GENERAL ASSEMBLY

SUPPORT VECTOR MACHINES

Joseph Nelson, Data Science Immersive

AGENDA

- To the basics: class separation and margins

- What is a Support Vector Machine?
- How SVM Works
- Nonlinear SVM (Different Kernels)
- Coding Implementation
- Sklearn documentation

AGENDA

- What is a Support Vector Machine?
- How SVM Works
- Nonlinear SVM (Different Kernels)
- Coding Implementation
- Sklearn documentation

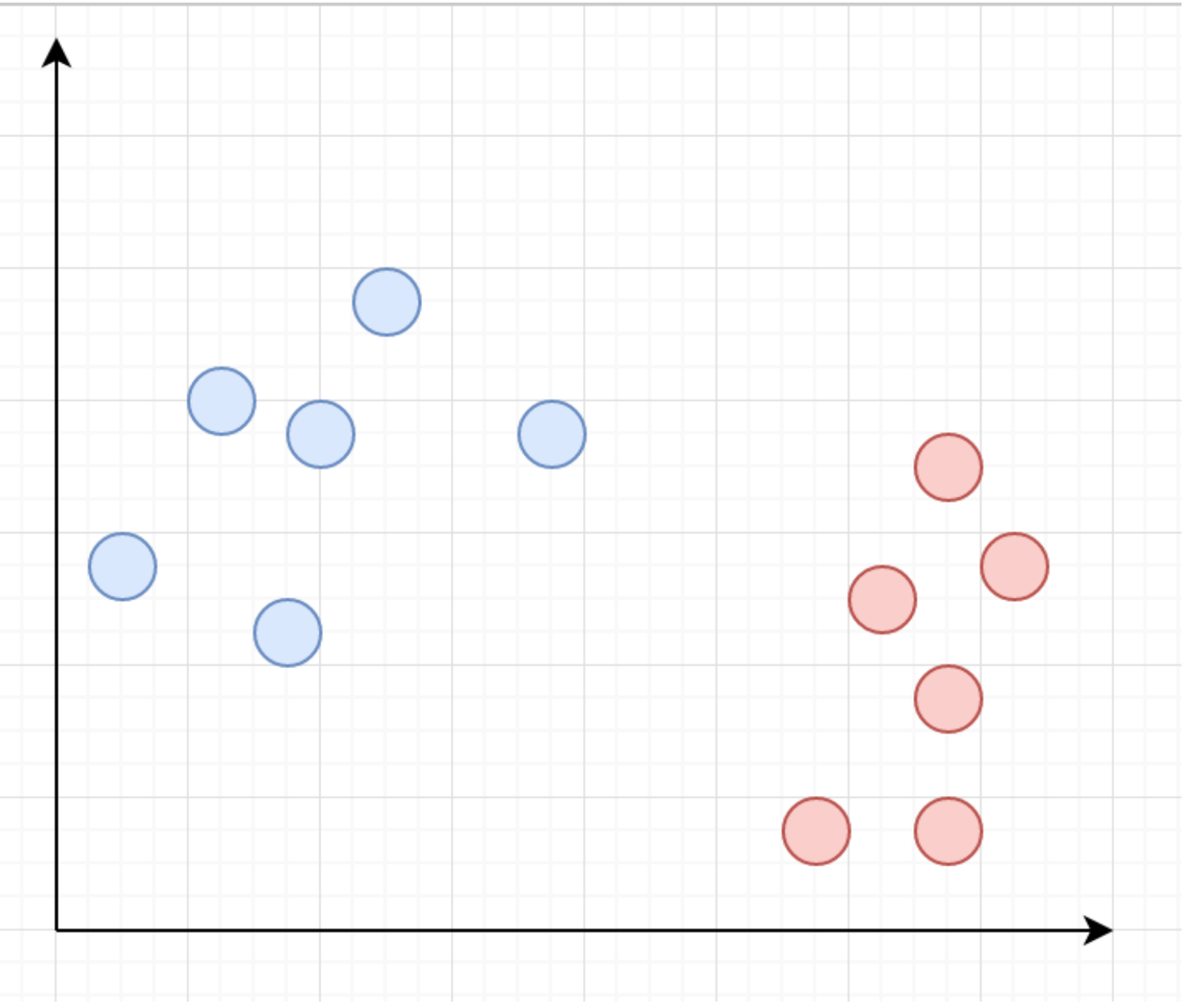
WHAT IS THE GOAL OF CLASSIFICATION?

- 1.) What is classification?
- 2.) What are we striving to do?
- 3.) What does this look like mathematically/visually?

WHAT IS THE GOAL OF CLASSIFICATION?

- Imagine we have read class and blue class, and plotted, they look like this graph.

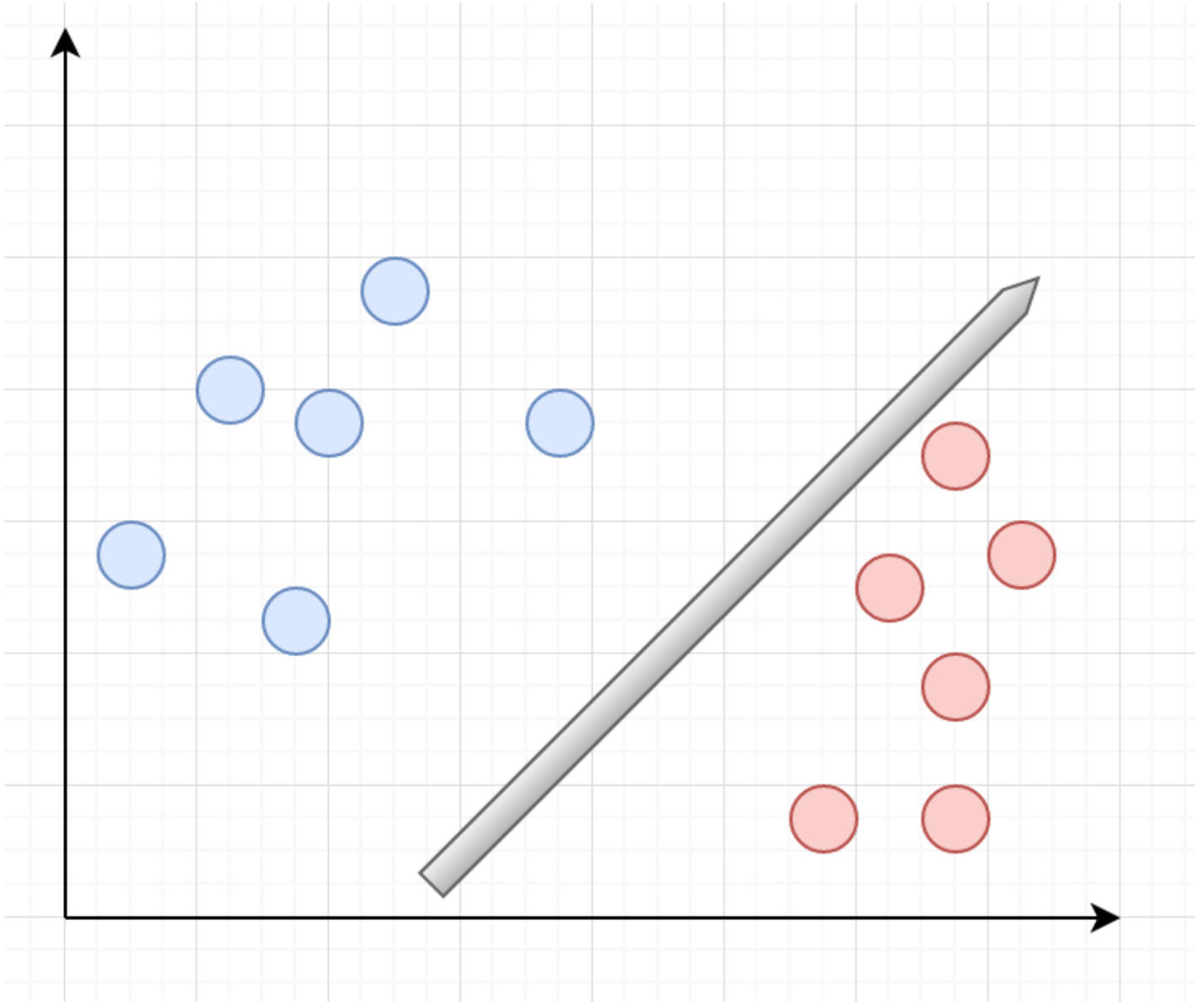
- What is the ideal separation boundary given this distribution?



WHAT IS THE GOAL OF CLASSIFICATION?

- Imagine we have read class and blue class, and plotted, they look like this graph.

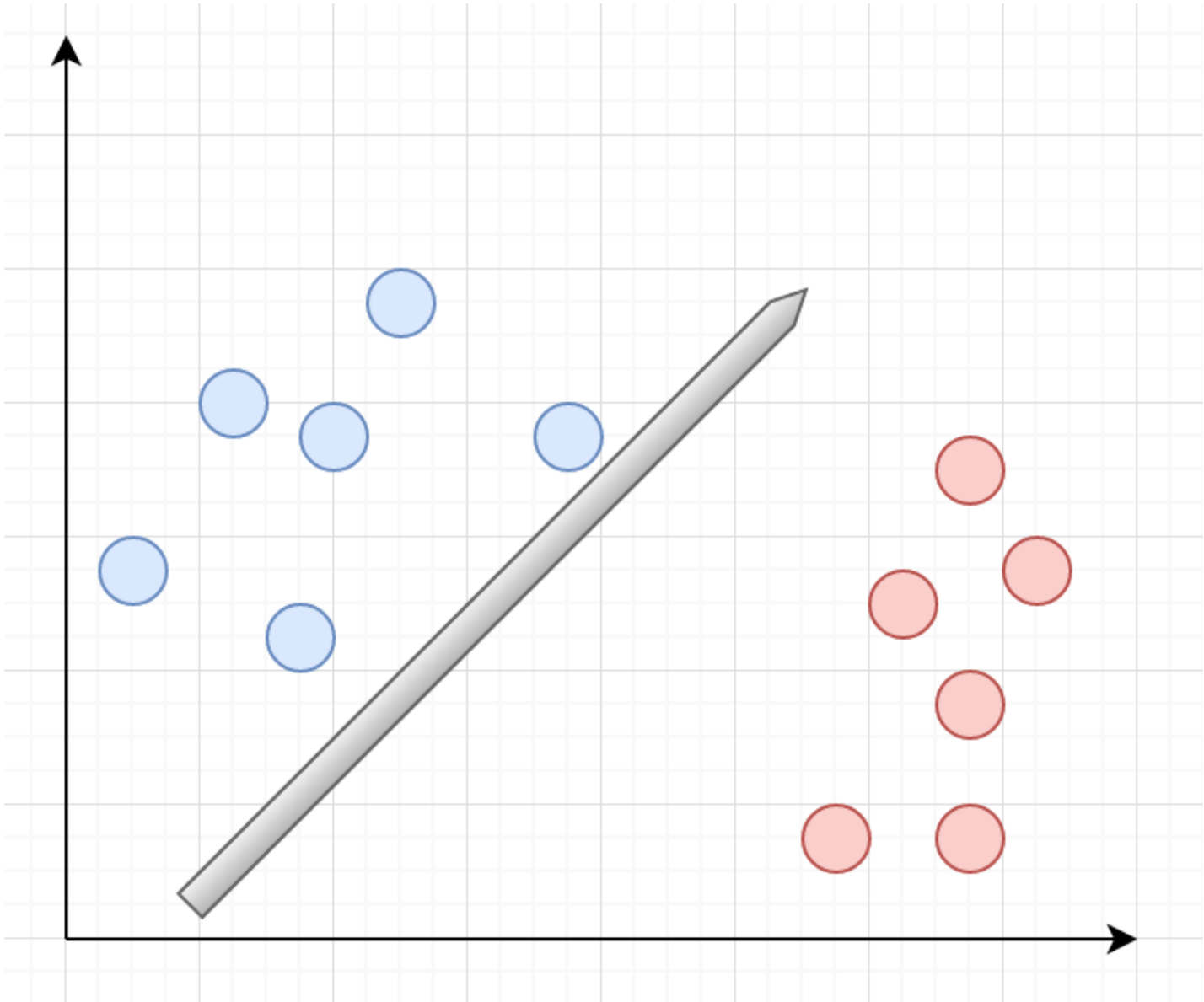
- Option 1



WHAT IS THE GOAL OF CLASSIFICATION?

- Imagine we have read class and blue class, and plotted, they look like this graph.

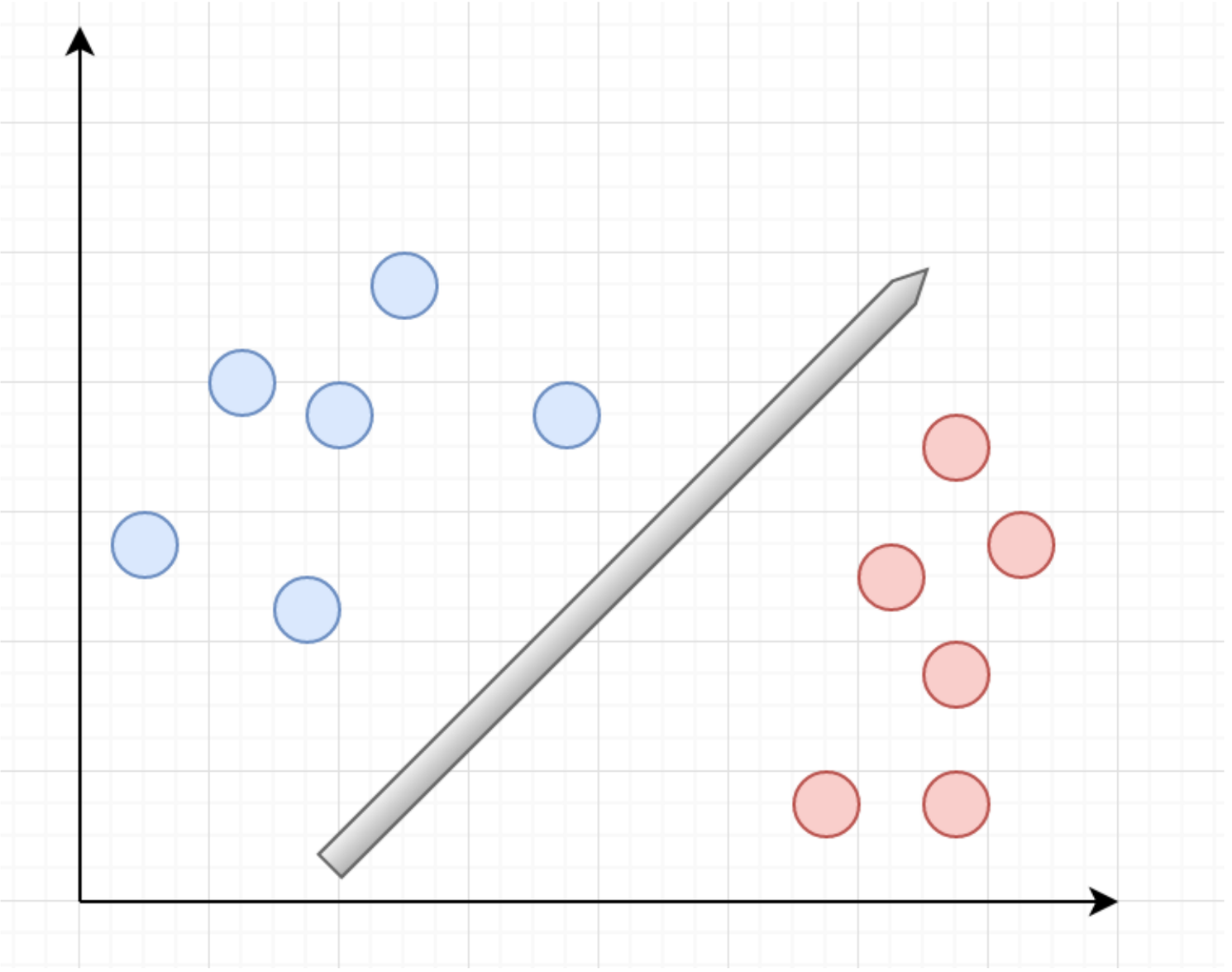
- Option 2



WHAT IS THE GOAL OF CLASSIFICATION?

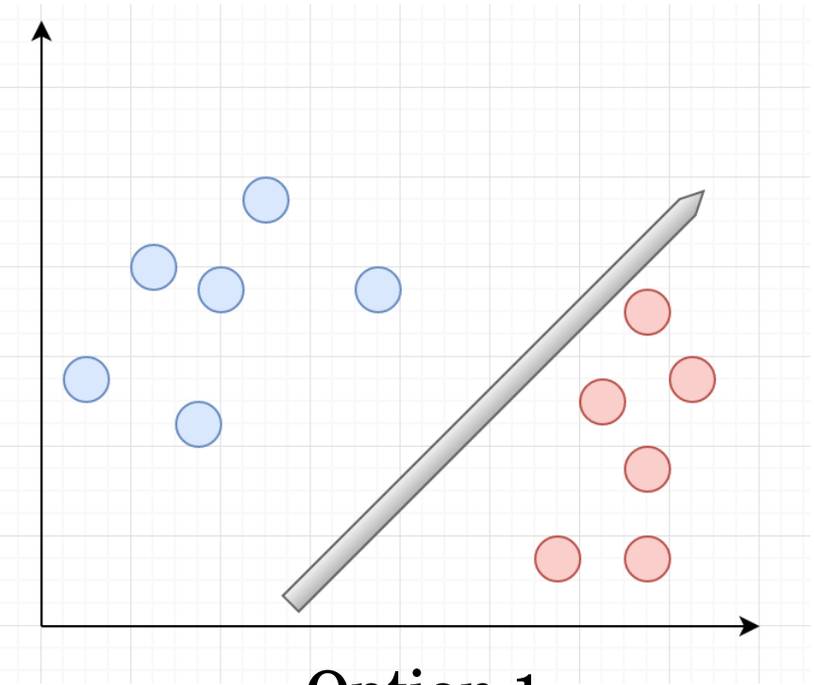
- Imagine we have read class and blue class, and plotted, they look like this graph.

- Option 3

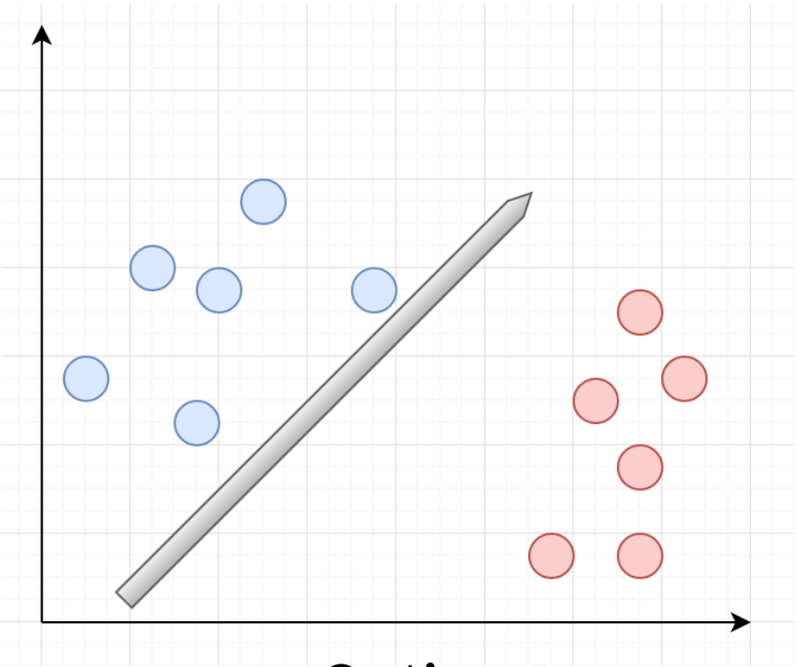


WHAT IS THE GOAL OF CLASSIFICATION?

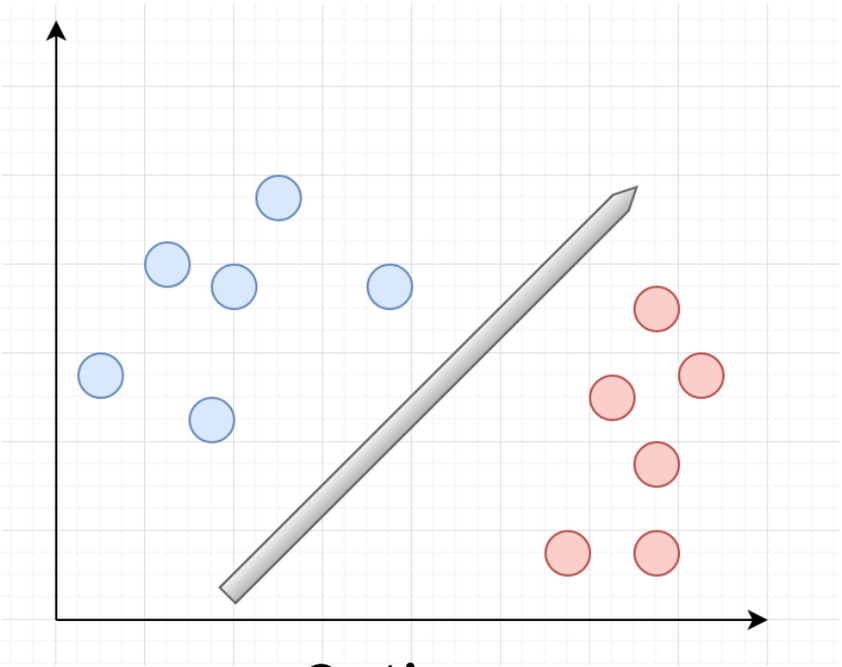
- Which is best?



- Option 1



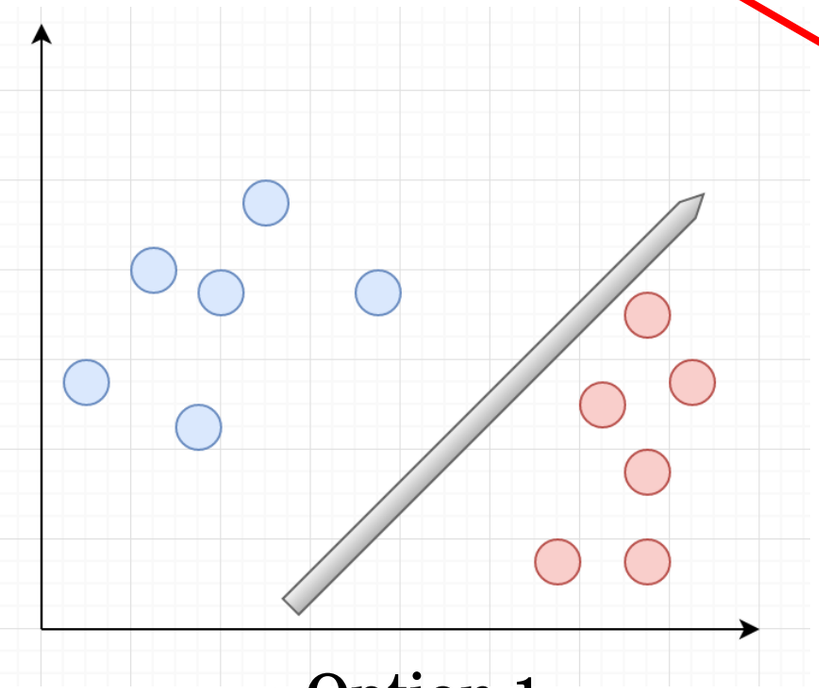
- Option 2



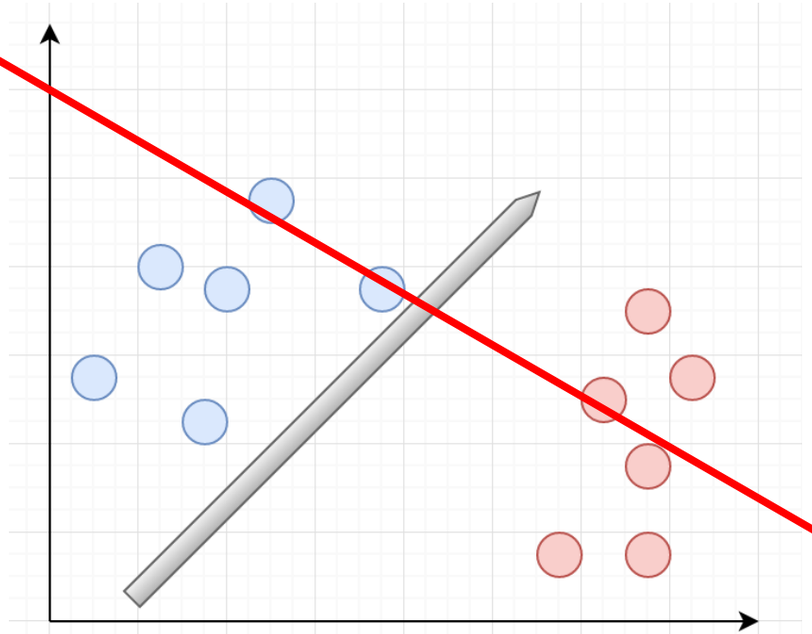
- Option 3

INTUITION

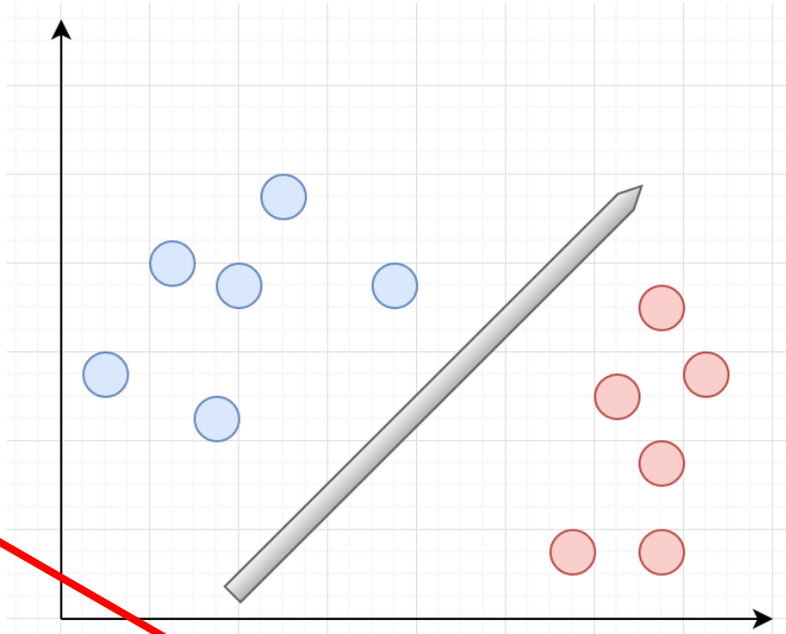
- Winner



- Option 1



- Option 2

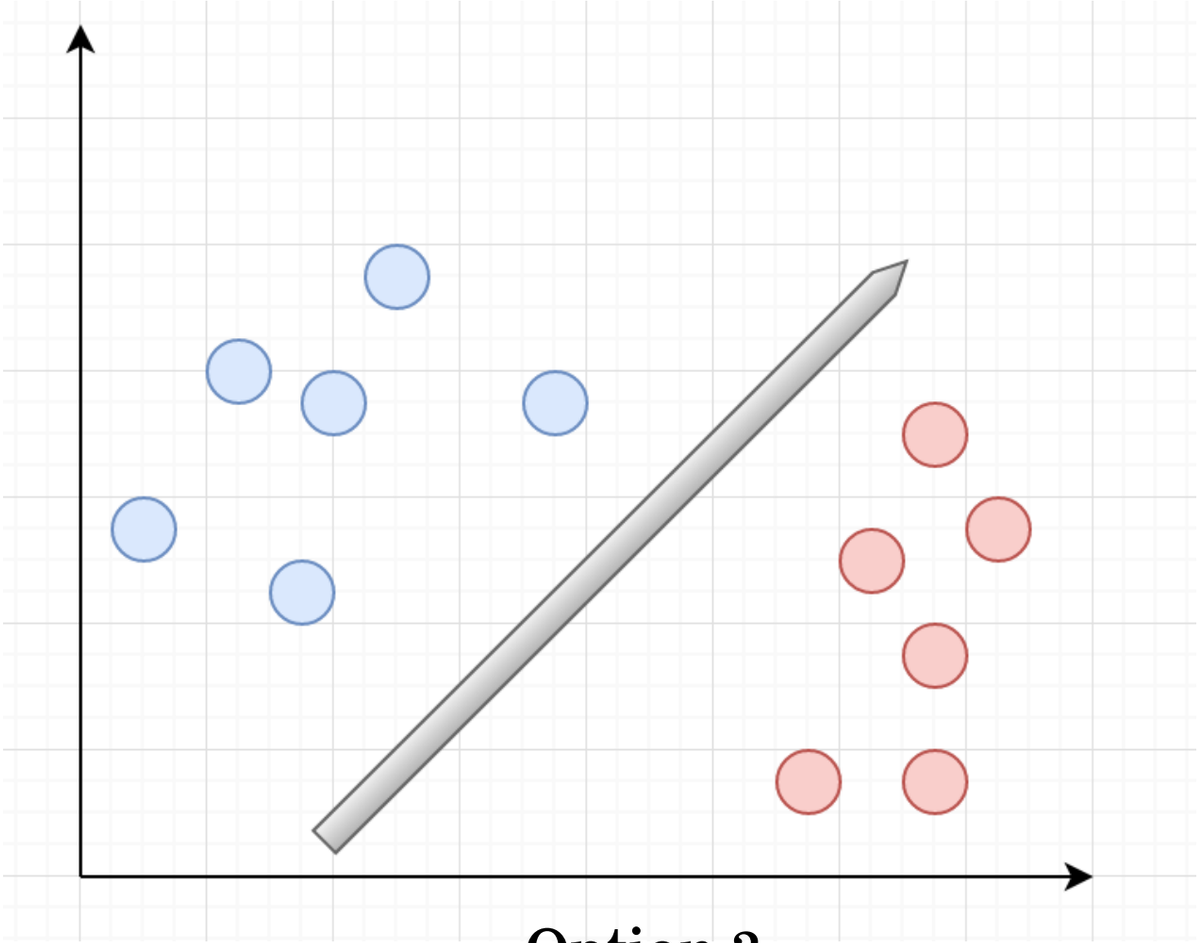


- Option 3

INTUITION

- Winner. Why?

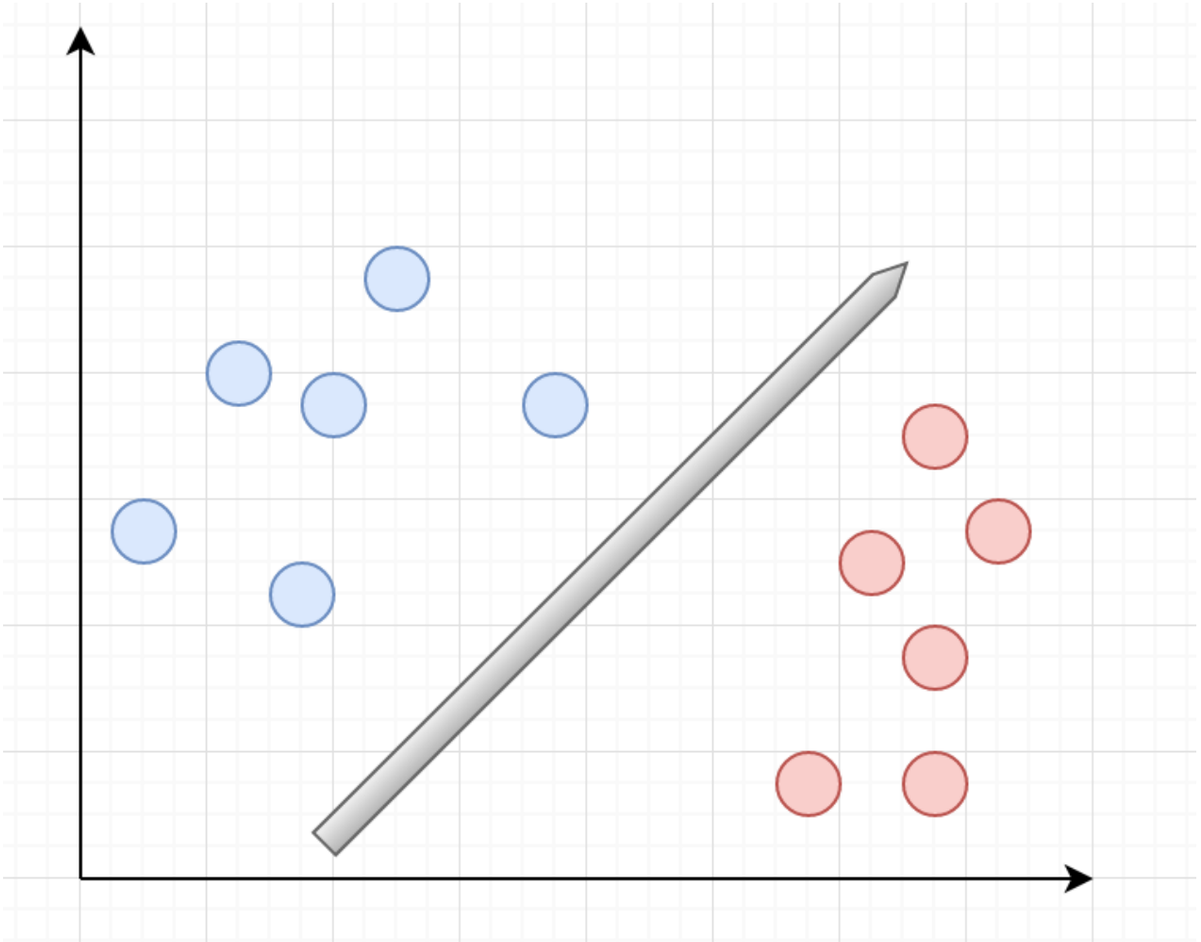
- What specific conditions can you define that makes this optimal?



- Option 3

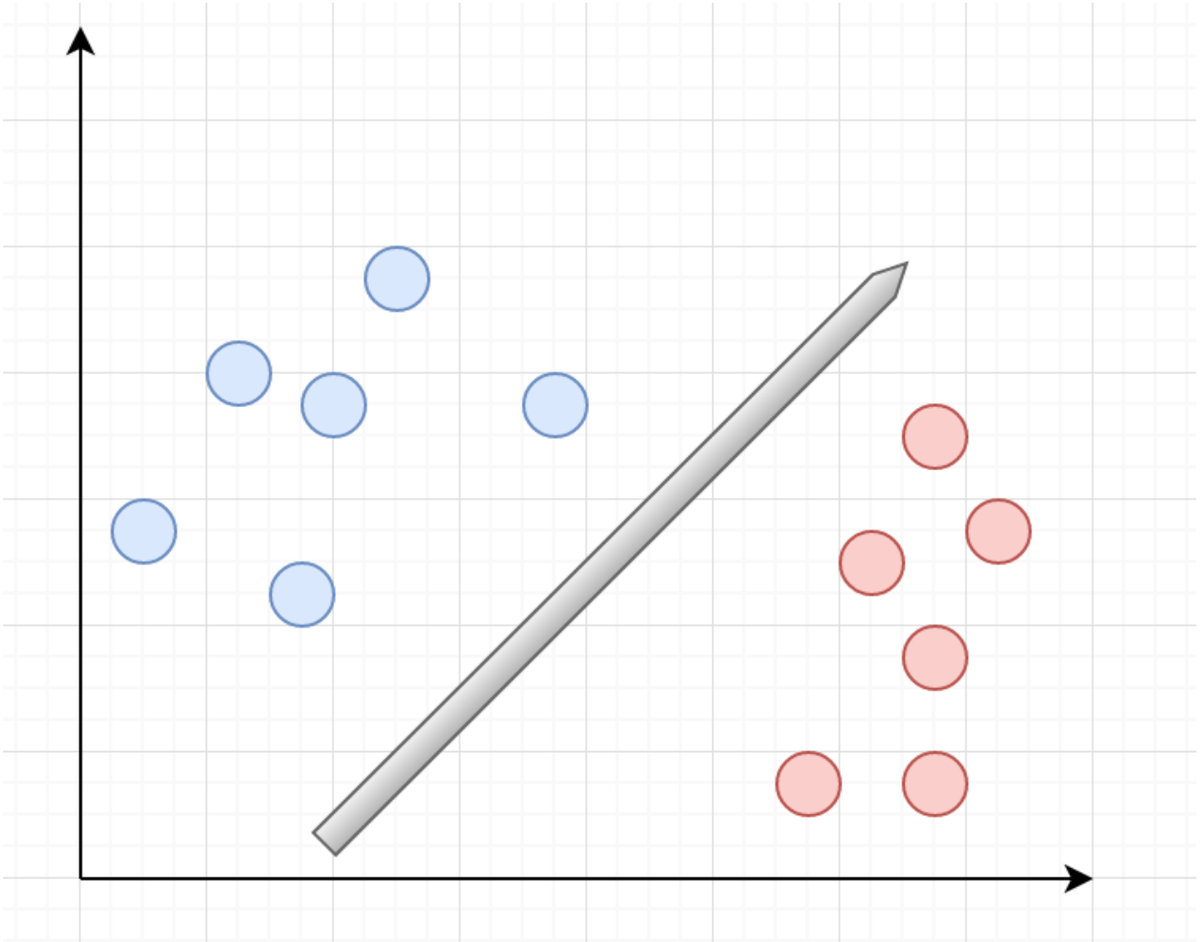
INTUITION

- In general, lots of possible solutions for a,b,c.
- Support vector machine finds an optimal solution (with respect to what cost?)



INTUITION

- In general, lots of possible solutions for a,b,c.
- Support vector machine finds an optimal solution (with respect to what cost?)

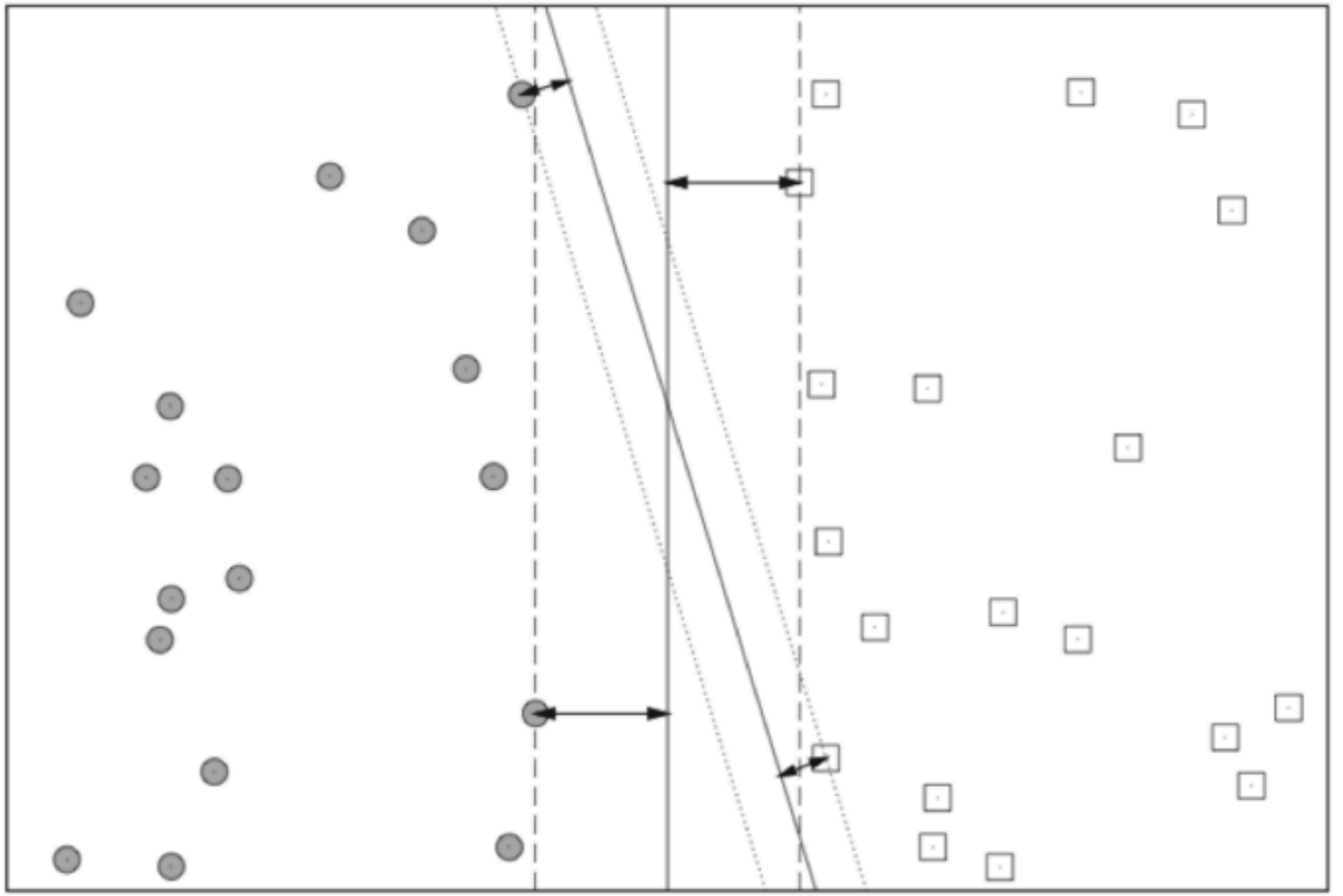


WHAT IS A SUPPORT VECTOR MACHINE?

- A Support Vector Machine is a binary linear classifier whose decision boundary is explicitly constructed to minimize generalization error
- Binary classifier: solves a two-class problem
- Linear classifier: creates a linear decision boundary

WHAT IS A SUPPORT VECTOR MACHINE?

- The decision boundary is derived using geometric reasoning (as opposed to the algebraic reasoning we've used to derive other classifiers). The generalization error is equated with the geometric concept of margin, which is the region along the decision boundary that is free of data points.



WHAT IS A SUPPORT VECTOR MACHINE?

- Finding the optimal hyperplane be like…

WHAT IS A SUPPORT VECTOR MACHINE?

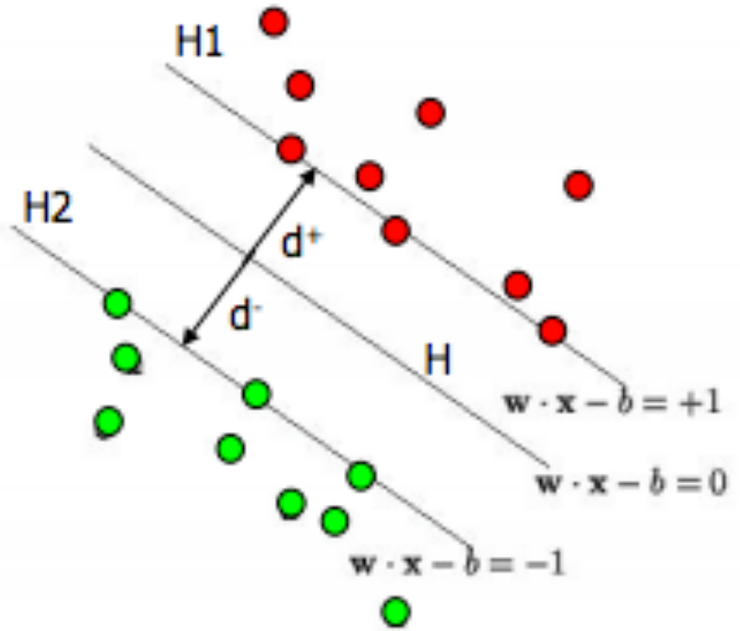
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane (UCF)
- The quest to find the optimal hyper plane is an optimization problem
- Because of this, support vector machines have discriminative solutions

MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- Let's set up some definitions

- $x_i * w + b \geq +1$ when $y_i = +1$
- $x_i * w + b \leq -1$ when $y_i = -1$

- H1: $x_i * w + b = +1$
- H2: $x_i * w + b = -1$



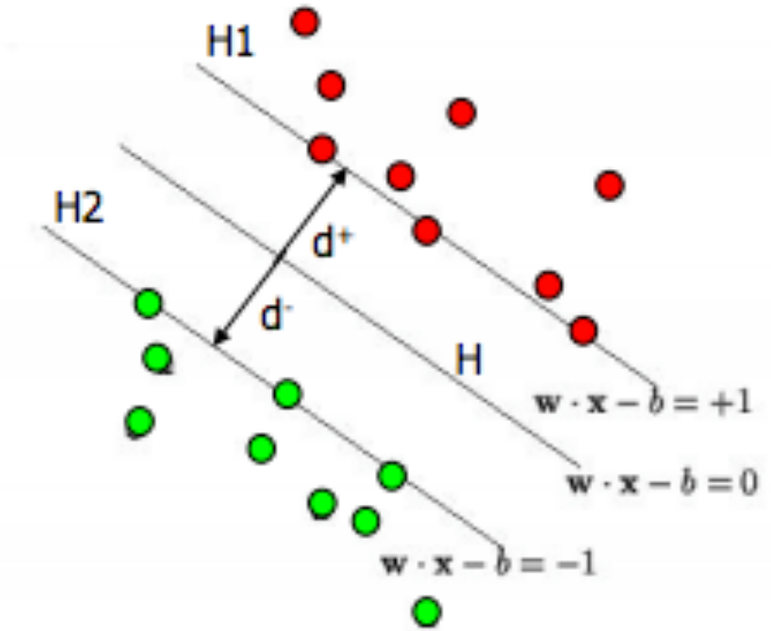
- d+ = the shortest distance to the closest positive point
- d- = the shortest distance to the closest negative point

SIDENOTE: REMEMBER LINEAR ALGEBRA REPRESENTS LINES

- Remember that $y = ax+b$ is the same as $y - ax - b = 0$
- Given two vectors $w\begin{pmatrix} -b \\ -a \\ 1 \end{pmatrix}$ and $x\begin{pmatrix} 1 \\ x \\ y \end{pmatrix}$
- $w^T x = -b\times(1)+(-a)\times x+1\times y$
- $w^T x = y-ax-b$

MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- Given these conditions, note:
- To have a closed form solution, our planes MUST be linearly separable.
- We will not arrive at a solution if this is not the case. (For now, to be resolved later)



MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- So, what about solving those two constraints to find the ideal hyperplane?

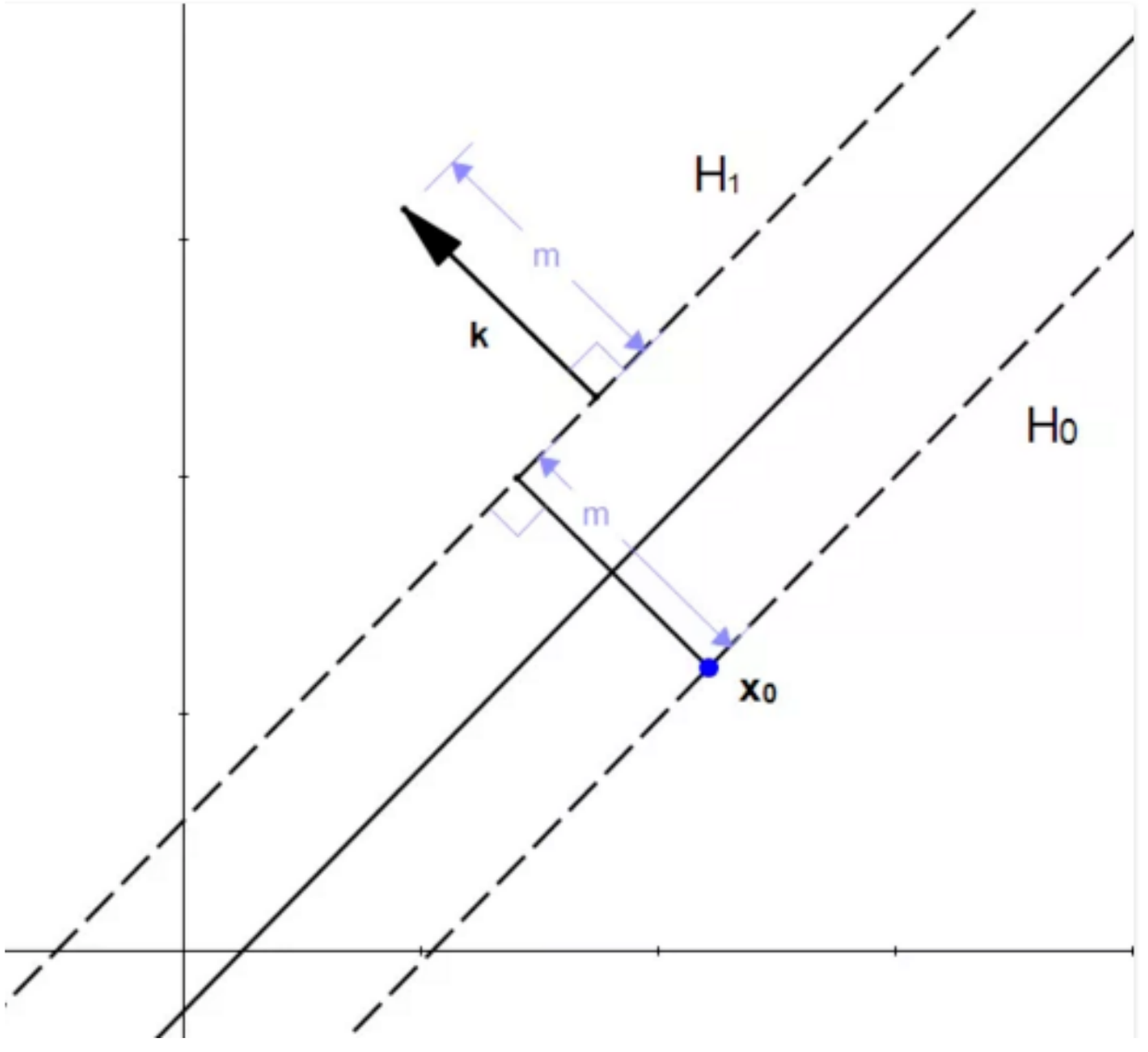
- We have: $w \cdot x_i + b \leq -1$ for class -1 (5)
- Multiply both sides by -1: $y_i(w \cdot x_i + b) \geq y_i(-1)$
- Which means we can write (5) as: $y_i(w \cdot x_i + b) \geq 1$ for x_i having the class -1 (6)
- Positive classes are identical: $y_i(w \cdot x_i + b) \geq 1$ for x_i having the class 1 (7)
- Combine (6) and (7): $y_i(w \cdot x_i + b) \geq 1$ for all $1 \leq i \leq n$

MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- We have the condition!
- $y_i(w \cdot xi + b) \geq 1$ for all $1 \leq i \leq n$
- Now we gotta solve this…

MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- $y_i(w \cdot xi + b) \geq 1$ for all $1 \leq i \leq n$
- We are producing our own unit vector (measuring stick) that is orthagonal (perpendicular) to our hyperplanes
- We want to maximize this length



MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- The math associated with solving this yields this condition: $$m = \frac{2}{\|\mathbf{w}\|}$$
- This norm measures the distance between the hyperplanes. That means we want to maximize it.

MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- The math associated with solving this yields this condition: $m = \frac{2}{\|\mathbf{w}\|}$
- This norm measures the distance between the hyperplanes. That means we want to maximize it.
- But minimization problems are nicer: ½ $w^T w$
- Solving this relies on a quadratic programming problem use Lagrangian multiplier method

MATHEMATICALLY, WHAT ARE WE DEALING WITH?

- “Solving this relies on a quadratic programming problem use Lagrangian multiplier method”
- Thanks, Joseph
- But for real: http://www.cs.ucf.edu/courses/cap6412/fall2009/papers/Berwick2003.pdf
- http://www.svm-tutorial.com/2015/06/svm-understanding-math-part-3/

BACK OUT OF MATH-LAND

- The goal of an SVM is to create the linear decision boundary with the largest margin. This is commonly called the maximum margin hyperplane (MMH).

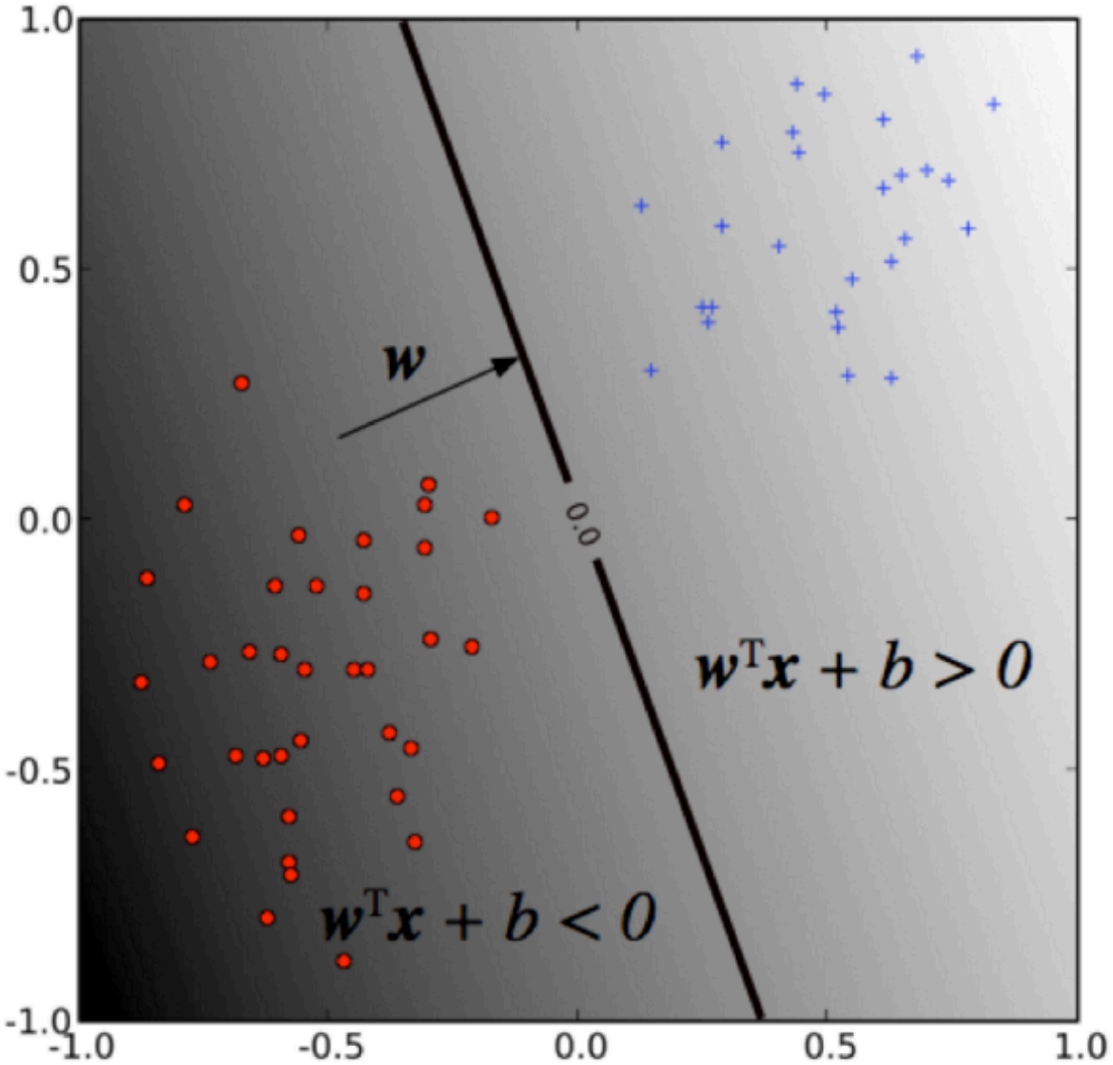
- Nonlinear applications of SVM rely on an implicit (nonlinear) mapping that sends vectors from the original feature space K into a higher-dimensional feature space K'. Nonlinear classification in K is then obtained by creating a linear decision boundary in K'. In practice, this involves no computations in the higher dimensional space, thanks to what is called the kernel trick.

DECISION BOUNDARY

- The decision boundary (MMH) is derived by the discriminant function:

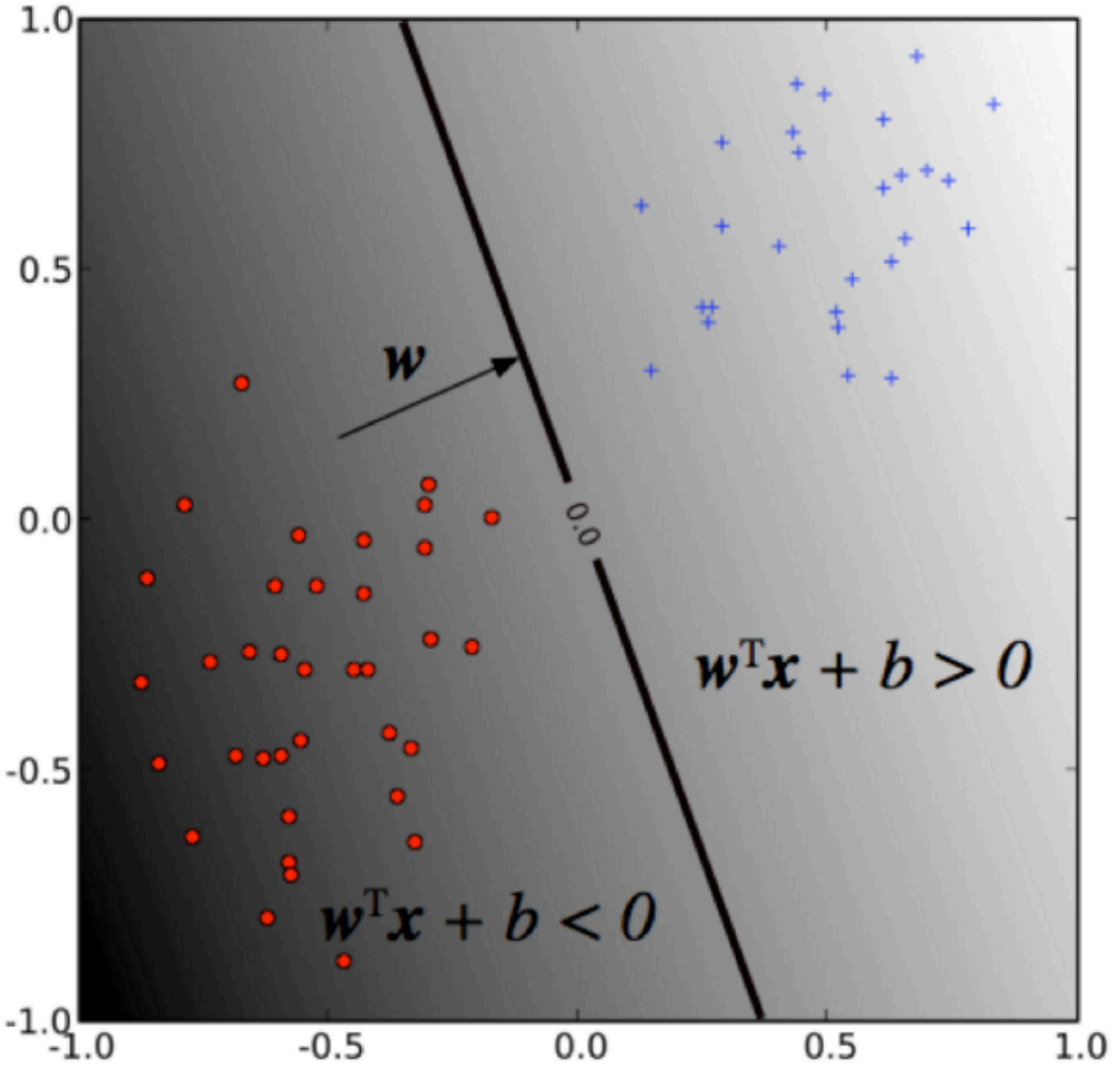
$f(x) = w^T x + b$

- where w is the weight vector and b is the bias. The sign of f(x) determines the (binary) class label of a record x.



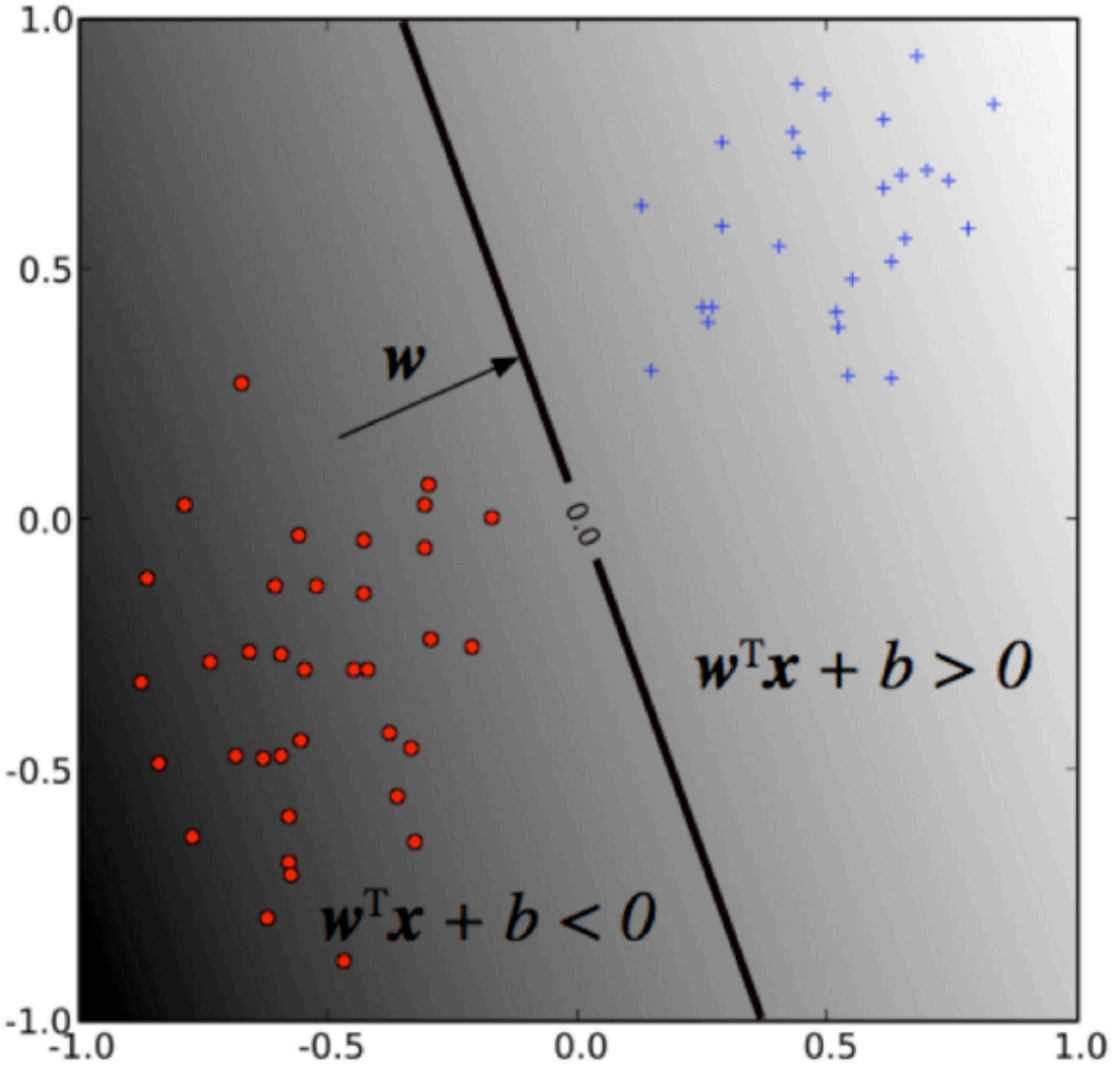
DECISION BOUNDARY

- Remember: SVM solves for the decision boundary that minimizes generalization error, or equivalently, that has the maximum margin. These are equivalent since using the MMH as the decision boundary minimizes the probability that a small perturbation in the position of a point produces a classification



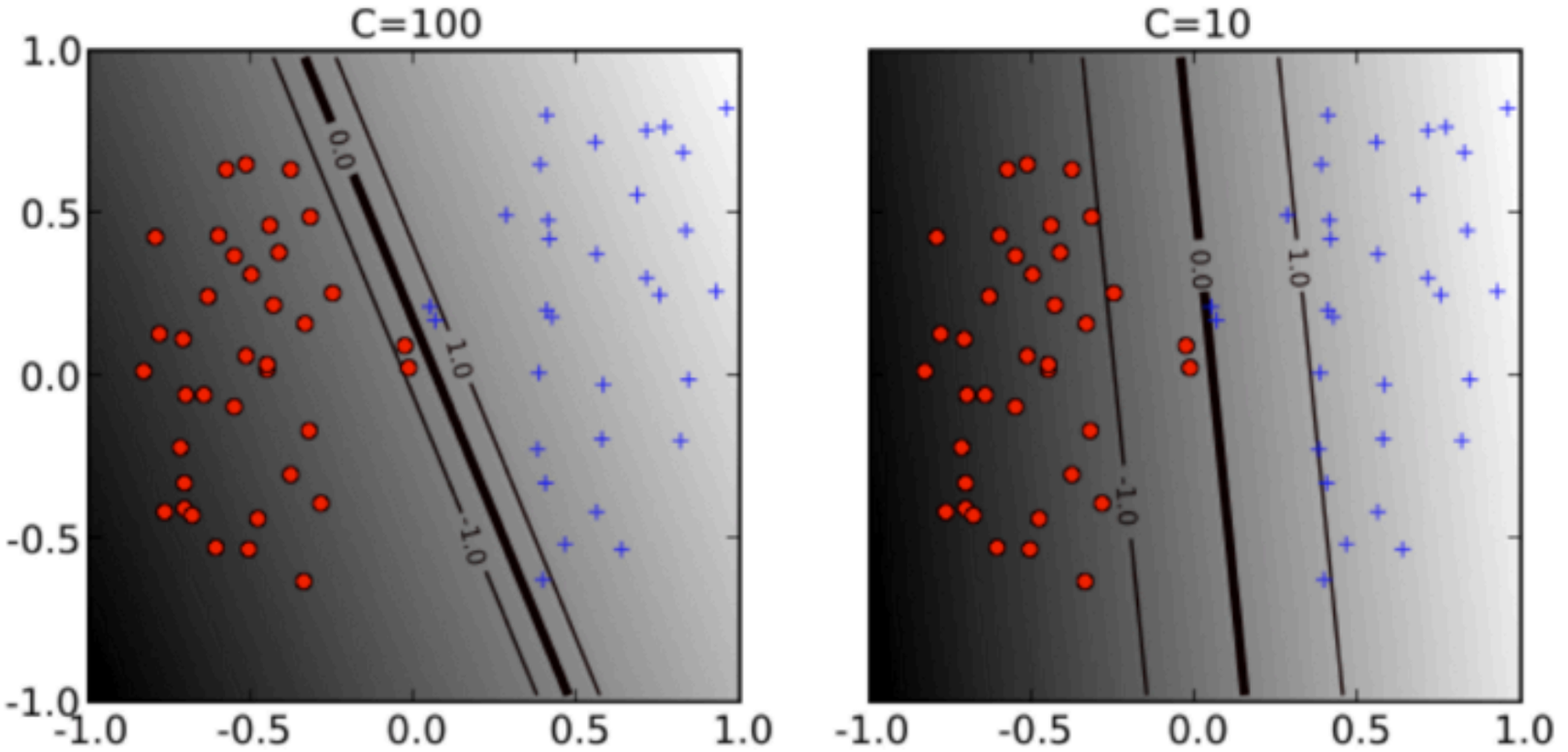
DECISION BOUNDARY

- Selecting the MMH is a straightforward exercise in analytic geometry
- The margin depends only on a subset of the training data; namely, those points that are nearest to the decision boundary. These points are called the support vectors. The other points (far from the decision boundary) don't affect the construction of the MMH at all.



SOFT MARGIN, SLACK VARIABLES

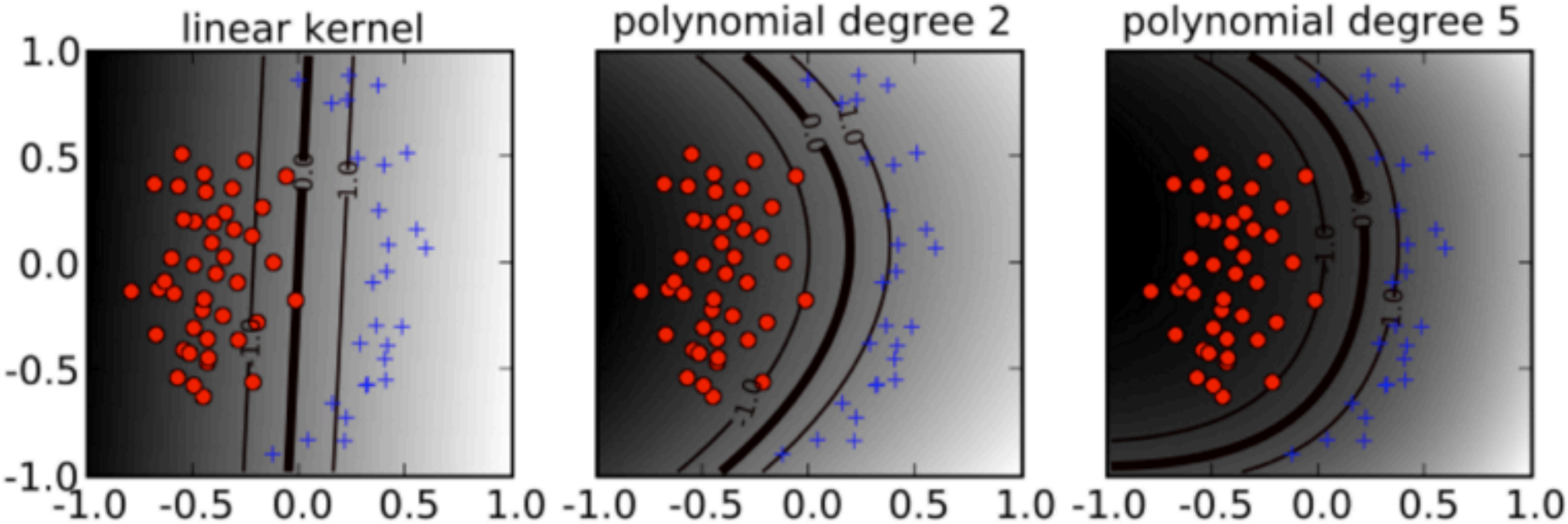
- Class overlap is achieved by relaxing the minimization problem or softening the margin.
- The hyper-parameter C (soft-margin constant) controls the overall complexity by specifying penalty for training error. This is yet another example of regularization.



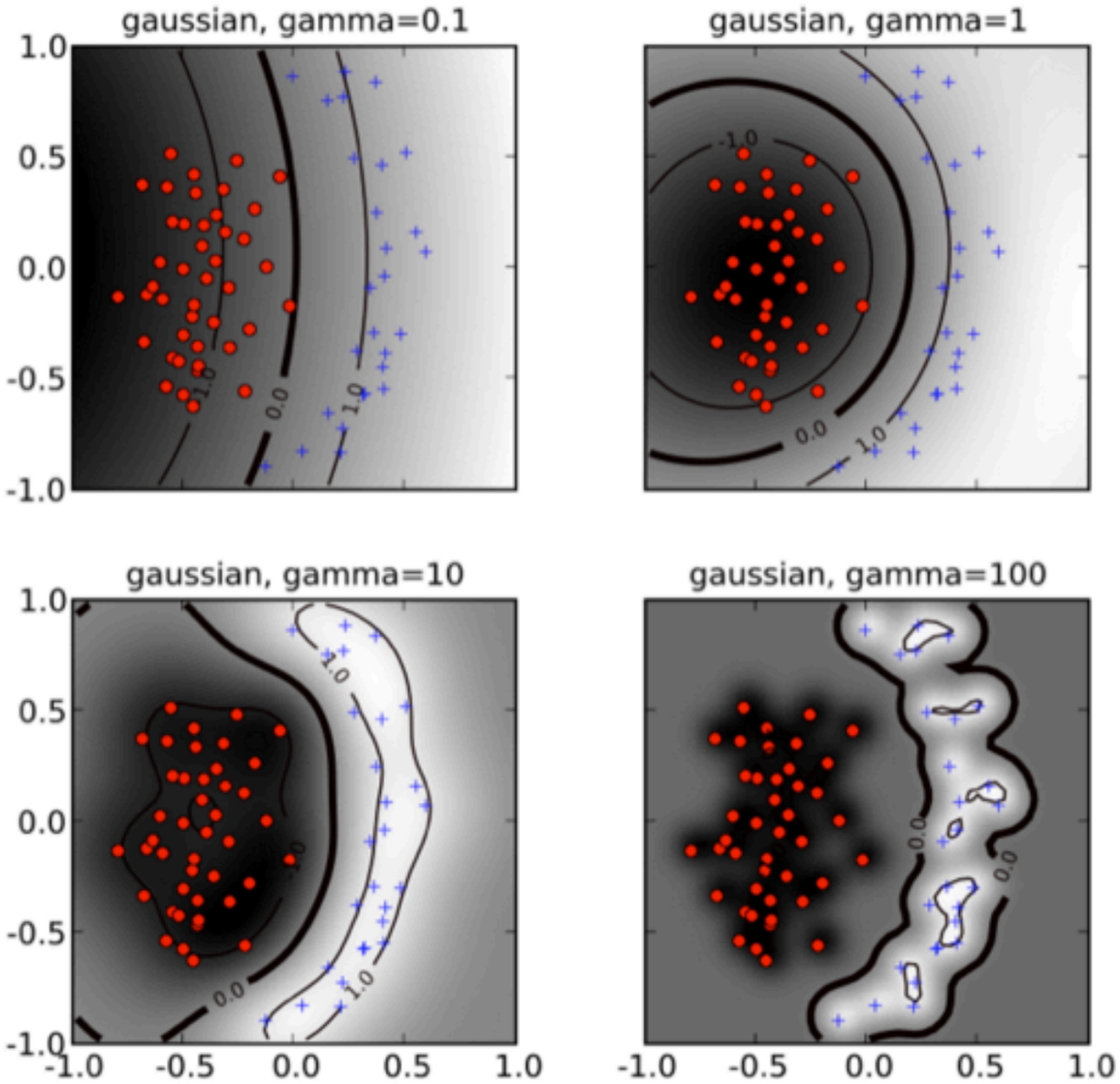
NONLINEAR SVM

- At its core, the optimization problem only has x as an inner product
- $f(x) = w^T x + b$
- We can replace this inner product with a more complex function: this is called the kernel trick
- Popular kernels:
- Linear k(x, x') = x ^ T x’
- Polynomial k(x, x') = (x ^ T x' + 1) ^ d
- Gaussian kernel (radial basis function) k(x, x') = \exp{(-\gamma||x - x'|| ^ 2)}

NONLINEAR SVM



NONLINEAR SVM



REMEMBER...

- “SVMs are among the best (and many believe are indeed the best) “off-the-shelf” supervised learning algorithm.”
  - Andrew Ng