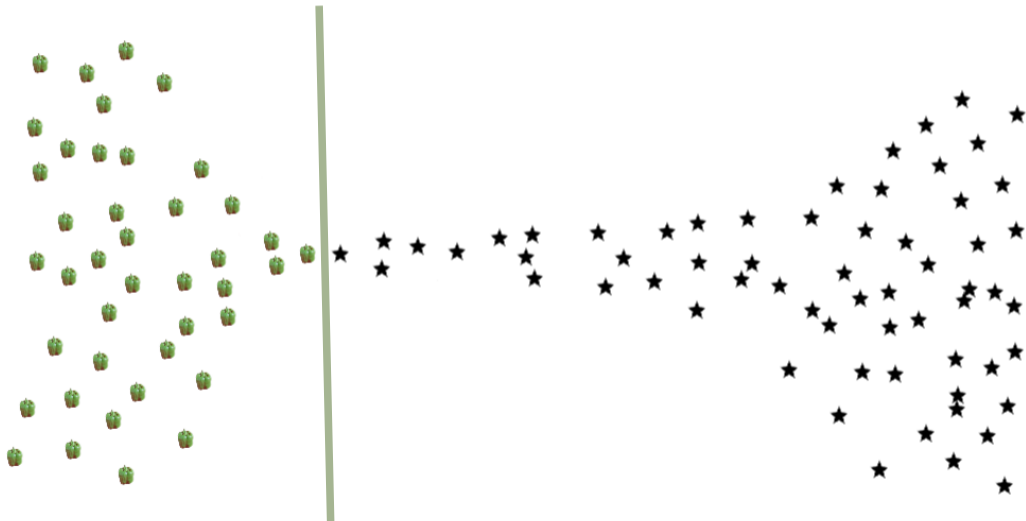


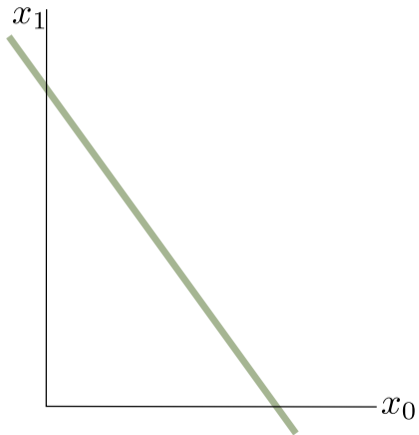
# **Support Vector Machine**

# Classification

Minimize the **number** of misclassified labels instead of probability.



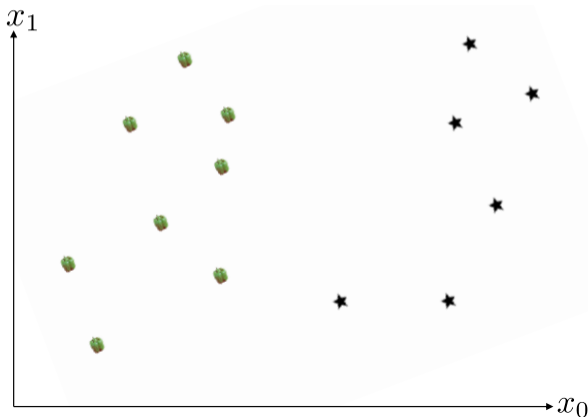
# Line equation



# Classification

Data:  $(X^{(1)}, y^{(1)}), (X^{(2)}, y^{(2)}), \dots, (X^{(n)}, y^{(n)})$ .

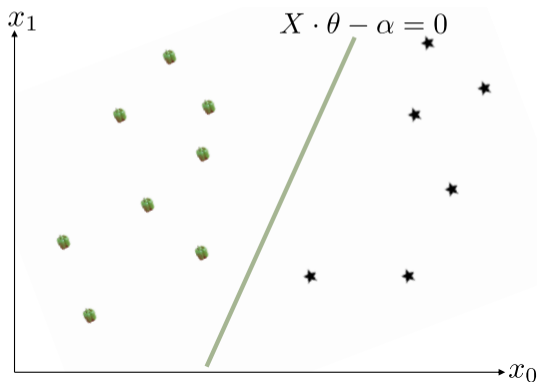
$d + 1$  variables:  $X^{(i)} = (x_0^{(i)}, x_1^{(i)}, \dots, x_d^{(i)})$ ,  $y^{(i)} \in \{+1, -1\}$



# Classification

The **Support Vector Machines** (SVM) is defined by the parameters  $\theta = (\theta_1, \dots, \theta_d)$  and  $\alpha$  which minimize the number of misclassified points

$$\hat{y}^{(i)} = \begin{cases} +1 & \text{if } X^{(i)} \cdot \theta - \alpha > 0 \\ -1 & \text{if } X^{(i)} \cdot \theta - \alpha \leq 0 \end{cases}$$



# Support Vector Machines

Data:  $(X^{(1)}, y^{(1)}), (X^{(2)}, y^{(2)}), \dots, (X^{(n)}, y^{(n)})$

Predictions:  $(X^{(1)}, \hat{y}^{(1)}), (X^{(2)}, \hat{y}^{(2)}), \dots, (X^{(n)}, \hat{y}^{(n)})$

$$\text{where } \hat{y}^{(i)} = \begin{cases} +1 & \text{if } X^{(i)} \cdot \theta - \alpha > 0 \\ -1 & \text{if } X^{(i)} \cdot \theta - \alpha \leq 0 \end{cases}$$

How do we **check** if a point is correctly classified?

# Support Vector Machines

Data:  $(X^{(1)}, y^{(1)}), (X^{(2)}, y^{(2)}), \dots, (X^{(n)}, y^{(n)})$

Predictions:  $(X^{(1)}, \hat{y}^{(1)}), (X^{(2)}, \hat{y}^{(2)}), \dots, (X^{(n)}, \hat{y}^{(n)})$

$$\text{where } \hat{y}^{(i)} = \begin{cases} +1 & \text{if } X^{(i)} \cdot \theta - \alpha > 0 \\ -1 & \text{if } X^{(i)} \cdot \theta - \alpha \leq 0 \end{cases}$$

How do we **check** if a point is correctly classified?

$$y^{(i)} = +1 \text{ and } X^{(i)} \cdot \theta - \alpha > 0$$

or

$$y^{(i)} = -1 \text{ and } X^{(i)} \cdot \theta - \alpha \leq 0$$

# Support Vector Machines

Data:  $(X^{(1)}, y^{(1)}), (X^{(2)}, y^{(2)}), \dots, (X^{(n)}, y^{(n)})$

Predictions:  $(X^{(1)}, \hat{y}^{(1)}), (X^{(2)}, \hat{y}^{(2)}), \dots, (X^{(n)}, \hat{y}^{(n)})$

$$\text{where } \hat{y}^{(i)} = \begin{cases} +1 & \text{if } X^{(i)} \cdot \theta - \alpha > 0 \\ -1 & \text{if } X^{(i)} \cdot \theta - \alpha \leq 0 \end{cases}$$

How do we **check** if a point is correctly classified?

$$y^{(i)}(X^{(i)} \cdot \theta - \alpha) \begin{cases} \geq 0 & \text{if correctly classified} \\ < 0 & \text{if misclassified} \end{cases}$$



# Support Vector Machines

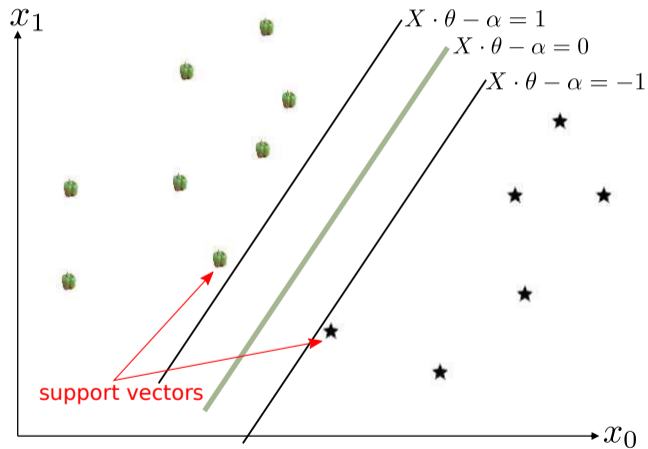
We want to **maximize** the **number** of correctly classified points

Find  $\theta = (\theta_1, \dots, \theta_d)$  and  $\alpha$  that maximize:

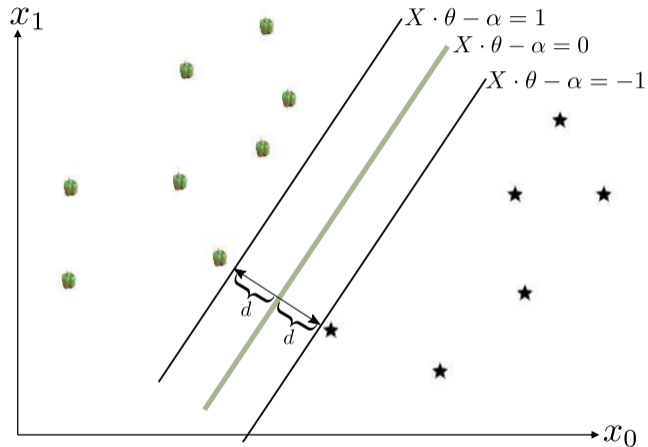
number of points with  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 0$

We will do this in a couple of steps...

# Step 1: construct margins

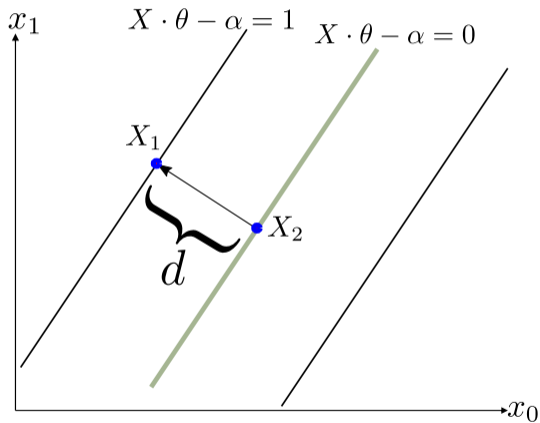


## Step 2: maximize the margins

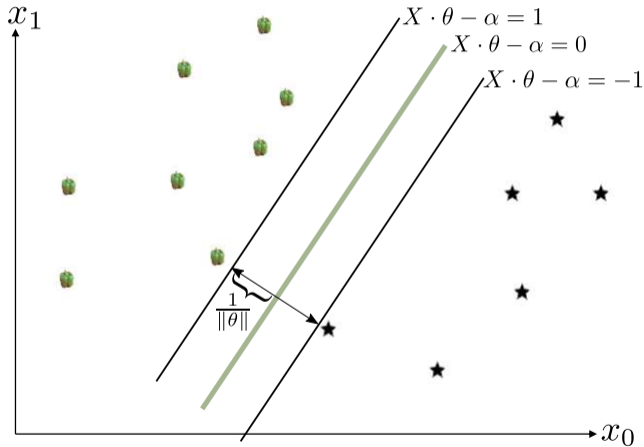


We want to maximize  $d$ , but what is  $d$ ?

# The size of margin



# Hard-margin SVM



Find  $\theta$  and  $\alpha$  that minimize

$$\min_{\theta} \|\theta\|_2^2$$

subject to

$$y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1$$

for all points  $(X_i, y_i)$ .

# Hard-margin SVM

Find  $\theta$  and  $\alpha$  that minimize  $\|\theta\|_2^2$

s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1$  for all  $i$

This is a **convex optimization problem**:

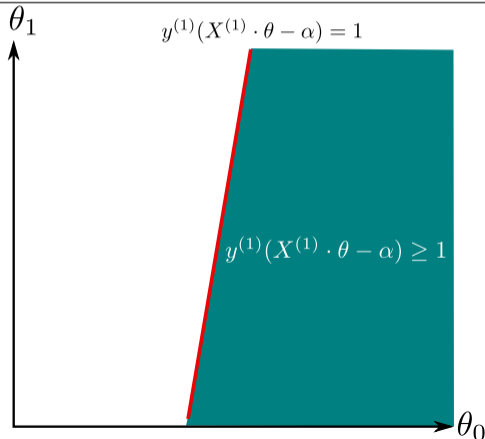
- Convex objective function
- Linear constraints

This means that the solution can be found efficiently.

# Minimization with constraint

Area of all  $\theta$  that satisfies the constraint:

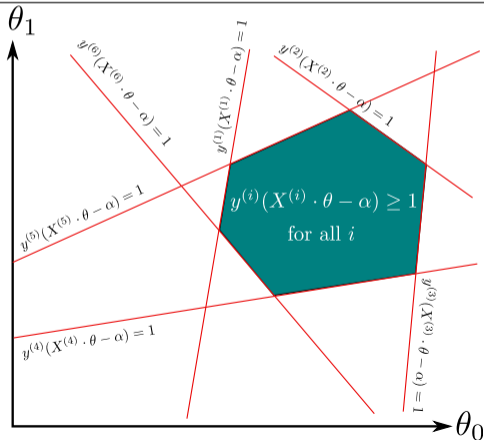
$$y^{(1)}(X^{(1)} \cdot \theta - \alpha) \geq 1$$



# Minimization with constraint

Area of all  $\theta$  that satisfies the constraint:

$$y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 \quad \text{for } i = 1, 2, 3, 4, 5, 6$$

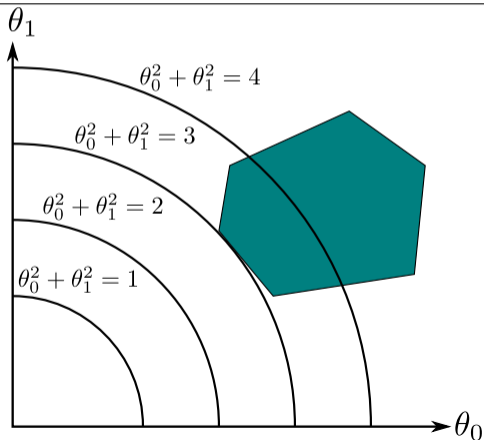




# Minimization with constraint

Find  $\theta$  that minimizes  $\|\theta\|_2^2$

s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1$  for  $i = 1, 2, 3, 4, 5, 6$



# Dual form

Find  $\theta$  and  $\alpha$  that minimizes  $\|\theta\|_2^2$

s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1$  for all  $i$

**Primal form**

is equivalent to

Find  $\alpha_1, \alpha_2, \dots, \alpha_n$  that maximize  $\sum_j \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (X^{(i)} \cdot X^{(j)})$

where  $\alpha_j \geq 0$  and  $\sum_j \alpha_j y_j = 0$ ,

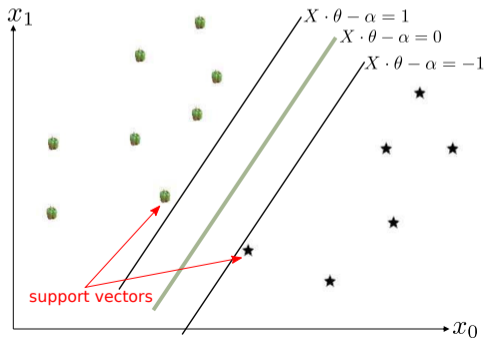
**Dual form**

then compute  $\theta = \sum_{i=1}^n \alpha_i y^{(i)} X^{(i)}$

# Interpretation of $\alpha_i$ 's

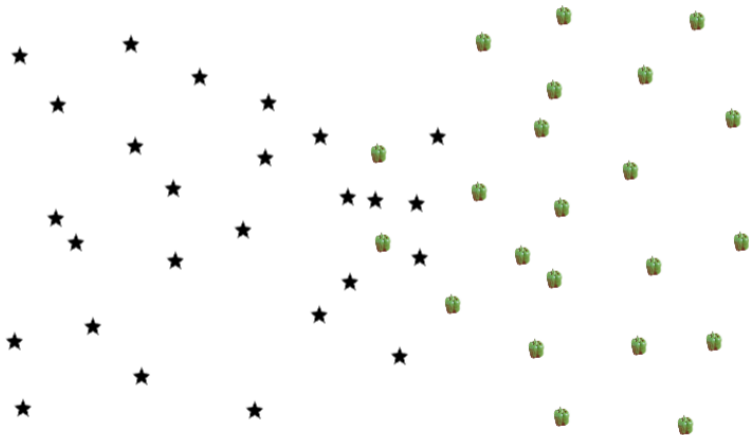
$$\theta = \sum_{i=1}^n \alpha_i y^{(i)} X^{(i)}$$

**Theorem.**  $\alpha_i$  is only non-zero when  $(X^{(i)}, y^{(i)})$  is a support vector!



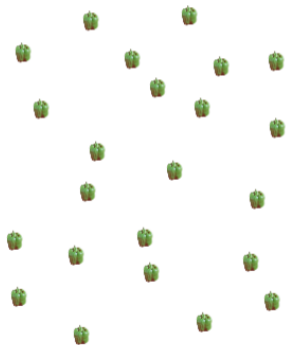
# Support Vector Machines

Is finding a separating hyperplane on this data possible?

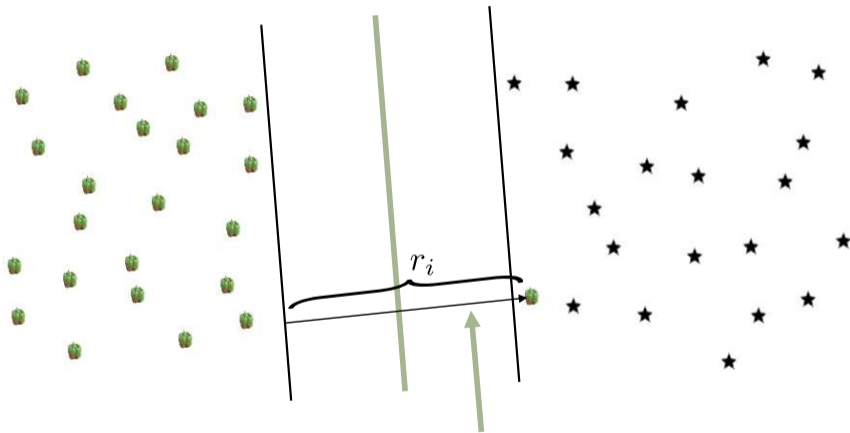


# Support Vector Machines

or is it always a good idea?



# Support Vector Machines

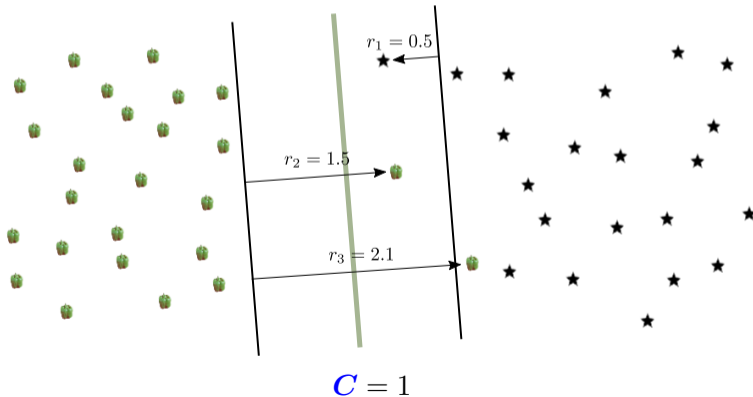


Introduce a **slack variable**  $r_i$

# Soft-margin SVM

Find  $\theta, \alpha$  and  $r_i$  that minimize  $\|\theta\|_2^2 + C \sum_i r_i$

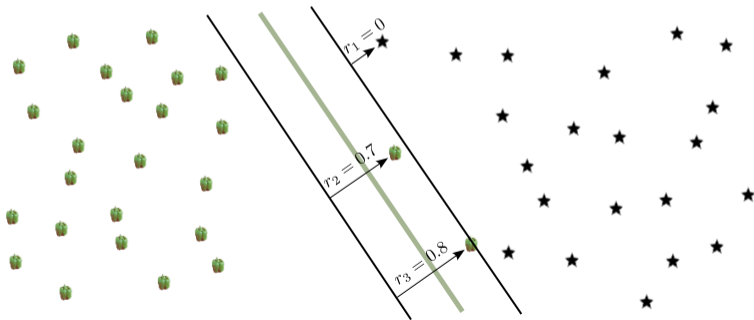
s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 - r_i$  for all  $i$



# Soft-margin SVM

Find  $\theta, \alpha$  and  $r_i$  that minimize  $\|\theta\|_2^2 + C \sum_i r_i$

s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 - r_i$  for all  $i$



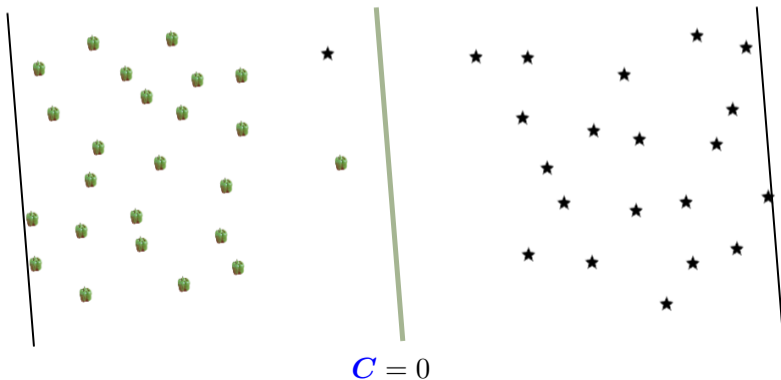
$C = 10$



# Examples

Find  $\theta, \alpha$  and  $r_i$  that minimize  $\|\theta\|_2^2 + C \sum_i r_i$

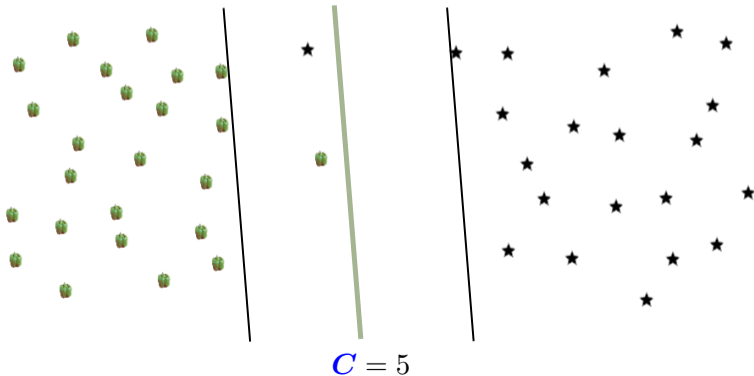
s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 - r_i$  for all  $i$



# Examples

Find  $\theta, \alpha$  and  $r_i$  that minimize  $\|\theta\|_2^2 + C \sum_i r_i$

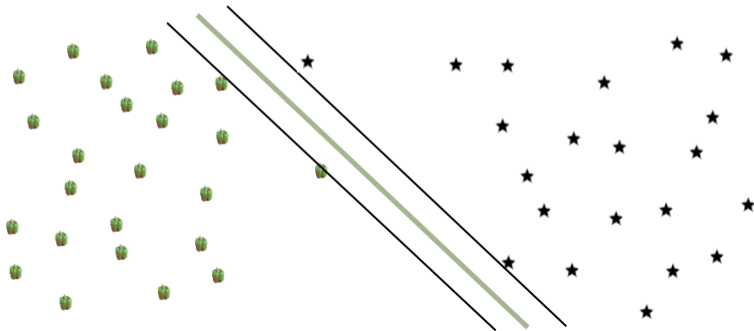
s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 - r_i$  for all  $i$



# Examples

Find  $\theta, \alpha$  and  $r_i$  that minimize  $\|\theta\|_2^2 + C \sum_i r_i$

s.t.  $y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 - r_i$  for all  $i$



$$C = 100$$

# Choosing $C$

$$\begin{aligned} &\text{Find } \theta, \alpha \text{ and } r_i \text{ that minimize } \|\theta\|_2^2 + C \sum_i r_i \\ &\text{s.t. } y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 - r_i \text{ for all } i \end{aligned}$$

$C$  has to be chosen prior to training SVM

How to choose  $C$ ?

# Dual form of soft-margin SVM

$$\begin{aligned} & \text{Find } \theta, \alpha \text{ and } r_i \text{ that minimize } \|\theta\|_2^2 + C \sum_i r_i \\ & \text{s.t. } y^{(i)}(X^{(i)} \cdot \theta - \alpha) \geq 1 - r_i \quad \text{for all } i \end{aligned} \quad \text{Primal form}$$

is equivalent to

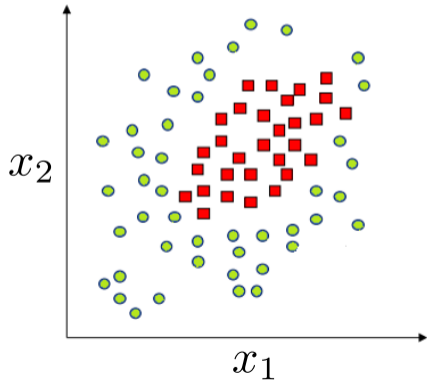
$$\begin{aligned} & \text{Find } \alpha_1, \alpha_2, \dots, \alpha_n \text{ that maximize } \sum_j \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (X^{(i)} \cdot X^{(j)}) \\ & \text{where } 0 \leq \alpha_j \leq C \text{ and } \sum_j \alpha_j y_j = 0, \end{aligned} \quad \text{Dual form}$$

then compute  $\theta = \sum_{i=1}^n \alpha_i y^{(i)} X^{(i)}$

# Nonlinear separability

How do we deal with nonlinear boundaries?

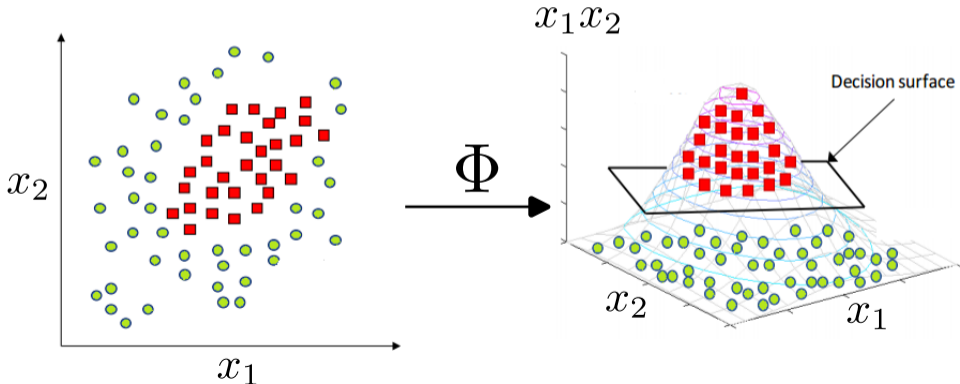
Example:  $X = (x_1, x_2), y \in \{+1, -1\}$



# Adding new features

Idea: From  $(x_1, x_2)$  we add more features:  $(x_1^2, x_2^2, x_1x_2)$

$$\Phi(x_1, x_2) = (x_1, x_2, x_1^2, x_2^2, x_1x_2)$$



# Adding new features

Suppose we originally have  $d$  features:  $X = (x_1, \dots, x_d)$

Idea: add more features:

$$x_1^2, x_2^2, \dots, x_d^2$$

$$x_1x_2, x_1x_3, \dots, x_{d-1}x_d$$



# Adding new features

Suppose we originally have  $d$  features:  $X = (x_1, \dots, x_d)$

Idea: add more features:

$$x_1^2, x_2^2, \dots, x_d^2$$

$$x_1x_2, x_1x_3, \dots, x_{d-1}x_d$$

New data vector:

$$\Phi(x_1, \dots, x_d) = (x_1, \dots, x_d, x_1^2, \dots, x_d^2, x_1x_2, \dots, x_{d-1}x_d)$$

## Quick question

$$\Phi(x_1, x_2, \dots, x_d) = (x_1, \dots, x_d, x_1^2, \dots, x_d^2, x_1x_2, \dots, x_dx_{d-1})$$

What is the dimension of  $\Phi(x)$ ?

# Kernel trick

Recall the dual form of the soft-margin SVM:

Find  $\alpha_1, \alpha_2, \dots, \alpha_n$  that maximize  $\sum_j \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (\mathbf{X}^{(i)} \cdot \mathbf{X}^{(j)})$

where  $0 \leq \alpha_j \leq C$  and  $\sum_j \alpha_j y_j \geq 0$ ,

**Dual form**

then compute  $\theta = \sum_{i=1}^n \alpha_i y^{(i)} X^{(i)}$

# Kernel trick

Replace with transformed vectors:

$$\text{Find } \alpha_1, \alpha_2, \dots, \alpha_n \text{ that maximize}$$
$$\sum_j \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (\Phi(\mathbf{X}^{(i)}) \cdot \Phi(\mathbf{X}^{(j)}))$$

$$\text{where } 0 \leq \alpha_j \leq C \text{ and } \sum_j \alpha_j y_j = 0,$$

**Dual form**

$$\text{then compute } \theta = \sum_{i=1}^n \alpha_i y^{(i)} \Phi(\mathbf{X}^{(i)})$$

Magic: we can compute  $\Phi(\mathbf{X}^{(i)}) \cdot \Phi(\mathbf{X}^{(j)})$  without ever writing out  $\Phi(\mathbf{X}^{(i)})$  and  $\Phi(\mathbf{X}^{(j)})$

# Computing dot product

Example in 2D

$$X = (x_1, x_2) \text{ and } \Phi(x) = (x_1, x_2, x_1^2, x_2^2, x_1x_2)$$

# Computing dot product

Example in 2D

$$X = (x_1, x_2) \text{ and } \Phi(x) = (x_1, x_2, x_1^2, x_2^2, x_1x_2)$$

$$\text{Actually, tweak a little: } \Phi(X) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$\text{and } Z = (z_1, z_2), \text{ so } \Phi(Z) = (1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, z_2^2, \sqrt{2}z_1z_2)$$

What is  $\Phi(X) \cdot \Phi(Z)$ ?

# Kernel trick

Suppose  $X = (x_1, x_2, \dots, x_d)$  and

$$\Phi(X) = (1, \sqrt{2}x_1, \dots, \sqrt{2}x_d, x_1^2, \dots, x_d^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{d-1}x_d)$$

Then

$$\begin{aligned} & \Phi(X) \cdot \Phi(Z) \\ &= (1, \sqrt{2}x_1, \dots, \sqrt{2}x_d, x_1^2, \dots, x_d^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{d-1}x_d) \cdot \\ & \quad (1, \sqrt{2}z_1, \dots, \sqrt{2}z_d, z_1^2, \dots, z_d^2, \sqrt{2}z_1z_2, \dots, \sqrt{2}z_{d-1}z_d) \\ &= 1 + 2 \sum_i x_i z_i + \sum_i x_i^2 z_i^2 + 2 \sum_{i \neq j} x_i x_j z_i z_j \\ &= (1 + X \cdot Z)^2 \end{aligned}$$

# MNIST example

$X^{(i)} = (x_1, \dots, x_{784})$ ,  $\Phi(X^{(i)})$  has 308,504 dimensions



Find  $\theta$  and  $\alpha$  that minimizes  $\|\theta\|_2^2$

s.t.  $y^{(i)}(\Phi(X^{(i)}) \cdot \theta - \alpha) \geq 1$  for all  $i$

**Primal form**

Find  $\alpha_1, \alpha_2, \dots, \alpha_n$  that maximize

$$\sum_j \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (\Phi(X^{(i)}) \cdot \Phi(X^{(j)}))$$

where  $0 \leq \alpha_j \leq C$  and  $\sum_j \alpha_j y_j = 0$ ,

**Dual form**

$$\text{then } \theta = \sum_{i=1}^n \alpha_i y^{(i)} \Phi(X^{(i)})$$

wait, looking at the formula of  $\theta$ ...do we have to compute  $\Phi(X^{(i)})$  after all?



# Kernel SVM

1. **Basis expansion.** Mapping  $X \mapsto \Phi(X)$
2. **Learning.** Solve the dual problem:

$$\text{Find } \alpha_1, \alpha_2, \dots, \alpha_n \text{ that maximize}$$
$$\sum_j \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (\Phi(\mathbf{X}^{(i)}) \cdot \Phi(\mathbf{X}^{(j)}))$$

$$\text{where } 0 \leq \alpha_j \leq C \text{ and } \sum_j \alpha_j y_j = 0$$

3. **Classification.** Given a new point  $X$ , classify as

$$\hat{y} = \begin{cases} +1 & \text{if } \sum_i \alpha_i y^{(i)} (\Phi(\mathbf{X}^{(i)}) \cdot \Phi(\mathbf{X})) - \alpha > 0 \\ -1 & \text{if } \sum_i \alpha_i y^{(i)} (\Phi(\mathbf{X}^{(i)}) \cdot \Phi(\mathbf{X})) - \alpha \leq 0 \end{cases}$$

# Kernel SVM (more general)

In general, we may use a kernel function  $k(X, Z)$  which **measures the similarity** between  $X$  and  $Z$

1. **Learning.** Solve the dual problem:

Find  $\alpha_1, \alpha_2, \dots, \alpha_n$  that maximize  $\sum_j \alpha_j - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j (k(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}))$

$$\text{where } 0 \leq \alpha_j \leq C \text{ and } \sum_j \alpha_j y_j = 0$$

2. **Classification.** Given a new point  $X$ , classify as

$$\hat{y} = \begin{cases} +1 & \text{if } \sum_i \alpha_i y^{(i)} k(\mathbf{X}^{(i)}, \mathbf{X}) - \alpha > 0 \\ -1 & \text{if } \sum_i \alpha_i y^{(i)} k(\mathbf{X}^{(i)}, \mathbf{X}) - \alpha \leq 0 \end{cases}$$

# Examples of kernel functions

- Polynomial

$$K(X^{(i)}, X^{(j)}) = (1 + X^{(i)} \cdot X^{(j)})^d$$

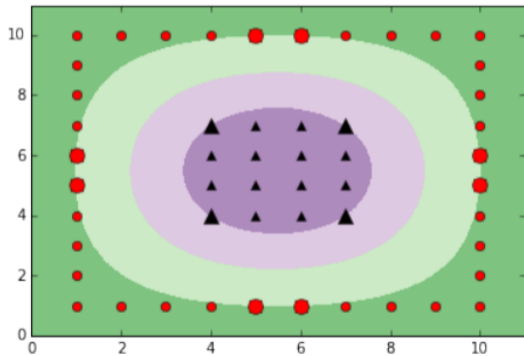
- RBF (Gaussian kernel)

$$K(X^{(i)}, X^{(j)}) = \exp\left(-\frac{\|X^{(i)} - X^{(j)}\|^2}{2\sigma^2}\right)$$

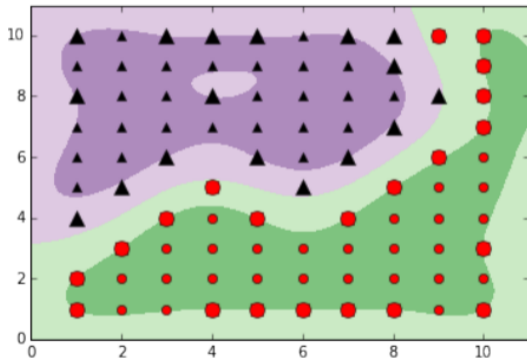
- Sigmoid

$$K(X^{(i)}, X^{(j)}) = \tanh(\eta X^{(i)} \cdot X^{(j)} + \nu)$$

# Example: RBF kernel



# Example: RBF kernel



# Conclusions

- Support Vector Machine
  - As a convex optimization problem
  - Hard-margin SVM
  - Soft-margin SVM
  - Primal and Dual forms
- Nonlinear boundaries
  - Mapping to a higher dimension
  - Kernel trick
  - Kernel SVM