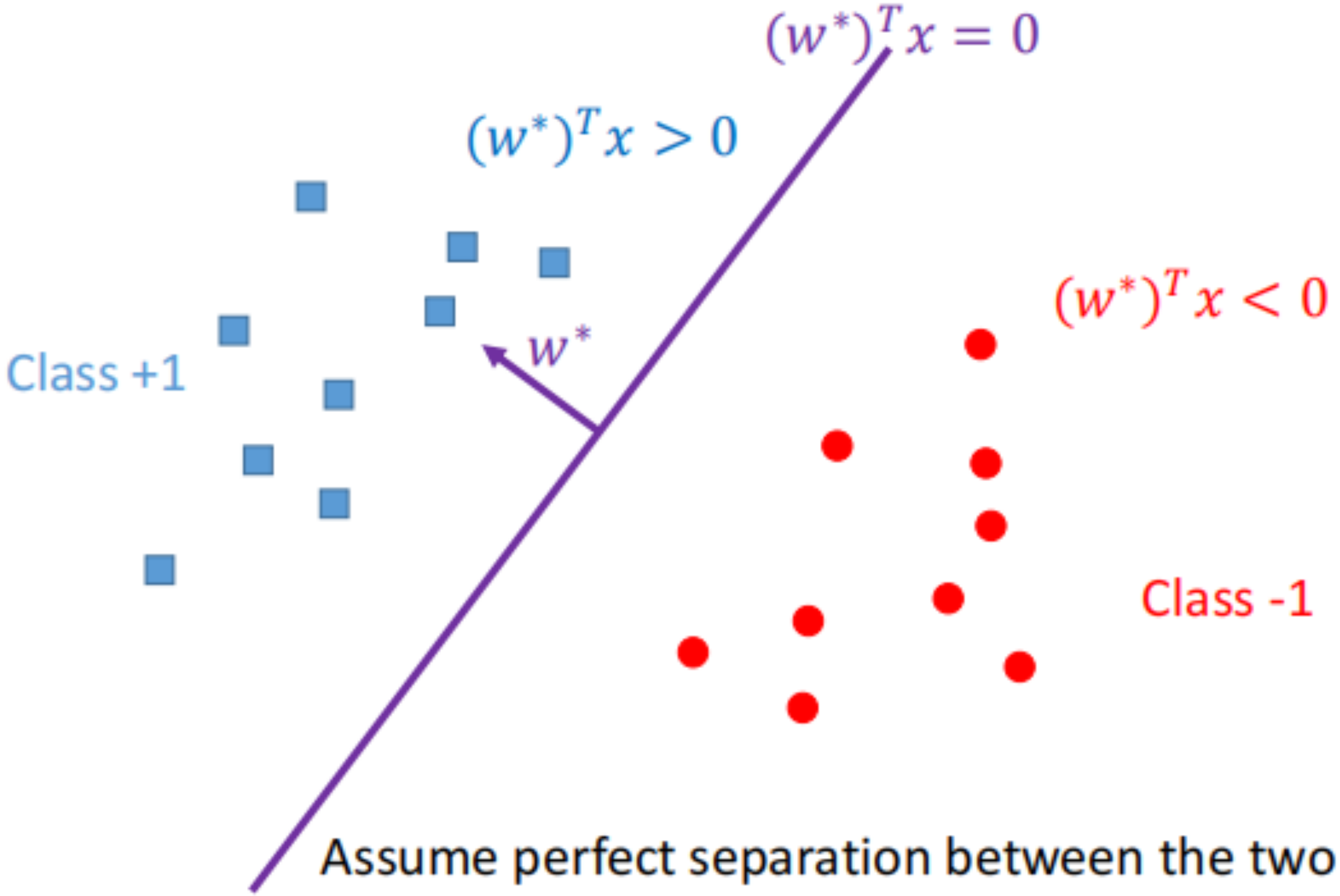
# 机器学习A
# 15.支持向量机
# SVM

王翔

中国科学技术大学
Lab for Data Science

# Motivation

# Linear classification
# 线性分类




Assume perfect separation between the two classes

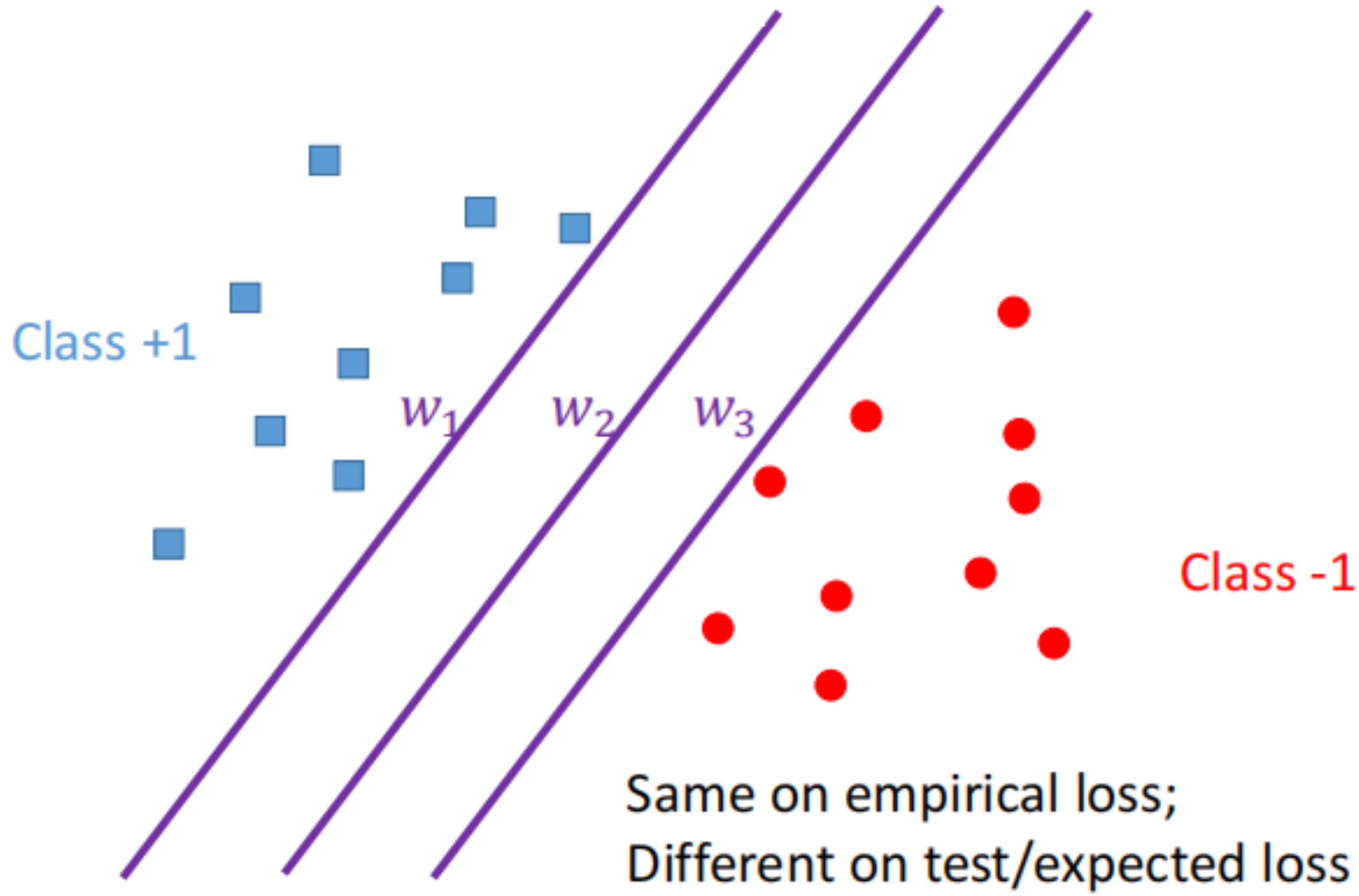
# Attempt
# 尝试

- Given training data $\{(x_i, y_i) : 1 \leq i \leq n\}$ i.i.d. from distribution $D$
给定训练数据 $\{(x_i, y_i) : 1 \leq i \leq n\}$,它们是从分布$D$独立同分布采样的

- Hypothesis $y = \text{sign}(f_w(x)) = \text{sign}(w^T x)$
    - $y = +1$ if $w^T x > 0$
    - $y = -1$ if $w^T x < 0$

- Let's assume that we can optimize to find w

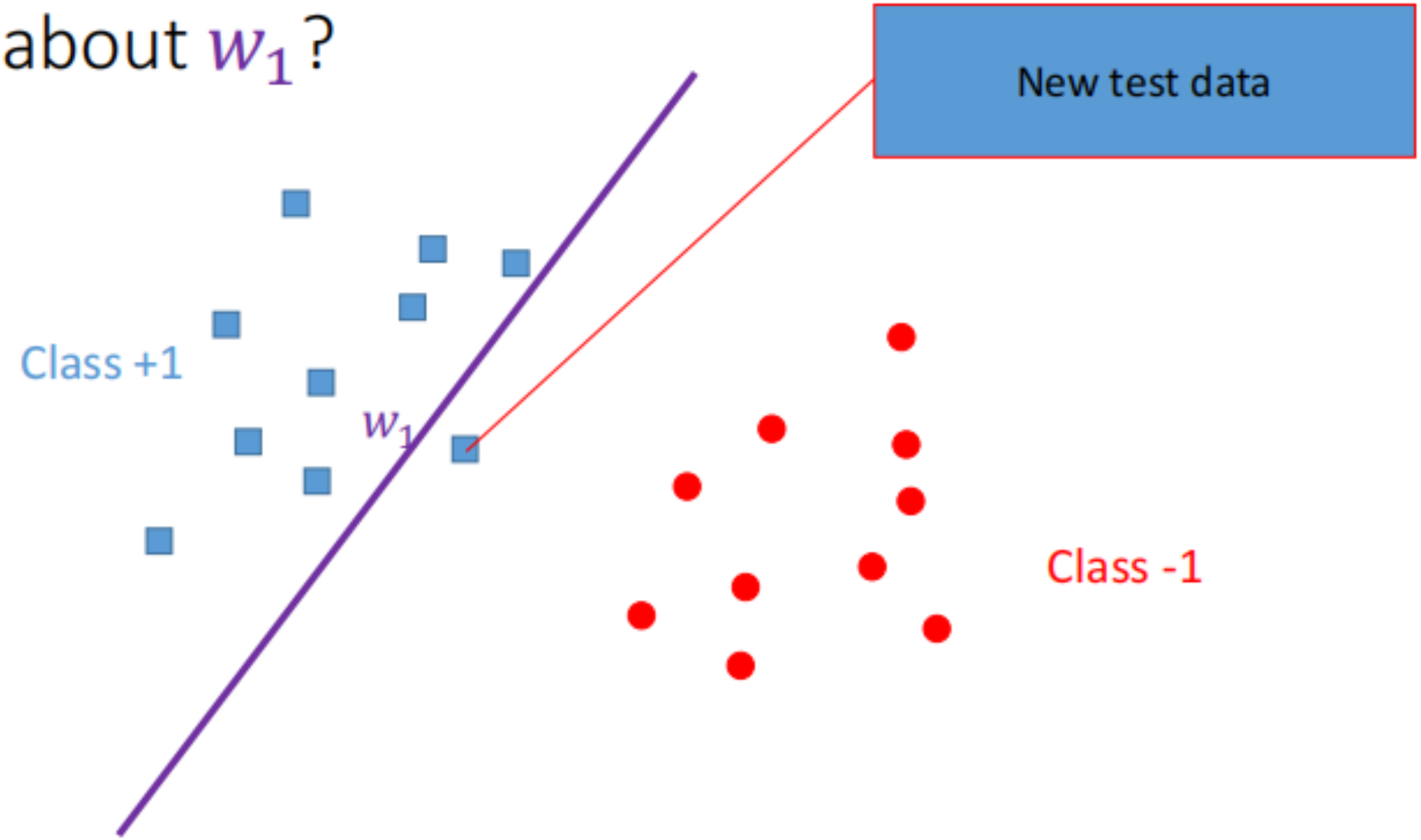
# Multiple optimal solutions?
# 多个最优解?




Same on empirical loss;
Different on test/expected loss

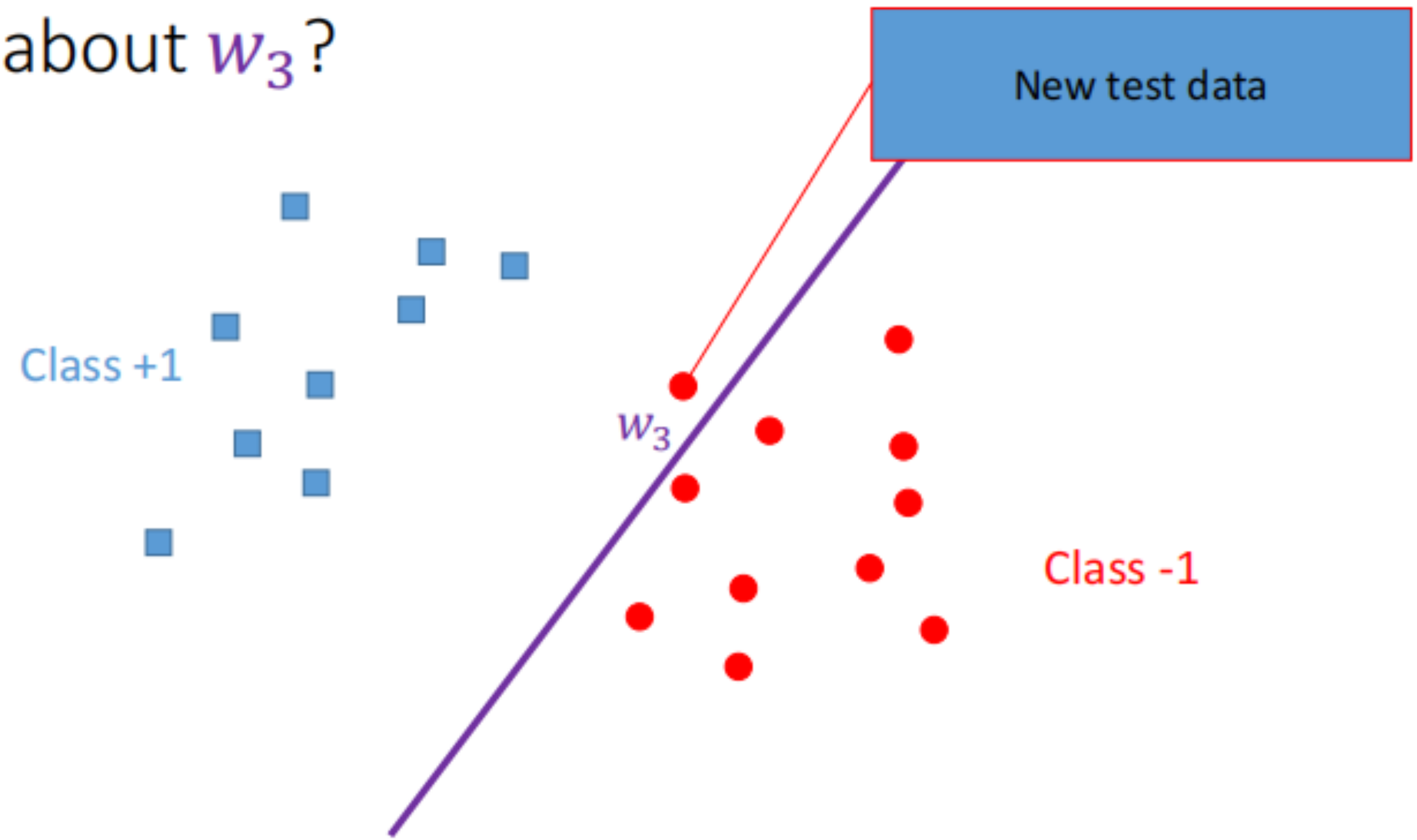
# Multiple optimal solutions?
# 多个最优解?

What about $w_1$?



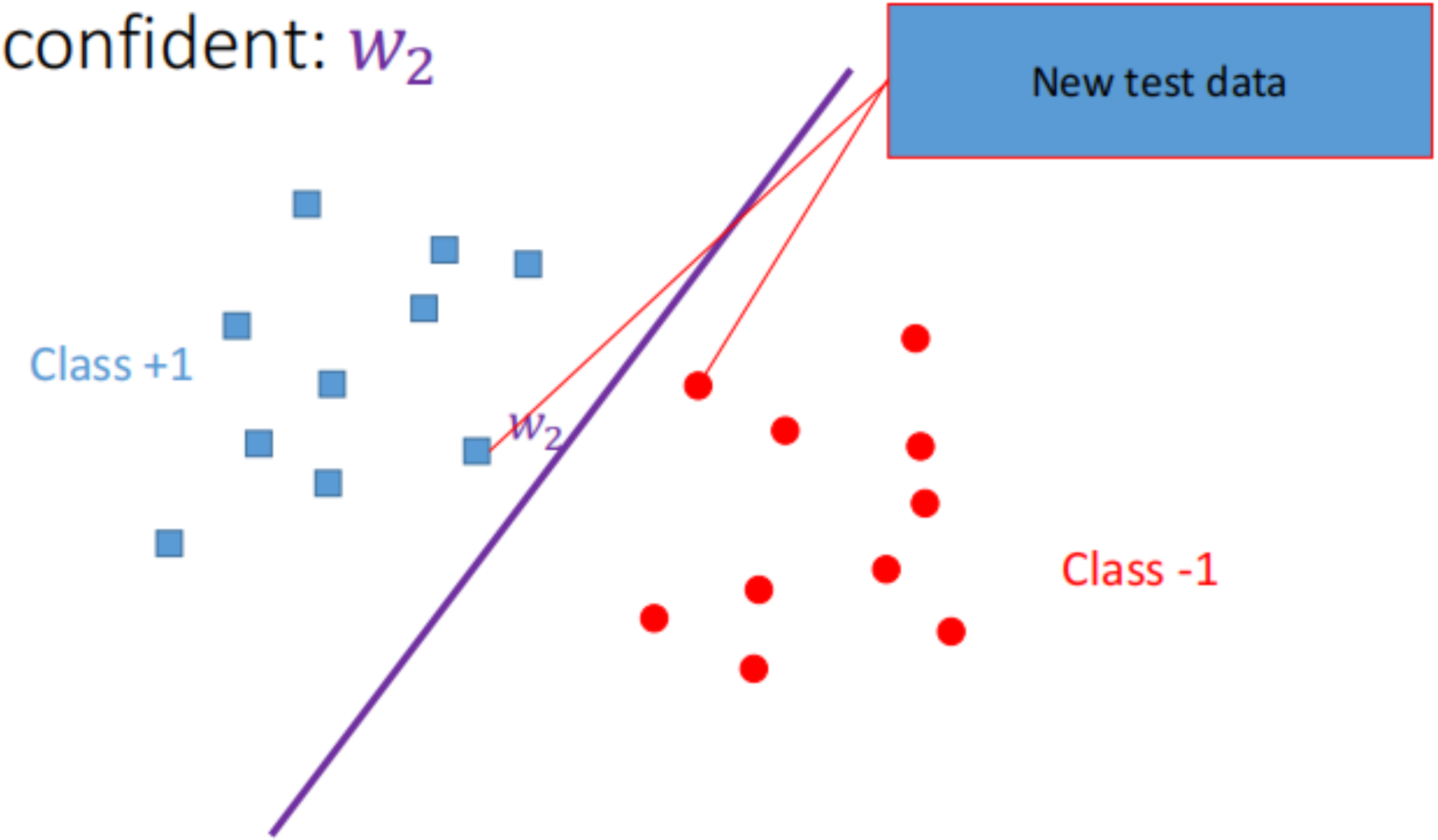
# Multiple optimal solutions?
# 多个最优解?

What about $w_3$?

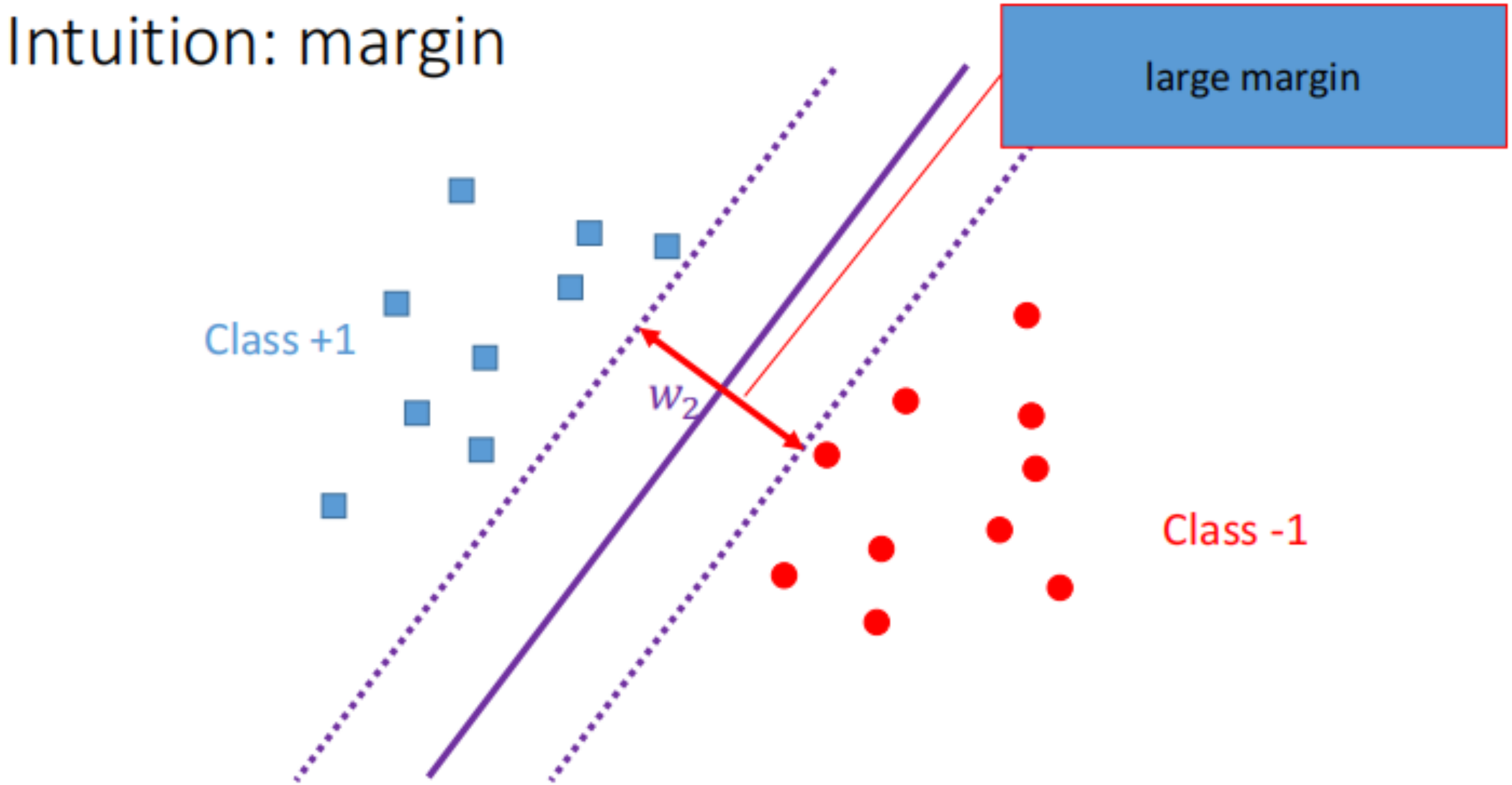


# Multiple optimal solutions?
# 多个最优解?

Most confident: $w_2$



# Multiple optimal solutions?
# 多个最优解?



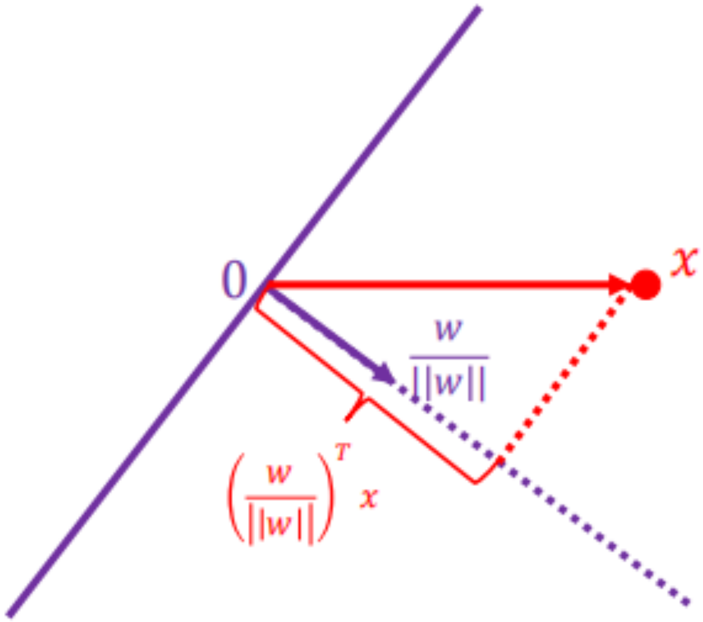
# Margin

# Margin
# 间隔

- Lemma 1: $x$ has distance $\frac{|f_w(x)|}{||w||}$ to the hyperplane $f_w(x) = w^T x = 0$

Proof:

- $w$ is orthogonal to the hyperplane
  $w$与超平面正交
- The unit direction is $\frac{w}{||w||}$
  单位方向
- The projection of $x$is $\left(\frac{w}{||w||}\right)^T x = \frac{f_w(x)}{||w||}$
  $x$的投影



# Margin: with bias
# 间隔：带偏置的情况

- Claim 1: $w$ is orthogonal to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

声明1：$w$与超平面正交

Proof:

- pick any $x_1$ and $x_2$ on the hyperplane

选择超平面上的任意$x_1$ 和$x_2$

- $w^T x_1 + b = 0$
- $w^T x_2 + b = 0$

- So $w^T(x_1 - x_2) = 0$

# Margin: with bias
# 间隔：带偏置的情况

- Claim 2: $0$ has distance $\frac{-b}{||w||}$ to the hyperplane $w^T x + b = 0$

Proof:

- pick any $x_1$ the hyperplane 选择超平面上的任意$x_1$
- Project $x_1$ to the unit direction $\frac{w}{||w||}$ to get the distance
- $\left(\frac{w}{||w||}\right)^T x_1 = \frac{-b}{||w||}$ since $w^T x_1 + b = 0$

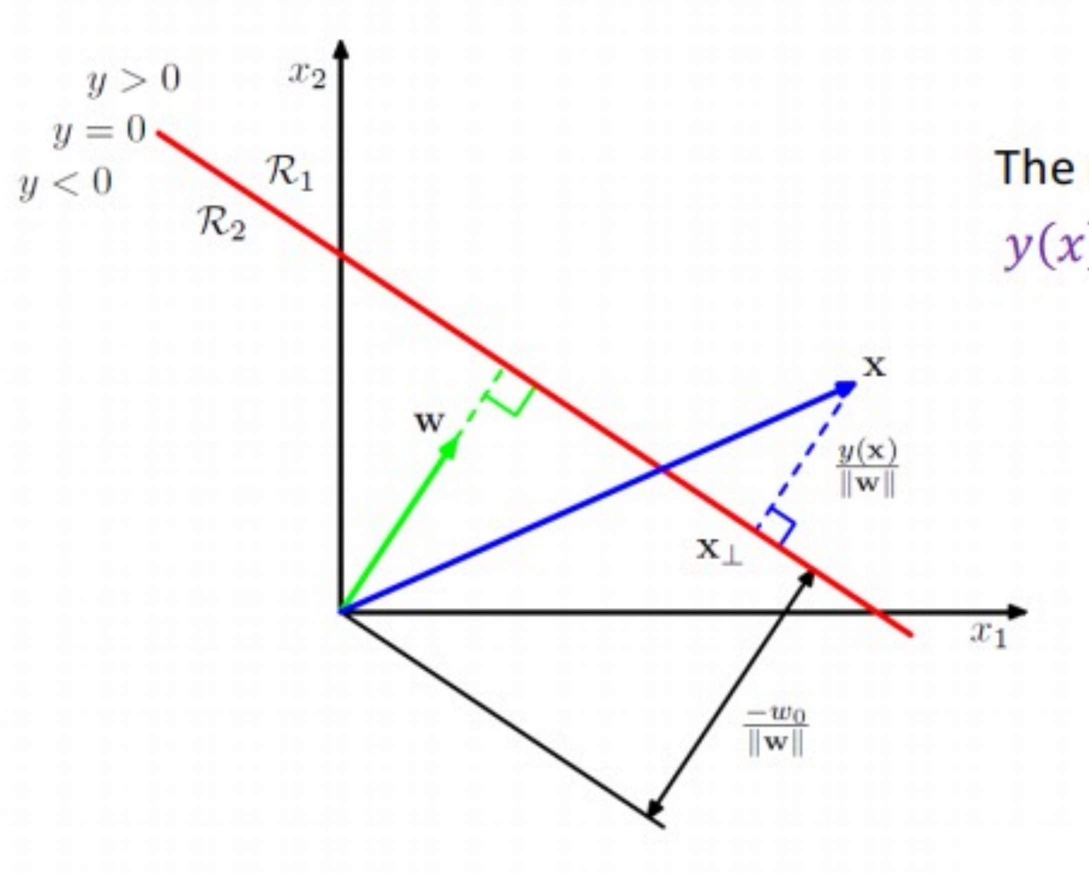
# Margin: with bias
# 间隔：带偏置的情况

- Lemma 2: $x$ has distance $\frac{|f_{w,b}(x)|}{||w||}$ to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

Proof:

- Let $x = x_{\perp} + r\frac{w}{||w||}$, then $|r|$ is the distance
- Multiply both sides by $w^T$ and add $b$ 两边都乘以$w^T$并加上b
- Left hand side: $w^T x + b = f_{w,b}(x)$
- Right hand side: $w^T x_{\perp} + r\frac{w^T w}{||w||} + b = 0 + r||w||$

# Margin: with bias
# 间隔：带偏置的情况




The notation here is:

$y(x) = w^T x + w_0$

# Support Vector Machine (SVM)

# SVM: objective
# SVM：优化目标函数

- Margin over all training data points:
所有训练数据点的边距：

$$\gamma = \min_i \frac{|f_{w,b}(x_i)|}{||w||}$$

- Since only want correct $f_{w,b}$, and recall $y_i \in \{+1, -1\}$, we have

$$\gamma = \min_i \frac{y_i f_{w,b}(x_i)}{||w||}$$

- If $f_{w,b}$ incorrect on some $x_i$, the margin is negative 边距为负

# SVM: objective
# SVM：优化目标函数

- Maximize margin over all training data points:
最大化所有训练数据点的边距

$$\max_{w,b} \gamma = \max_{w,b} \min_{i} \frac{y_i f_{w,b}(x_i)}{||w||} = \max_{w,b} \min_{i} \frac{y_i(w^T x_i + b)}{||w||}$$

- A bit complicated … 有点复杂

# SVM: simplified objective
# SVM：简化目标函数

- Observation: when $(w, b)$ scaled by a factor $c$, the margin unchanged
  当 $(w, b)$被一个因子$c$缩放时，边距不变

$$\frac{y_i(cw^T x_i + cb)}{||cw||} = \frac{y_i(w^T x_i + b)}{||w||}$$

- Let’s consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

where $x_{i*}$ is the point closest to the hyperplane
其中$x_{i*}$是离超平面最近的点。

# SVM: simplified objective
# SVM：简化目标函数

- Let’s consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

where $x_{i*}$ is the point closest to the hyperplane
其中$x_{i*}$是离超平面最近的点。

- Now we have for all data

$$y_i(w^T x_i + b) \geq 1$$

and at least for one $i$ the equality holds　　至少对某个$i$成立等式

- Then the margin is $\frac{1}{||w||}$

# SVM: simplified objective
# SVM：简化目标函数

- Optimization simplified to　　　优化问题简化为

$$\min_{w,b} \frac{1}{2}||w||^2$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- How to find the optimum $\hat{w}^*$?

# SVM: principle for hypothesis class

# Thought experiment
# 思想实验

- Suppose pick an $R$, and suppose can decide if exists $w$ satisfying
假设选择一个$R$，并假设可以确定是否存在满足条件的$w$

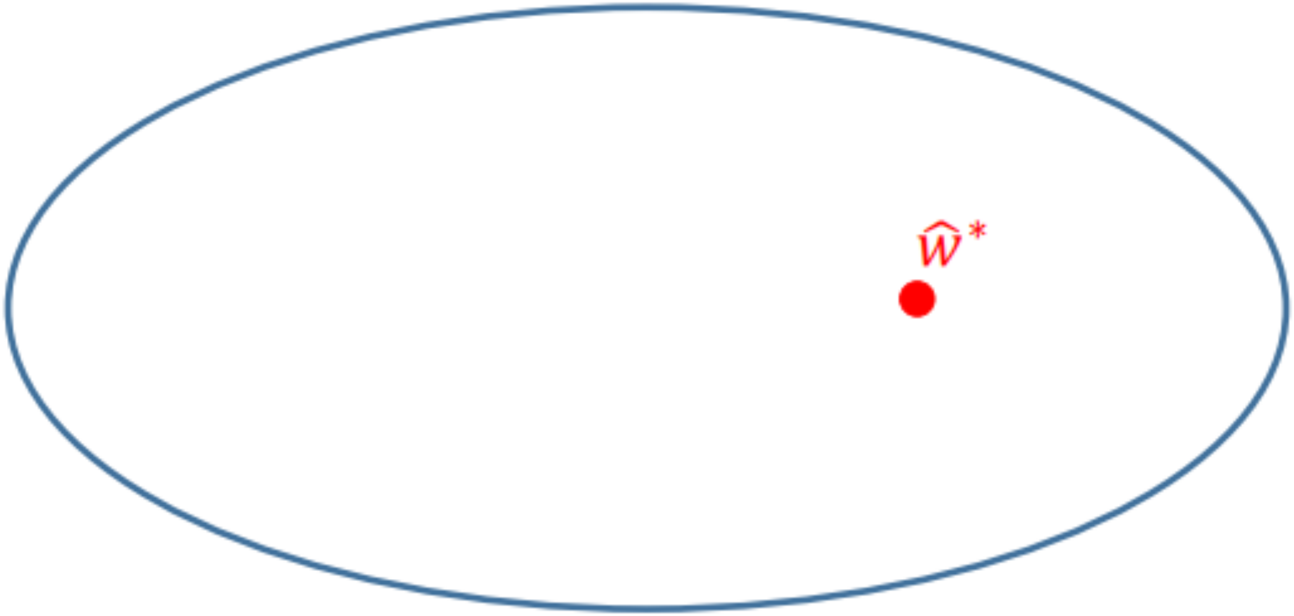
$$\frac{1}{2}||w||^2 \leq R$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- Decrease $R$ until cannot find $w$ satisfying the inequalities
减小$R$，直到找不到满足不等式的$w$为止。

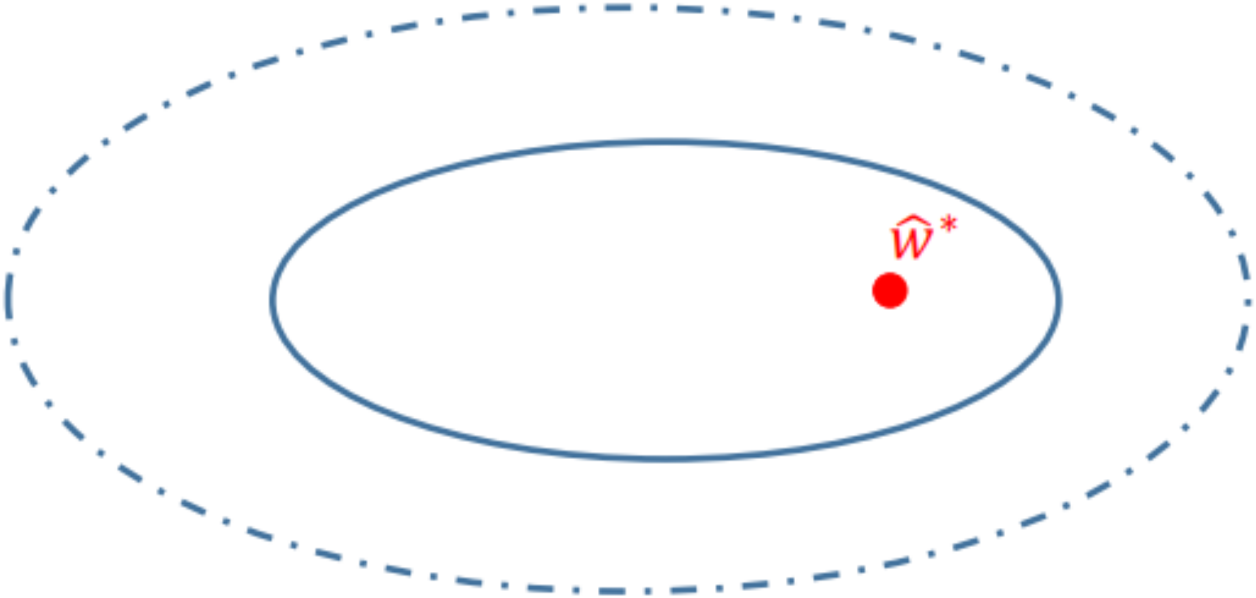
# Thought experiment
# 思想实验

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest $R$)



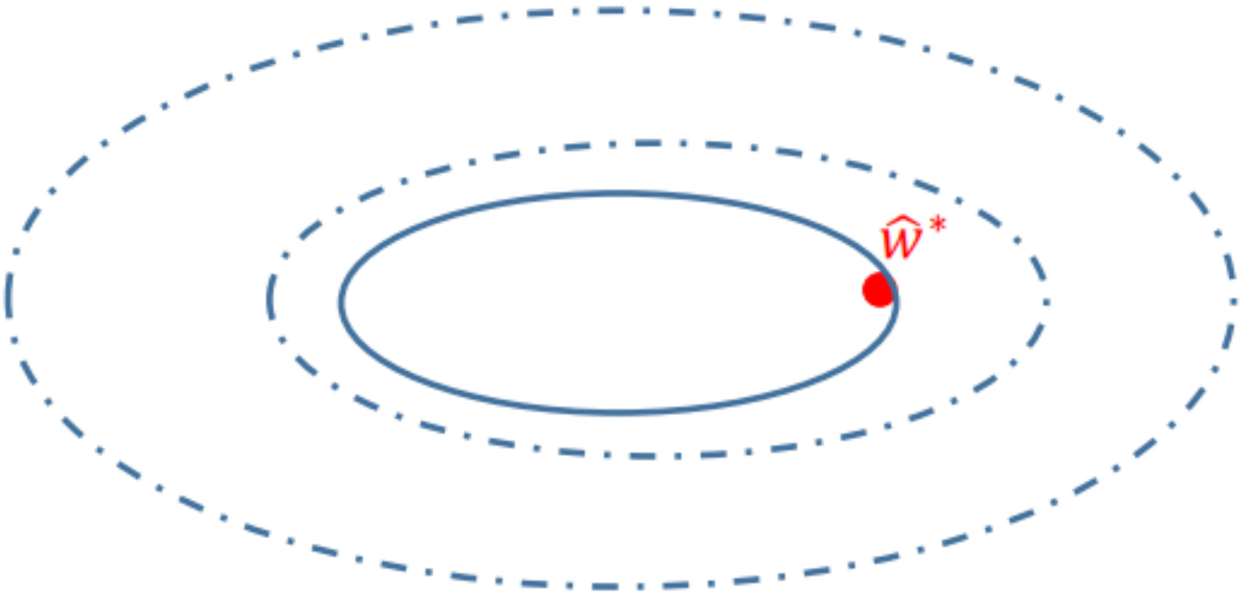
# Thought experiment
# 思想实验

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest $R$)



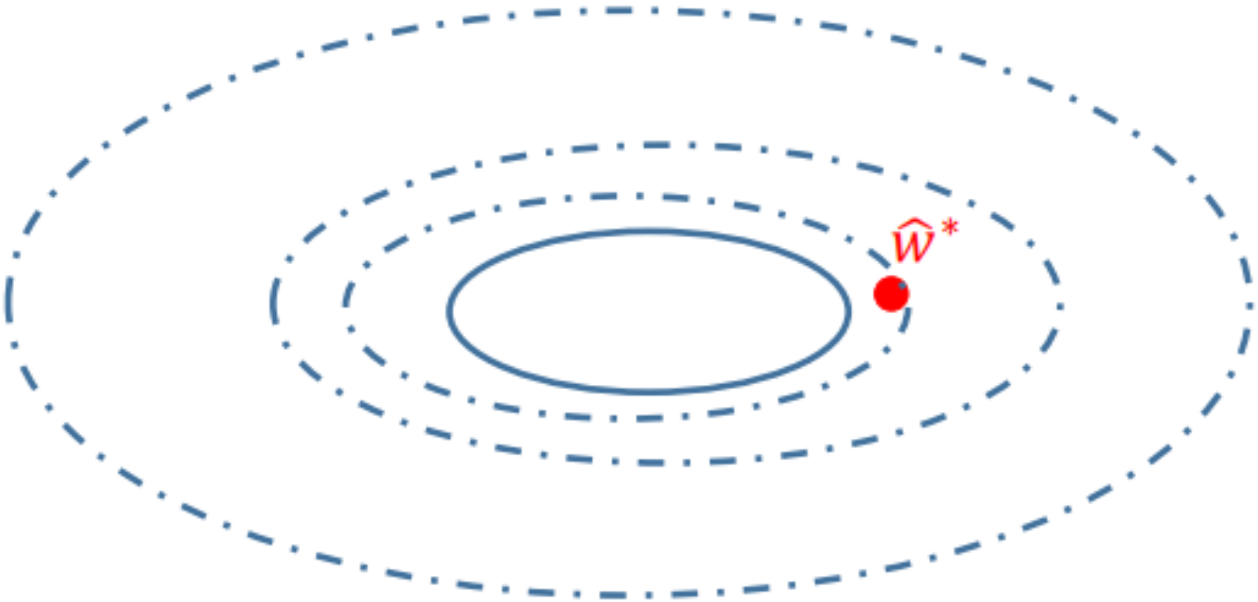
# Thought experiment
# 思想实验

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest $R$)



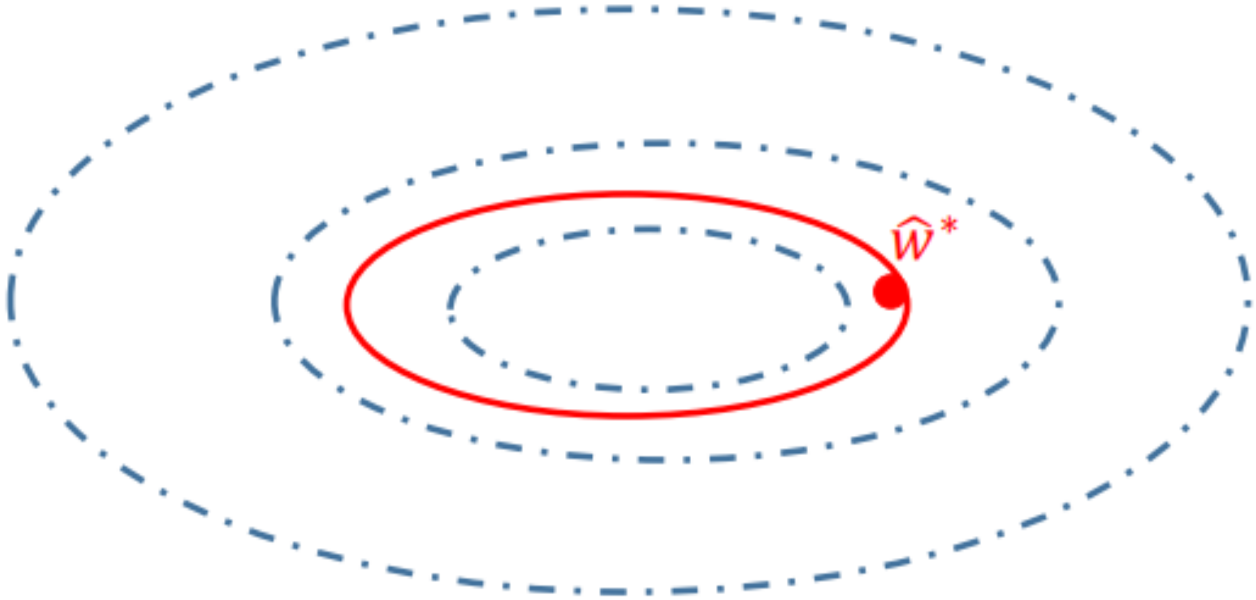
# Thought experiment
# 思想实验

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest $R$)



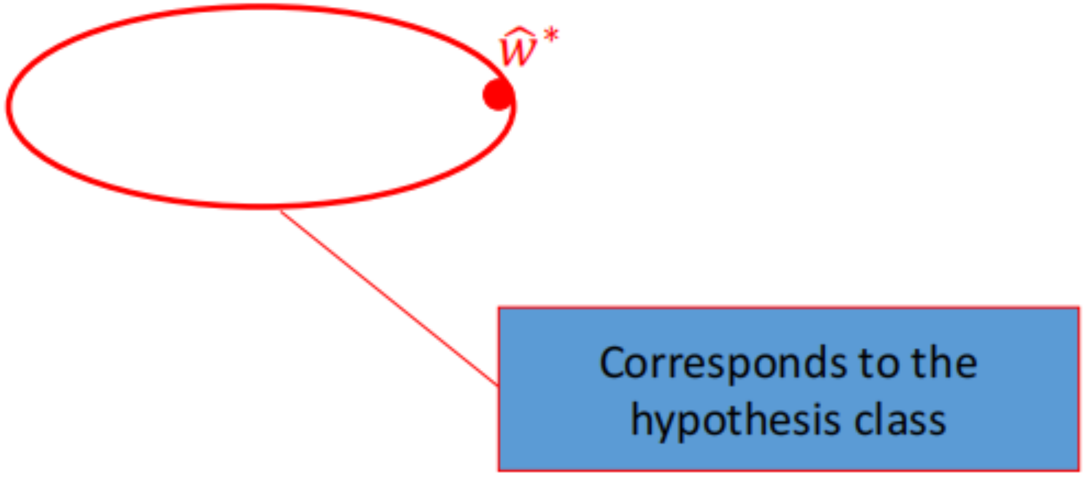
# Thought experiment
# 思想实验

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest $R$)



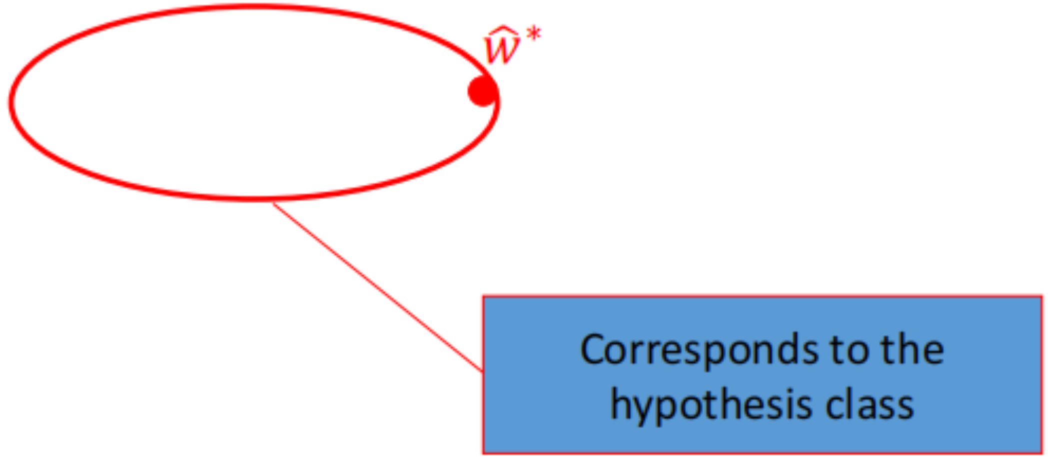
# Thought experiment
# 思想实验

- To handle the difference between empirical and expected losses →
为了处理经验损失和期望损失之间的差异
- Choose large margin hypothesis (high confidence) →
选择大边距假设（高置信度）
- Choose a small hypothesis class
选择一个较小的假设类



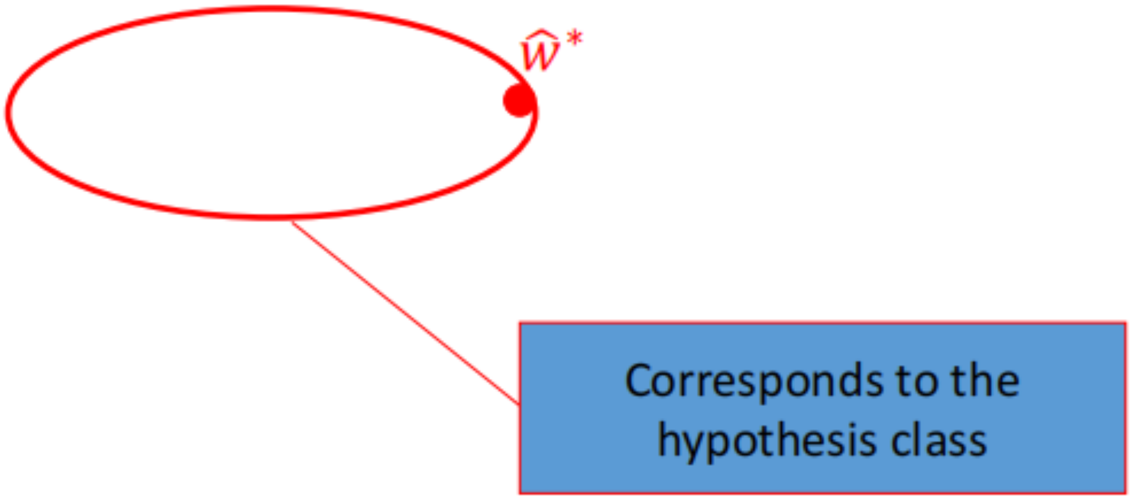
# Thought experiment
# 思想实验

- Principle: use smallest hypothesis class still with a correct/good one
  使用仍包含正确/良好假设的最小假设类
  - Also true beyond SVM
  - Also true for the case without perfect separation between the two classes
    即使在两类之间无法完全分离的情况下也同样适用
  - Math formulation: VC-dim theory, etc.
    数学形式化：如 VC 维理论等



# Thought experiment
# 思想实验

- Principle: use smallest hypothesis class still with a correct/good one
  使用仍包含正确/良好假设的最小假设类
  - Whatever you know about the ground truth, add it as constraint/regularizer
    将其添加为约束或正则项



# SVM: optimization
# SVM：优化器

- Optimization (Quadratic Programming): （二次规划）

$$\min_{w,b} \frac{1}{2}||w||^2$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- Solved by Lagrange multiplier method: 通过拉格朗日乘数法求解

$$\mathcal{L}(w, b, \boldsymbol{\alpha}) = \frac{1}{2}||w||^2 - \sum_i \alpha_i [y_i(w^T x_i + b) - 1]$$

where $\alpha$ is the Lagrange multiplier
其中$\alpha$ 是拉格朗日乘数
- Details in next lecture

# Lagrange multiplier

# Lagrangian
# 拉格朗日函数

- Consider optimization problem: 考虑优化问题

$$\min_{w} f(w)$$

$$h_i(w) = 0, \forall 1 \le i \le l$$

- Lagrangian:

$$\mathcal{L}(w, \boldsymbol{\beta}) = f(w) + \sum_i \beta_i h_i(w)$$

where $\beta_i$ ’s are called Lagrange multipliers
其中$\beta_i$’被称为拉格朗日乘数

# Lagrangian
# 拉格朗日函数

- Consider optimization problem: 考虑优化问题

$$\min_{w} f(w)$$

$$h_i(w) = 0, \forall 1 \leq i \leq l$$

- Solved by setting derivatives of Lagrangian to 0

通过将拉格朗日函数的导数设为 0 来求解

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0$$

# Generalized Lagrangian
# 广义拉格朗日函数

- Consider optimization problem: 考虑优化问题

$$\min_{w} f(w)$$

$$g_i(w) \le 0, \forall 1 \le i \le k$$

$$h_j(w) = 0, \forall 1 \le j \le l$$

- Generalized Lagrangian: 广义拉格朗日函数

$$\mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(w) + \sum_{i} \alpha_i g_i(w) + \sum_{j} \beta_j h_j(w)$$

where $\alpha_i$ , $\beta_j$ 's are called Lagrange multipliers

# Generalized Lagrangian
# 广义拉格朗日函数

- Consider the quantity:

$$\theta_P(w) := \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- Why?

如果w满足所有约束条件

$$\theta_P(w) = \begin{cases} f(w), & \text{if } w \text{ satisfies all the constraints} \\ +\infty, & \text{if } w \text{ does not satisfy the constraints} \end{cases}$$

- So minimizing $f(w)$ is the same as minimizing $\theta_P(w)$

$$\min_w f(w) = \min_w \theta_P(w) = \min_w \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

# Lagrange duality
# 拉格朗日对偶性

- The primal problem  原问题

$$p^* := \min_w f(w) = \min_w \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- The dual problem  对偶问题

$$d^* := \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \geq 0} \min_w \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- Always true:

$$d^* \leq p^*$$

# Lagrange duality
# 拉格朗日对偶性

- The primal problem 原问题

$$p^* := \min_w f(w) = \min_w \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \ge 0} \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- The dual problem 对偶问题

$$d^* := \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i \ge 0} \min_w \mathcal{L}(w, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- Interesting case: when do we have

$$d^* = p^*?$$

# Lagrange duality
# 拉格朗日对偶性

- Theorem: under proper conditions, there exists ($w^*$ , $\boldsymbol{\alpha}^*$ ,$\boldsymbol{\beta}^*$) such that

在适当条件下，存在($w^*$ , $\boldsymbol{\alpha}^*$ ,$\boldsymbol{\beta}^*$), 使得

$$d^* = \mathcal{L}(w^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = p^*$$

Moreover, ($w^*$ , $\boldsymbol{\alpha}^*$ ,$\boldsymbol{\beta}^*$)satisfy Karush-Kuhn-Tucker (KKT) conditions:

($w^*$ , $\boldsymbol{\alpha}^*$ ,$\boldsymbol{\beta}^*$)满足Karush-Kuhn-Tucker (KKT) 条件：

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0, \qquad \alpha_i g_i(w) = 0$$

$$g_i(w) \leq 0,\ \ h_j(w) = 0, \qquad \alpha_i \geq 0$$

dual complementarity

对偶互补性

primal constraints

原始约束

dual constraints

对偶约束

# Lagrange duality
# 拉格朗日对偶性

- What are the proper conditions?

什么是适当的条件?

- A set of conditions (Slater conditions):
- $f, g_i$ convex, $h_j$ affine $f, g_i$ 是凸的，$h_j$ 是仿射的
- Exists $w$ satisfying all $g_i(w) < 0$
- There exist other sets of conditions

还有其他条件集

  - Search Karush–Kuhn–Tucker conditions on Wikipedia

# SVM: optimization

# SVM: optimization
# SVM：优化器

- Optimization (Quadratic Programming): （二次规划）

$$\min_{w,b} \frac{1}{2}||w||^2$$
$$y_i(w^T x_i + b) \geq 1, \forall i$$

- Generalized Lagrangian: 广义拉格朗日函数

$$\mathcal{L}(w,b,\boldsymbol{\alpha}) = \frac{1}{2}||w||^2 - \sum_i \alpha_i [y_i(w^T x_i + b) - 1]$$

where $\alpha$ is the Lagrange multiplier
其中$\alpha$ 是拉格朗日乘数

# SVM: optimization
# SVM：优化器

- KKT conditions: KKT条件：

$$\frac{\partial \mathcal{L}}{\partial w} = 0, \rightarrow w = \sum_i \alpha_i y_i x_i \quad (1)$$
$$\frac{\partial \mathcal{L}}{\partial b} = 0, \rightarrow 0 = \sum_i \alpha_i y_i \quad (2)$$

- Plug into $\mathcal{L}$: 代入$\mathcal{L}$:

$$\mathcal{L}(w, b, \boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (3)$$

combined with $0 = \sum_i \alpha_i y_i, \alpha_i \geq 0$

# SVM: optimization
# SVM：优化器

Only depend on inner products

- Reduces to dual problem:

$$\mathcal{L}(w,b,\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

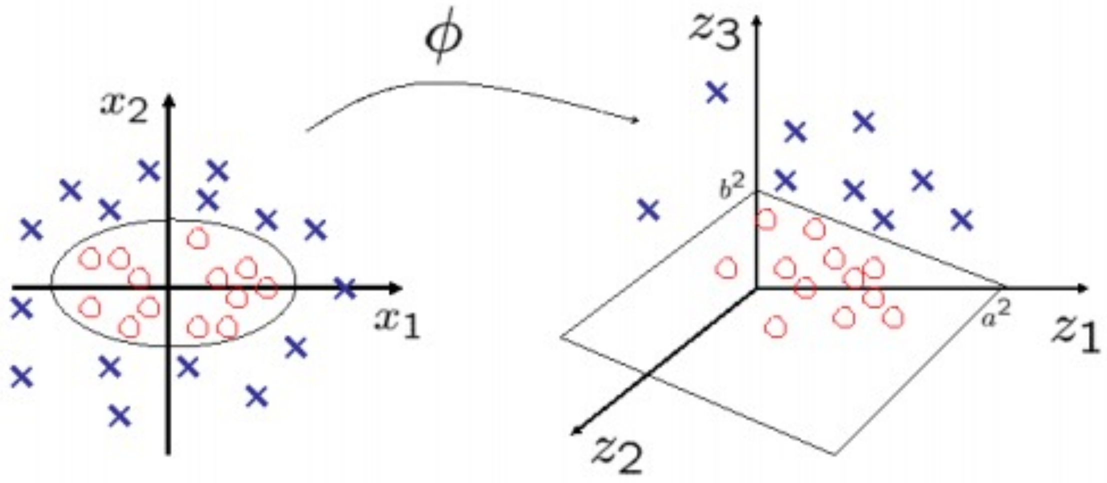
$$\sum_i \alpha_i y_i = 0, \alpha_i \geq 0$$

- Since $w = \sum_i \alpha_i y_i x_i$, we have $w^T x + b = \sum_i \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$

# Kernel methods

# Features
# 特征

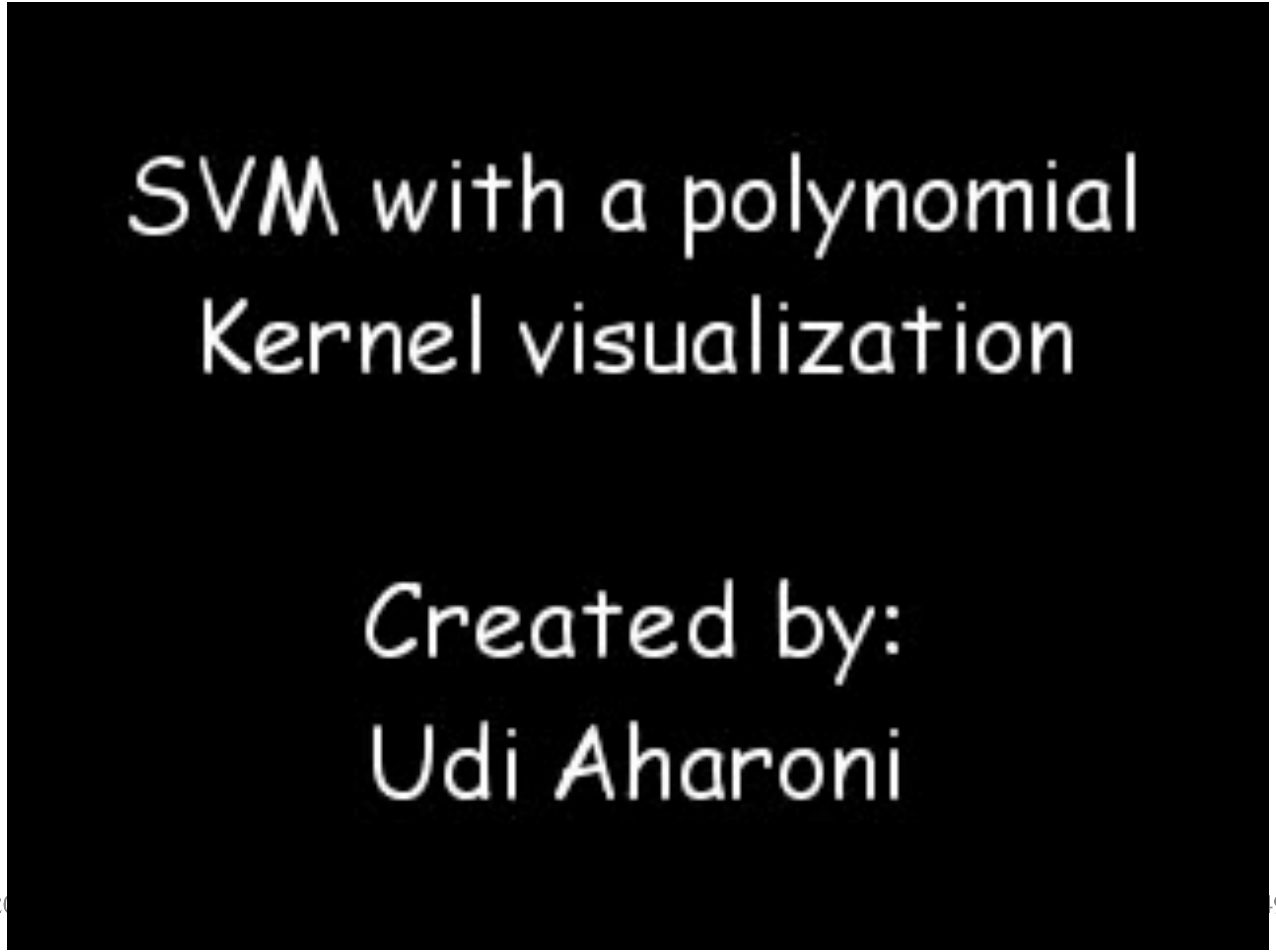
# Features
# 特征




$$\phi : (x_1, x_2) \longrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1 \longrightarrow \frac{z_1}{a^2} + \frac{z_3}{b^2} = 1$$

# Features
## 特征


SVM with a polynomial
Kernel visualization
Created by:
Udi Aharoni

# Features
# 特征

- Proper feature mapping can make non-linear to linear
适当的特征映射可以将非线性转化为线性

- Using SVM on the feature space $\{\phi(x_i)\}$: only need $\phi(x_i)^T\phi(x_j)$

- Conclusion: no need to design $\phi(\cdot)$, only need to design

$$k(x_i, x_j) = \phi(x_i)^T\phi(x_j)$$

# Polynomial kernels
# 多项式核

- Fix degree $d$ and constant $c$:

$$k(x, x') = (x^T x' + c)^d$$

- What are $\phi(x)$?
- Expand the expression to get $\phi(x)$

展开表达式以得到$\phi(x)$

# Polynomial kernels
# 多项式核

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^2, \quad K(\mathbf{x}, \mathbf{x}') = (x_1 x_1' + x_2 x_2' + c)^2 = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2c}\, x_1 \\ \sqrt{2c}\, x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} {x'}_1^2 \\ {x'}_2^2 \\ \sqrt{2}\, x_1' x_2' \\ \sqrt{2c}\, x_1' \\ \sqrt{2c}\, x_2' \\ c \end{bmatrix}$$

# Polynomial kernels
# 多项式核

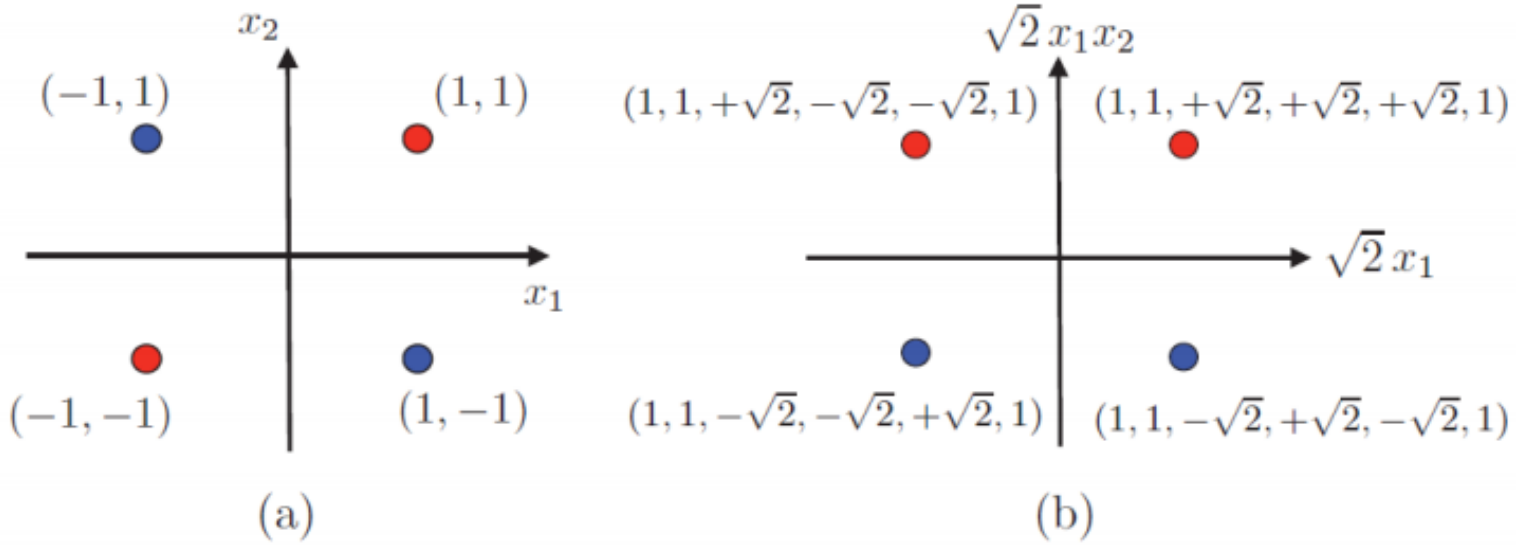



**Figure 5.2** Illustration of the XOR classification problem and the use of polynomial kernels. (a) XOR problem linearly non-separable in the input space. (b) Linearly separable using second-degree polynomial kernel.

# Gaussian kernels
# 高斯核

- Fix bandwidth $\sigma$:

$$k(x,x') = \exp(-||x - x'||^2/2\sigma^2)$$

- Also called radial basis function (RBF) kernels

也称为径向基函数（RBF）核

- What are $\phi(x)$? Consider the un-normalized version

考虑未归一化的版本

$$k'(x,x') = \exp(x^T x'/\sigma^2)$$

- Power series expansion: 幂级数展开：

$$k'(x,x') = \sum_{i}^{+\infty} \frac{(x^T x')^i}{\sigma^i i!}$$

# Mercer's condition for kenerls
# 核的 Mercer 条件

- Theorem: $k(x, x')$ has expansion

$k(x, x')$具有展开式

$$k(x, x') = \sum_{i}^{+\infty} a_i \phi_i(x) \phi_i(x')$$

if and only if for any function $c(x)$,

当且仅当对于任意函数$c(x)$,

$$\int \int c(x) c(x') k(x, x') dx dx' \geq 0$$

(Omit some math conditions for $k$ and $c$)

（省略一些关于$k$和$c$的数学条件）

# Constructing new kernels
# 构造新核函数

- Kernels are closed under positive scaling, sum, product, pointwise limit, and composition with a power series $\sum_{i}^{+\infty} a_i k^i(x, x')$

  核函数在正向缩放、求和、乘积、逐点极限和与幂级数的合成下是封闭的

- Example: $k_1(x, x'), k_2(x, x')$ are kernels, then also is

$$k(x, x') = 2k_1(x, x') + 3k_2(x, x')$$

- Example: $k_1(x, x')$ is kernel, then also is

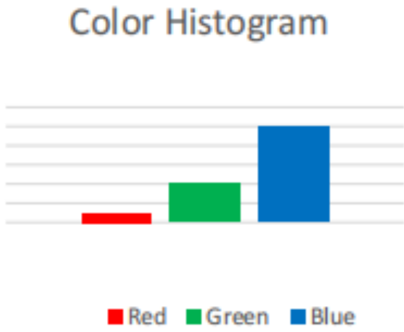
$$k(x, x') = \exp(k_1(x, x'))$$

# Kernels v.s. Neural networks

# Features
# 特征

$x$

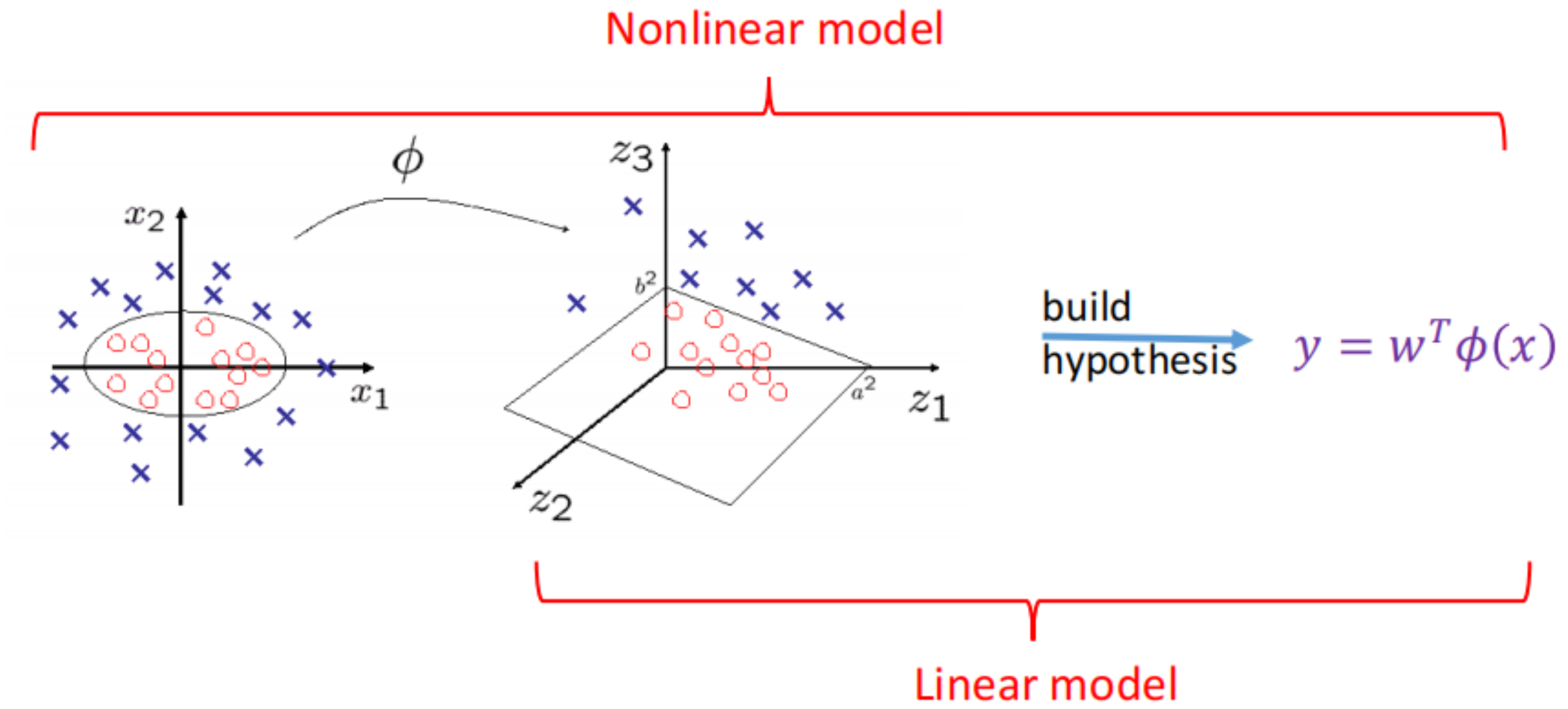
Extract features




build hypothesis

$y = w^T \phi(x)$

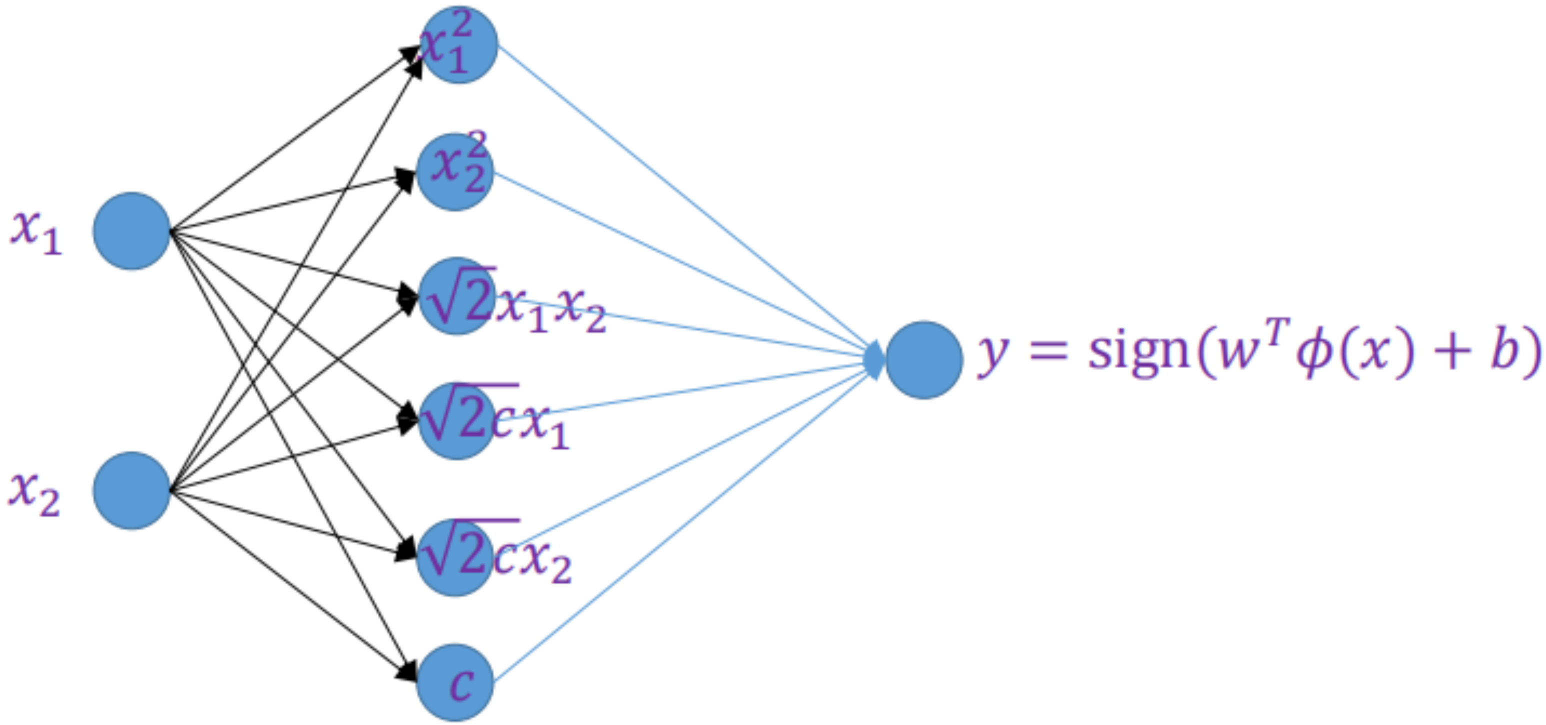
# Features: part of the model
# 特征：模型的一部分



# Polynomial kernels
# 多项式核

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^2, \quad K(\mathbf{x}, \mathbf{x}') = (x_1 x_1' + x_2 x_2' + c)^2 = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \\ \sqrt{2c}\, x_1 \\ \sqrt{2c}\, x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} {x'}_1^2 \\ {x'}_2^2 \\ \sqrt{2}\, x_1' x_2' \\ \sqrt{2c}\, x_1' \\ \sqrt{2c}\, x_2' \\ c \end{bmatrix}$$

# Polynomial kernel SVM as two layer neural network
# 多项式核SVM作为二层神经网络




First layer is fixed. If also learn first layer, it becomes two layer neural network