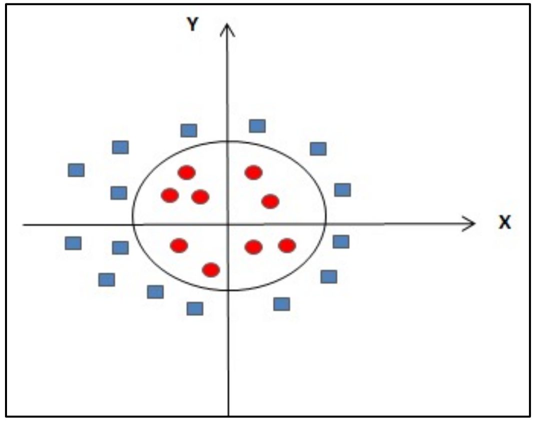
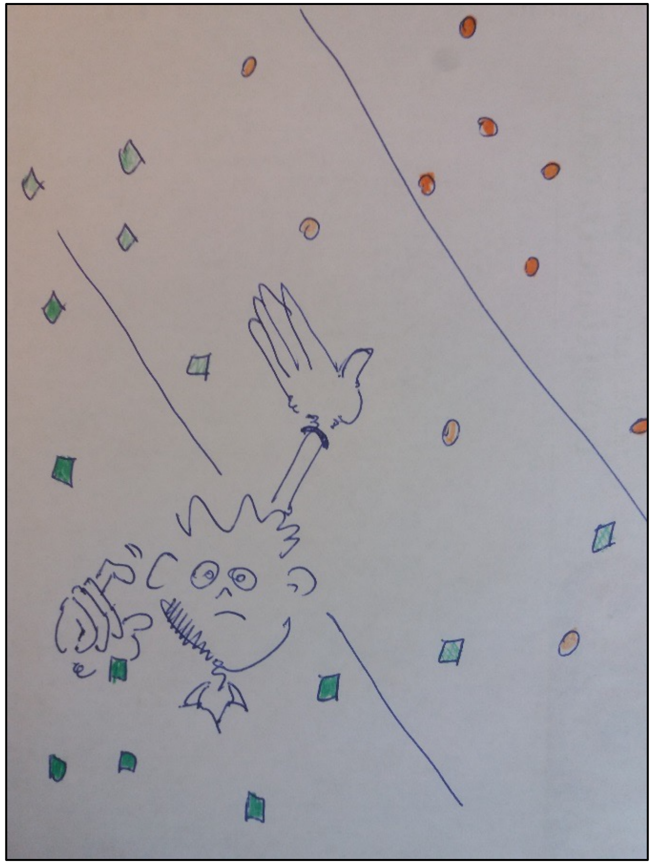
Linear classifiers in higher dimensions



Ariel Shamir
Zohar Yakhini

Outline



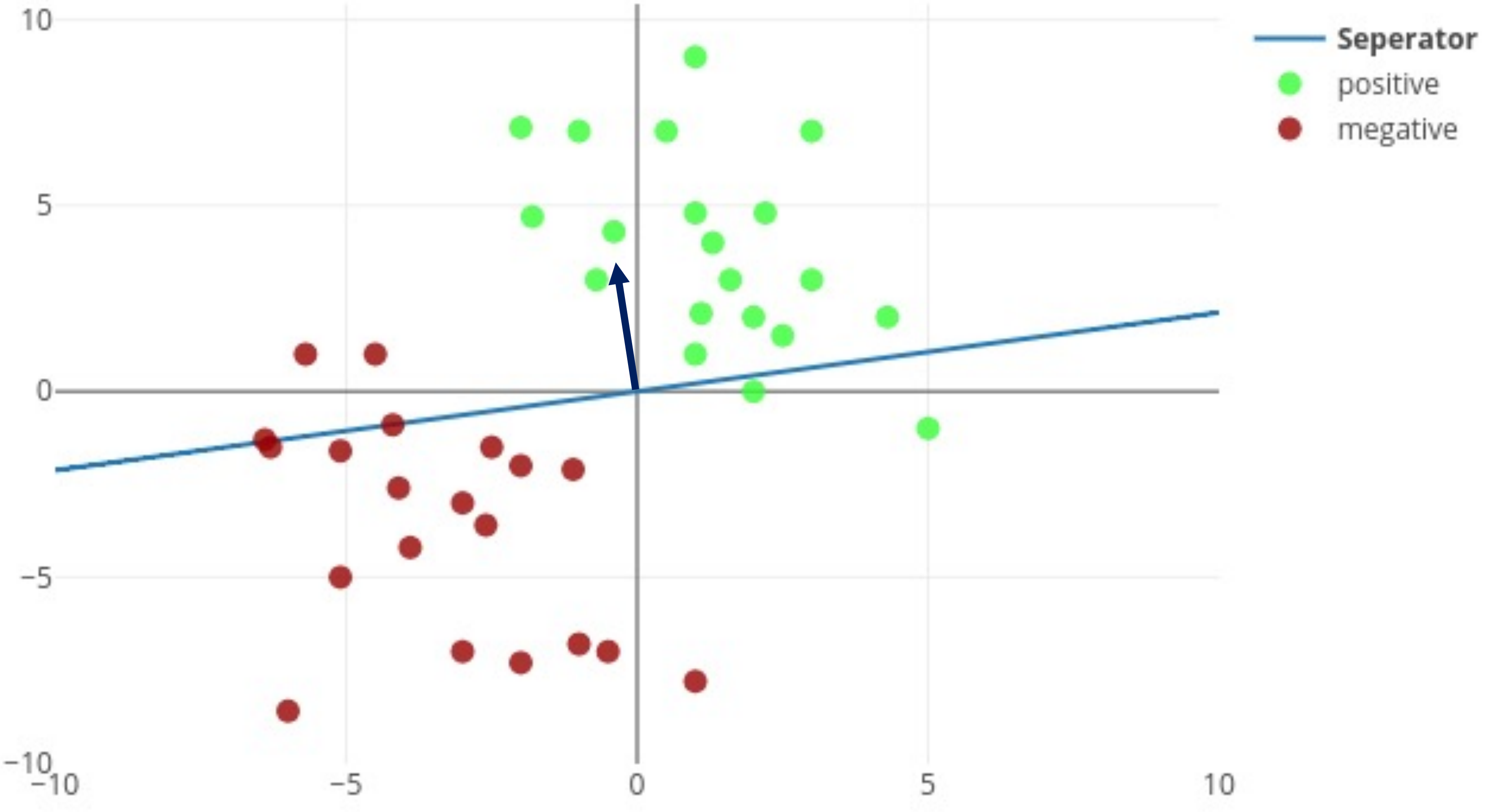
- Linear Separability
- Linear decision boundaries
- The perceptron algorithm
- Non linear mapping of features
- Mapping to higher dimension
- Cover's Thm
- Kernels
- Dual perceptron
- Kernel perceptron
- SVMs

A hyperspace as a linear separator

Consider the decision function

$$h(\vec{x}) = \mathrm{sgn}(\vec{x} \cdot \vec{w} - \mathrm{b})$$

In this figure, who is w ? what is b? what values does h take for various points?



Linear separability

Data in $\mathbb{R}^d$ is linearly separable iff we can find a hyperplane so that:

$$y_i(\vec{x}_i \cdot \overrightarrow{w} - \mathrm{b}) \geq 0\,, 1 \leq \forall i \leq m$$

The Perceptron Learning Algorithm

- Assume the target value for classification are $t \in \{-1, +1\}$
- The perceptron seeks to find a linear separator with NO ERRORs.
- Is this always possible?

Initialize each w_i to some small random number.

Until termination do

For each $\vec{x}_d$ in D compute

$$o_d = \mathrm{sgn}(\vec{w} \cdot \vec{x}_d)$$

For each linear unit weight w_i, Do

$$\Delta w_i = -\eta(o_d - t_d)x_{id}$$

$$w_i = w_i + \Delta w_i$$

n+1 weights to be updated in the normal vector

For any misclassified (at the present iteration) training instance, x, with $C(x) = +1$ we update the weights as:

$$w = w + 2\eta x$$

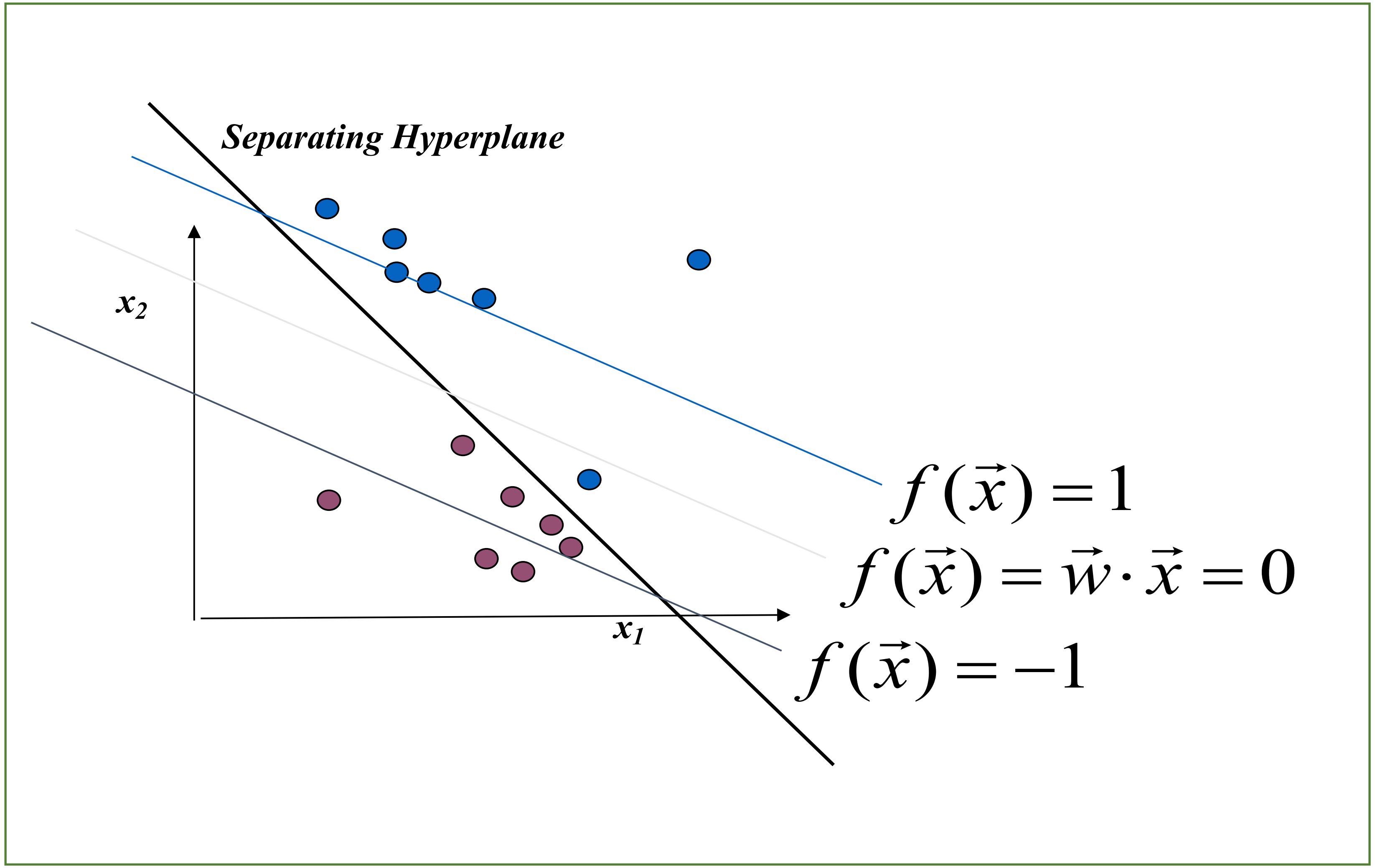
Rosenblatt's Perceptron Theorem

Frank Rosenblatt
Cornell Univ, NY, US
1928-1971

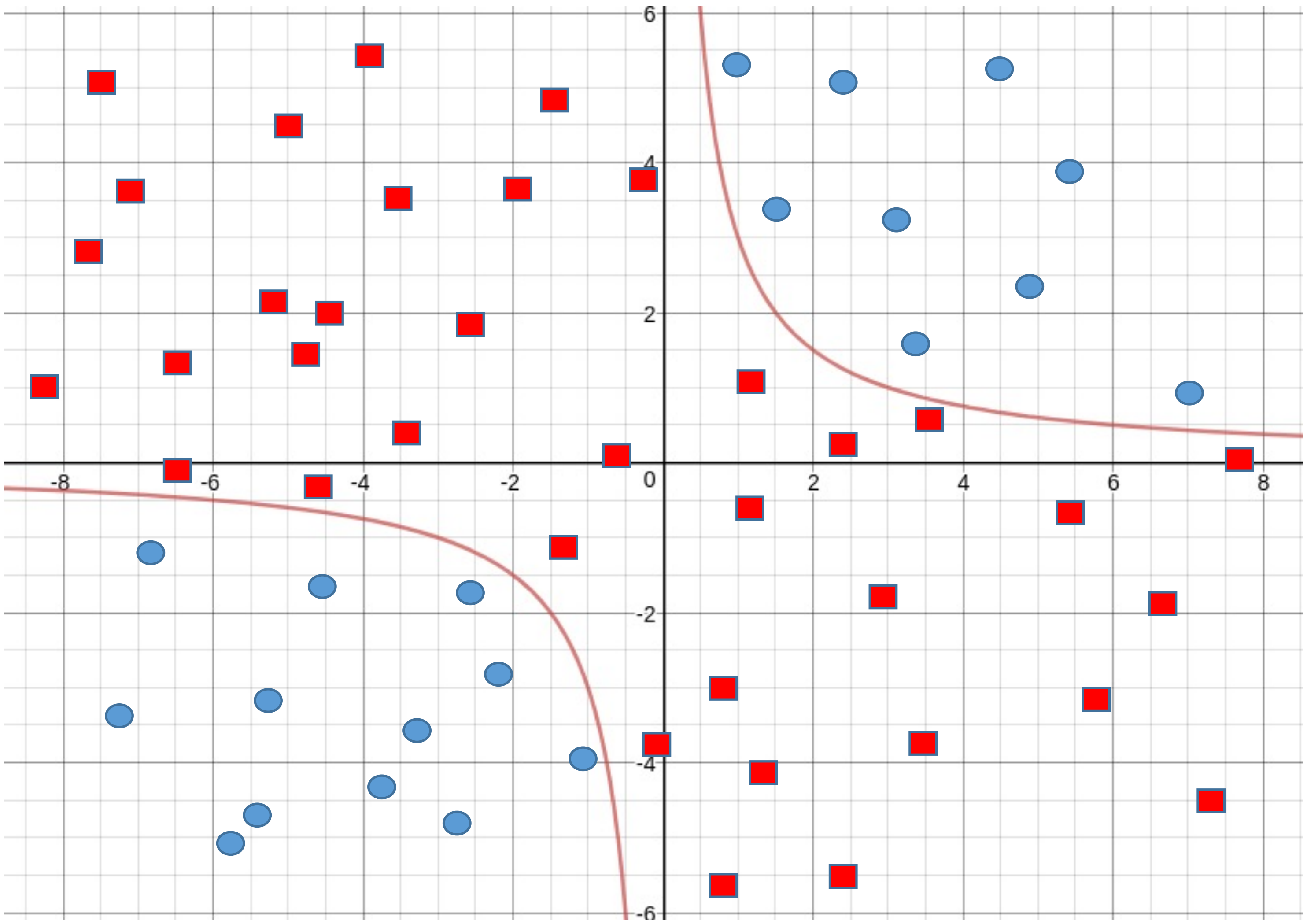
The Perceptron learning algorithm converges to a perfect classifier (no errors on the training data) iff the training data, D, is linearly separable.

- Note: we also need to control η to really guarantee convergence
  (if its too big we may overshoot the perfect classifier)
- Some results on the rate of convergence were proven and can be useful in the context of ANNs (and deep learning)
- The Perceptron itself is not a practical learning approach but is an important component of many modern learning approaches.

What would the Perceptron do here?



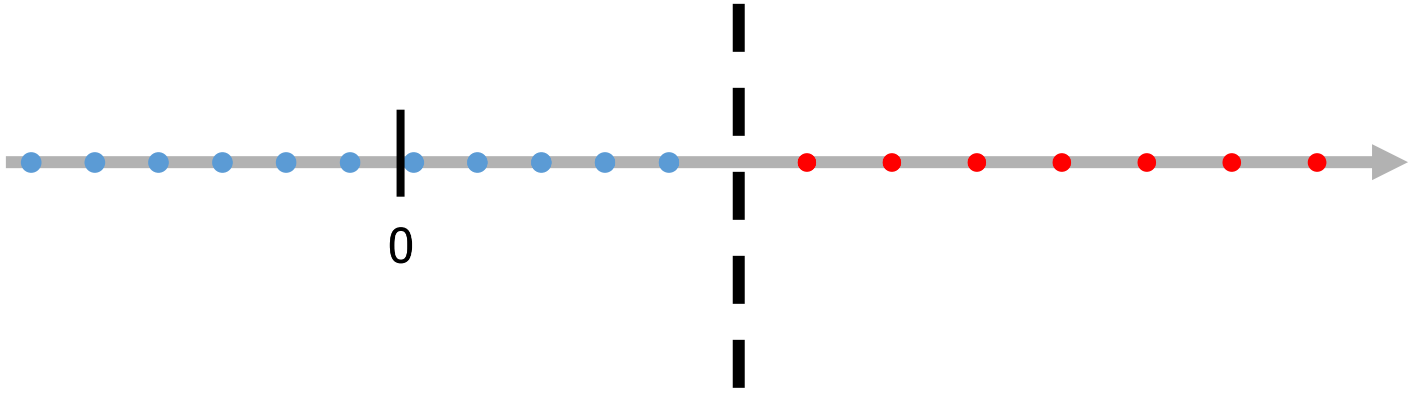
Data not always linearly separable



The Perceptron and Linear Separability

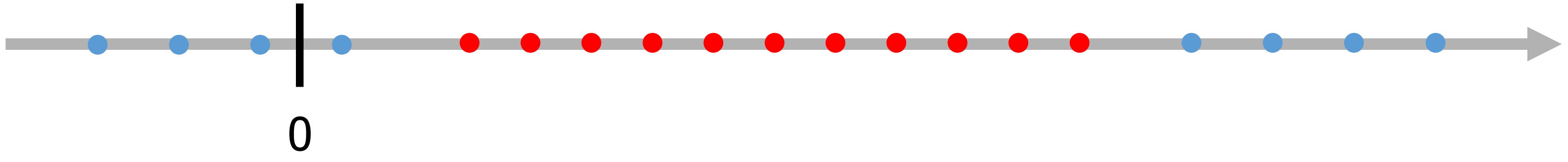
In 1D: $w_1 \cdot x + w_0 > 0$ OR $C(x) = \text{sgn}(w, x)$

$$w = (w_1, w_0)$$



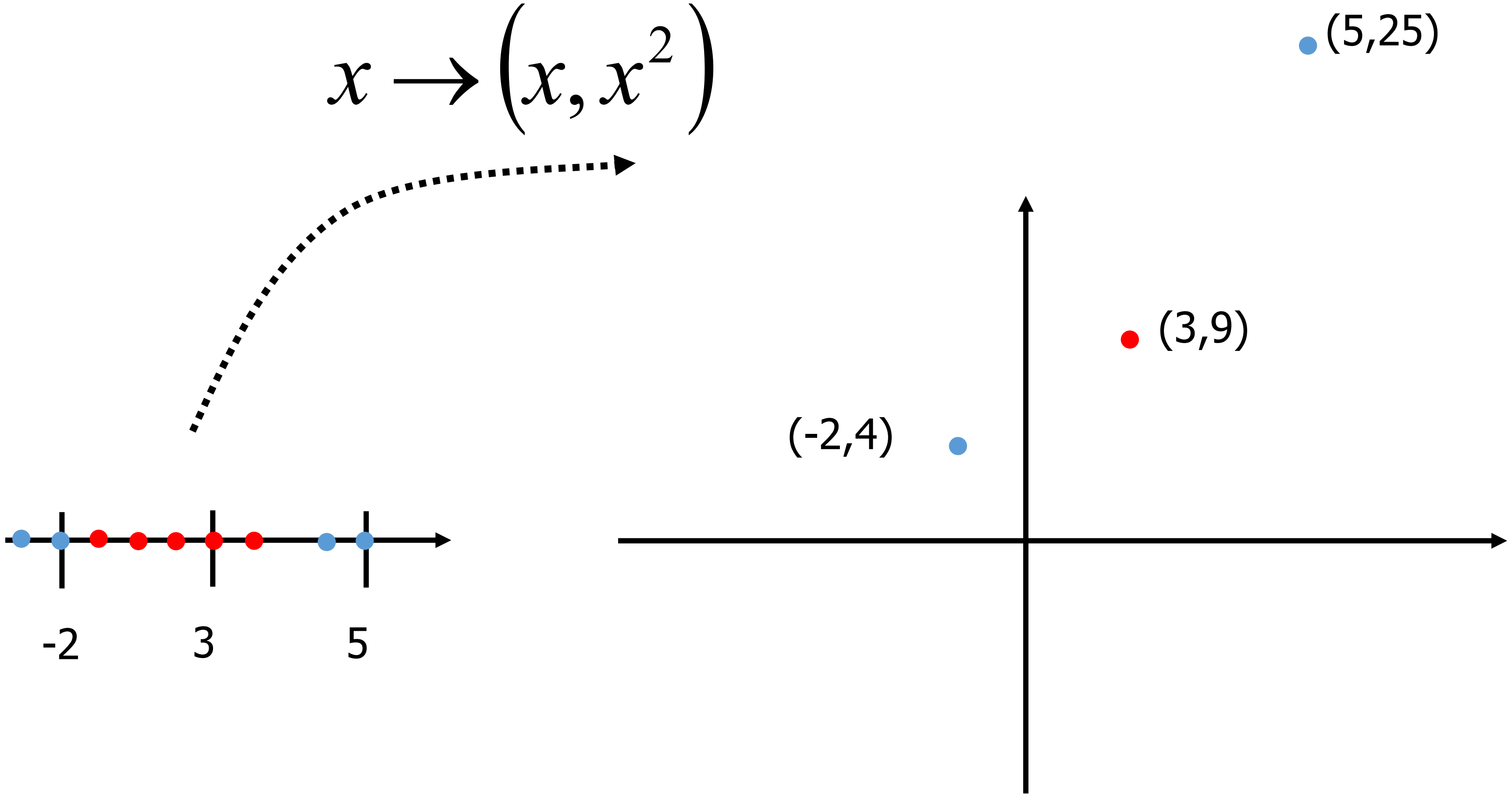
Can We Build a 1D Perceptron for This?

- Red: C(x) = -1
- Blue: C(x) = +1



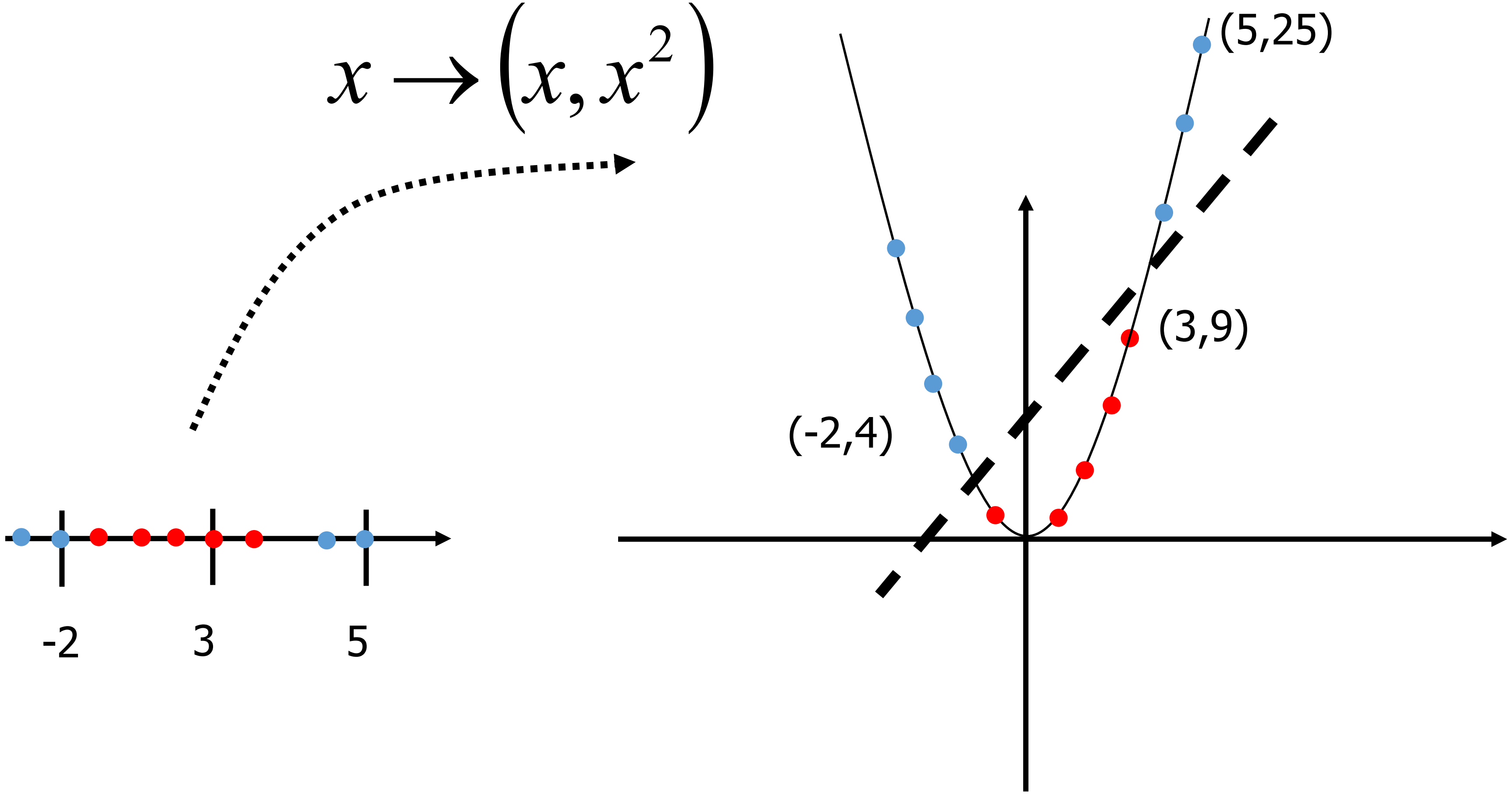
Mapping to Higher Dimension

$$x \rightarrow \left(x, x^2\right)$$

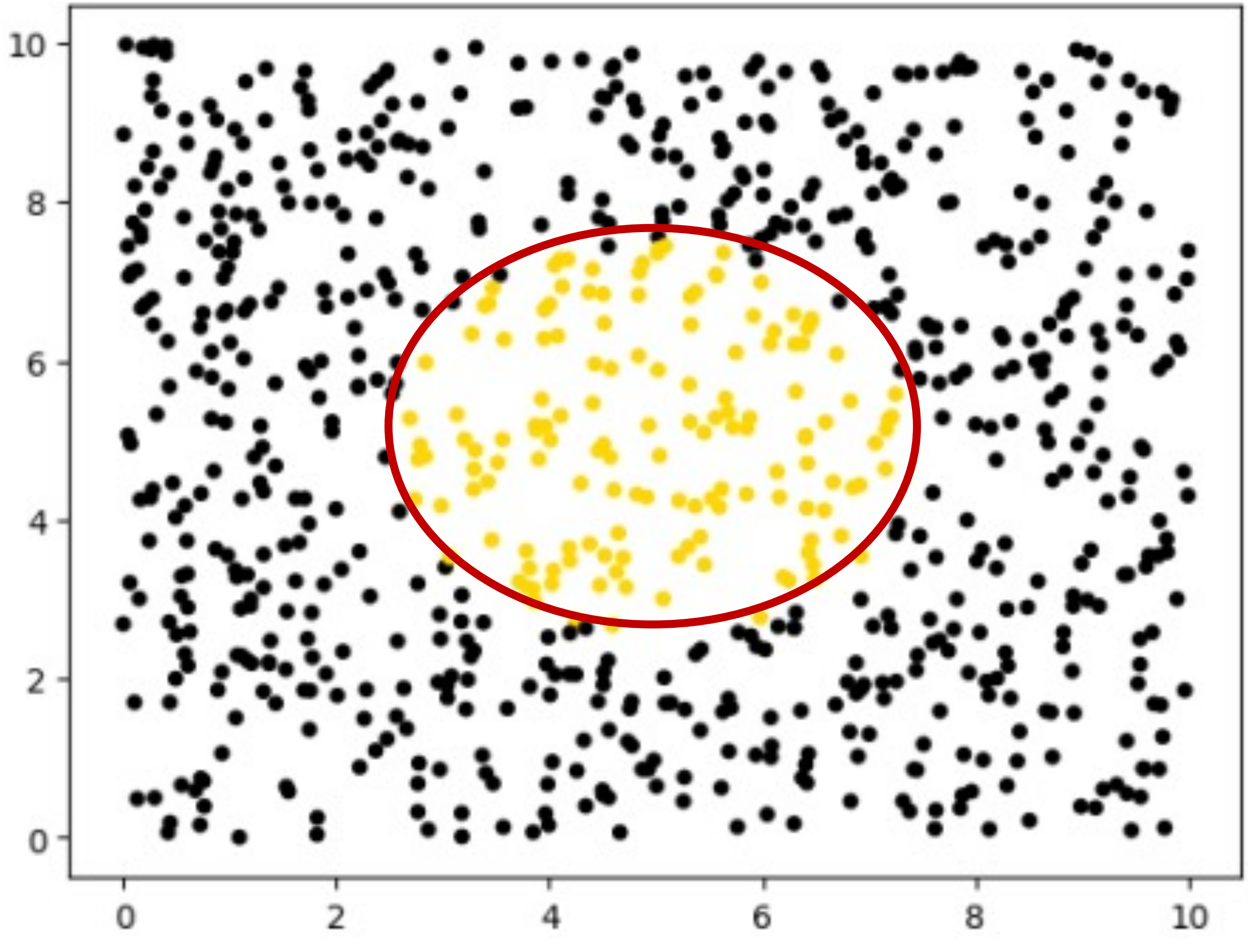


Linear Separability in the Target Space

$$x \rightarrow \left(x, x^2\right)$$

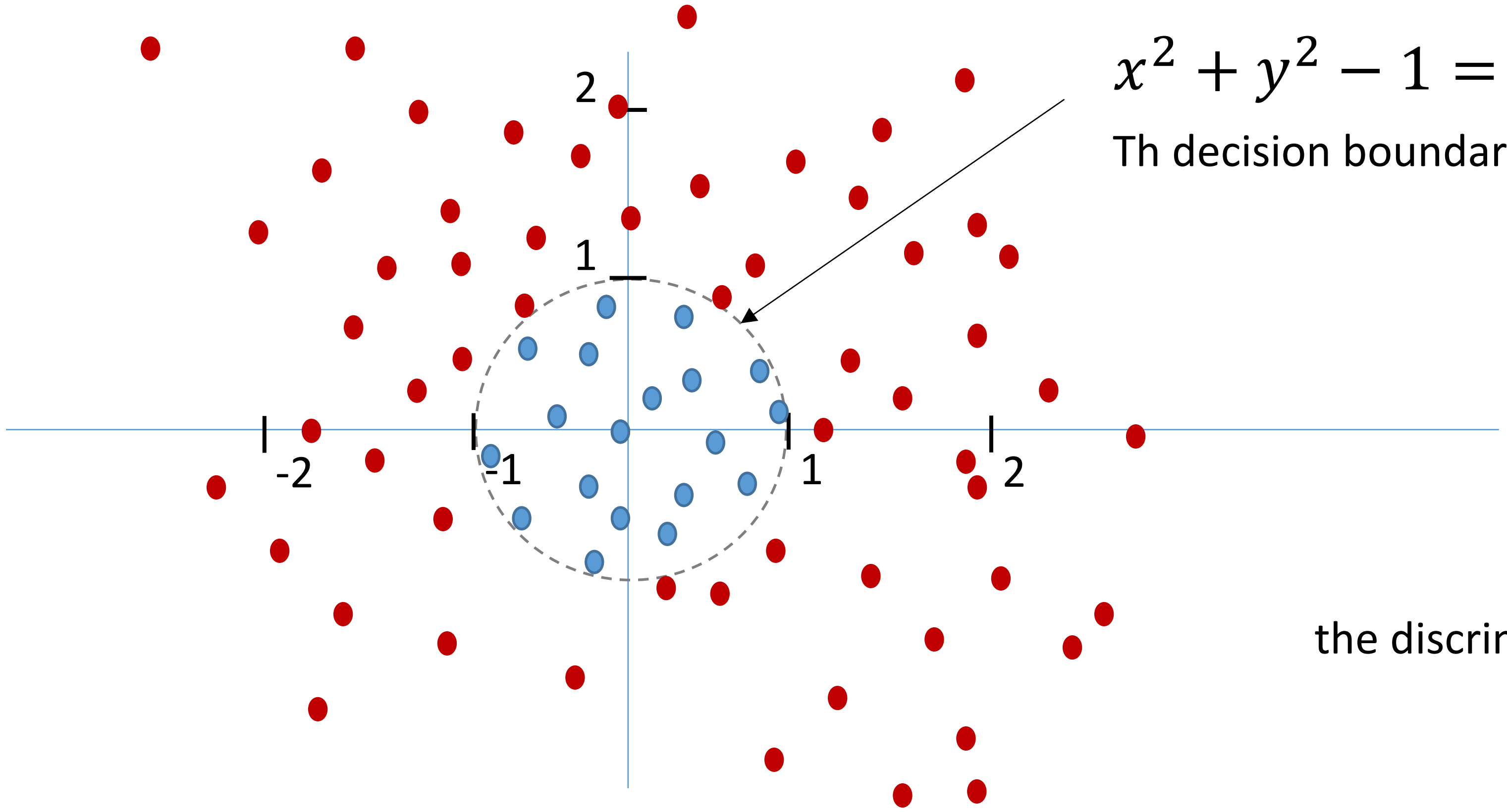


Data is not always linearly separable ....



$$x^2 + y^2 - 1 = 0$$

Th decision boundary



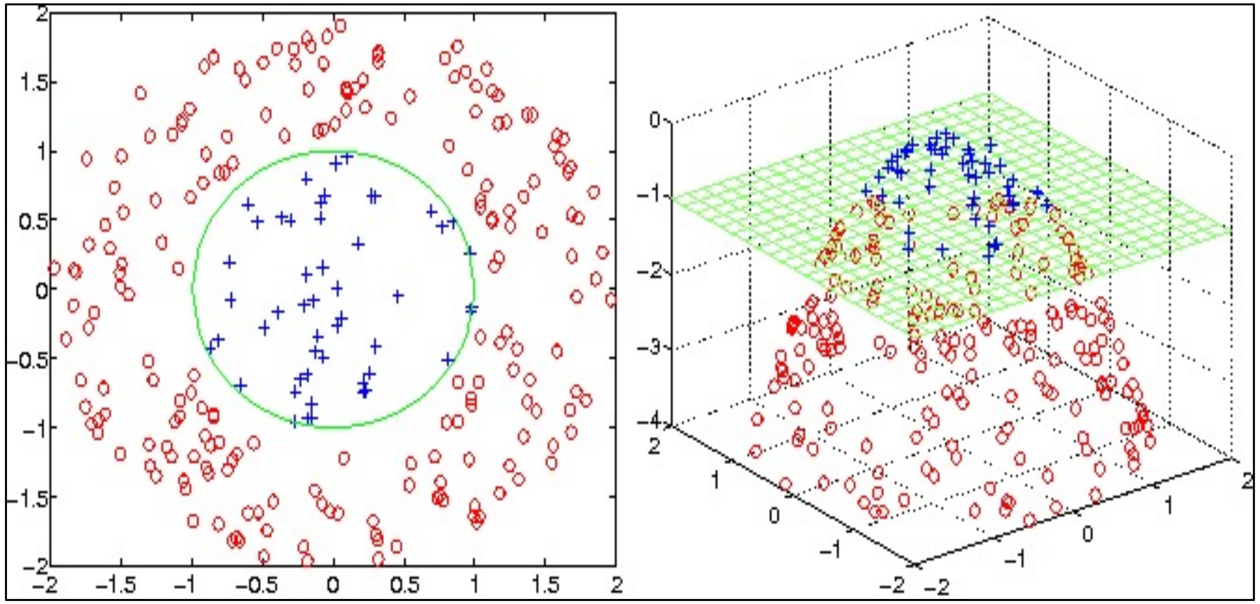
the discriminant function

$$(x, y) \longmapsto \varphi(x, y) = (1, x, y, x^2 + y^2)$$

$$w = (-1, 0, 0, 1)$$

(x, y) is BLUE iff

$$F(x, y) = \operatorname{sgn}(w \cdot \varphi(x, y)) < 0$$

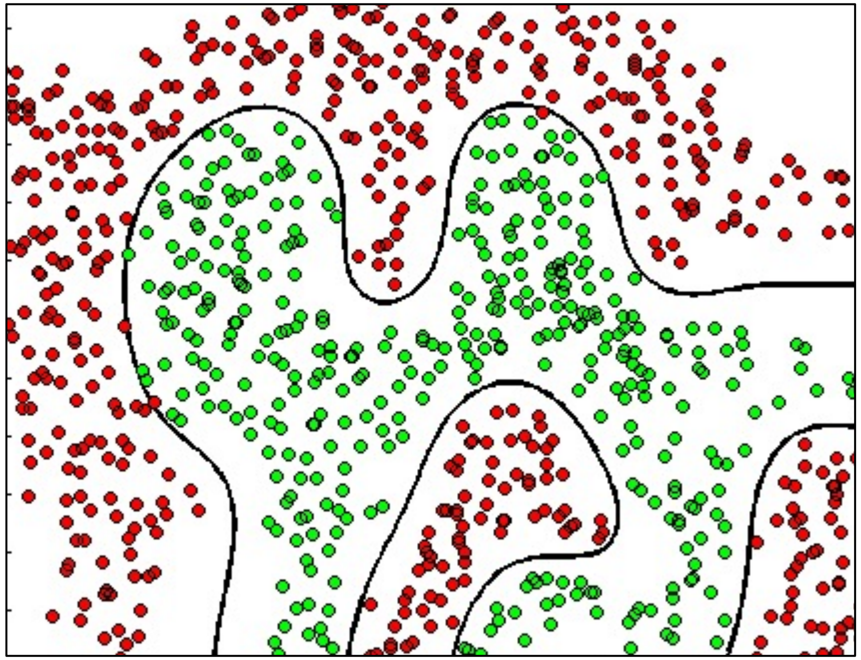


$$(x, y) \longmapsto \varphi(x, y) = (x, y, -x^2 - y^2)$$

$$w = (1, 0, 0, 1)$$

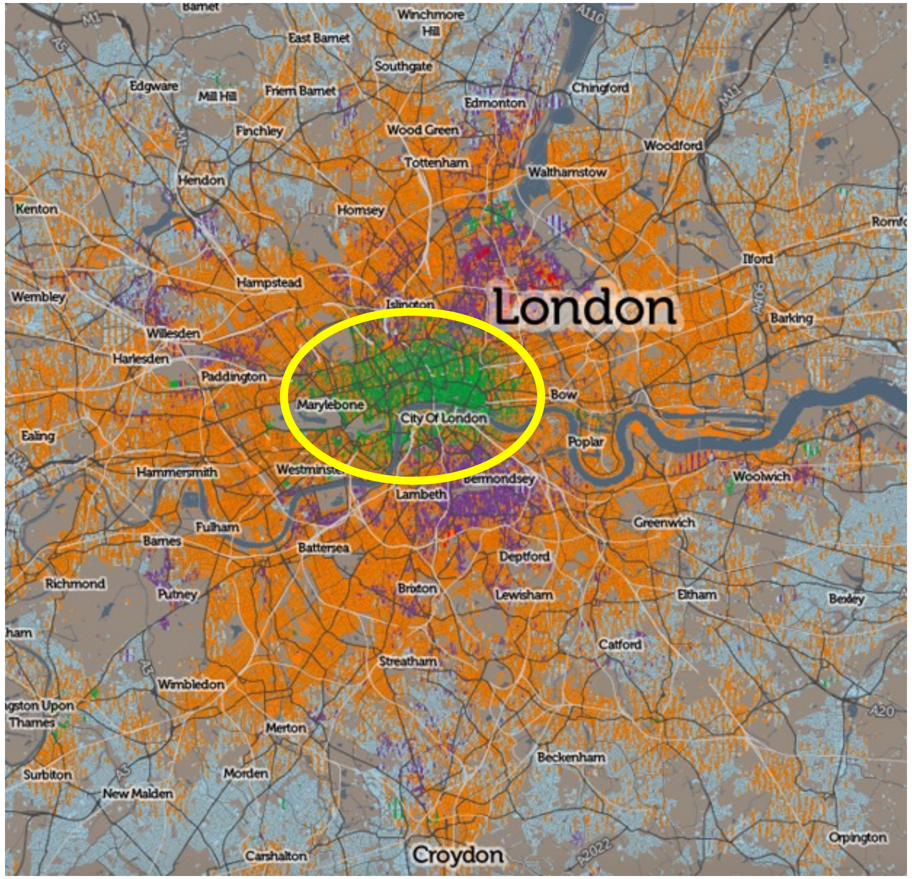
Non-Linear Decision Boundaries

- Decision boundaries which separate between classes may not always be linear
- In fact, they will sometimes be very complex boundaries and therefore may sometimes require the use of highly non-linear discriminant functions

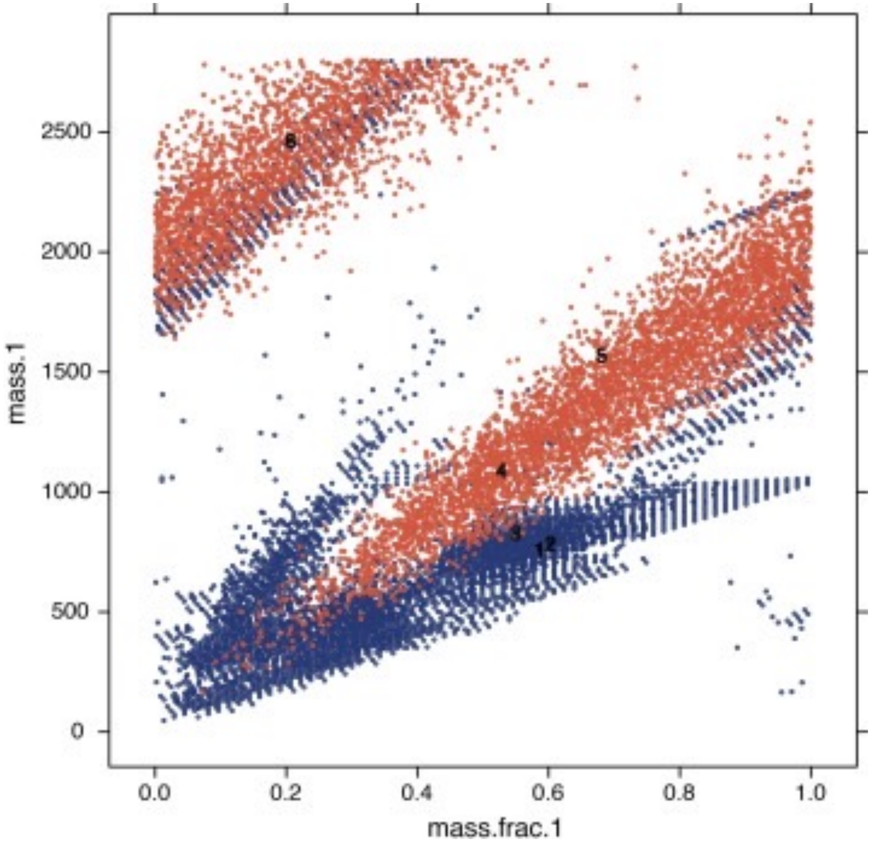


Data is not always linearly separable ….

London walking commuters



Lipids vs peptides (Dittwald et al)



Generalized Linear Discriminant Functions

A possible approach to generalizing the concept of linear decision functions is to consider a generalized decision function as:

$$F(\vec{x}) = w_0 + w_1\varphi_1(\vec{x}) + w_2\varphi_2(\vec{x}) + \ldots + w_N\varphi_N(\vec{x})$$

where

$\varphi_i(\vec{x})\colon \mathbb{R}^n \to \mathbb{R},\ \ 1 \leq i \leq N$

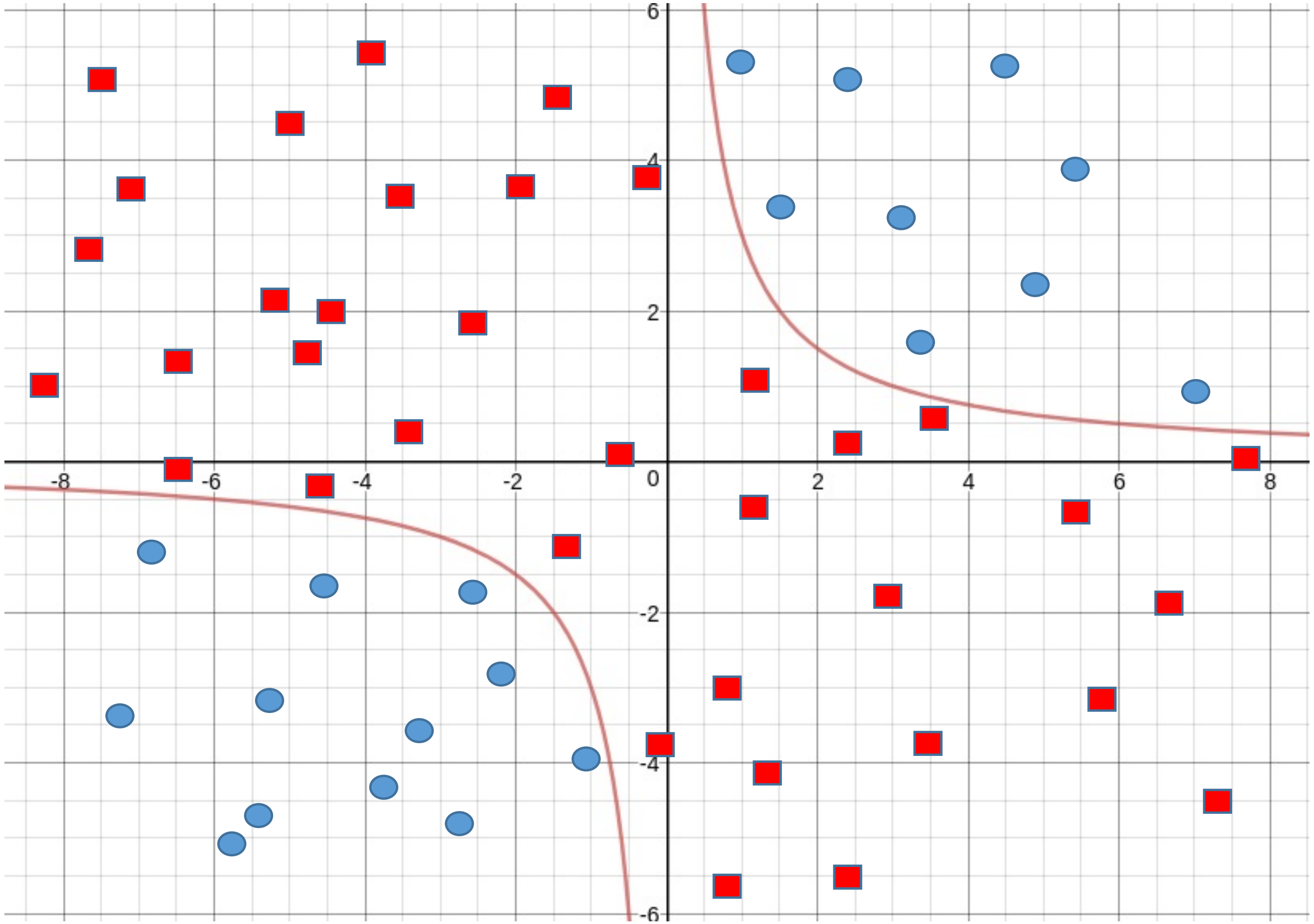
are scalar functions of the input $\vec{x} \in \mathbb{R}^n$

(all spaces here are Euclidean)

The <u>ambient dimension</u> N is typically larger than n (but not necessarily)

Classification will be based on the Boolean $F(x) \geq 0$

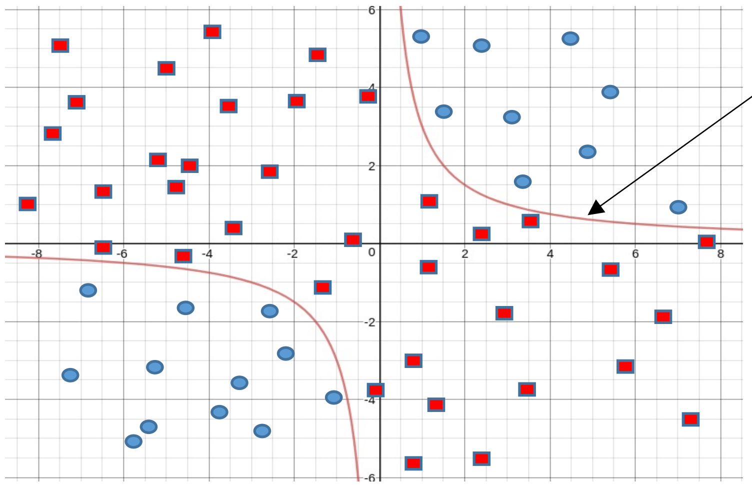
Linear separation after mapping into a different space?



What φ_i s can we use in this example?

Mapping into 3D

$x_1 x_2 = 3$



??

$$(x_1, x_2) \rightarrow$$
$$\varphi(x_1, x_2) = (1, x_1, x_2, x_1 x_2)$$

$$F(x) = w \cdot \varphi(x) > 0 \Rightarrow$$

What are the ws?

Higher Dimension Linear Separability

Classifier: $\mathrm{sgn}(F(\vec{x}))$, wherein

$$F(x) = \sum_{i=0}^{N} w_i \varphi_i(x) = \vec{w} \cdot \varphi(\vec{x})$$

consists of a vector of coefficeints

$$\vec{\mathrm{w}} = (\mathrm{w}_0, \mathrm{w}_1, w_2, \ldots, w_N)$$

and of the mapping

$$\varphi(\vec{x}) = (\varphi_0(x), \varphi_1(x), \varphi_2(x), \ldots, \varphi_N(x))^T$$

- We can assume $\varphi_0(x) = 1$
- This representation of $\varphi(x)$ implies that any decision function defined by the weight equation can be treated as linear in the *N*-dimensional space (*where possibly N > n*)
- Note that the components of $\varphi(x)$ may be non-linear in the input space, $\mathbb{R}^n$, e.g polynomial or exponential terms

Non-linear Mapping Idea

- Map data from low dimensional Input Space to high dimensional Mapped Space (aka Ambient Space) and hope that the data is linearly separable there:
- Example: quadratic mapping

$$\vec{x} = (x_1, x_2) \longmapsto (1, x_1, x_2, x_1 \cdot x_2, x_1^2, x_2^2)$$

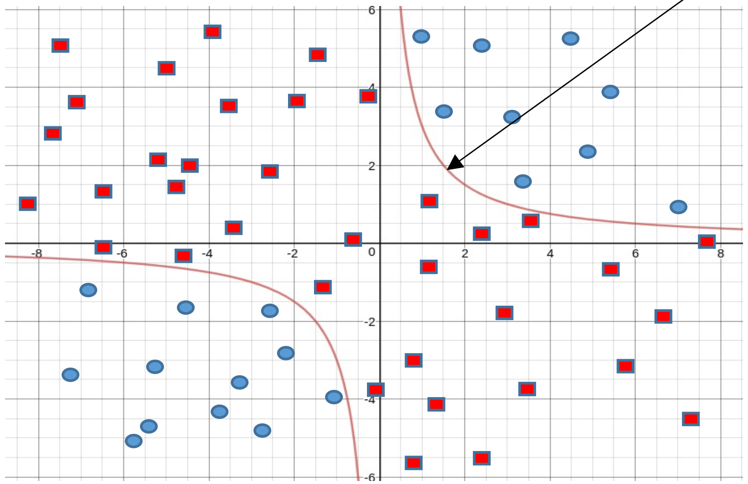
- The discriminant function would then be:

$$F(\vec{x}) = \sum_{i=0}^{5} w_i \varphi_i(\vec{x})$$

$$= w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2$$

Mapping into 3D

$x_1 x_2 = 3$



$$(x_1, x_2) \rightarrow$$
$$\varphi(x) = (1, x_1, x_2, x_1 \cdot x_2, \; x_1^2, \; x_2^2)$$

$$F(x) = w \cdot \varphi(x) > 0 \;\Rightarrow$$

Polynomial Discriminant Functions

- The most commonly used generalized decision function is $F(\vec{x})$ for which $\varphi_i(\vec{x})$ $(1 \ \leq \ i \leq N)$ are multi-dimensional monomials.
- Examples:

$$\varphi_1(\vec{x}) = 5x_4^3$$

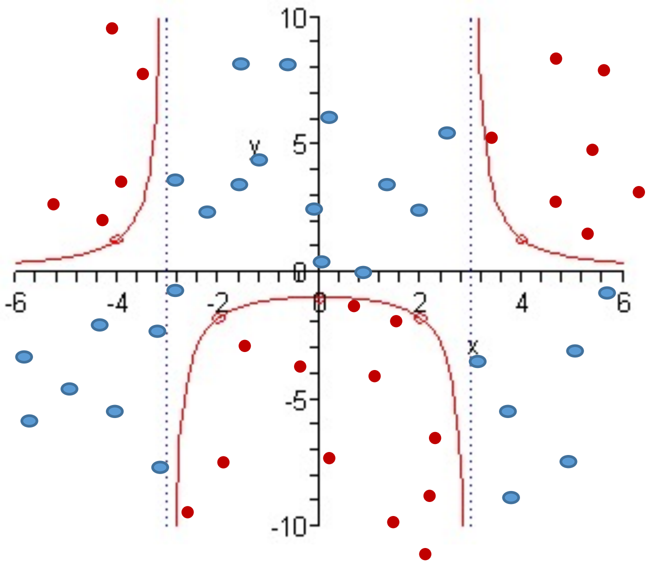
$$\varphi_2(\vec{x}) = 7x_4^3x_5^2$$

$$\varphi_2(\vec{x}) = -x_2x_3^2$$

The discriminant function might then look something like:

$$F(\vec{x}) = 100 + 3\,x_1 \ - \pi x_3^2 x_4^2 x_5^2 + \ x_2^7$$

Mapping to higher dimensional space, revisited – full rational varieties



But how would we engineer a map for this?

We could do this but a more principled approach is to try all quadratic forms by mapping into the full quadratic variety:

$$(x, y) \mapsto (1, \sqrt{2}x, \sqrt{2}y, \sqrt{2}xy, x^2, y^2)$$

And then apply the Perceptron in $\mathbb{R}^6$.

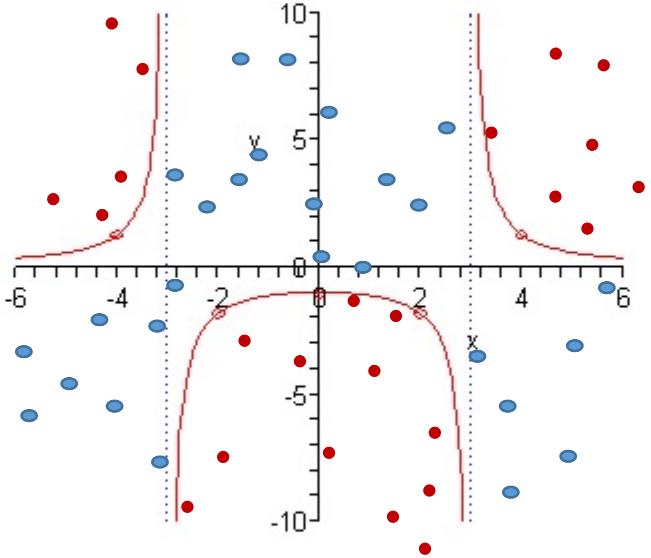
That does not work …
What's next?

Mapping to higher dimensional space, revisited – full rational varieties

We now map
into the full cubic variety:

$$(x, y) \longmapsto (1, c_1x, c_1y, c_2xy, c_3x^2, c_3y^2, c_4x^2y, c_4xy^2, x^3, y^3)$$

And then apply the Perceptron in $\mathbb{R}^{10}$.



Mapping to higher dimensional space, revisited – full rational varieties

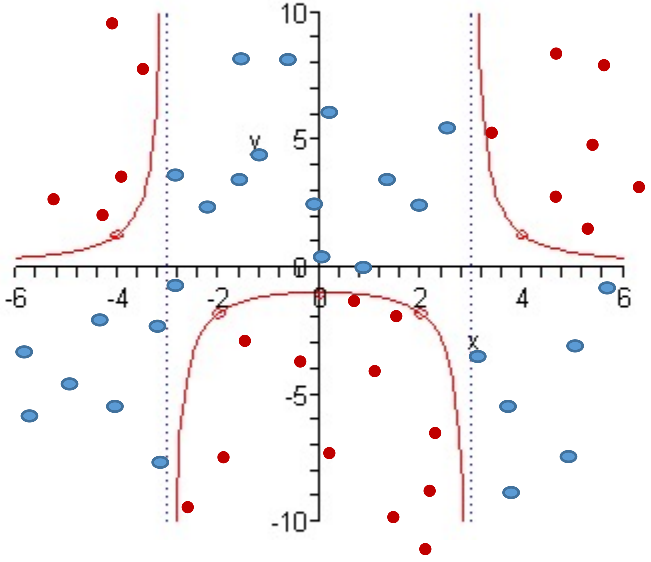
The equation of the red curves is

$$y = \frac{9}{x^2 - 9}$$

And we therefore now get a perfect classifier with

$$w = (-9, 0, \frac{-9}{c_2}, 0,0,0, \frac{1}{c_4}, 0,0,0) ,$$

namely $C(x, y) = \mathrm{sgn}(x^2 y - 9y - 9)$



Rational Varieties

- A full rational variety of order r in an input space of dimension n is described by all r-th degree monomials of the input variables in x:

$$\varphi_i(\vec{x}) = 1^{r_0} {x_1}^{r_1} {x_2}^{r_2} \ \ldots \ {x_n}^{r_n}$$

where $\sum_{j=0}^{n} r_j = r$

- The number of different monomer terms in such expressions is:

What is the benefit?

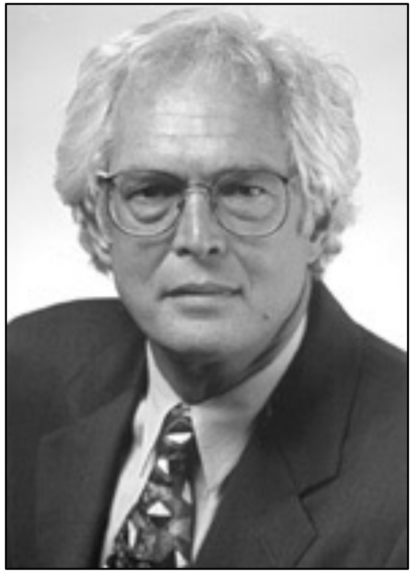
- Why should we assume that in higher dimensions the classes are more likely to be linearly separable?
- What can we do if they are?

Cover's Pattern Counting Theorem (1965)

A complex pattern classification problem cast in a high dimensional space nonlinearly is increasingly more likely to be linearly separable.

Cover, T.M. , 1965

Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition.



Thomas M Cover
US, 1938-2012
Stanford University
World leader in Statistics and Information Theory

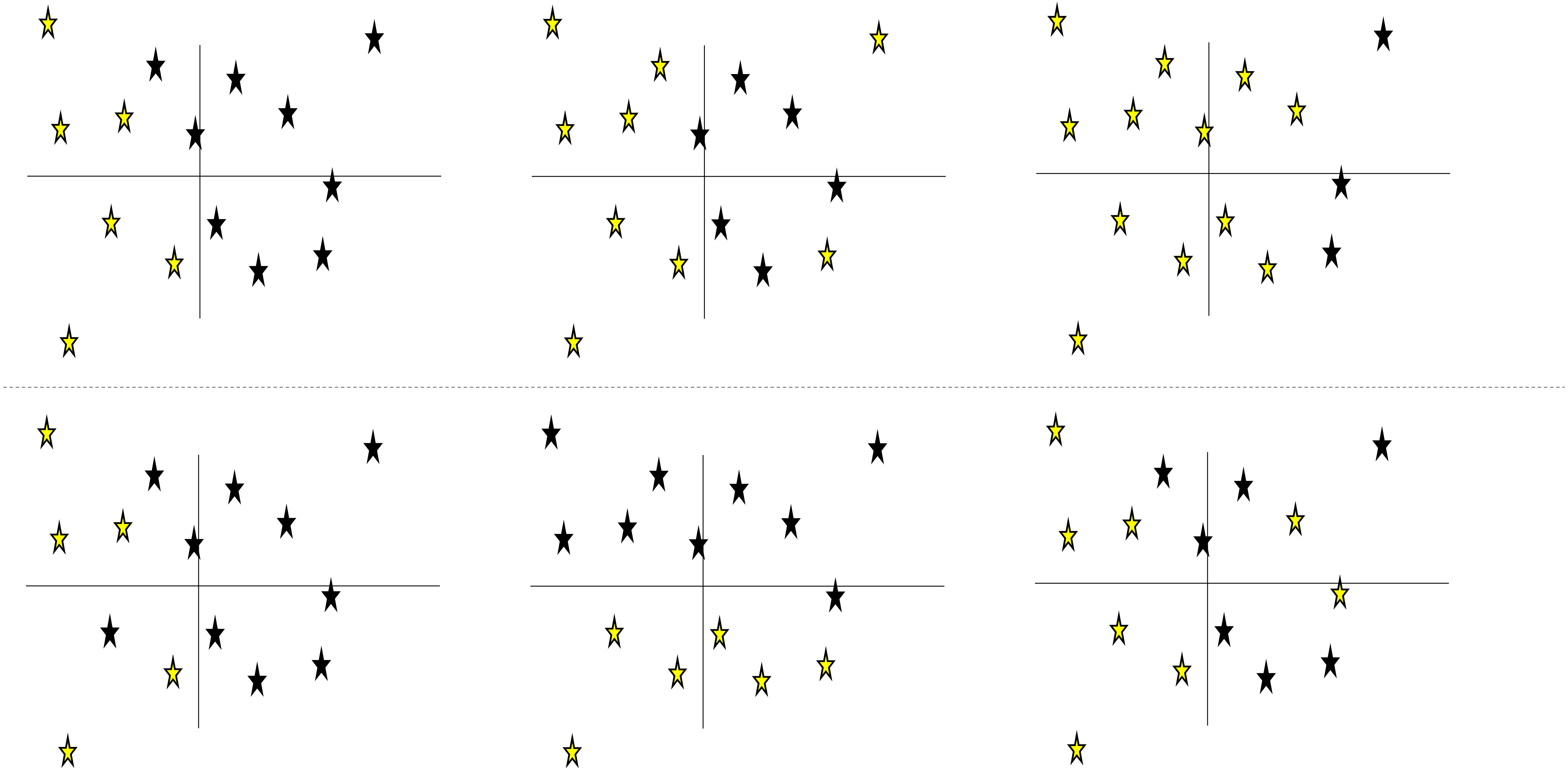
Counting Dichotomies

- A dichotomy of a set S is a partition of S into two disjoint subsets.
- Assume we have k samples in a set of instances S.
- We then have 2^k possible dichotomies over these instances
- Each dichotomy defines a classification task (separate between the two classes)

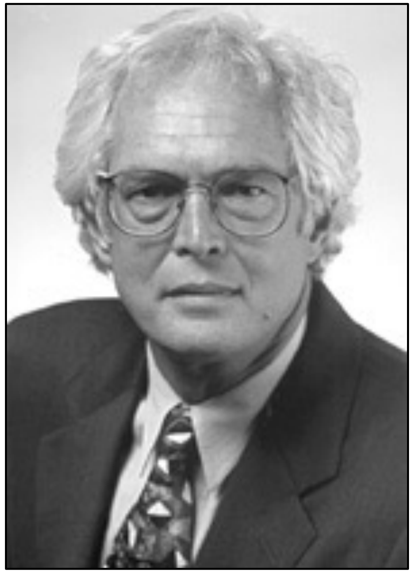
Linearly separable/non-separable dichotomies in 1D

How many separable dichotomies for k points in $\mathbb{R}$?

Linearly separable/non-separable dichotomies in 2D



Cover's Pattern Counting Theorem (1965)



Thomas M Cover
US, 1938-2012
Stanford University
World leader in Statistics and Information Theory

A complex pattern classification problem cast in a high dimensional space nonlinearly is increasingly more likely to be linearly separable.

Cover, T.M. , 1965

Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition.

Cover's Thm: counting separable dichotomies

- How many dichotomies of a set of points S are linearly separable?
- Cover's Counting Thm: in N dimensional space the number of linearly separable dichotomies of k samples is:

$$2\sum_{i=0}^{N}\binom{k-1}{i}$$

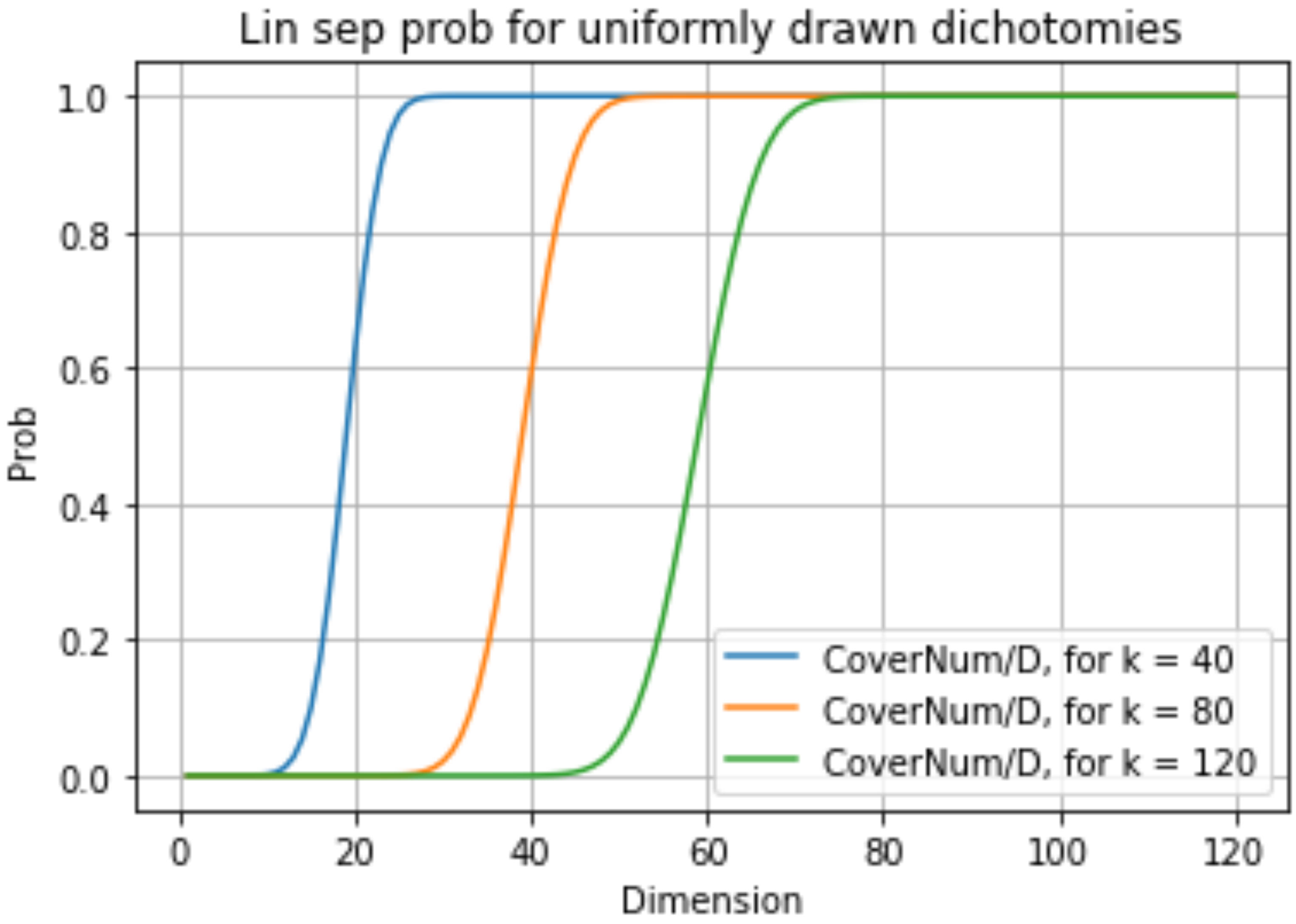
- Hence, the probability that a dichotomy, uniformly drawn at random, is linearly separable is:

$$P(k,N) = \frac{1}{2^{k-1}}\sum_{i=0}^{N}\binom{k-1}{i}$$

- This gets larger as N grows

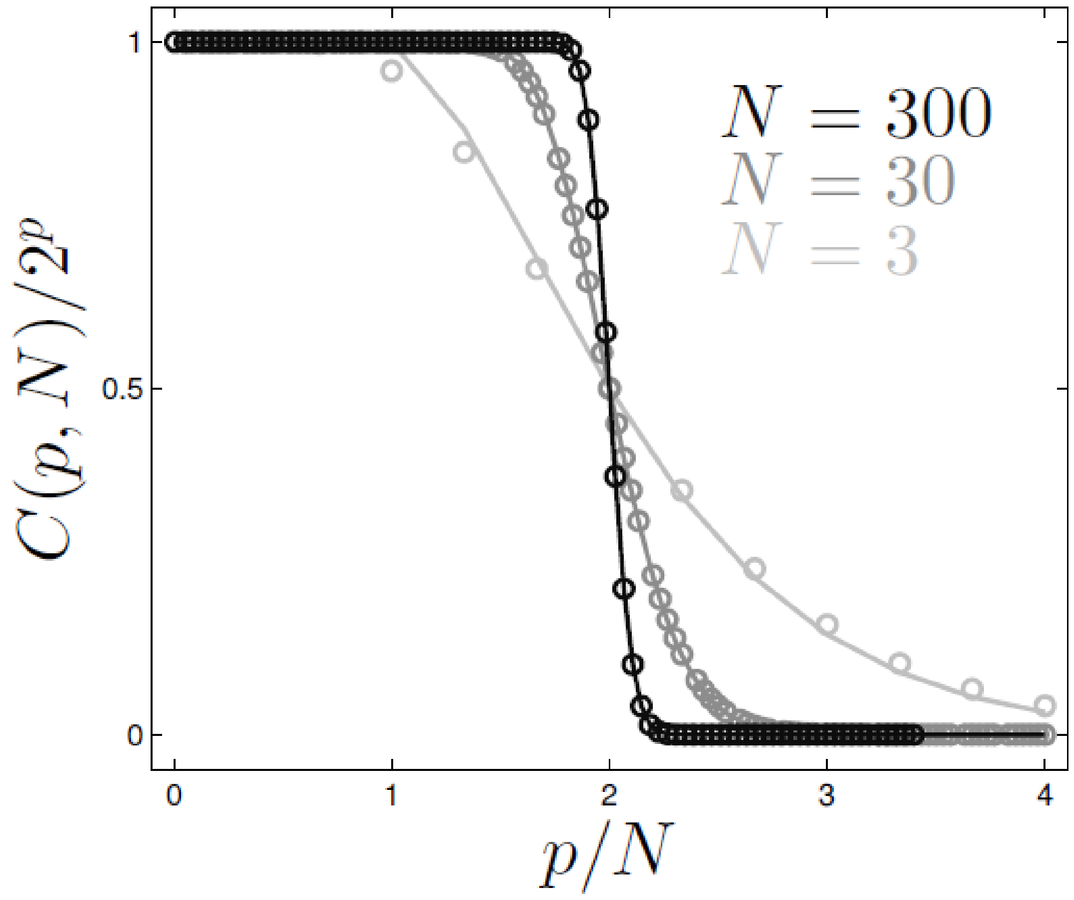
  (Assume that k is fixed at a number typically much larger than N)

Spuriously separable in higher dimensions …



Proof of Cover's Counting Theorem

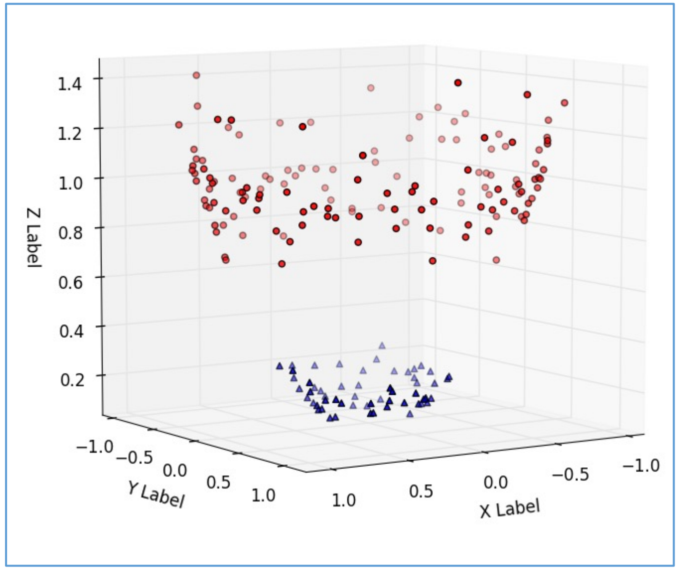
Fixed N, as a function of k



From Emin Orhan, NYU

Full rational varieties – what is the dimension?

Summary so far



- The Perceptron converges to a perfect linear classifier for linearly separable data.
- We can often translate non linear decision boundaries to linear ones by mapping the original instance space, in a non linear manner, into higher dimensional space – the ambient space.
  We then run e.g the Perceptron in the ambient space
- We saw examples with cleverly engineered mapping and mentioned rational varieties as a general approach, to bypass the clever engineering.
- Cover's Thm: in higher dimensions, an increasingly high fraction of dichotomies of k points are linearly separable;
- Implication: linear separability in higher dimension is easier to achieve.
  Caveat: it can be spurious!