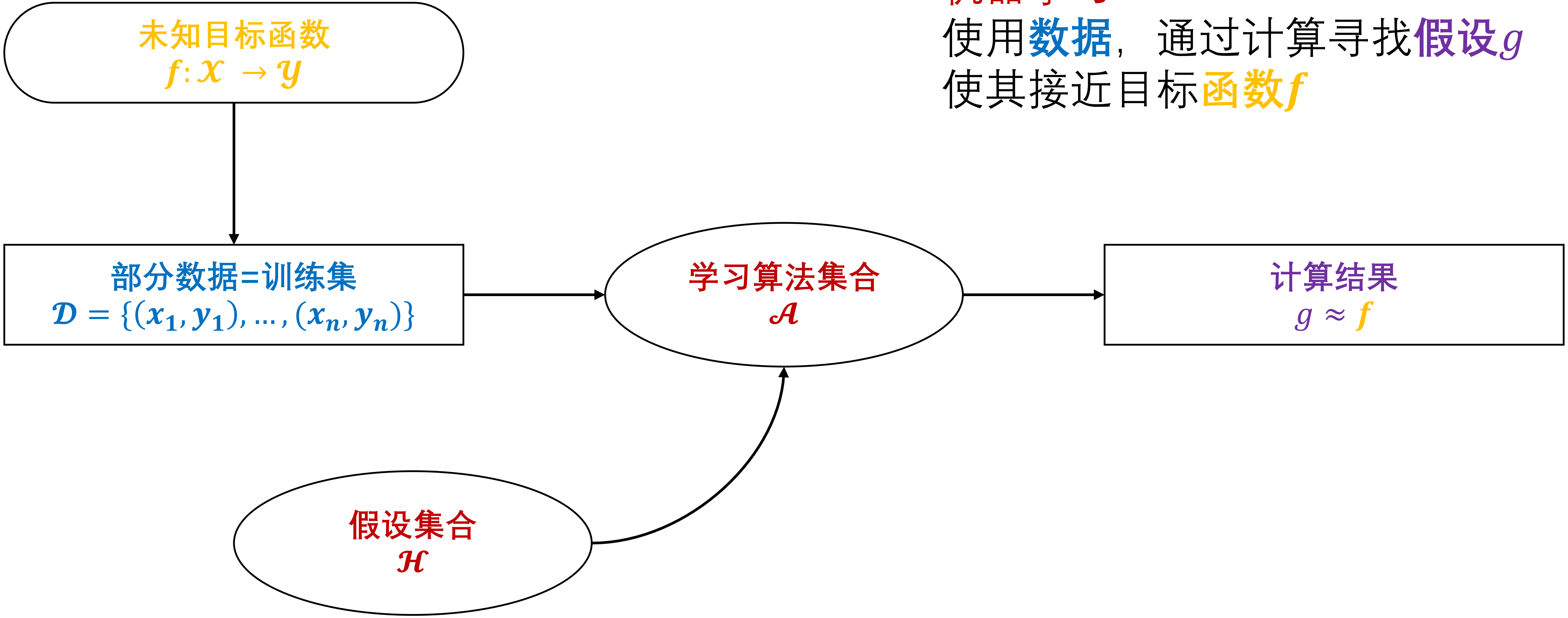
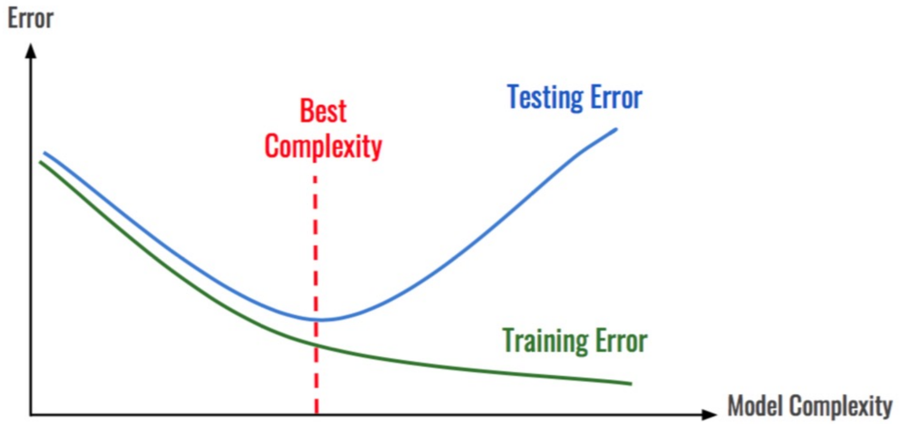
0 机器学习的抽象

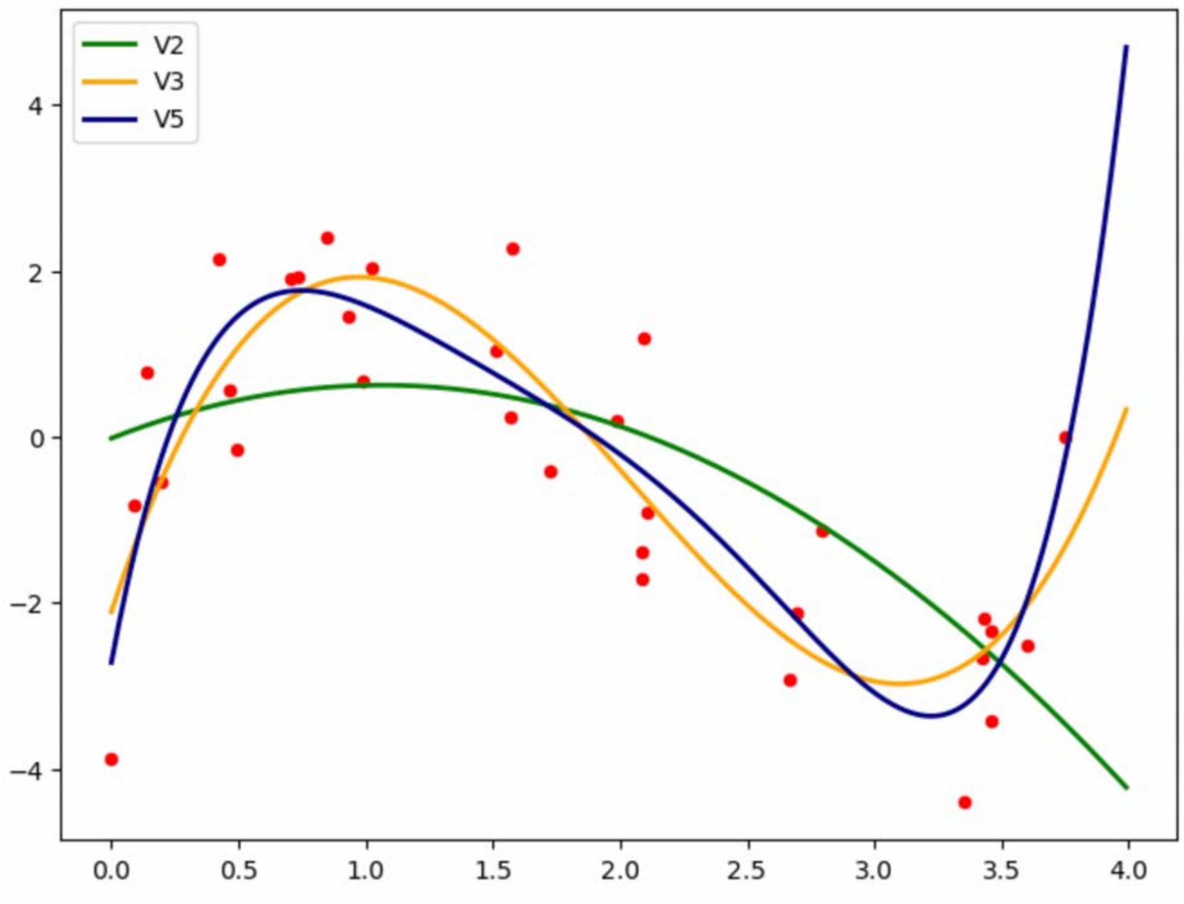
机器学习：
使用**数据**，通过计算寻找**假设**g
使其接近目标**函数**f



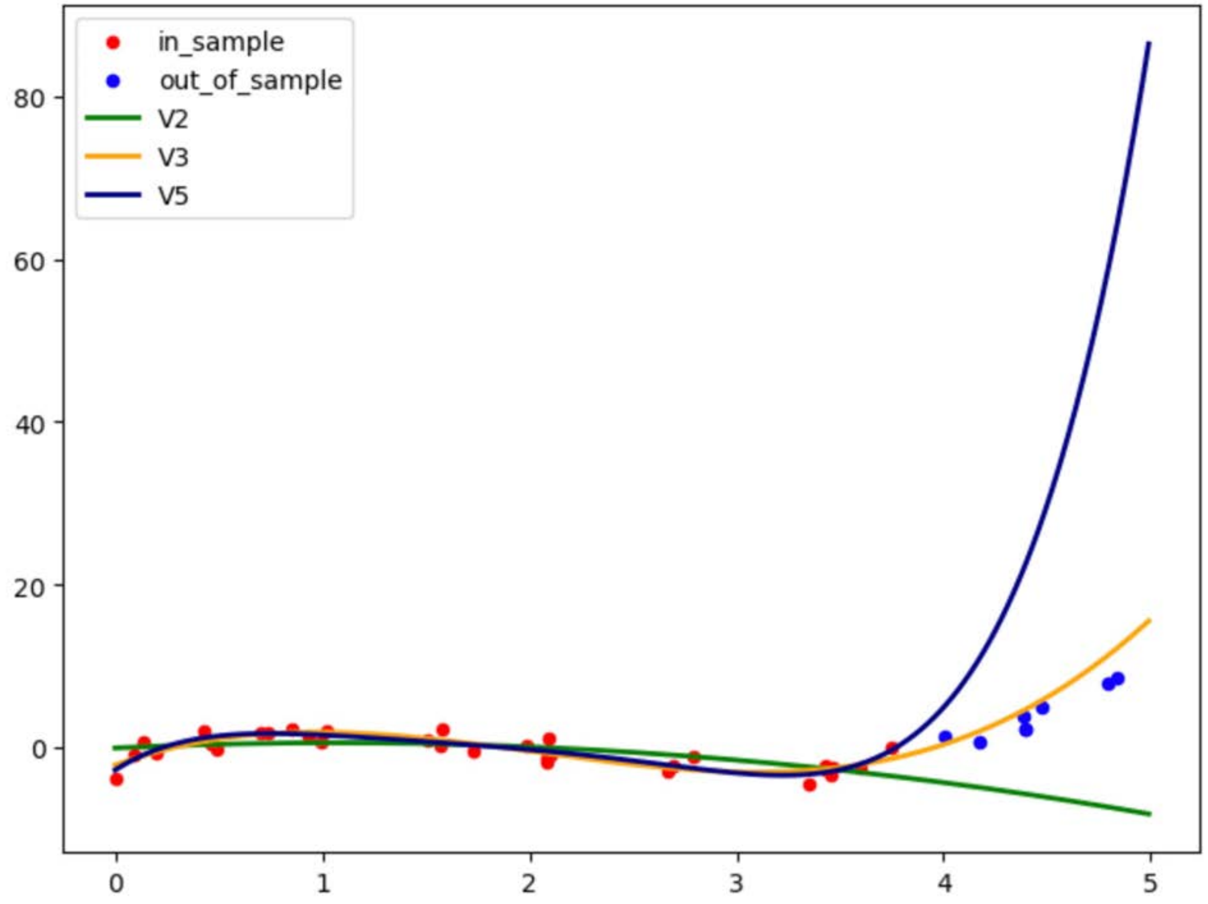
0 样本内与样本外



Error
Best
Complexity
Testing Error
Training Error
Model Complexity



V2
V3
V5



in_sample
out_of_sample
V2
V3
V5

0 正则化

OLS

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j.$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2\right\}.$$

Ridge Regression

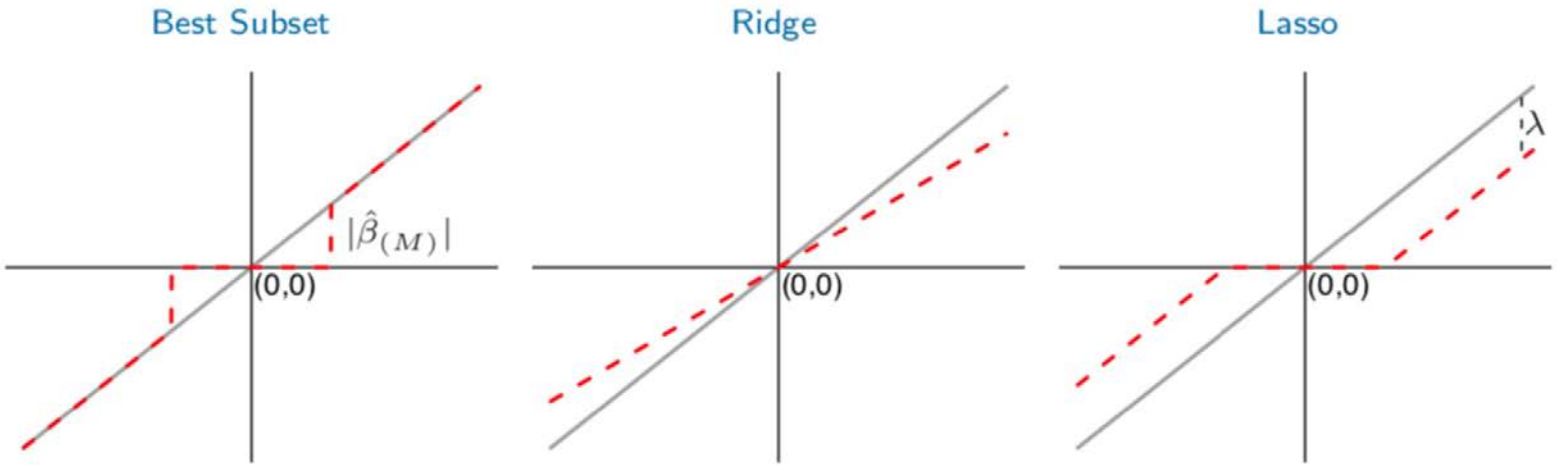
$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j.$$

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}}\Big\{\sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \boxed{+ \lambda \sum_{j=1}^{p} \beta_j^2 \Big\}.}$$

Lasso Regression

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j.$$

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}}\Big\{\frac{1}{2}\sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \boxed{+ \lambda \sum_{j=1}^{p} |\beta_j| \Big\}.}$$



2023

支持向量机

目 录

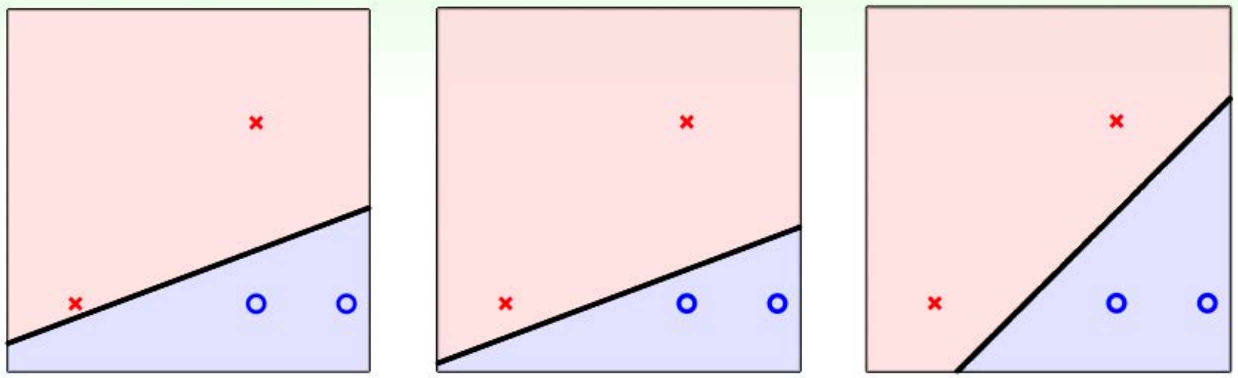
CONTENT

Part 01

从直觉到实践

- 好的分割
- 支持向量

1.1 直觉：哪种分割比较好?

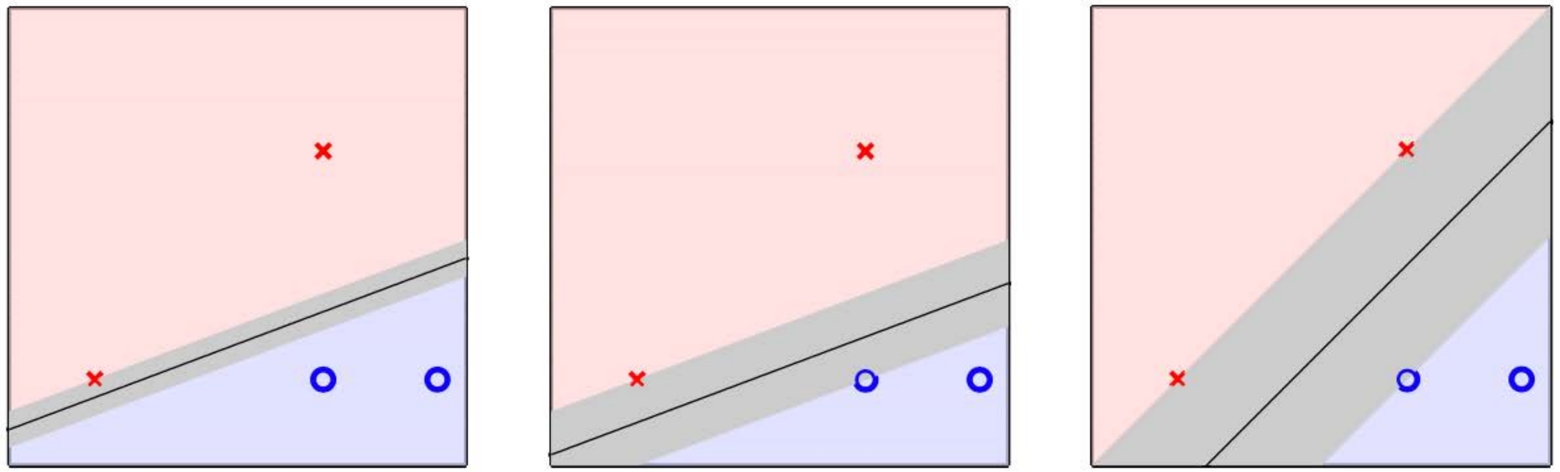


未知目标函数

$f: \mathcal{X} \rightarrow \mathcal{Y}$

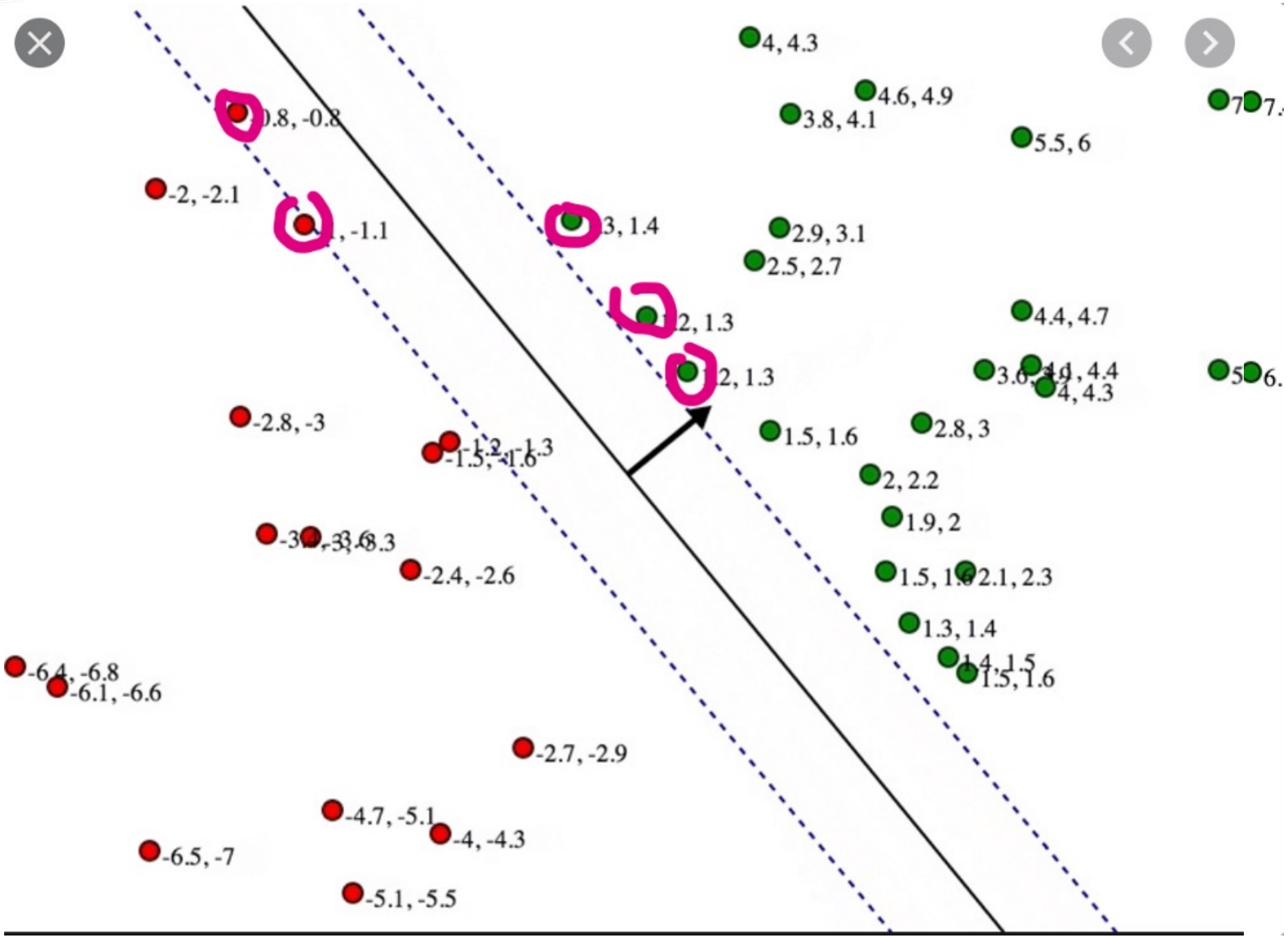
部分数据=训练集

$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$



- 大量的数据我们看不到
- 尽力追求“看不到的正确”
- 尽量大的间隔
- 两条与样本相切的“平行线”
- 找到各种可能中最宽的一组

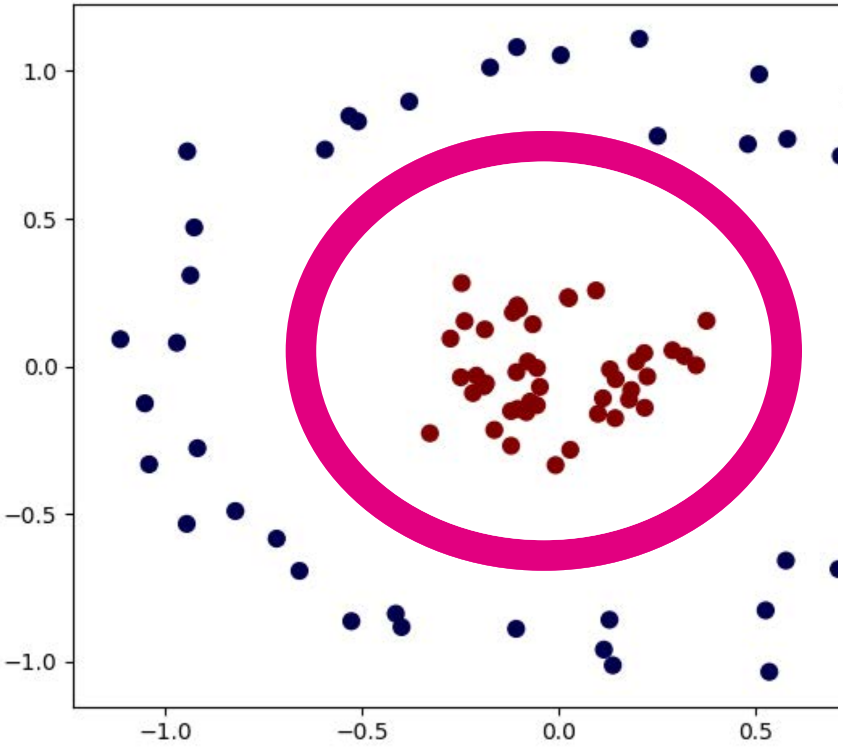
1.2 实践

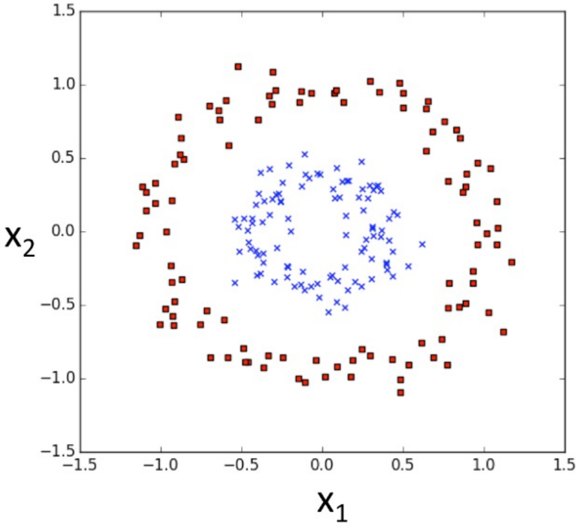


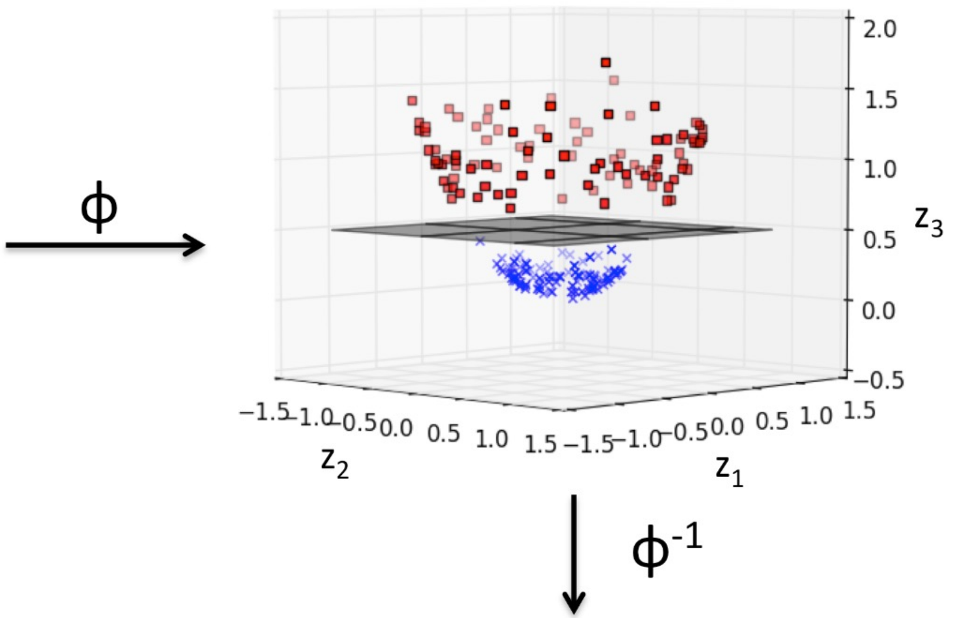
- 左侧的“搜寻”只是一种想象
- 真实求解过程是基于数学的
- 有颇具“艺术性”的数学支撑

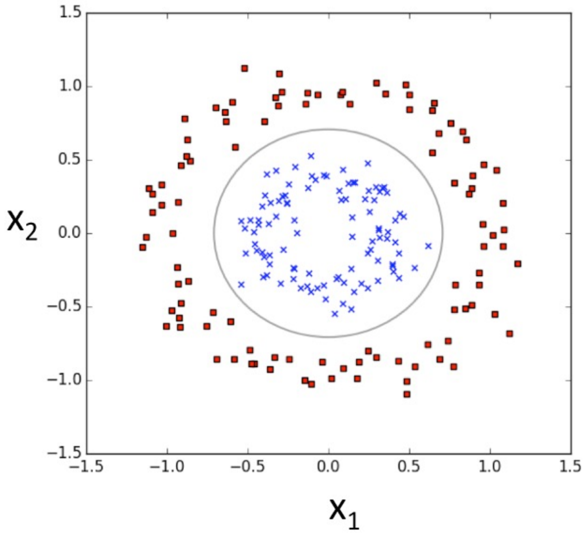
- 最终的结果被少数样本决定
- 我们称之为**支撑向量**
- 稀疏性
  - 结果的稳定性
  - 数据的需求低
  - 算法复杂度与支持向量相关

1.3 我们真的能够分得开这些数据么?









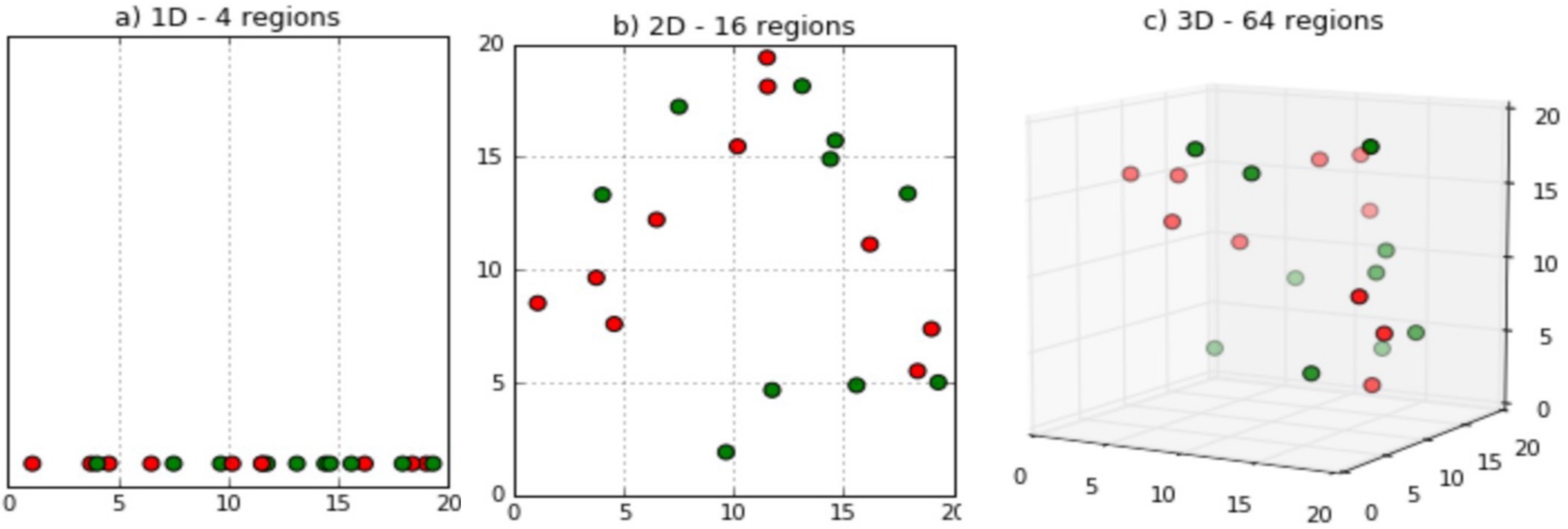
- 有没有理论保障?
- 有没有高效执行?

Part 02

维度与核方法

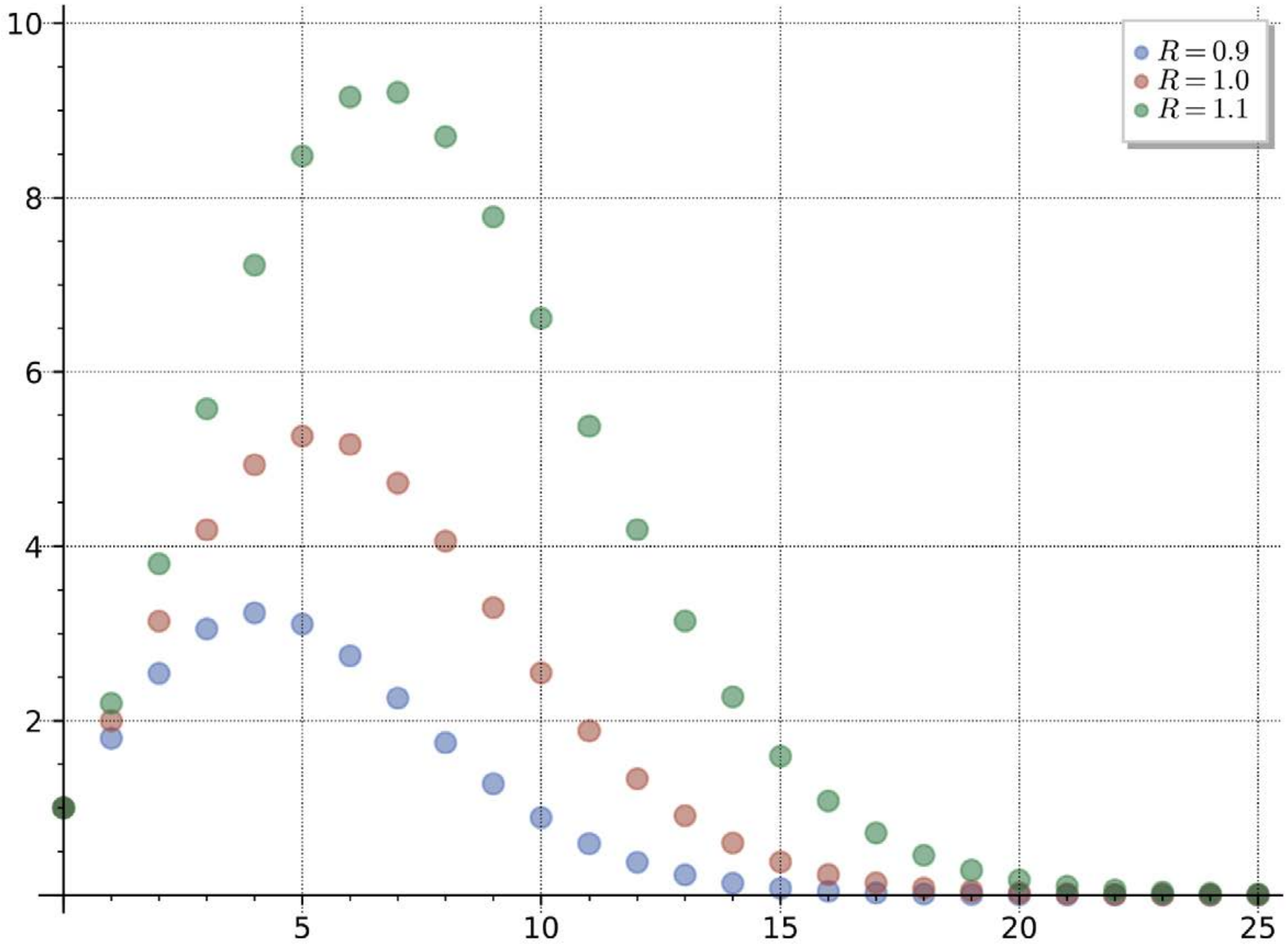
- 维度祝福
- 维度诅咒
- 核方法

2.1 从熟悉的维度开始



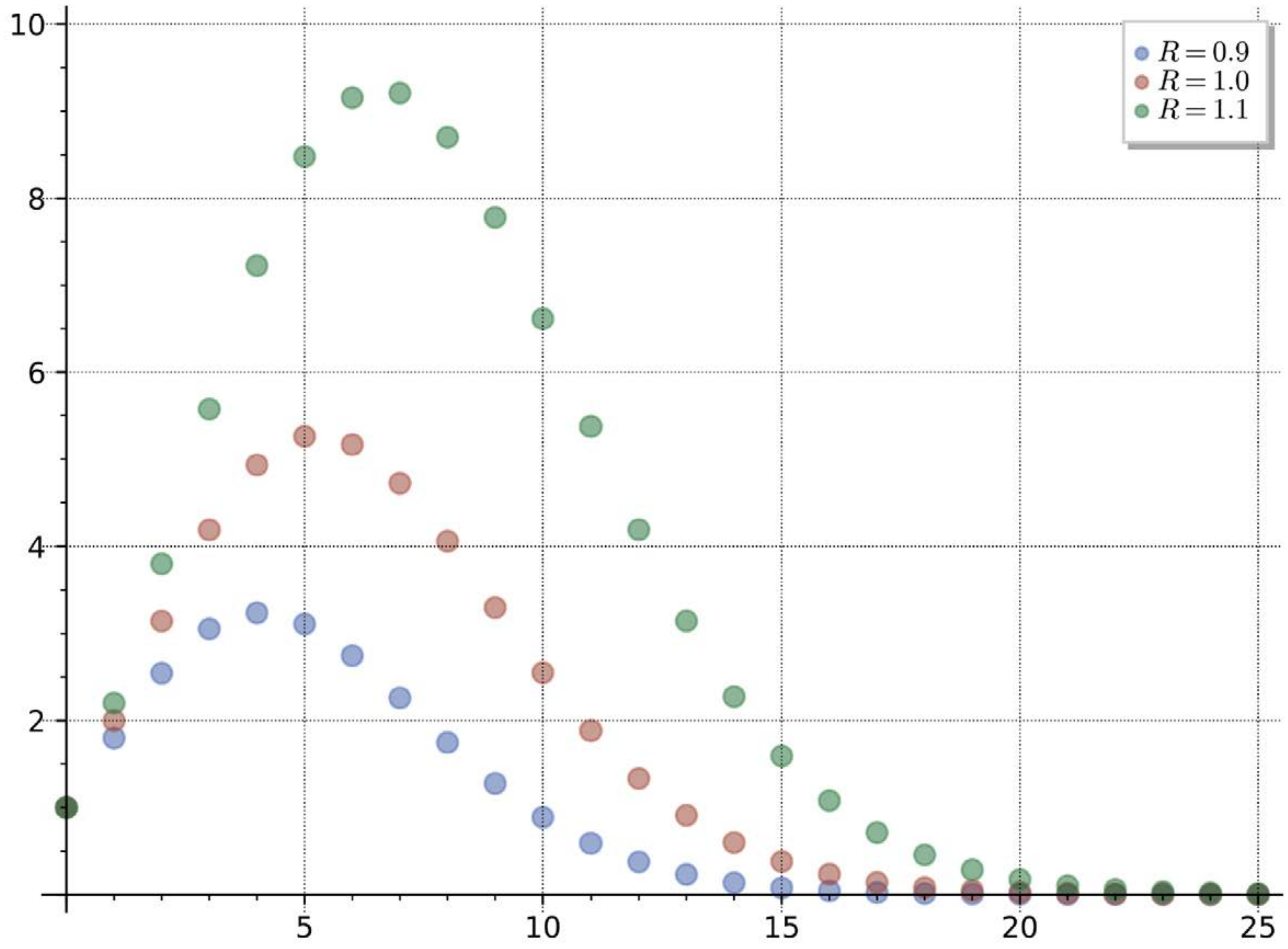
- 同样的样本点，在越高维度越显得稀疏
- 等距划分，随着维度增加而指数上升

2.1 走向更高维度



Dimension	Volume of a ball of radius R
0	1
1	$2R$
2	$\pi R^2 \approx 3.142 \times R^2$
3	$\frac{4\pi}{3} R^3 \approx 4.189 \times R^3$
4	$\frac{\pi^2}{2} R^4 \approx 4.935 \times R^4$
5	$\frac{8\pi^2}{15} R^5 \approx 5.264 \times R^5$
6	$\frac{\pi^3}{6} R^6 \approx 5.168 \times R^6$
7	$\frac{16\pi^3}{105} R^7 \approx 4.725 \times R^7$
8	$\frac{\pi^4}{24} R^8 \approx 4.059 \times R^8$
9	$\frac{32\pi^4}{945} R^9 \approx 3.299 \times R^9$
10	$\frac{\pi^5}{120} R^{10} \approx 2.550 \times R^{10}$
11	$\frac{64\pi^5}{10395} R^{11} \approx 1.884 \times R^{11}$
12	$\frac{\pi^6}{720} R^{12} \approx 1.335 \times R^{12}$
13	$\frac{128\pi^6}{135135} R^{13} \approx 0.911 \times R^{13}$
14	$\frac{\pi^7}{5040} R^{14} \approx 0.599 \times R^{14}$
15	$\frac{256\pi^7}{2027025} R^{15} \approx 0.381 \times R^{15}$
n	$V_n(R)$

2.1 维度祝福 Blessing of Dimension

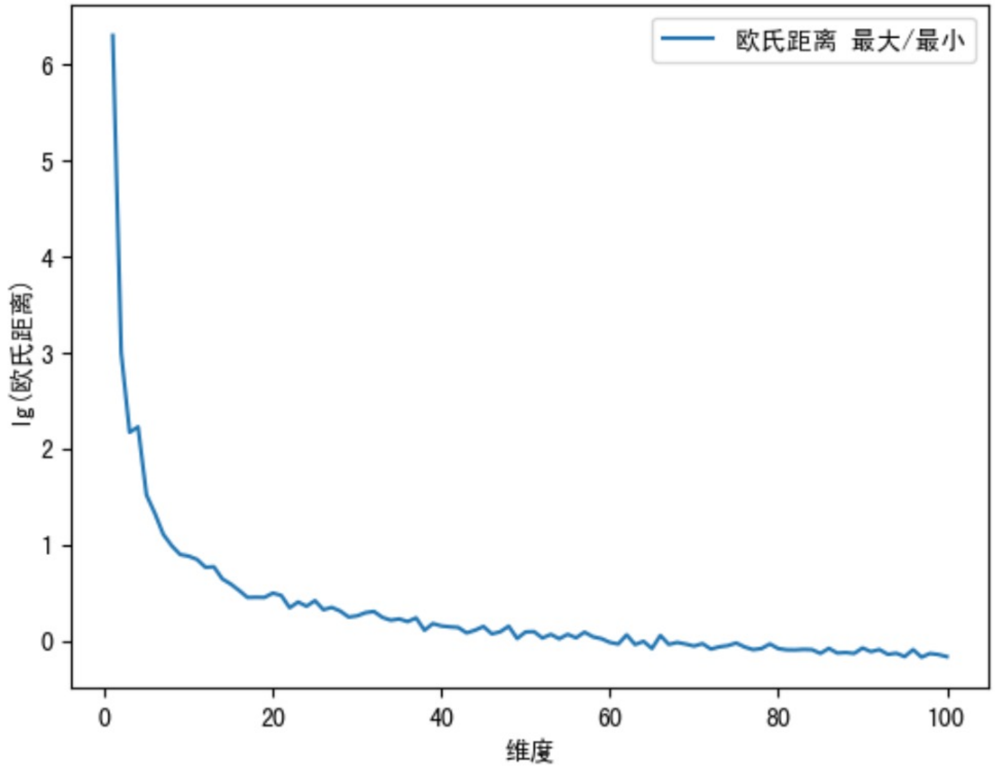


- 随着维度增加:
- 相同数量的点所占体积小
- 点与点之间越来越稀疏
- 数据容易被区分出来

2.2

故事的另外一面：维度诅咒 Curse of Dimension

- 空间中的距离度量:
- D维空间的欧几里得（欧氏）距离 $Dis_{Euc}^{D} = \sqrt{\sum_{d=1}^{D}(x_d - y_d)^2}$
- 我们尝试2-100维空间中，随机散布500个点，计算点之间的$\frac{\text{最大距离}}{\text{最小距离}}$



- 定理 Beyer et al. 1999
- 给定$\varepsilon > 0, N$。如果数据真的是高维，
- 那么基于一些很宽松的数据分布假设
- 我们可以证明
- $\lim_{D\to\infty} \Pr[d_{max}(N,D) \le (1+\varepsilon) d_{min}(N,D)] = 1$
- **高维空间中的距离失效**
- 需要特别的数据处理
- 大数定理（大样本的祝福）

2.3 带着维度的Buff，向更高维迈进!

- 我们已经有一种感觉，增加维度可以方便求解
- 如何增加维度?
  - 之前线性方法的启示：交乘变量
  - 有没有更方便的方法？（代码、存储、时间）
- 增加维度的上限是什么?
  - 二次项、三次项……这变量有意义么?
  - 数下去的上限是什么？有没有可能到无穷?
- 会不会带来很大的开销?
  - 时间复杂度？空间复杂度?
  - 如果d>N了怎么办?
  - 记得支撑向量?

$d = 1 \quad (1, x_1, x_2)$

$d = 2 \quad (x_1^2, x_2^2, x_1x_2, x_1, x_2, 1)$

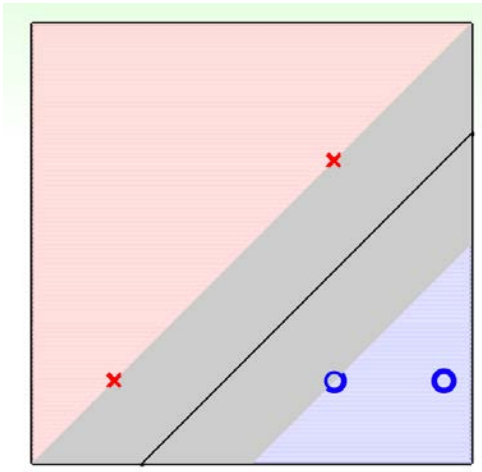
$d = 3 \quad (x_1^3, x_2^3, x_1^2x_2, x_1x_2^2 \; x_1^2, x_2^2, x_1x_2, x_1, x_2, 1)$

2.3 核方法

- 核方法：kernel　是指一种映射方法，可以高效地增加维度，合理开销内求解
- 但不妨保持一种误解：这个确实是核弹级的性质
  - 核弹的能力、核弹的重要性、核弹的研发难度、原子弹的过时
- 应用很方便，只需要指定几个超参数，就可以自动高效执行。

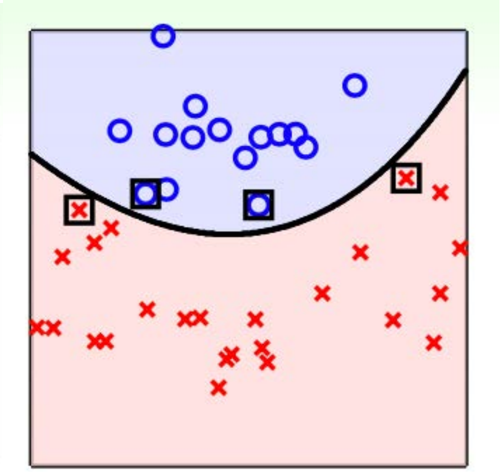
线性核

$K(X,X') = X^T X'$



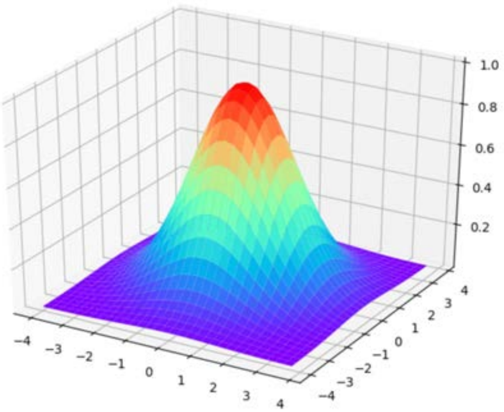
多项式核

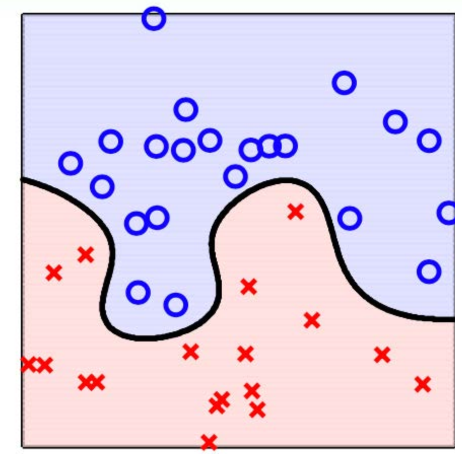
$K(X,X') = (\xi + \gamma X^T X')^Q$



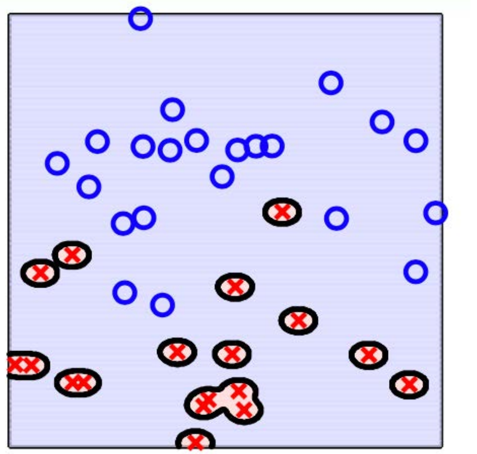
高斯核 RBF

$K(X,X') = \exp(-\gamma \|X - X'\|^2)$





$\exp(-1\|\mathbf{x} - \mathbf{x}'\|^2)$



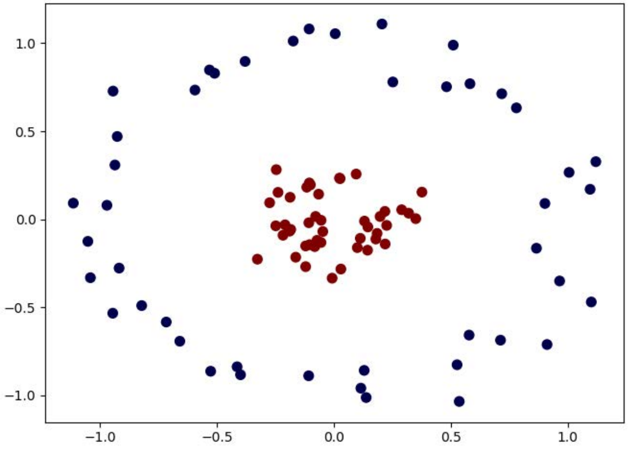
$\exp(-100\|\mathbf{x} - \mathbf{x}'\|^2)$

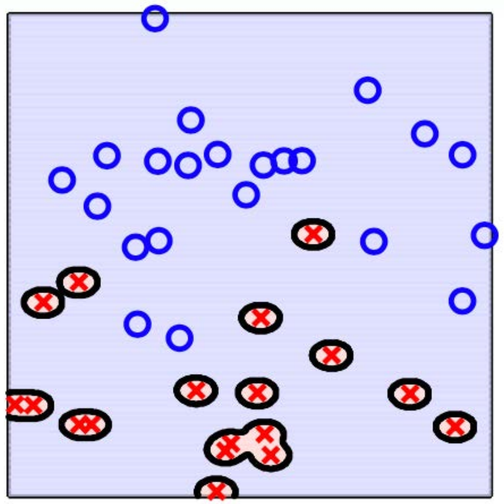
Part 03

软间隔与回归

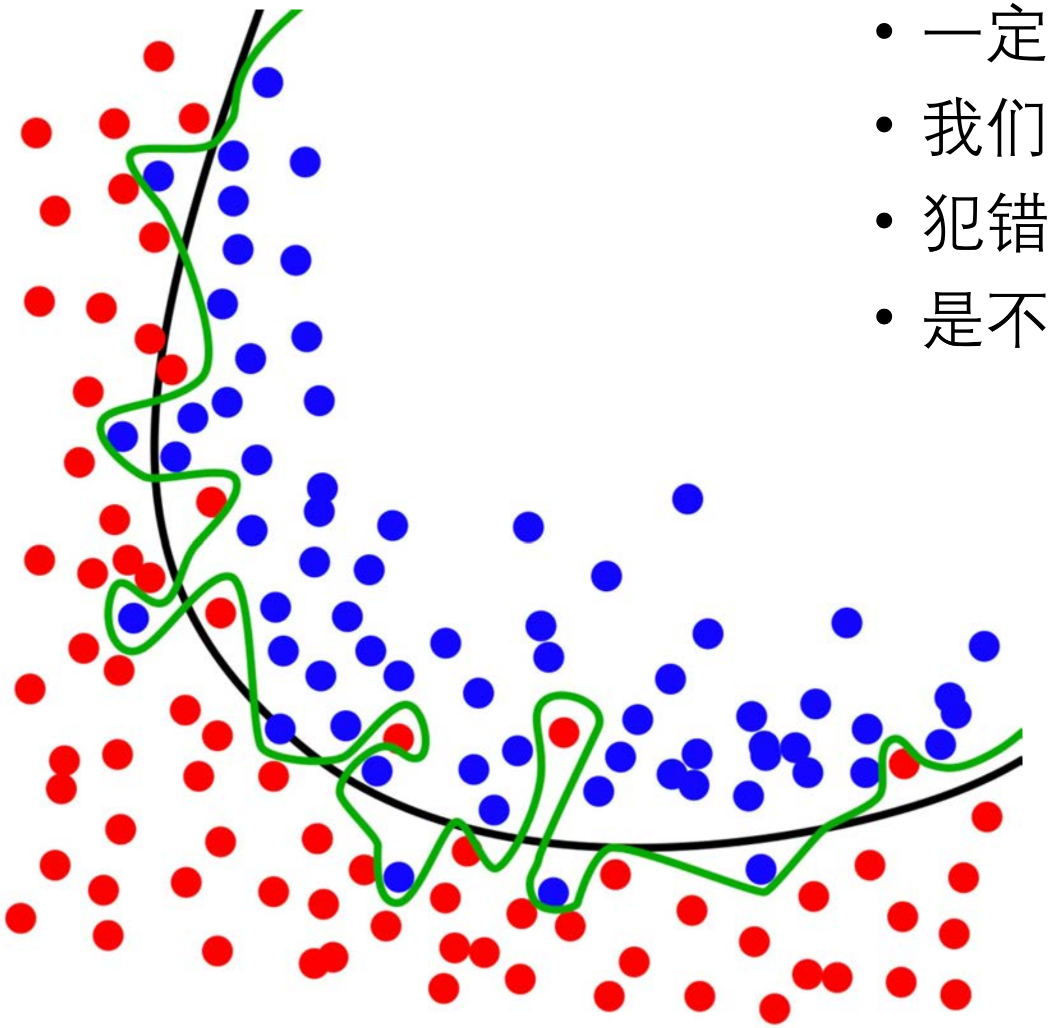
- 过拟合的威胁
- 软间隔策略
- 从分类到回归

3.1 一统山河？



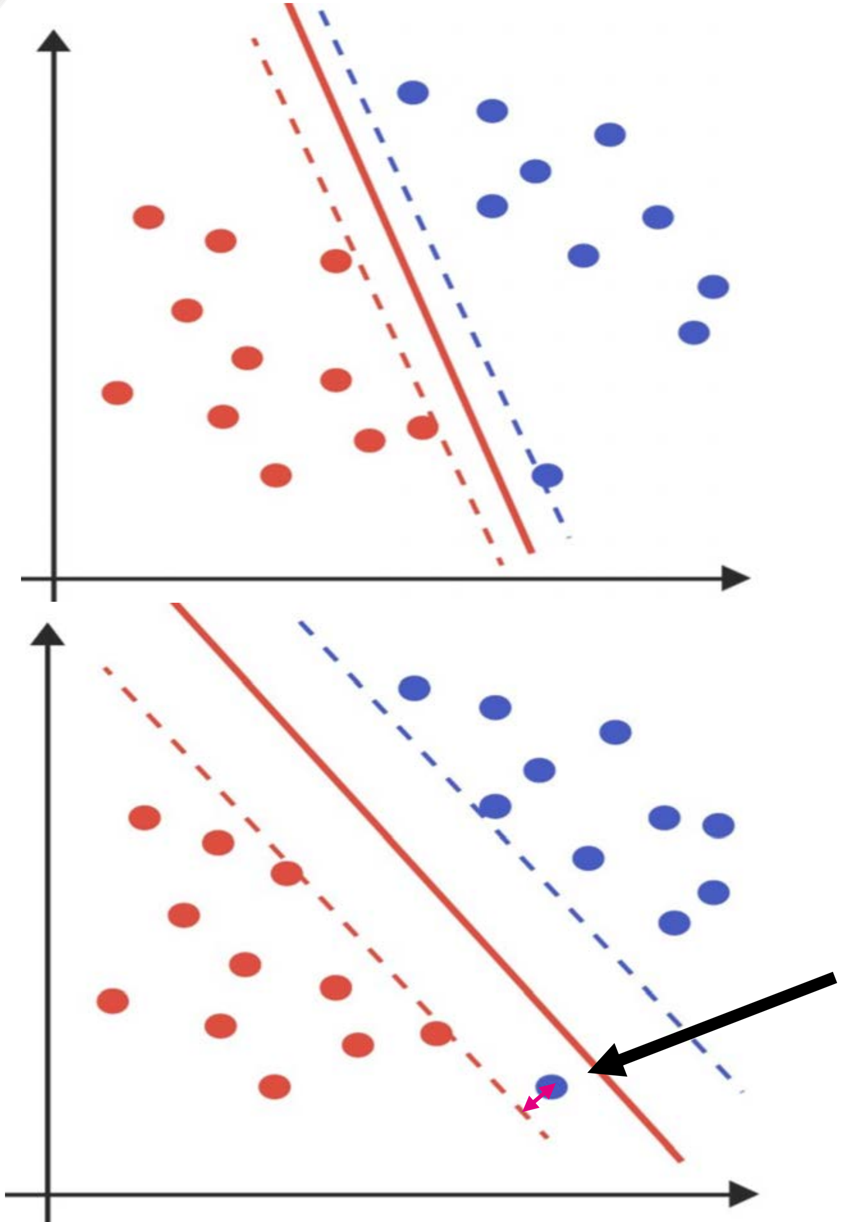


$\exp(-100\|\mathbf{x}-\mathbf{x}'\|^2)$



- 一定要不犯错么？
- 我们为什么介意犯错？
- 犯错到底意味着什么？
- 是不是该宽容一些？

3.2 如何把模型改的宽容些？



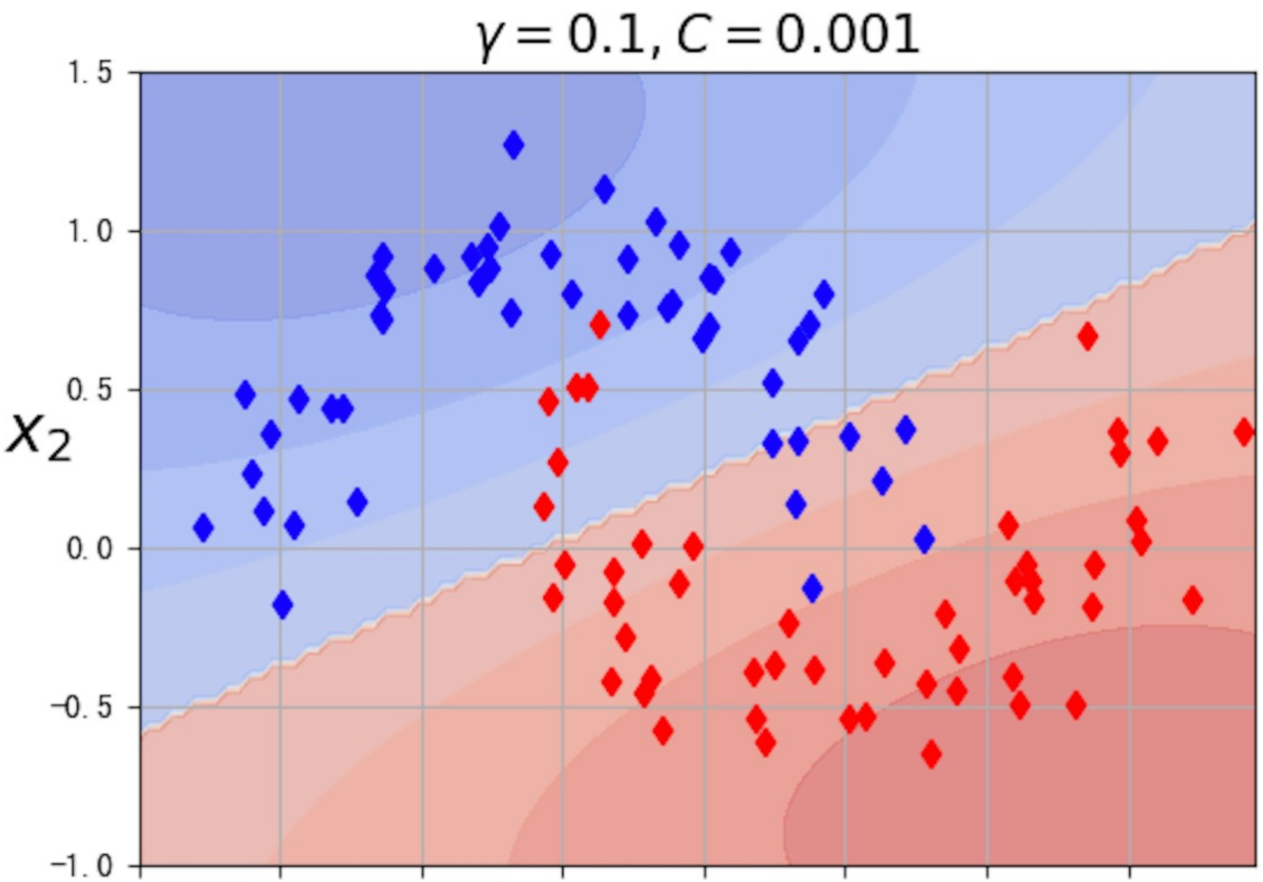
- 之前的问题
  - 最大化 分割间隔 $\max(Gap)$
  - 限制：所有变量做对
- 现在的问题
  - 最大化 分割间隔
  - 直接去掉限制？
- 交给市场解决
  - 从数量管理到价格管理，从租到税

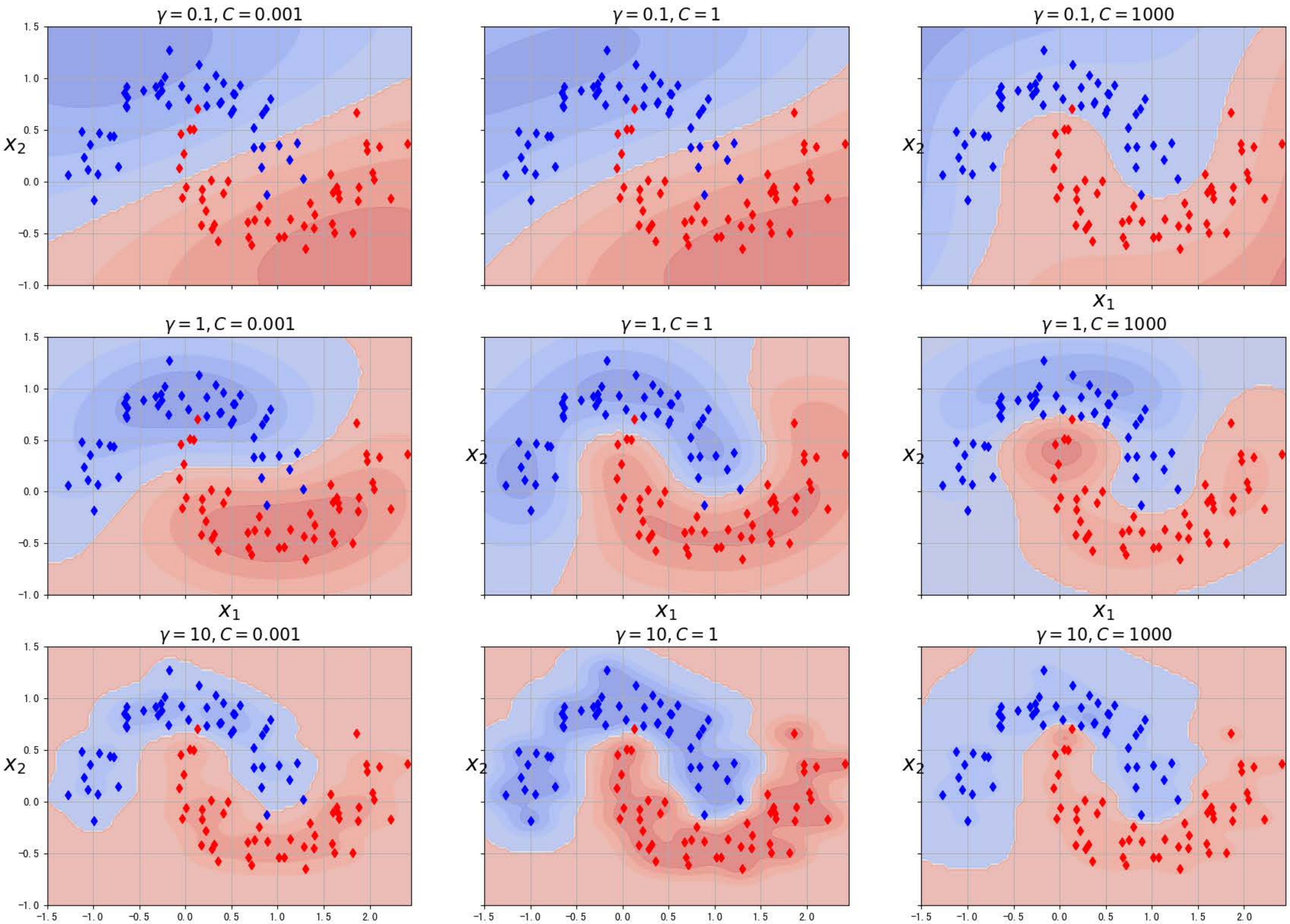
$$\min\left(\frac{1}{Gap} + C * distance\ if\ violate\right)$$

3.2 价格C的意义

$$\min\left(\frac{1}{Gap} + C * distance\ if\ violate\right)$$

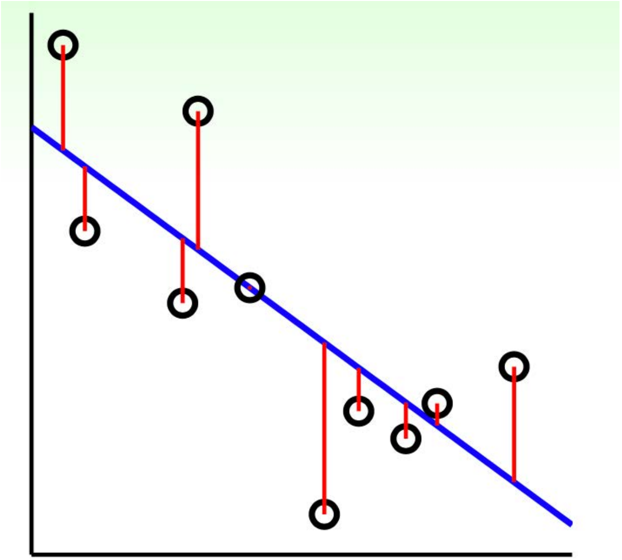
- Gap 越大， 违反的点可能越多
- Gap越大， distance也会更大
- C实际上是在权衡宽边界和违反
- C越大， 则我们越在意犯错， 模型可能越复杂来拟合数据
- C越小， 越在意Gap， 模型可能越稳健来扩大Gap



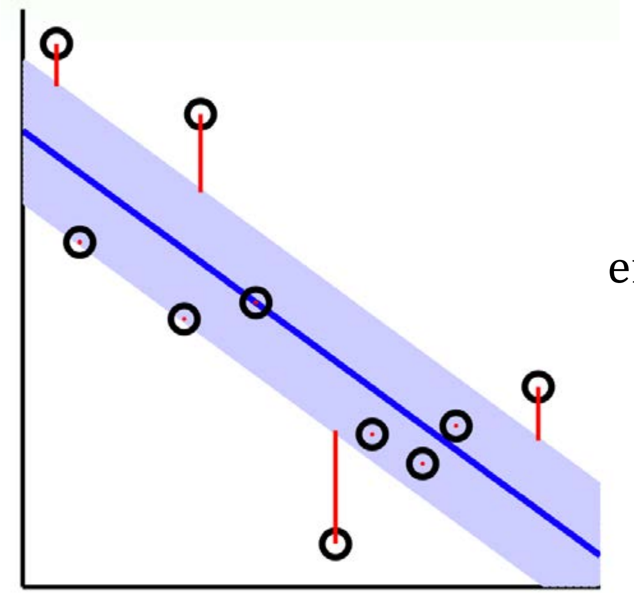


$\gamma = 0.1, C = 0.001$
$\gamma = 0.1, C = 1$
$\gamma = 0.1, C = 1000$
$\gamma = 1, C = 0.001$
$\gamma = 1, C = 1$
$\gamma = 1, C = 1000$
$\gamma = 10, C = 0.001$
$\gamma = 10, C = 1$
$\gamma = 10, C = 1000$
x_1
x_2

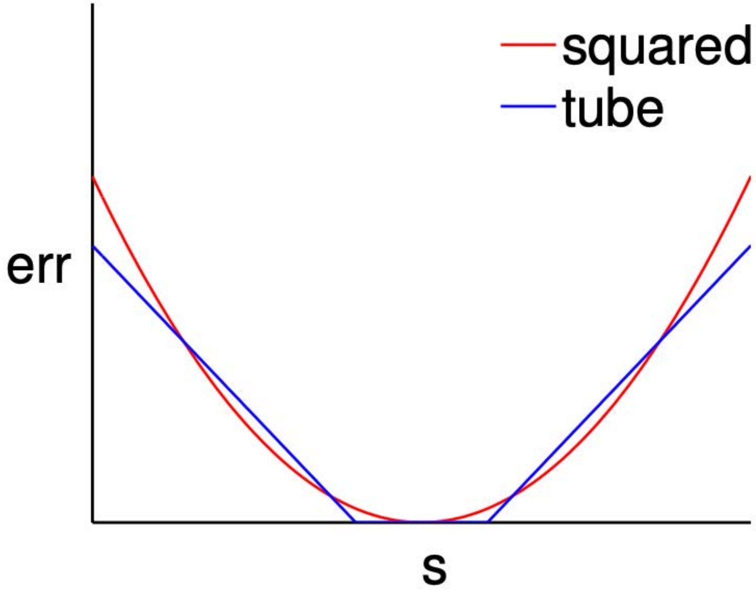
3.3 可否用到回归?



$$\text{error} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

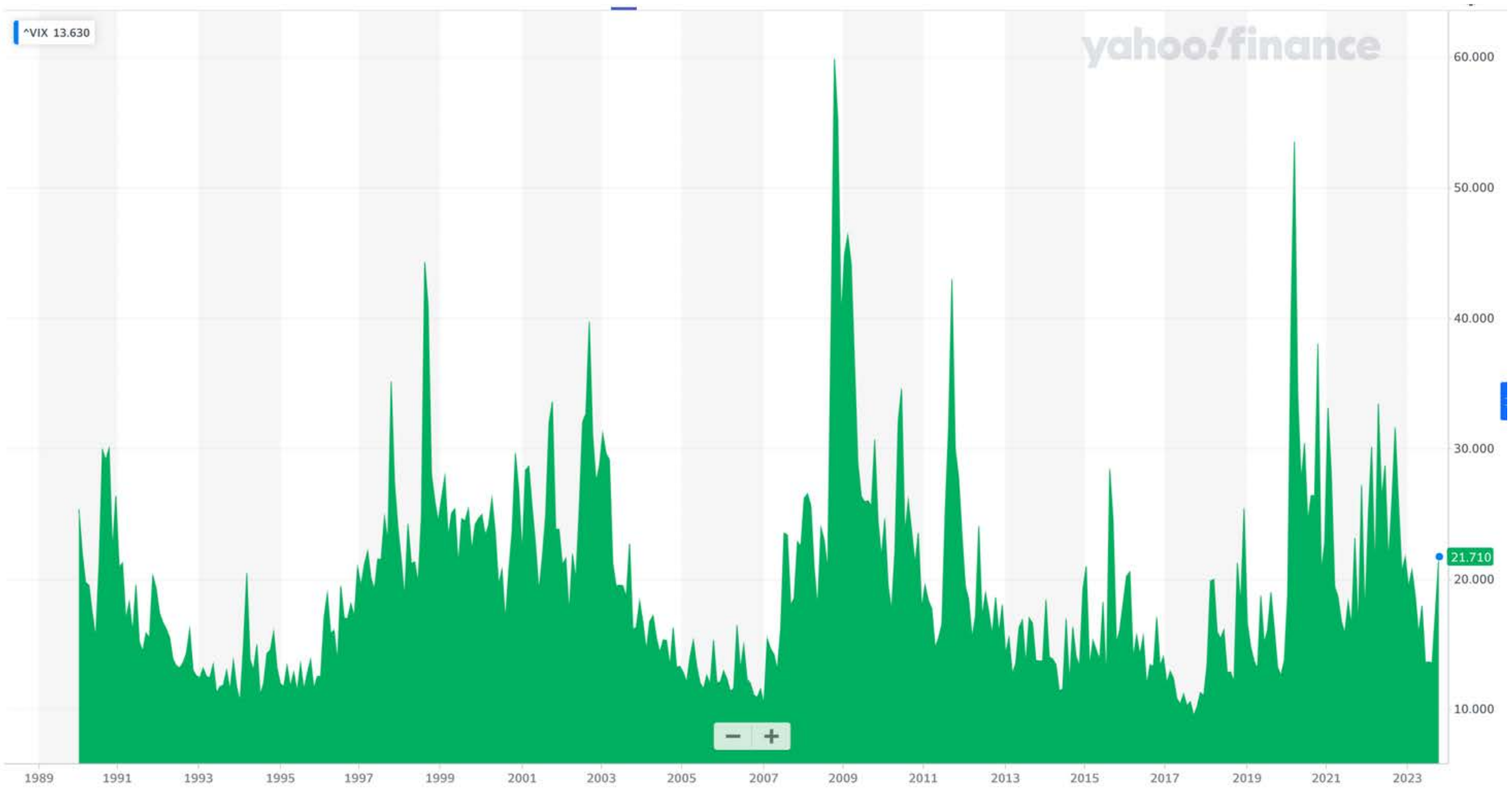


$$\text{error} = \sum_{i=1}^{N}\max(0, |y_i - \hat{y}_i| - \boldsymbol{\varepsilon})$$



- L1惩罚带来的好处
  - 稀疏性、稳健性、b>N
- SVM的好处
  - 数学解、核方法
- 在高维度、稀疏数据、复杂问题非常好用
- 一套超参 **C** 而可以不是$\boldsymbol{\varepsilon}$

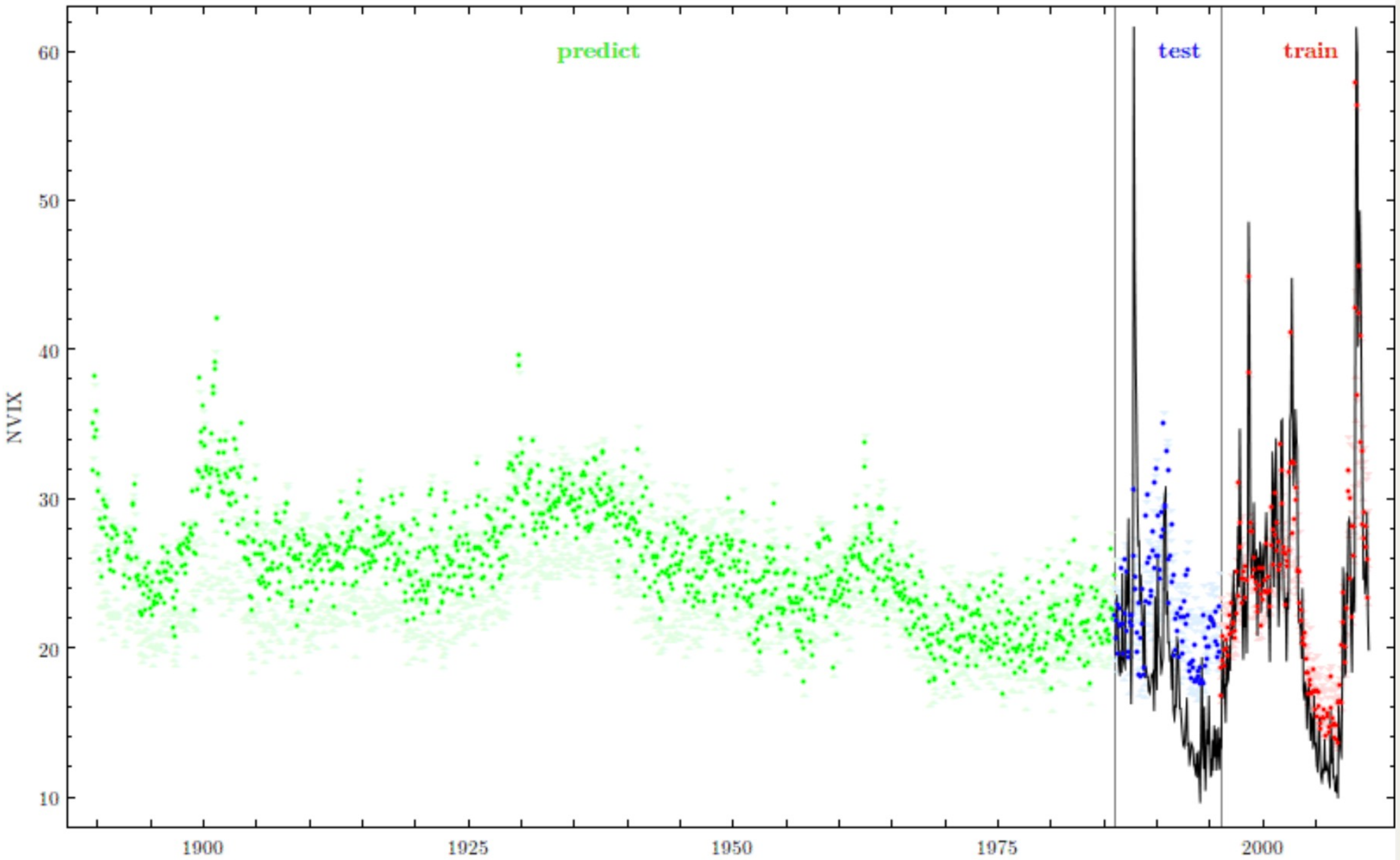
3.3 SVR的应用



- VIX恐慌指数
- 基于期权构建
- 学界业界好评如潮
- 1990年开始发布
- 不可回溯

3.3 SVR应用：基于金融刊物构建并回溯VIX

Manela, Asaf, and Alan Moreira. "News implied volatility and disaster concerns." *Journal of Financial Economics* 123.1 (2017): 137-162.



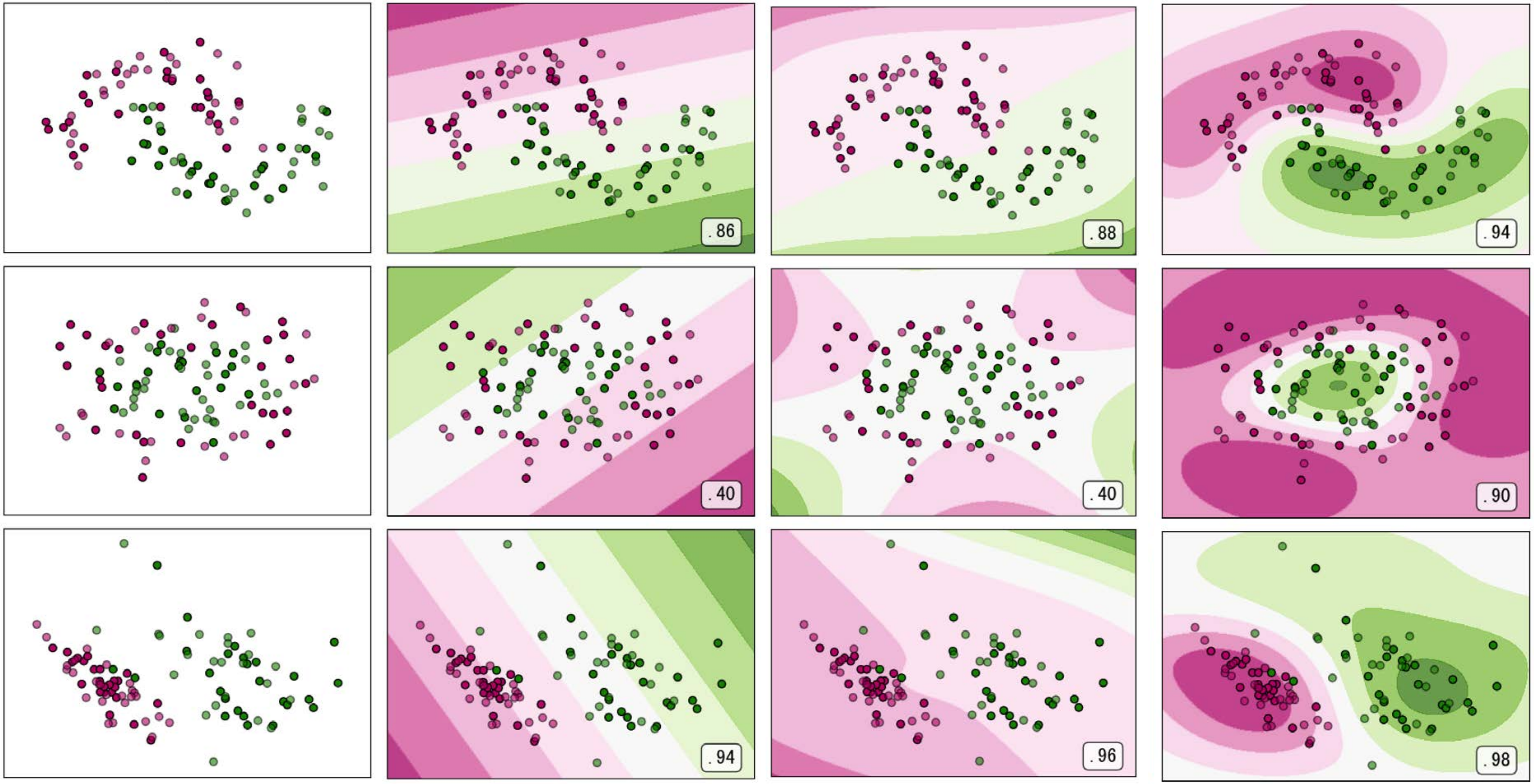
- 作者使用五本金融刊物的标题与摘要的文本进行拟合
- 用one-hot编码表示每个词
- 用一个向量维护，向量维数等于词典的规模，约2w
- 每天的文本变成当天所有词向量的加和，也是2w维，大于数据天数（b>N）
- 使用SVR，y为VIX
- 进行train、test，并前溯
- 有良好的指标意义
- 中国版本？

Part 04

模型选择

- 不同模型设定的表现
- 没有免费的午餐
- 奥卡姆剃刀
- 实操建议
- 深度学习时代的SVM

4.1 不同数据分布下不同模型的表现



4.2

没有免费的午餐 No Free Lunch

- 没有免费的午餐 定理
- 如果我们不对特征空间（数据分布）有先验的假设，则所有算法的平均表现是一样的。

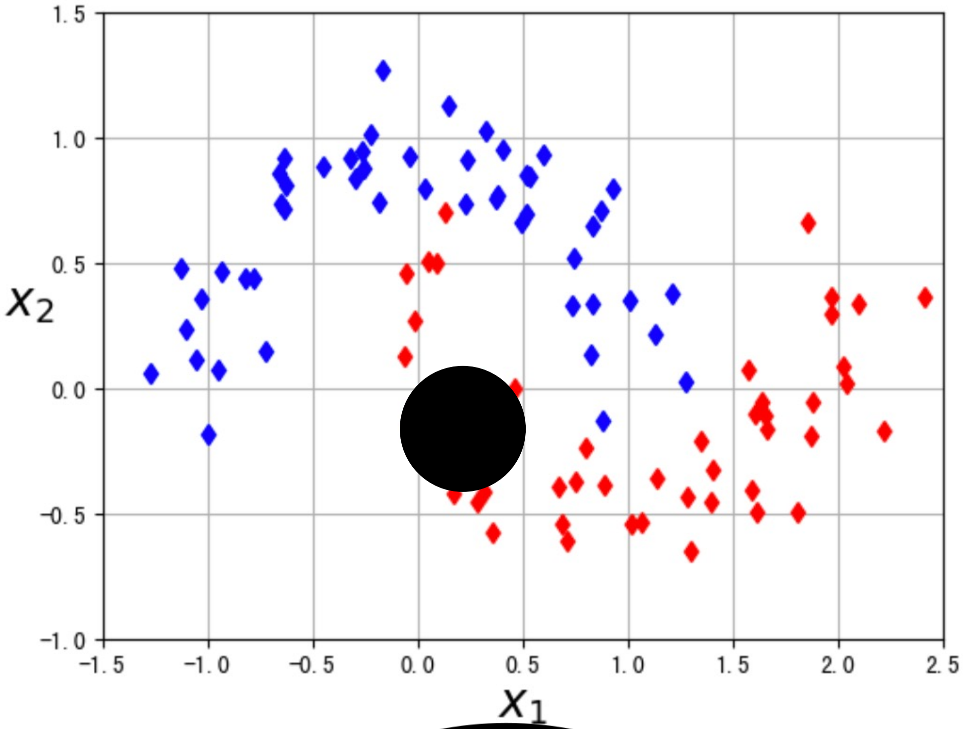
未知目标函数

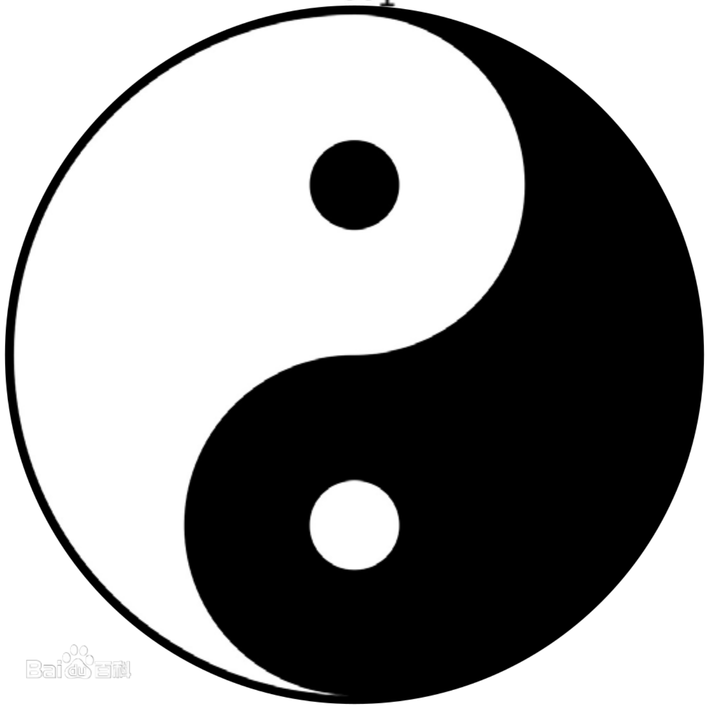
$f: \mathcal{X} \rightarrow \mathcal{Y}$

↓

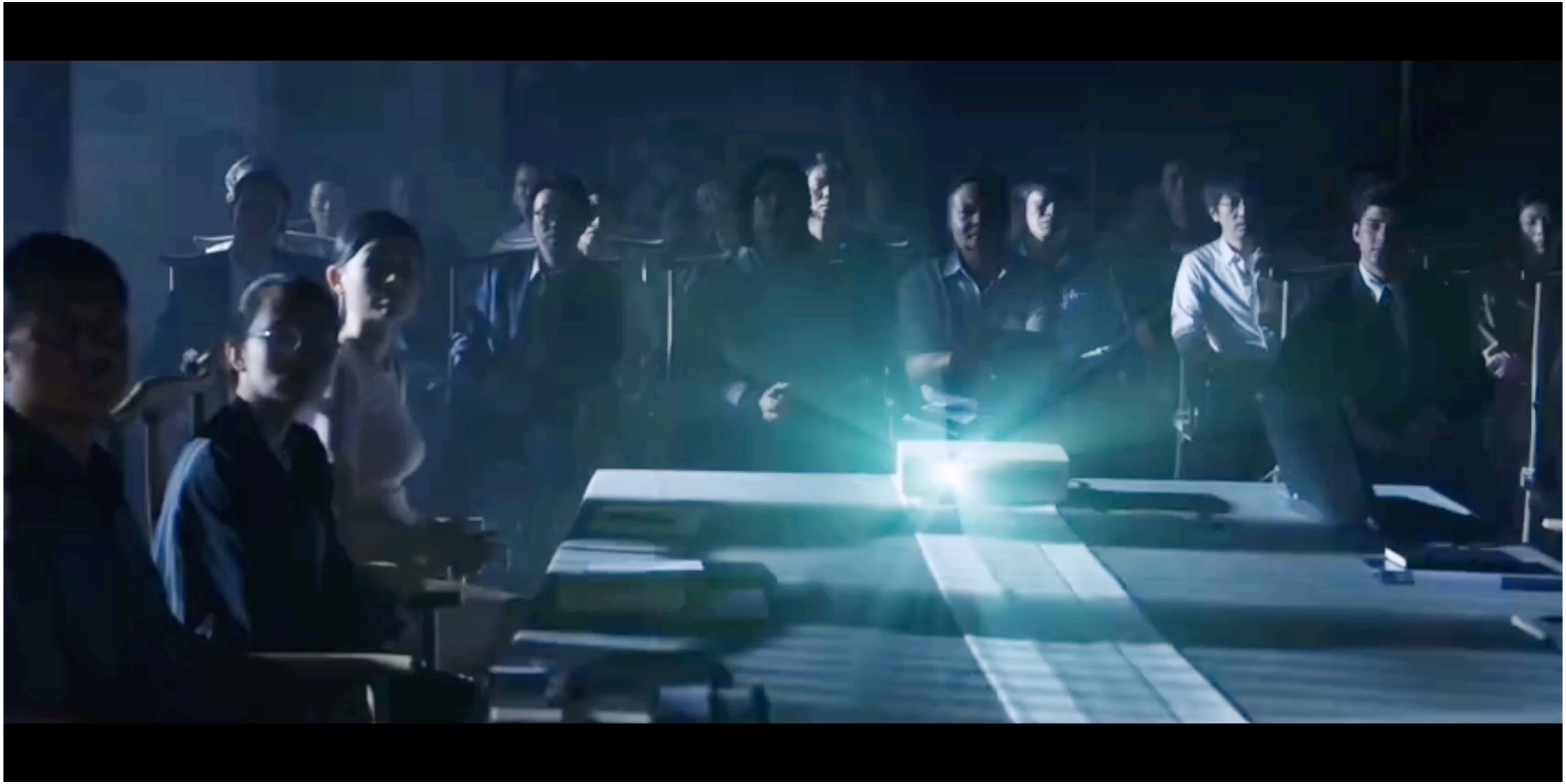
部分数据=训练集

$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$





4.2 你的人生中发生过重大变故么?



4.3 六便士的选择：买一把剃刀

- **奥卡姆剃刀原则**
- 如无必要，勿增实体
- 如果能用A说明一件事儿，就不要去用B再解释
- 如果多个方法都可以达到相似的结果，那么最简单的最好
- 我们永远先从最简单的模型开始尝试

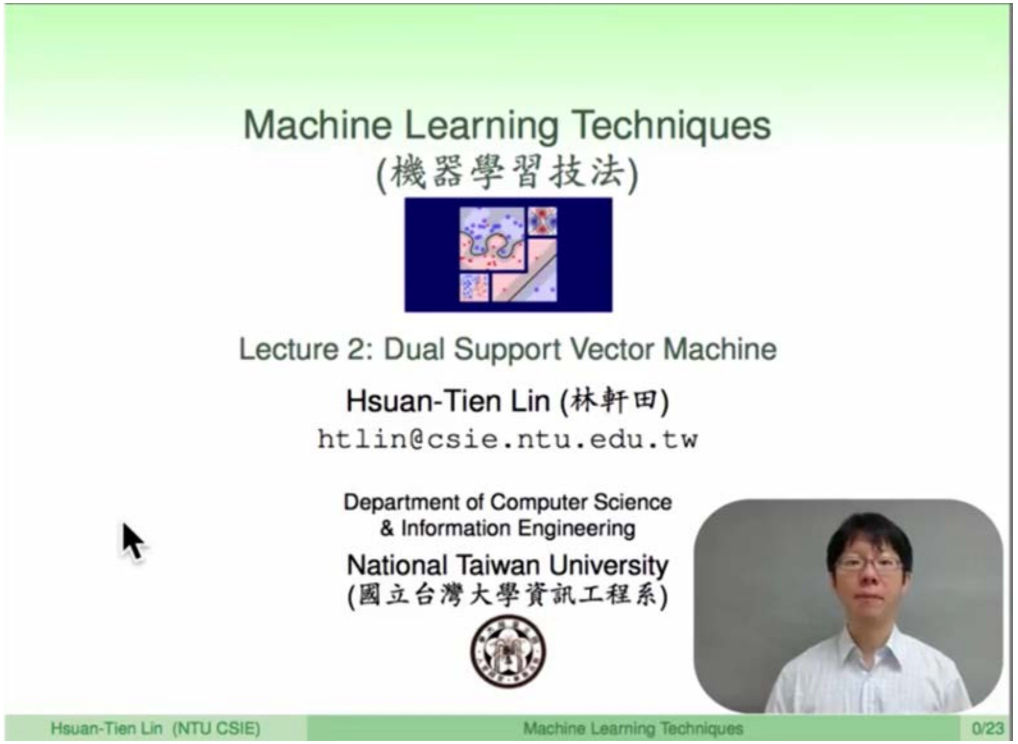
4.4 SVM实操建议

- 核相关
  - 线性核、二三维多项式核、高斯核
- 适用数据
  - 高维空间，b>n依然work
  - 样本量不是特别大
  - 不需要预测的置信区间
- 编程相关
  - 使用模型之前先将数据进行标准化，不同变量同规模
  - 注意linearSVC的参数dual
  - 注意缓存大小，cache_size = 200MB
  - C的尝试空间（超参搜索）
  - 样本不平衡问题

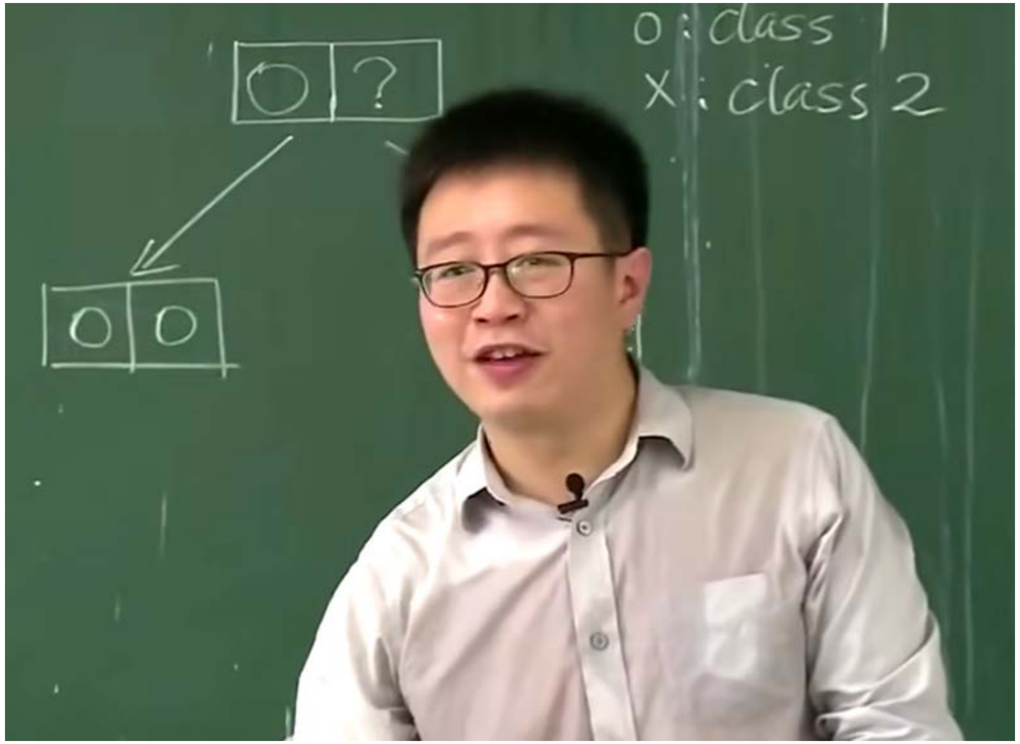
4.5 进一步学习建议

- 本节课丧心病狂地忽略了几乎所有的**数学细节**
- 然而SVM可能是最具有数学美感的常见机器学习方法

台大林轩田教授



浙大胡浩基教授



4.5 深度学习时代的SVM

- SVM = **S**uper **V**ainglorious **M**ath？
- SVM不只是强在漂亮的数学支撑，更是在图像、文本上面的（曾经）统治力
  - 高维空间、稀疏feature、b>N，确实和图像、文本非结构化数据是天造地设一对
  - 辉煌时刻谁都有 别拿一刻当永久
- 数学特性在面试上面是个很好的Vainglorious
- 对于经济金融来说呢?
  - 维度、feature-> 真的高维，真的好度量，真的无需feature提取
  - 我们对于数学支撑、随机性低、白盒模型的追求
- SVM的未来?

2023

THANKS