CS 4342: Class 15

Jacob Whitehill

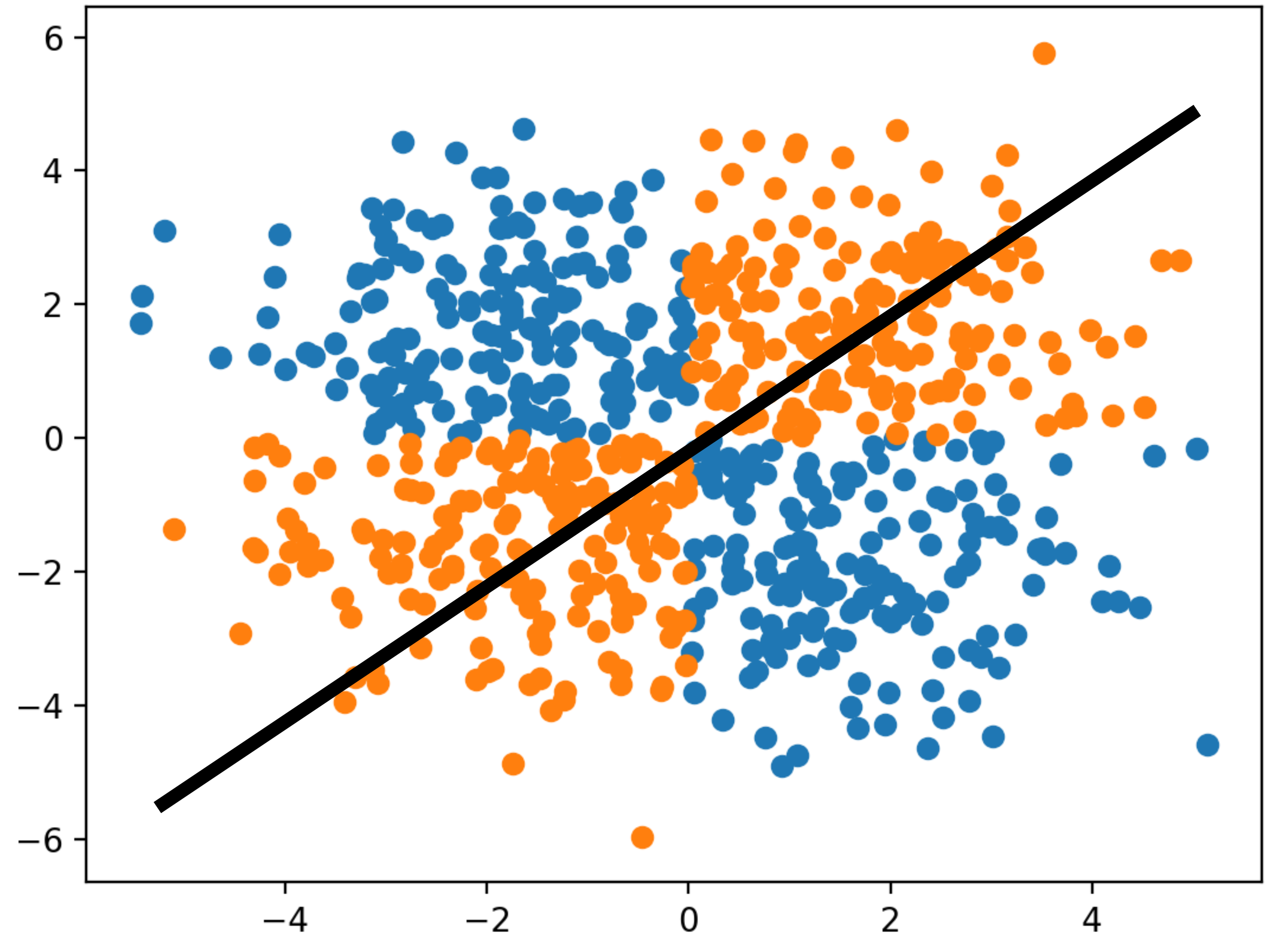
Feature transformations

Linearly inseparable data

- SVMs use a hyperplane to separate data in two classes.

- But what if the data are **linearly inseparable**, e.g.:

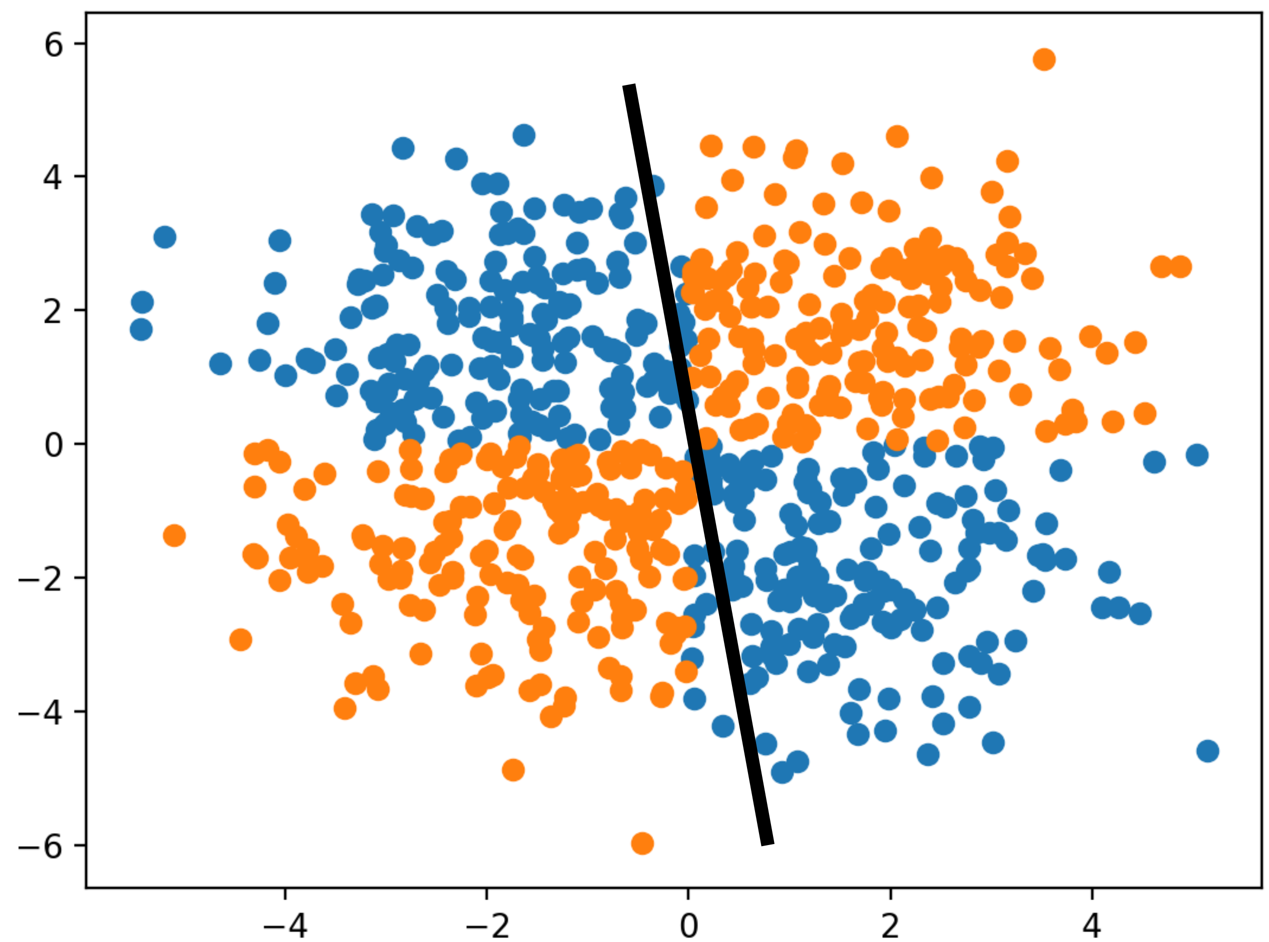
- No matter what **w**, *b* we choose, the SVM will never do a good job of classifying the data.



"XOR" problem

Linearly inseparable data

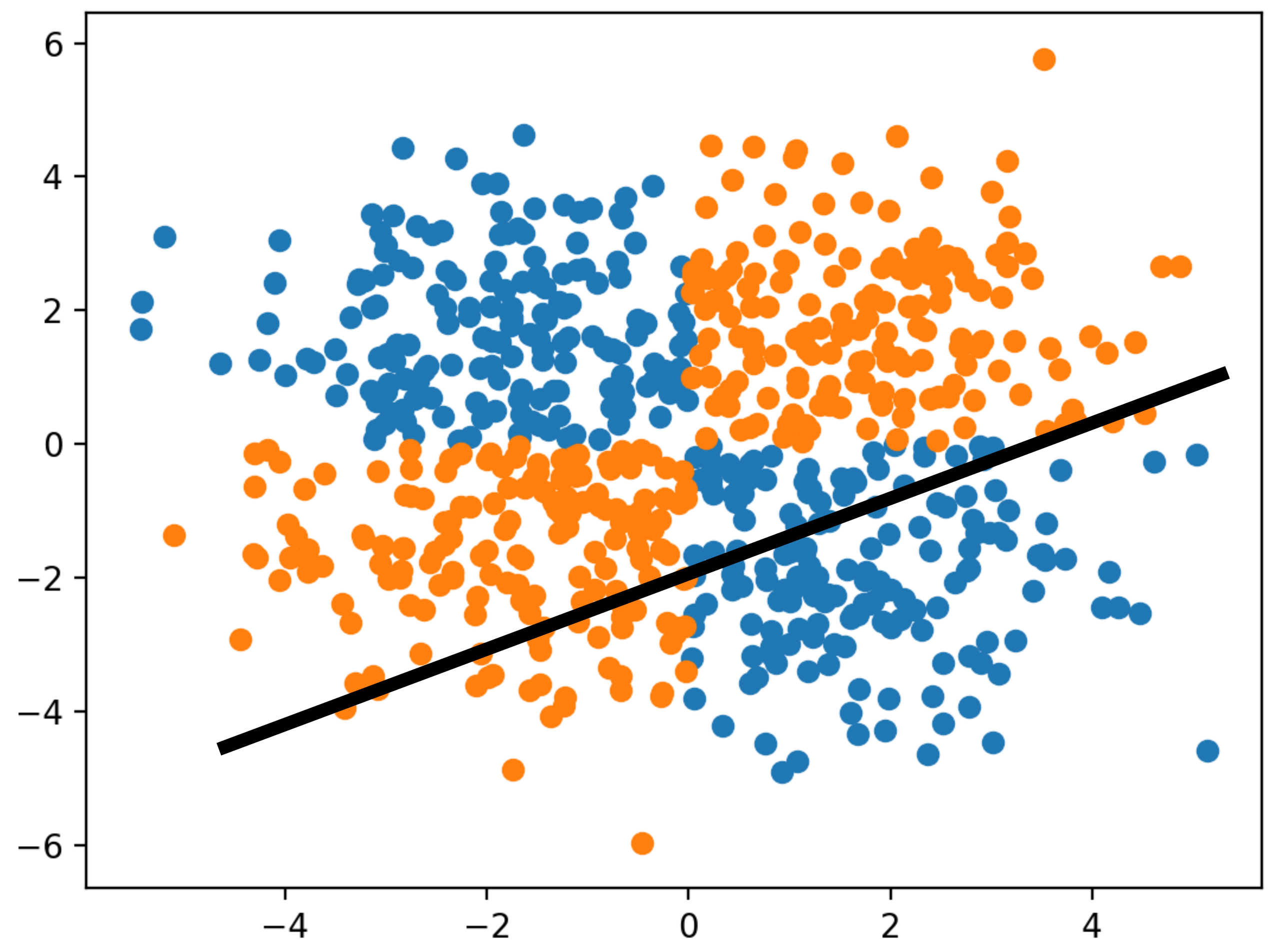
- SVMs use a hyperplane to separate data in two classes.
- But what if the data are **linearly inseparable**, e.g.:
- No matter what **w**, *b* we choose, the SVM will never do a good job of classifying the data.



“XOR” problem

Linearly inseparable data

- SVMs use a hyperplane to separate data in two classes.
- But what if the data are **linearly inseparable**, e.g.:
- No matter what **w**, *b* we choose, the SVM will never do a good job of classifying the data.

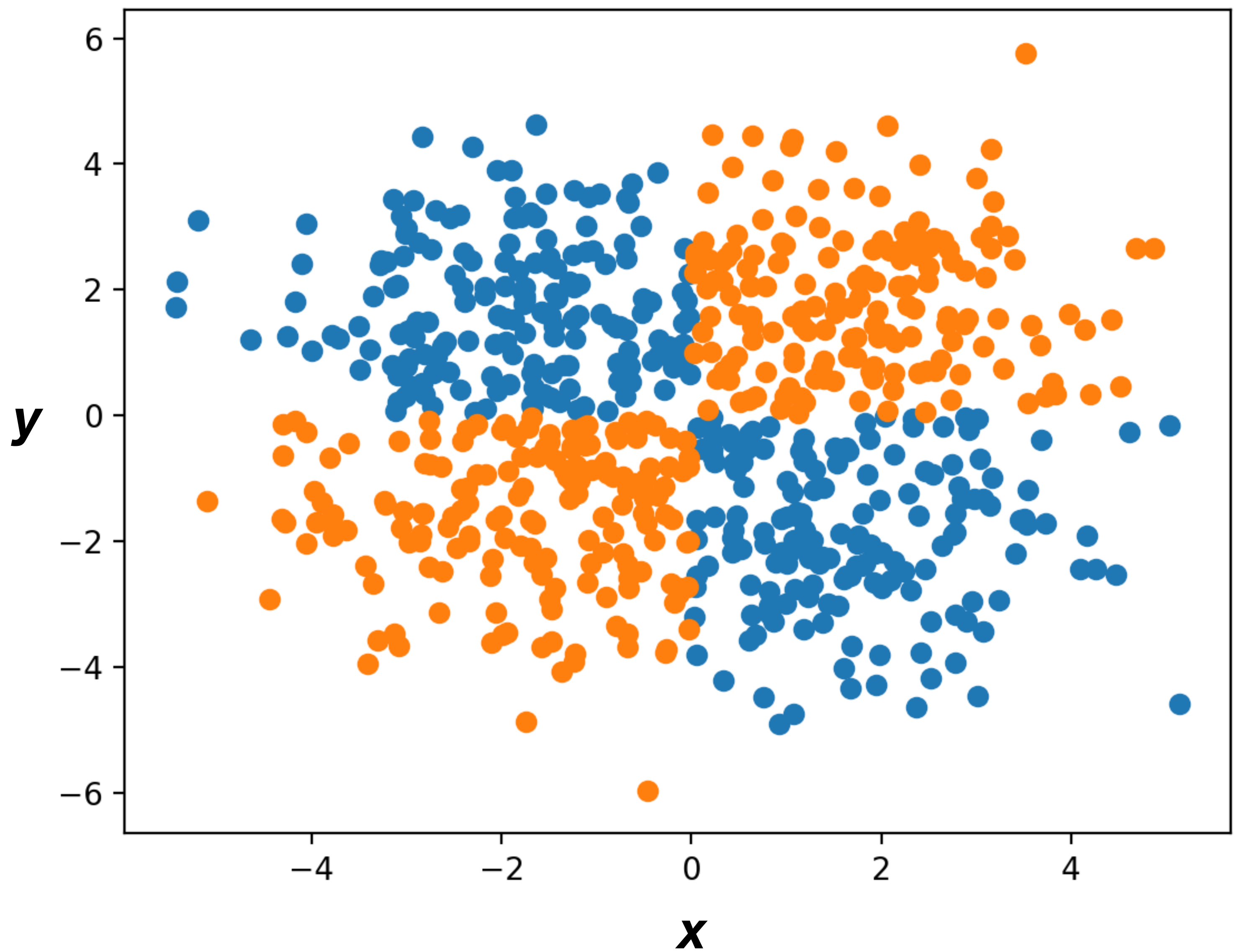


“XOR” problem

XOR problem

- We can use a non-linear transformation to make these data linearly separable, e.g.:

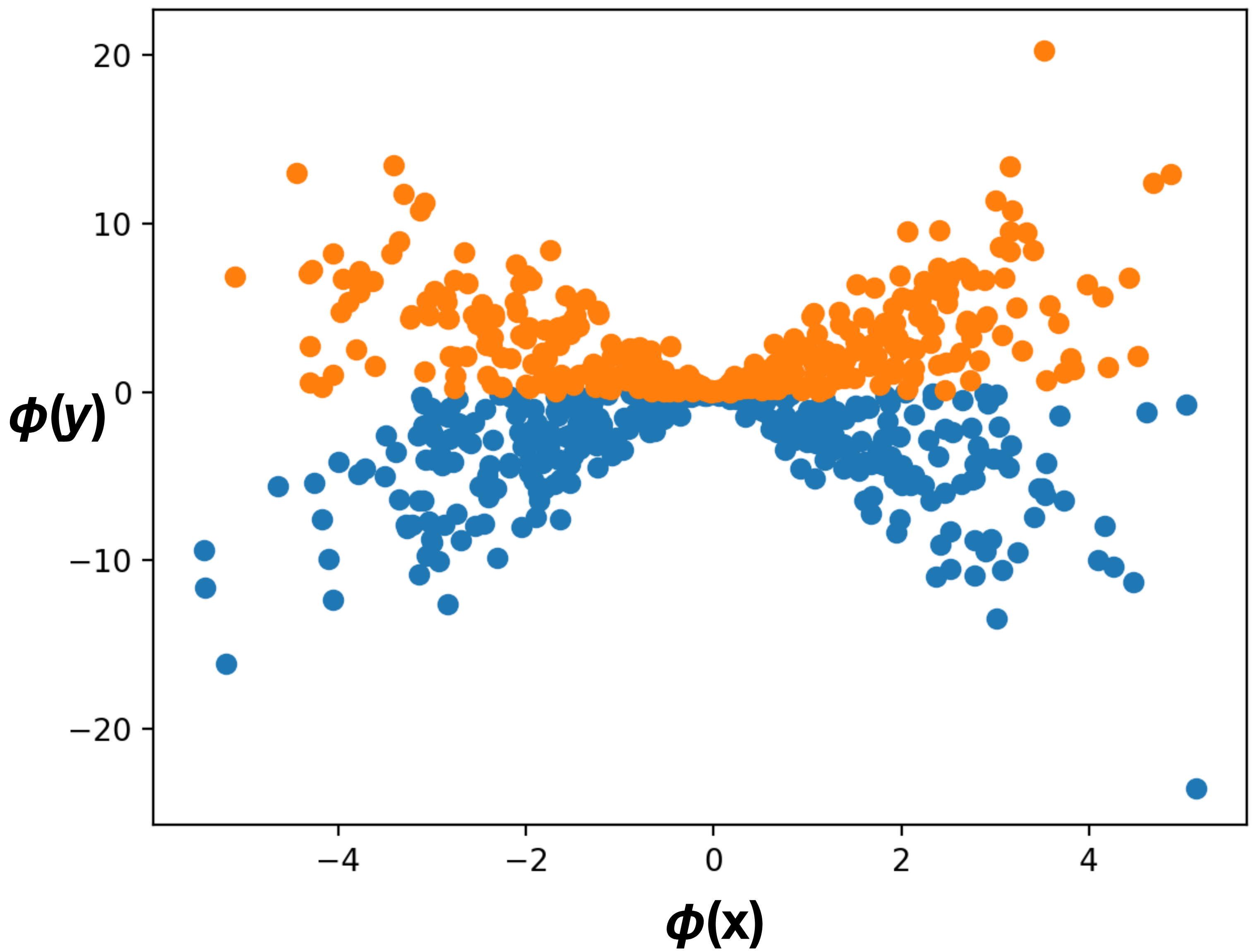
$$\phi(x, y) = \left[\begin{array}{c} x \\ xy \end{array} \right]$$



XOR problem

- We can use a non-linear transformation to make these data linearly separable, e.g.:

$$\phi(x,y) = \left[\begin{array}{c} x \\ xy \end{array} \right]$$



XOR problem

- There are many other transformations we could use. While not visualizable in 2-D, we could use:

$$\phi\left(\left[\begin{array}{c} x \\ y \end{array}\right]\right) = \left[\begin{array}{c} 1 \\ \sqrt{2}x \\ \sqrt{2}y \\ \sqrt{2}xy \\ x^2 \\ y^2 \end{array}\right]$$

(6-dimensional plot goes here)

XOR problem

- It turns out that, through a process known as **kernelization** (more next week), these transformations ϕ can be computed **implicitly**.

$$\phi\left(\left[\begin{array}{c} x \\ y \end{array}\right]\right) = \left[\begin{array}{c} 1 \\ \sqrt{2}x \\ \sqrt{2}y \\ \sqrt{2}xy \\ x^2 \\ y^2 \end{array}\right]$$

(6-dimensional plot goes here)

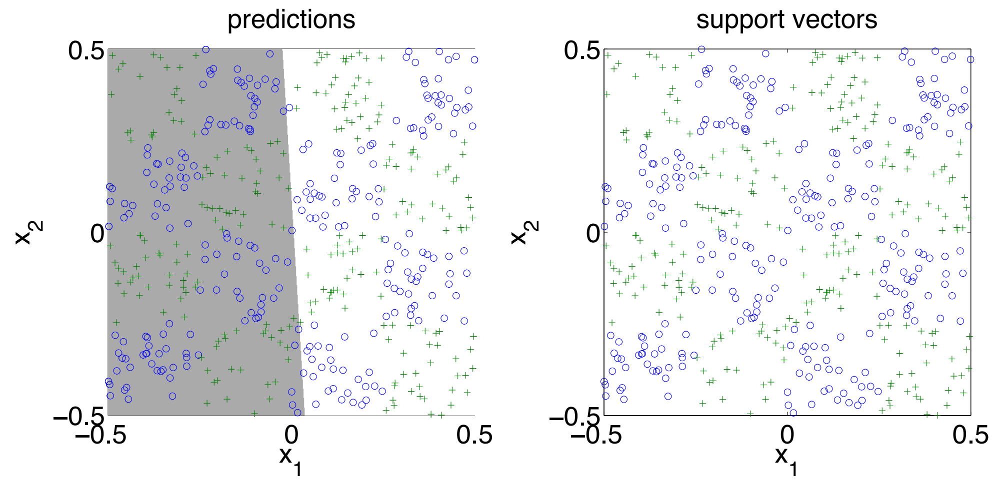
Equivalent to a polynomial kernel of degree 2.

Kernelization

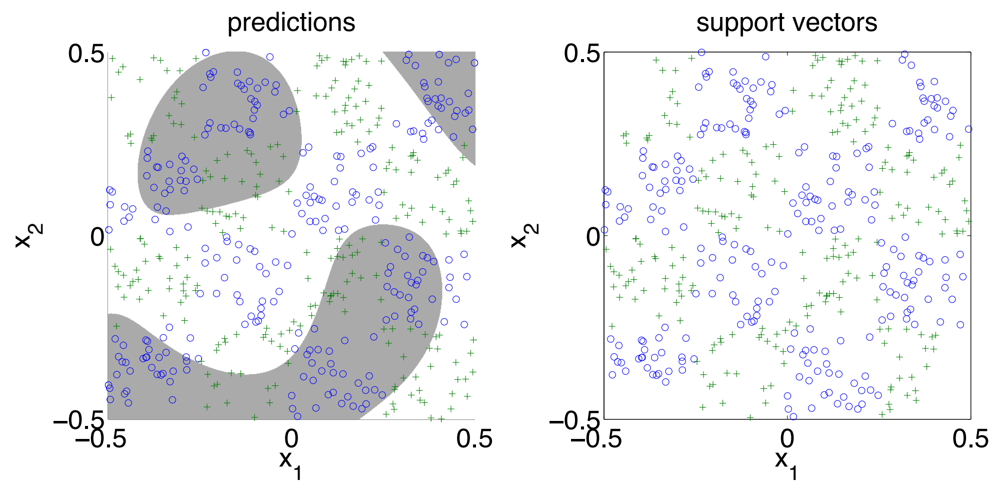
- SVMs **always** try to separate the positive from the negative examples using a **hyperplane** — a linear decision boundary.
- But the hyperplane might exist in a very different (transformed) space than the raw input data.
- In the original input space, the decision boundary can be non-linear.

Non-linear decision boundaries

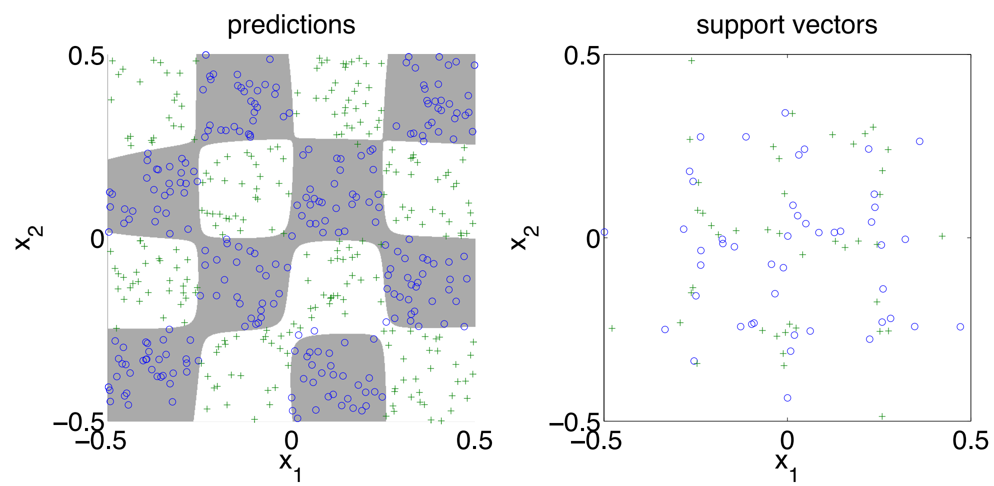
Dataset B, $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = 1 + \mathbf{x} \cdot \mathbf{v}$.



Dataset B, $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = (1 + \mathbf{x} \cdot \mathbf{v})^5$.



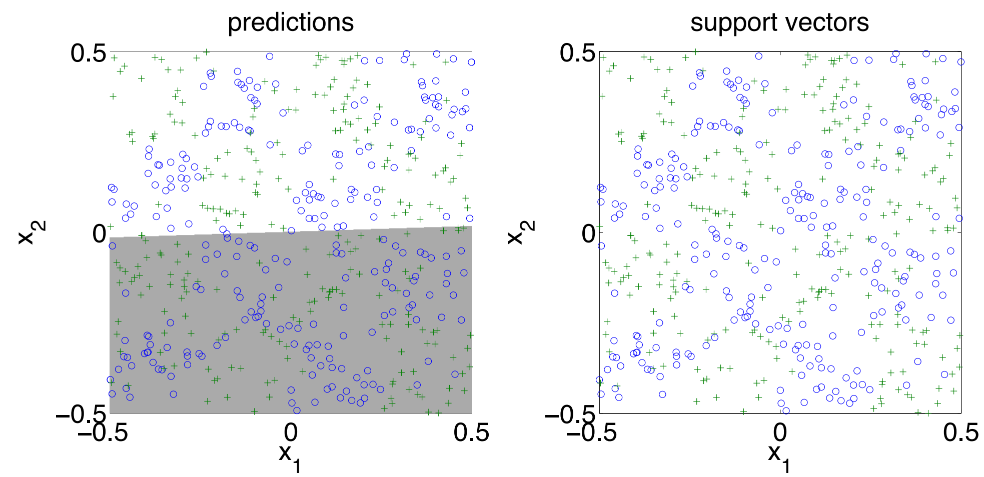
Dataset B, $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = (1 + \mathbf{x} \cdot \mathbf{v})^{10}$.



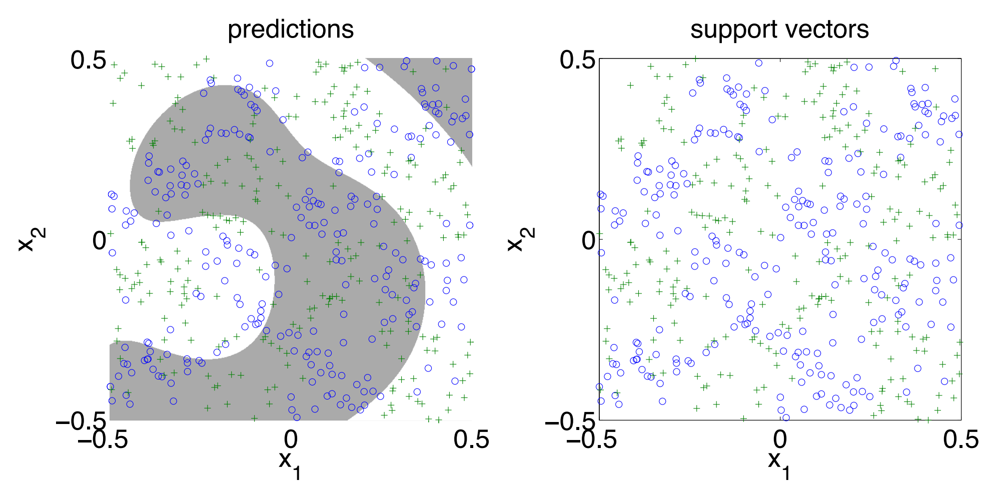
https://people.cs.umass.edu/~domke/courses/sml2010/06kernels.pdf

Non-linear decision boundaries

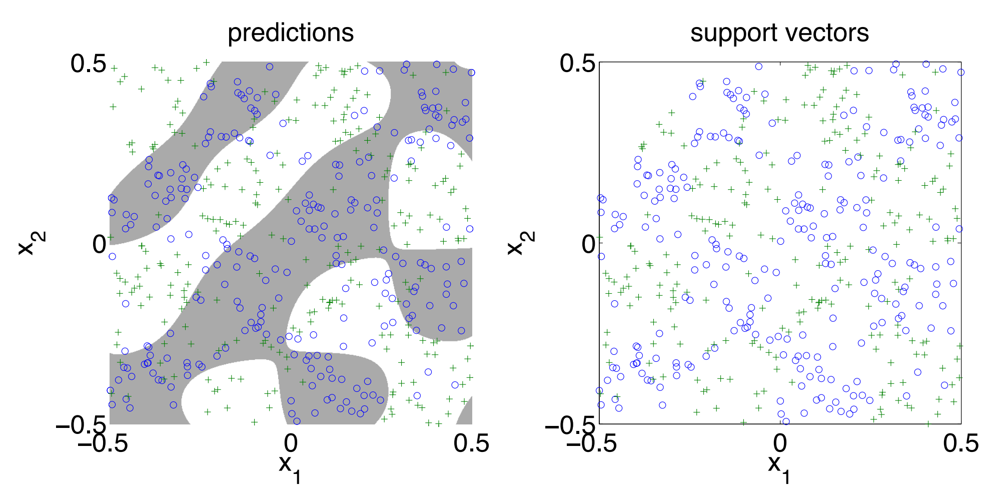
Dataset C (dataset B with noise), $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = 1 + \mathbf{x} \cdot \mathbf{v}$.



Dataset C, $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = (1 + \mathbf{x} \cdot \mathbf{v})^5$.



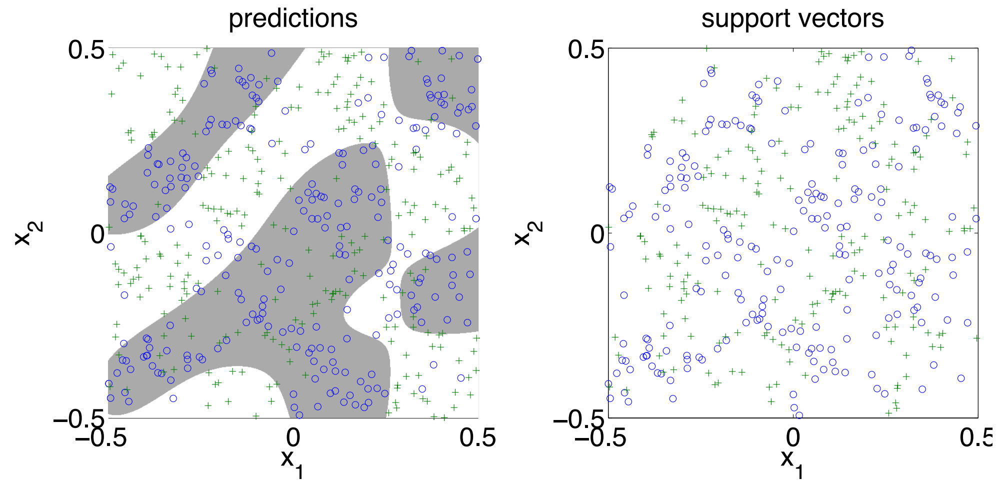
Dataset C, $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = (1 + \mathbf{x} \cdot \mathbf{v})^{10}$.



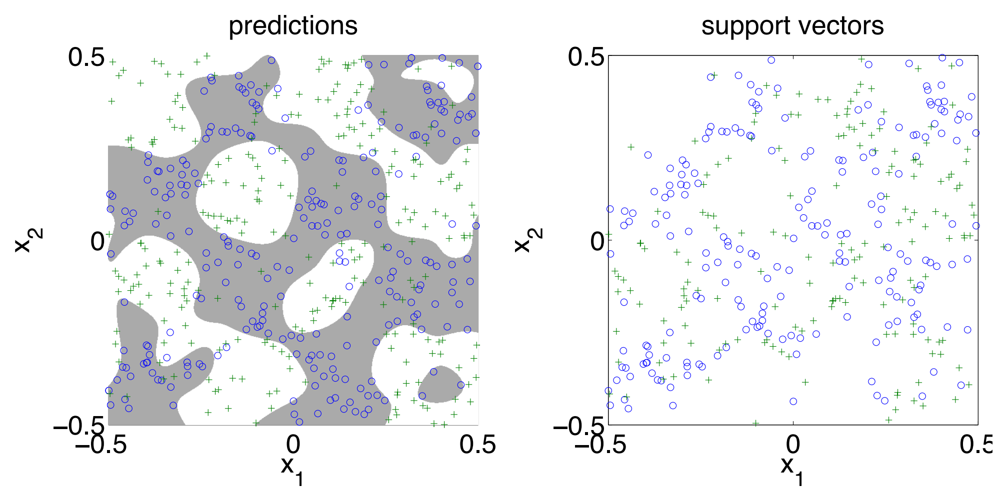
https://people.cs.umass.edu/~domke/courses/sml2010/06kernels.pdf

Non-linear decision boundaries

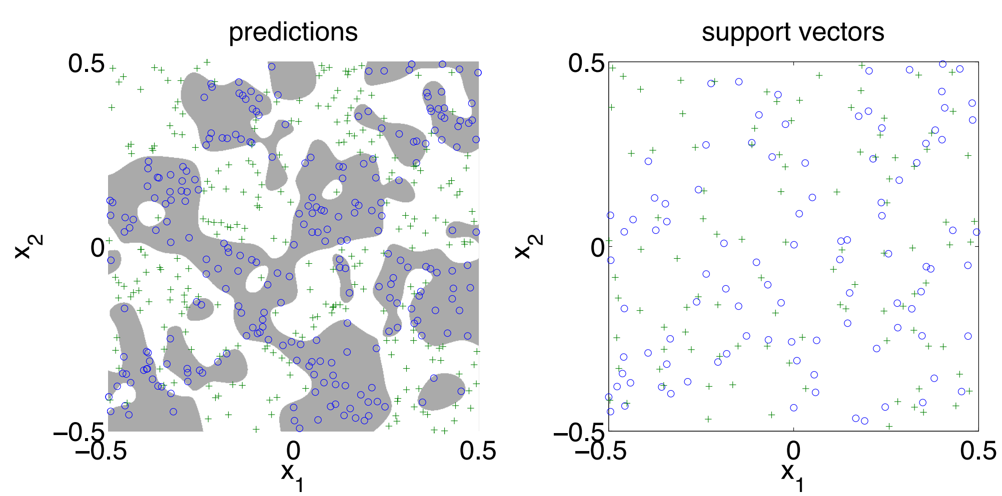
Dataset C (dataset B with noise), $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = \exp\big(-2||\mathbf{x} - \mathbf{v}||^2\big)$.



Dataset C, $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = \exp\big(-20||\mathbf{x} - \mathbf{v}||^2\big)$.



Dataset C, $c = 10^5$, $k(\mathbf{x}, \mathbf{v}) = \exp\big(-200||\mathbf{x} - \mathbf{v}||^2\big)$.



https://people.cs.umass.edu/~domke/courses/sml2010/06kernels.pdf