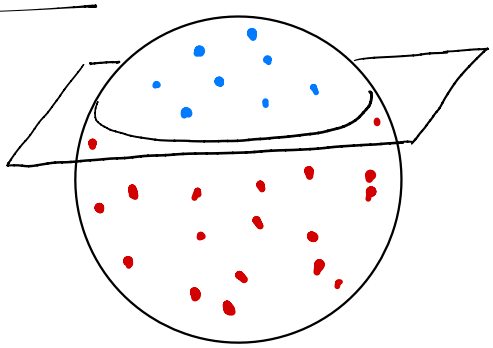
- What if there is NO linear separation? Transform:



$\mathbb{R}^d$

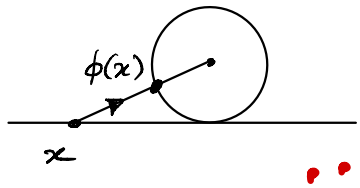
ϕ

"Feature map"



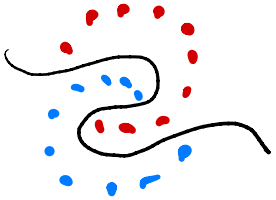
$\mathbb{R}^D$: "feature space"

e.g. stereographic projection



Transformed data is linearly separated

$\Rightarrow$ Proceed as Before.

- What if  ? How do we find ϕ?

NO NEED:

KERNEL TRICK

- Rewrite the learning algorithm in terms of inner products $\langle x_i, x_j \rangle$:

WLOG $w = \sum_{j=1}^{n} \alpha_j x_j$. $\quad \langle w, x_i \rangle = \sum_{j=1}^{n} \alpha_j \langle x_i, x_j \rangle$

Find $(\alpha_j)_{j=1}^n$: $\sum_{j=1}^{n} y_i \alpha_j \langle x_i, x_j \rangle > 0 \quad \forall i$	$\xrightarrow{\phi}$ transform	Find $(\alpha_j)_{j=1}^n$: $\sum_{j=1}^{n} y_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle > 0 \; \forall i$
Prediction : $f(x) = \text{sign}\left(\sum_{j=1}^{n} \alpha_j \langle x, x_j \rangle \right)$		Prediction : $f(x) = \text{sign}\left(\sum_{j=1}^{n} \alpha_j \langle \phi(x), \phi(x_j) \rangle \right)$
↖ Linear program		still a linear program ↗

- Denote $\boxed{K(x,y) := \langle \phi(x), \phi(y) \rangle}$ "Kernel" $\Rightarrow$

Find (α_j) : $\sum_j y_j \alpha_j K(x_i, x_j) > 0 \quad \forall i$	(*)
Prediction : $\text{sign}\left(\sum_j \alpha_j K(x, x_j) \right)$	

(still a linear program).

- Forget ϕ. Just choose a suitable kernel $K(\cdot,\cdot)$ and solve (*). "Kernel SVM"

- Ex : $K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2} \right)$ "Radial Basis Function" Kernel (RBF)

$\|$ nontrivial

$\langle \phi(x), \phi(y) \rangle$ but we don't need to know ϕ! Just use in (*)

$\|\phi(x)\|_2^2 = K(x,x) = \exp(0) = 1 \Rightarrow$ maps $\mathbb{R}^d \to$ sphere.

- "Kernelizing" a ML algorithm:

1. Express it in terms of inner products of data $\langle x_i, x_j \rangle$
2. Replace all instances $\langle x, y \rangle \mapsto K(x,y)$ for a suitable choice of the kernel K.