

支持向量机

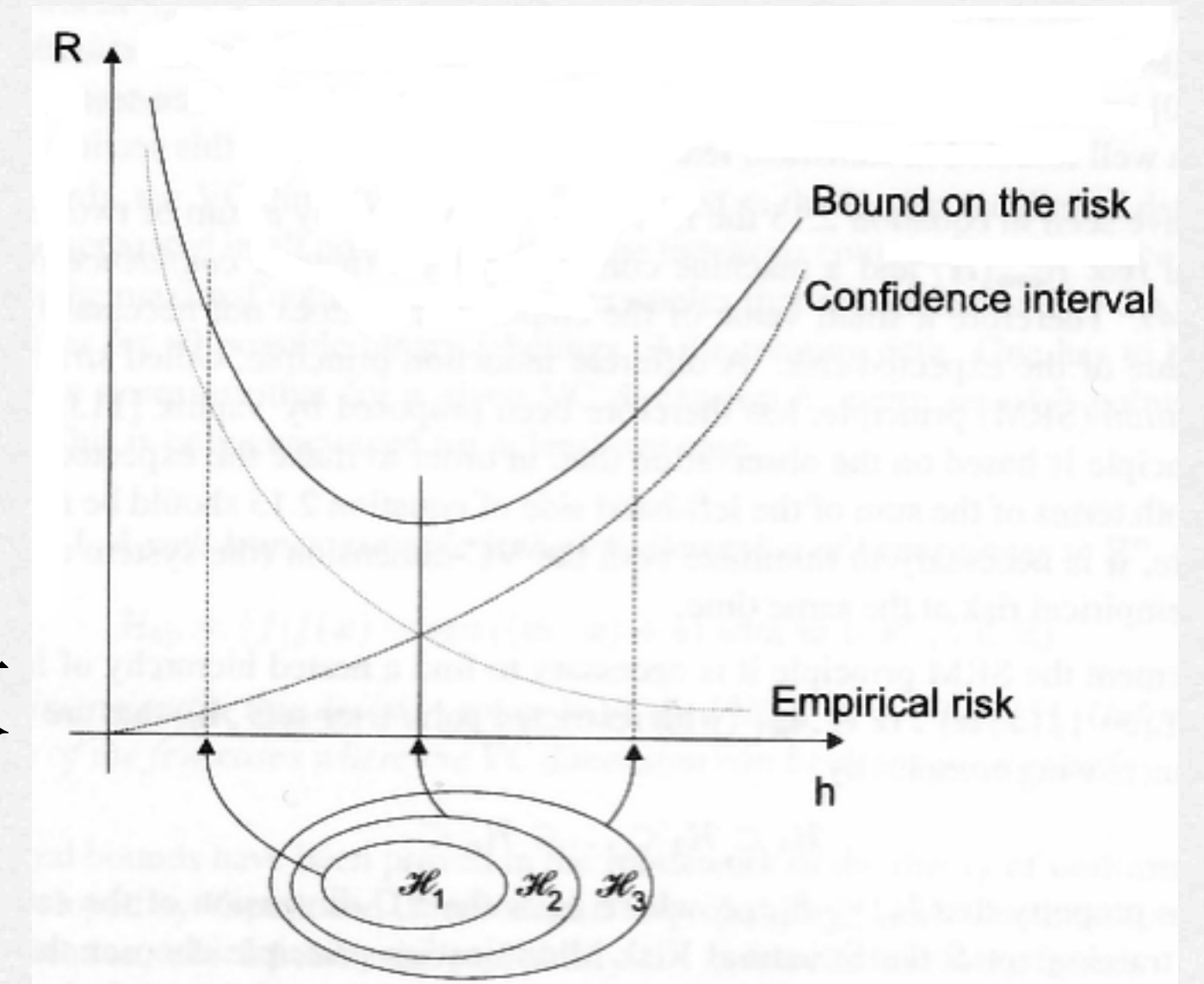
张超@CSLT, RIIT, Tsinghua University

支持向量机(SVM)

- 迄今为止最好的二分类分类器之一
 - 可靠完整的理论基础: (在最坏情况下)最优的分类器
 - 强大的性能: (及其变种)在很多问题上取得最好的效果
 - 只需要较少的训练数据量
- 应用纯理论(统计学习理论)成功设计实用模型的范例
 - 大边距准则; 基于支持向量的稀疏性; 软间隔; 核函数映射
- 大量变种模型及工具包

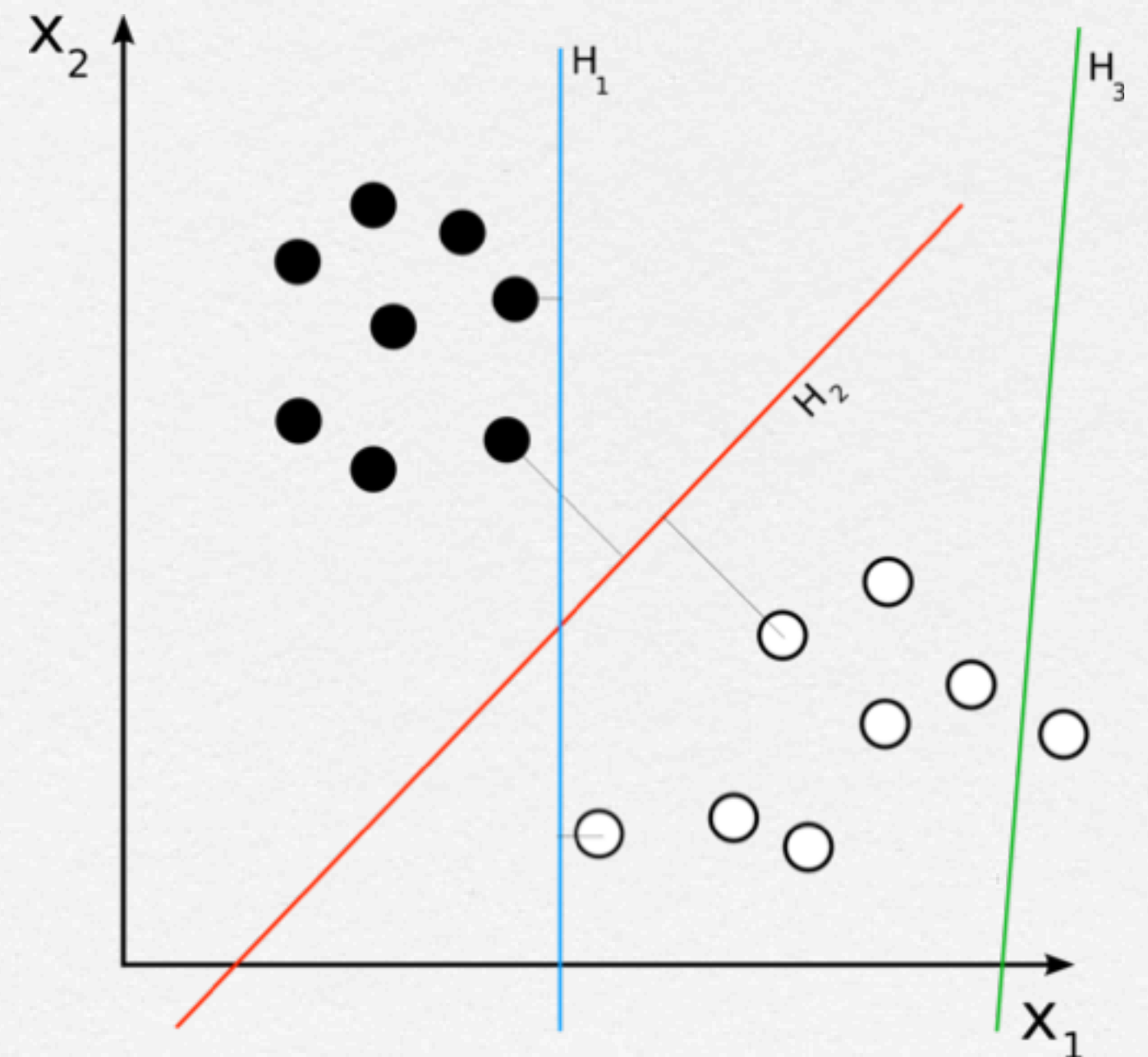
统计学习理论回顾

- 一种分析(在最坏情况下)统计学习中所需的数据量和在真实数据上所达到的错误率(模型的期望风险、推广能力)的理论
- 期望风险的边界: 结构风险最小化
 - $R(\alpha) \leq R_{\text{emp}}(\alpha) + \Phi\left(\frac{h}{l}\right)$
 - $R_{\text{emp}}(\alpha)$ 是经验风险, $\Phi\left(\frac{h}{l}\right)$ 是置信区间
 - h 是机器函数集的VC维, l 是训练样本数
 - 在损失函数有界时 $\Phi\left(\frac{h}{l}\right)$ 与 $\frac{h}{l}$ 同增减



线性二分类问题

- 假设训练数据 $(x^1, y^1), \dots, (x^l, y^l), x \in R^n, y \in \{+1, -1\}$ 可被一个(超)平面 $(w \cdot x) - b = 0$ 无错误地分开
- 决策函数为 $f(x) = \text{sgn}\{w \cdot x - b\}$
- 选择哪一个分类平面? 为什么?
 - 使训练数据无错误率 ($R_{\text{emp}}(\alpha) = 0$)
 - 继续选择(红、蓝)? 红!



最优分类超平面

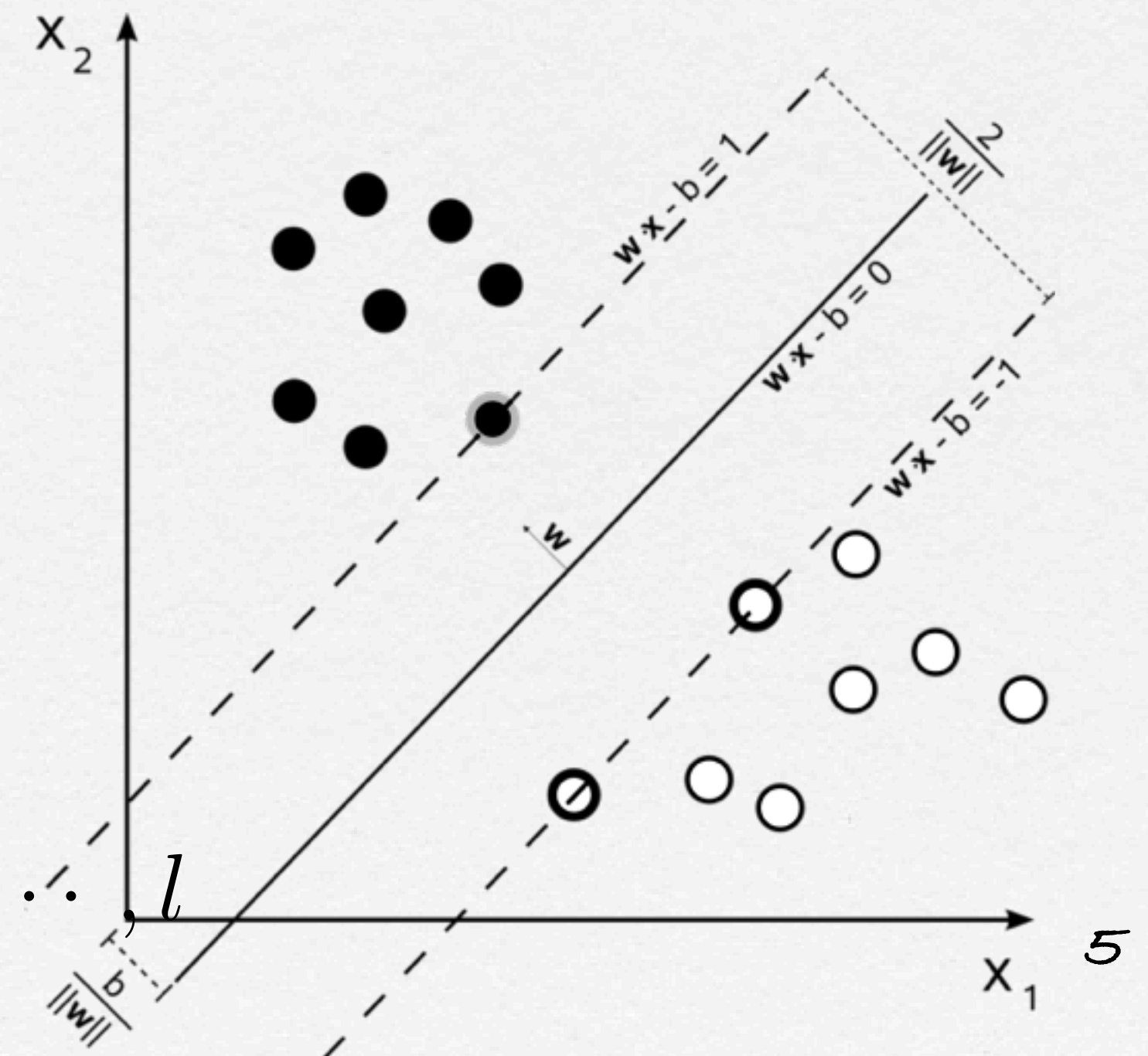
- 若训练数据被无错误分开，且距离超平面最近的数据与超平面间的距离（间隔）最大，则说数据集被最优超平面分开(红)
- 超平面参数可以任意放缩：无穷多解 $\rightarrow$ 强制把离分类面距离最近的样本的决策函数值归一化

$$\square (w \cdot x) - b \begin{cases} \geq 1 & \text{if } y_i = 1, \\ \leq -1 & \text{if } y_i = -1. \end{cases}$$

$$\square \text{ 边距 } \frac{1}{\|w\|}$$

$$\square \text{ 最优分类超平面 } \min_{w,b} \Phi(w) = \|w\|^2,$$

$$\text{subject to } y^i [(w \cdot x) - b] \geq 1, i = 1, \dots, l$$



为什么选最优分类超平面

- Δ 间隔分类超平面的VC维

- 设向量 $x \in X$ 属于一个半径为 R 的 n 维超球体中, 那么 Δ 分类超平面集合的VC维 h 的界为 $h \leq \min \left\{ \frac{R^2}{\Delta^2}, n \right\} + 1$

- 最小化结构风险 $R(\alpha) \leq R_{\text{emp}}(\alpha) + \Phi\left(\frac{h}{l}\right)$

- 假定了所有训练样本正确分类, 有 $R_{\text{emp}}(\alpha) = 0$

- 最大化间隔 $\Delta \rightarrow$ 最小化 h 的上界 $\rightarrow$ 最小化 $\Phi\left(\frac{h}{l}\right)$ 的上界

- 利用控制间隔实现对结构风险最小化准则中推广能力的控制

如何求得最优分类超平面

$$\min_{w,b} \Phi(w) = \|w\|^2,$$

□ 最优分类超平面 subject to $y^i[(w \cdot x) - b] \geq 1, i = 1, \dots, l$

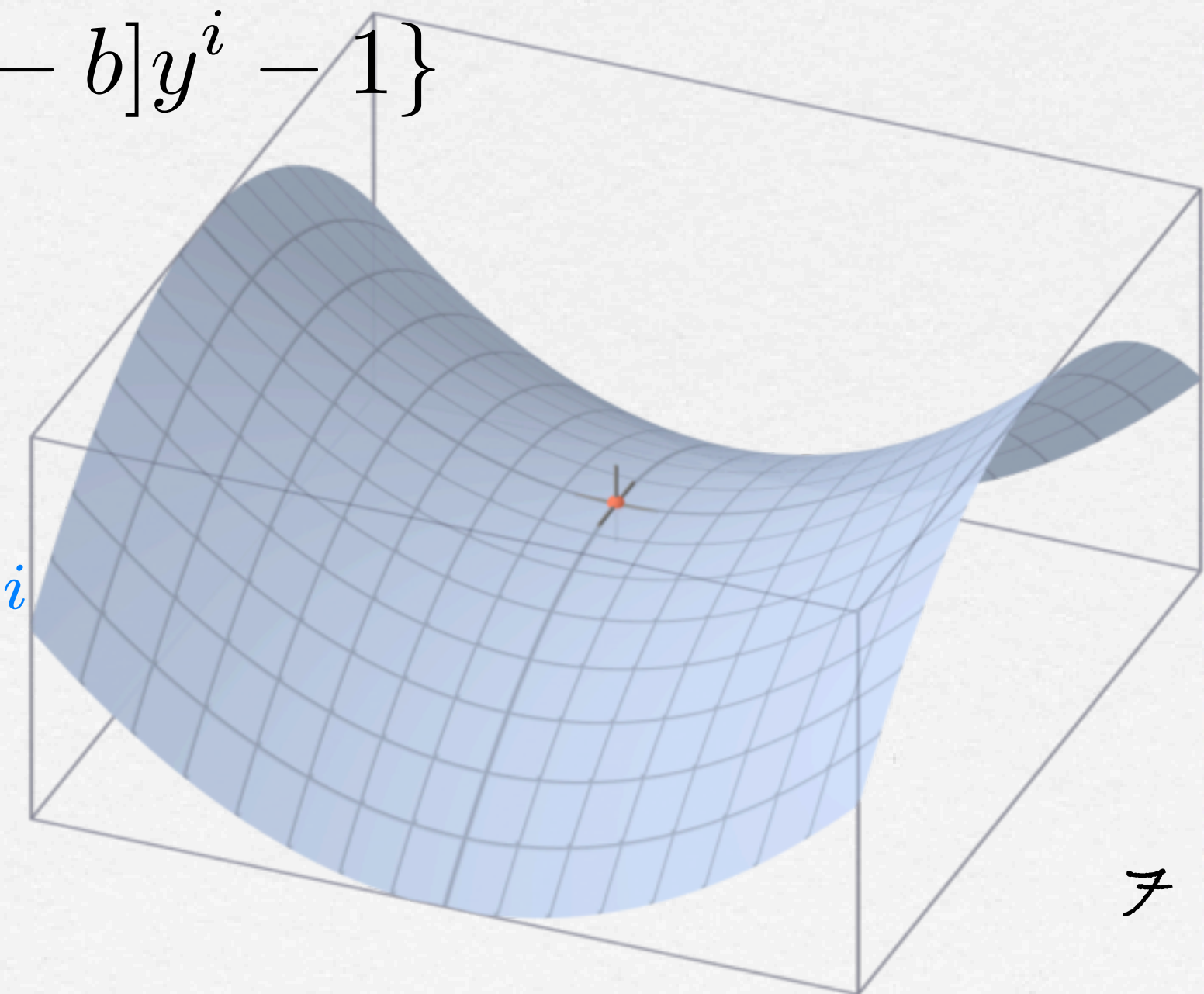
□ 由拉格朗日方法，上式解为拉格朗日函数的鞍点 $\min_{w,b} \max_{\alpha} L(w, b, \alpha)$

$$\square L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^l \alpha_i \{[(x^i \cdot w) - b]y^i - 1\}$$

□ 鞍点的解满足

$$\square \frac{\partial L(w_0, b_0, \alpha^0)}{\partial b} = 0 \rightarrow \sum_{i=1}^l \alpha_i^0 y^i = 0$$

$$\square \frac{\partial L(w_0, b_0, \alpha^0)}{\partial w} = 0 \rightarrow w_0 = \sum_{i=1}^l x^i \alpha_i^0 y^i$$



如何求得最优分类超平面

□ 由非线性优化问题的最优解的一阶必要条件(KT条件):

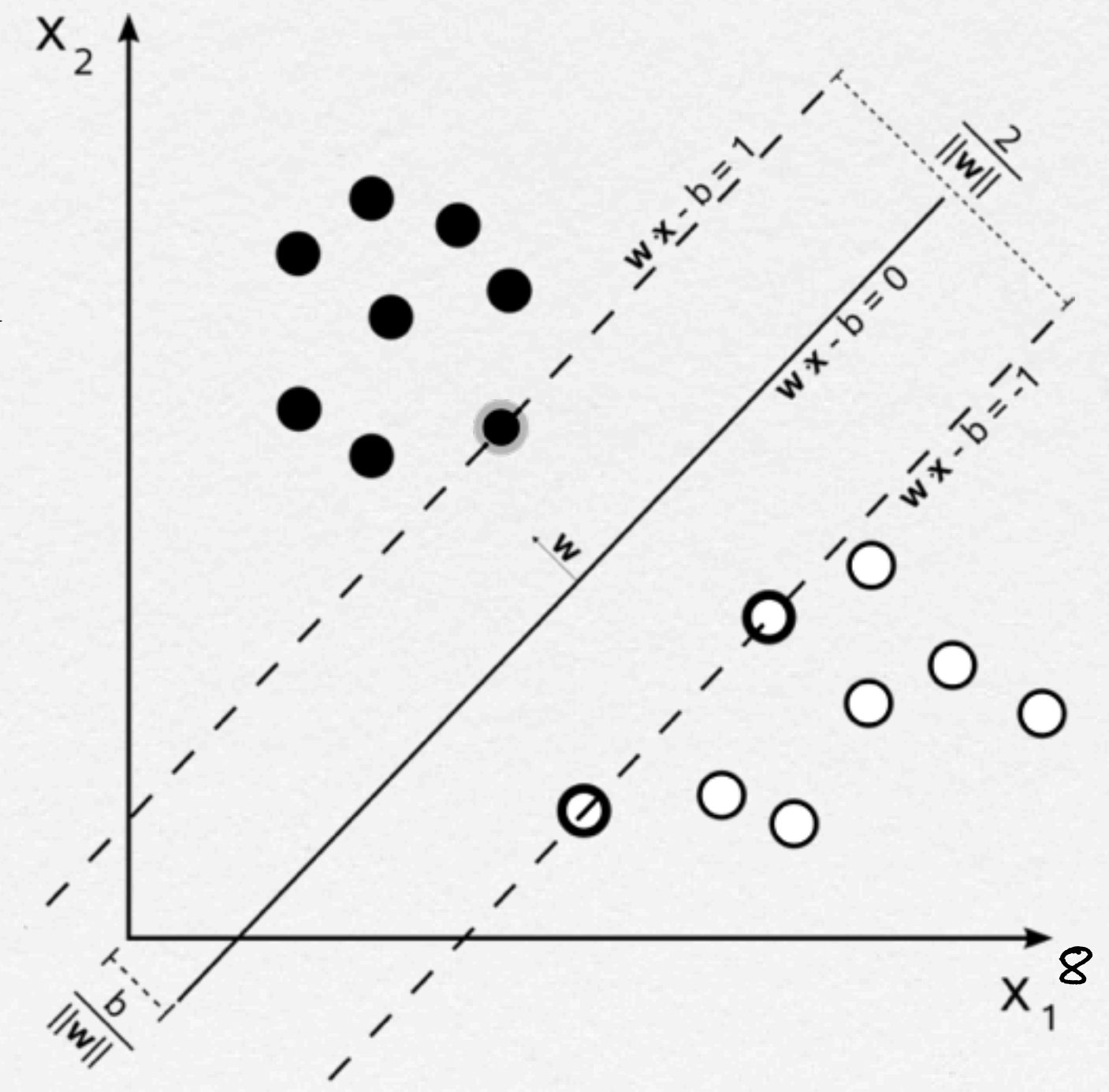
□ $\alpha_i^0 \{[(x^i \cdot w) - b_0]y^i - 1\} = 0, i = 1, \dots, l$

□ 又 $y^i[(w \cdot x) - b] \geq 1$, 且 $\alpha_i \geq 0, i = 1, \dots, l$

□ 仅满足 $y^i[(w \cdot x) - b] = 1$ 的样本系数非零

□ 称这些样本为支持向量, 集合记做有 SV

□ 有 $w_0 = \sum_{i=1}^l x^i \alpha_i^0 y^i = \sum_{i \in SV} x^i \alpha_i^0 y^i$



如何求得最优分类超平面

□ 将 $w_0 = \sum_{i=1}^l x^i \alpha_i^0 y^i$ 代入 $L(w, b, a)$, 得求解最优超平面的对偶问题(训练)

$$\square \max_{\alpha} W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=0}^l \alpha_i \alpha_j y^i y^j (x^i \cdot x^j),$$

subject to $\alpha_i \geq 0, i = 1, \dots, l,$

$$\sum_{i=1}^l \alpha_i y^i = 0$$

□ 是一个简单的二次规划问题: $L\text{-BFQS}, SQP, \dots$; 正则化

如何求得最优分类超平面

□ 设 α^0 为前述二次优化问题的解，则对于最优超平面有

□ 阈值用支持向量待定系数求得：

$$b_0 = \frac{1}{2}[(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))]$$

□ 则最优分类超平面的分类规则为(测试)

□ $f(x) = \text{sgn} \left(\sum_{i \in SV} y^i \alpha_i^0 (x^i \cdot x) - b_0 \right)$

□ 故SVM需要存储所有支持向量

训练样本线性不可分的情况

- 训练样本线性不可分 $\rightarrow$ 经验风险 $R_{\text{emp}}(\alpha) \neq 0$
- 此时原最优超平面的约束条件不能全部满足
 - 引入非负松弛变量 $\xi_i \geq 0$ ；指定错份样本惩罚度 C
 - 则优化问题变为 (软边距技术)
$$\min \Phi(w, \xi) = \frac{1}{2}(w \cdot w) + C \left(\sum_{i=1}^l \xi_i \right)$$
$$\text{subject to } y^i((w \cdot x^i) - b) \geq 1 - \xi_i, i = 1, 2, \dots, l$$
- 仍使用求解其对偶问题的技术求最优分类超平面

非线性问题和内积回旋

- 非线性分类问题

- 即通过预先选定的非线性映射将 X 空间映射为高维空间 Z

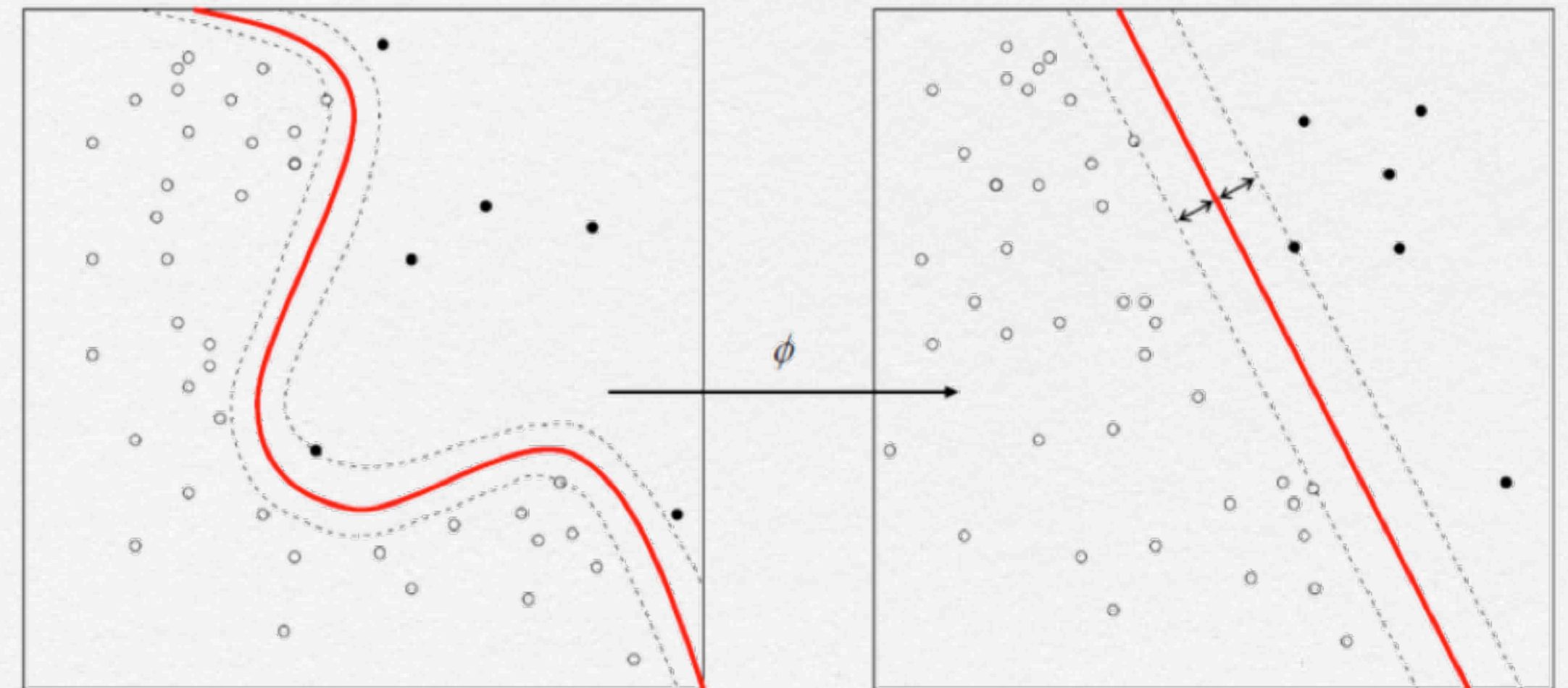
- 在高维空间 Z 中构造最优分类超平面

- 直接升维可能带来维度灾难

- X 空间: $3x^3 + 2x^2 + 0x + 1$

- Z 空间:

$$\begin{bmatrix} 3 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$



非线性问题和内积回旋

- 内积的回旋

- Z 空间中内积可转换为原 X 空间中的核函数 (Mercer 准则)

- 故变换空间中的最优分类超平面只需在原空间中求解

- 概念性变换 (不实现变换) $(z^i, z^j) = (\phi(x^i), \phi(x^j)) = K(x^i, x^j)$

- 训练
$$W(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=0}^l \alpha_i \alpha_j y^i y^j K(x^i, x^j),$$

- 测试
$$f(x) = \text{sgn} \left(\sum_{i \in SV} y^i \alpha_i^0 K(x, x^i) - b_0 \right)$$

非线性问题和内积回旋

- 常用核函数

- 多项式核函数 (Polynomial)

- d 阶多项式核函数 $K_p(x, x^i) = [(x \cdot x_i) + 1]^d$

- 径向基核函数 (RBF)

- 核函数 $K_\gamma(x, x^i) = \exp\{-\gamma\|x - x^i\|^2\}$

- Sigmoid核函数

- 核函数 $K_s(x, x^i) = \tanh(v(x \cdot x_i) + c)$, 包含隐藏的MLP

N分类问题

- 基于二分类SVM

- *one vs. rest*

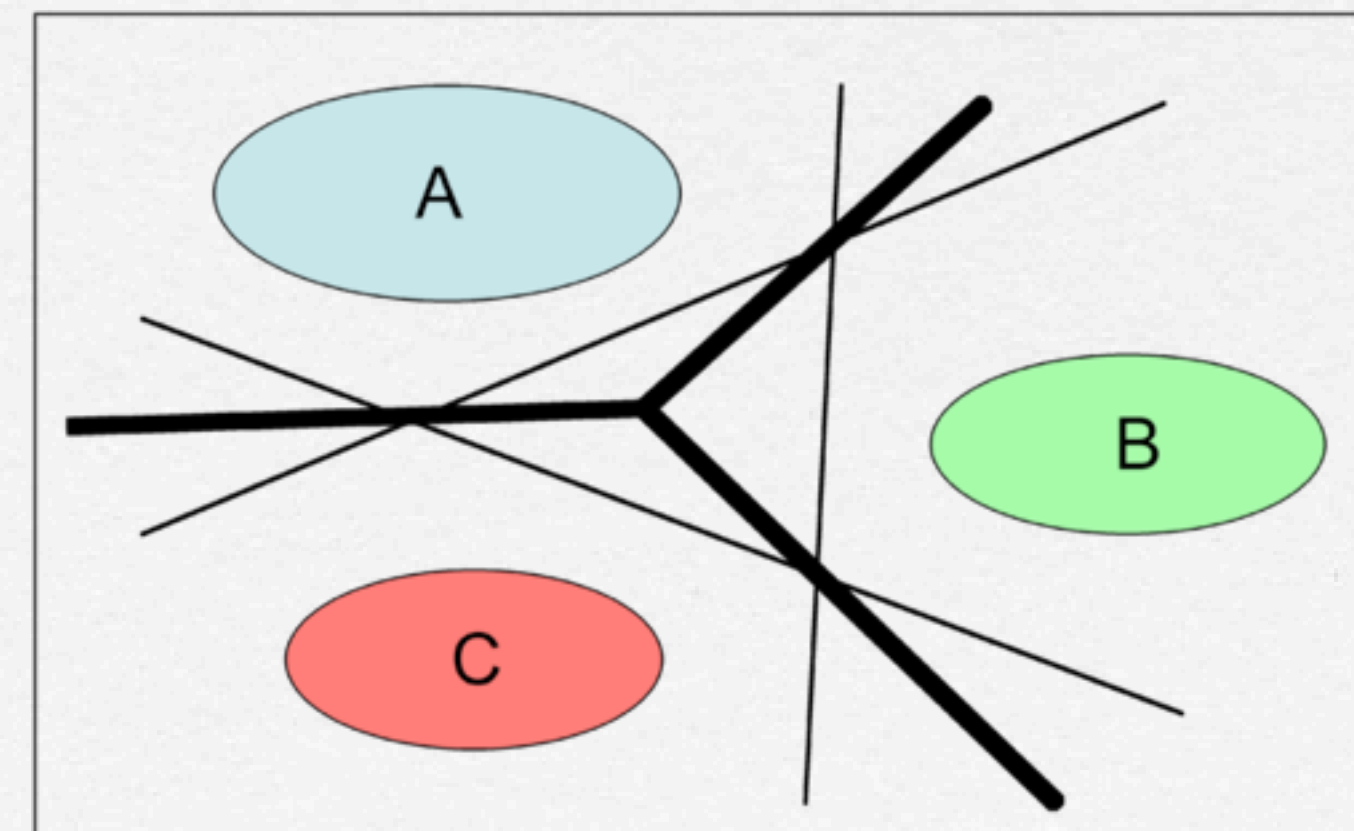
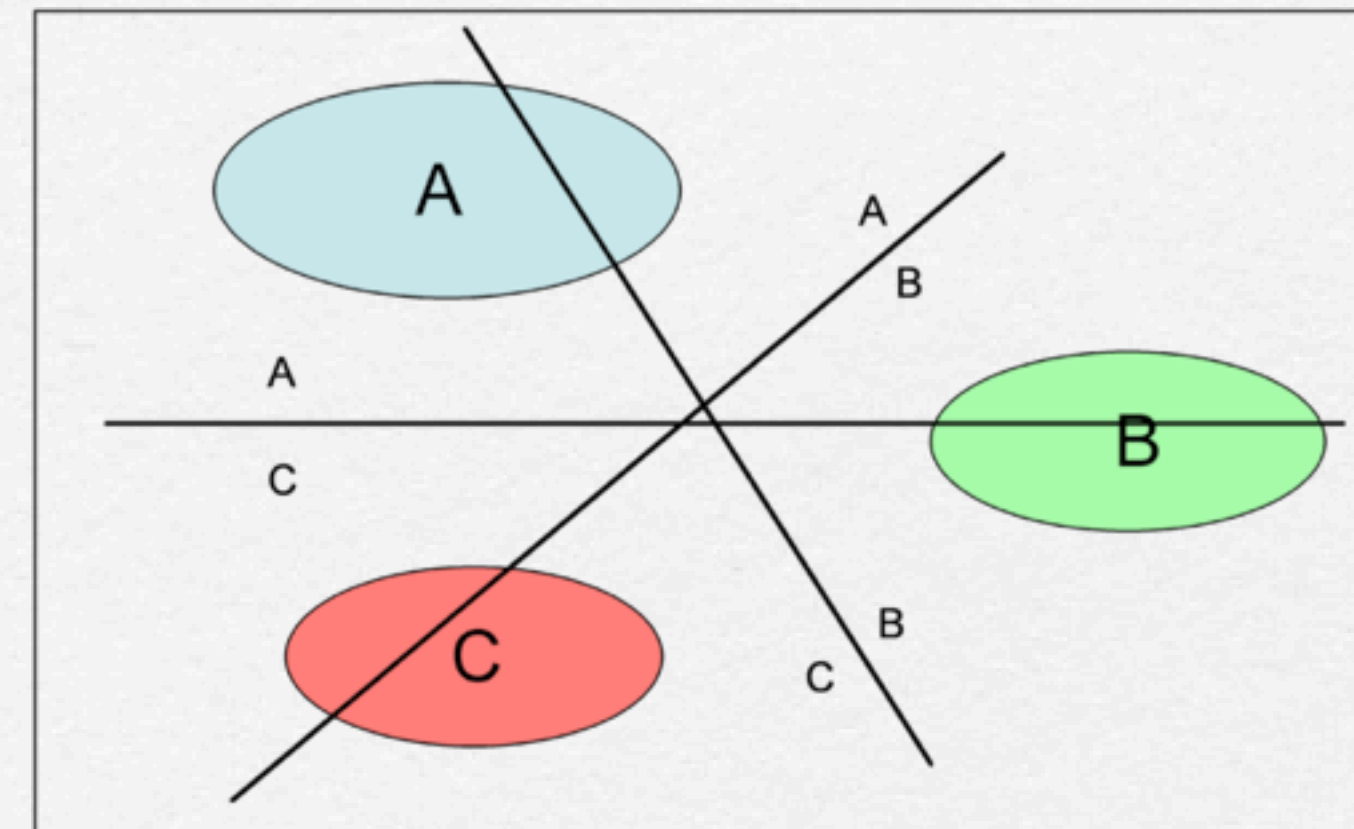
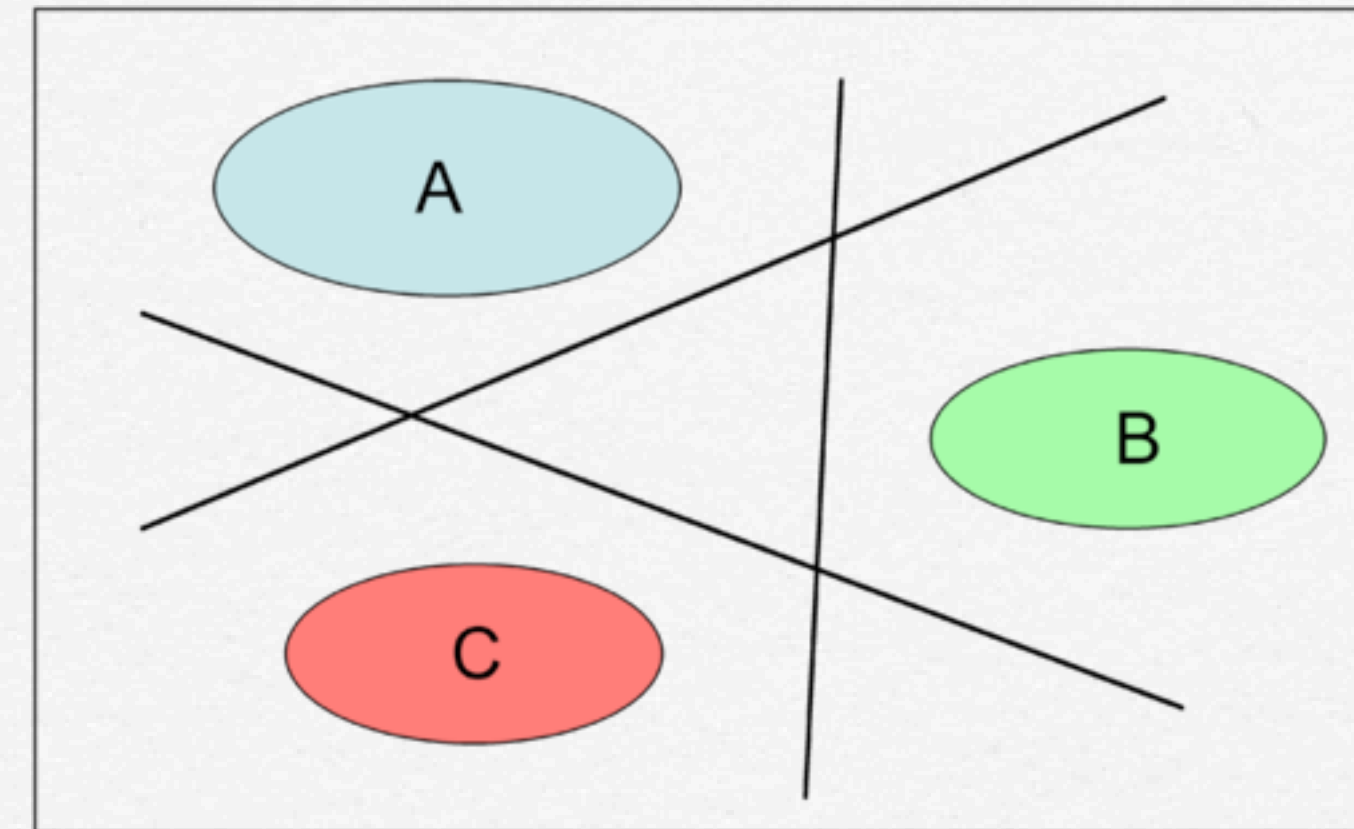
- N 个分类器

- *one vs. one*

- C_N^2 个分类器

- 多分类SVM

- 1个分类器



工具包及其使用

- SVM Light: <http://svmlight.joachims.org/>
- Lib Linear: <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>
- ☑ Lib SVM: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>
- 使用: 选择核函数 -> 给定1阶、2阶正则化惩罚 -> 选定参数
 - 核函数的选择与训练速度和效果密切相关(采样、均值等方法)
 - 正则化惩罚用于避免过拟合
 - 通常数据量越大, 得到的模型中支持向量越多, 分类速度越慢
 - 二次优化的SVM对于非对称数据(正负例)敏感

SVM相关技术在ASR中的应用

- 应用SVM进行音子分类任务
 - Clarkson and Moreno: 在帧级别进行分类任务
- 应用SVM准则训练声学模型
 - Berkeley, MIT, MSR, Gatech: 利用大边距、软边距等准则作为HMM声学模型的训练标准
- 结构化的SVM进行语音识别
 - Cambridge: 使用HMM相关技术进行分段，再用SVM区分易混淆单元

Thank for your listening~!