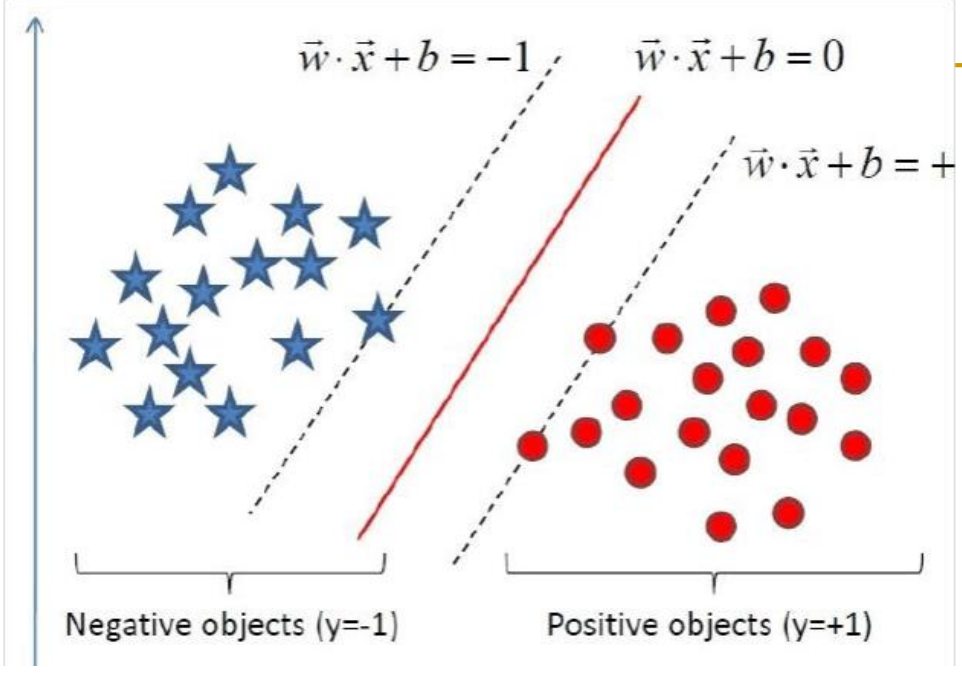
《机器学习及其在化学中的应用》2025年课程

Sec. 6

核方法

（近邻法；支持向量机）



刘志荣 (LiuZhiRong@pku.edu.cn)

北京大学化学学院

2025.10.27

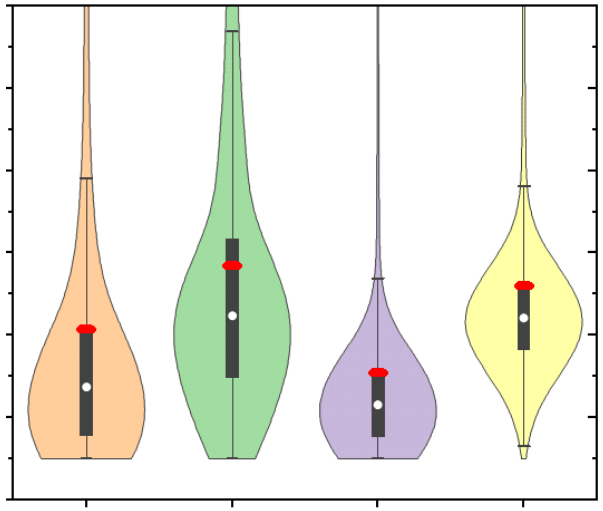
内容提要

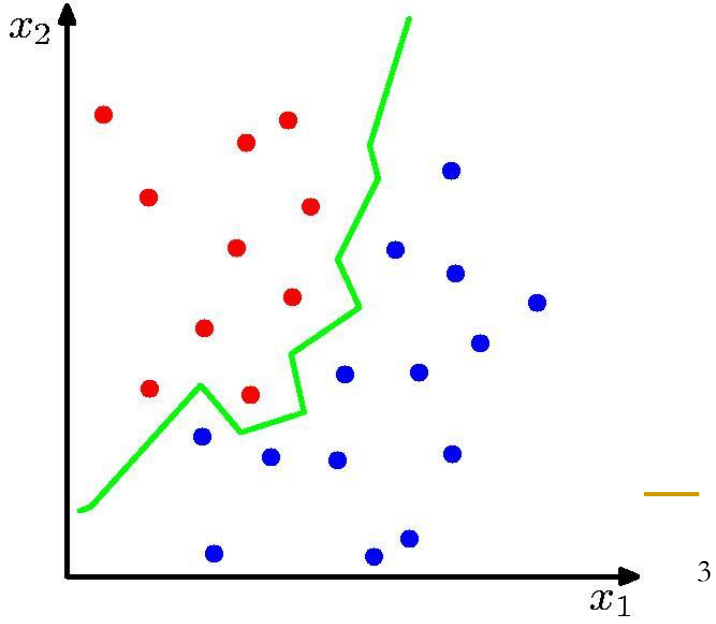
- 密度估计的非参数法
  - 直方图方法
  - 核密度估计法
  - 近邻法
- 核方法：主要想法
- 支持向量机
- 应用例子

1. 密度估计的非参数法...

（nonparametric approaches to density estimation）

- 直方图方法
- 核密度估计法
- 近邻法



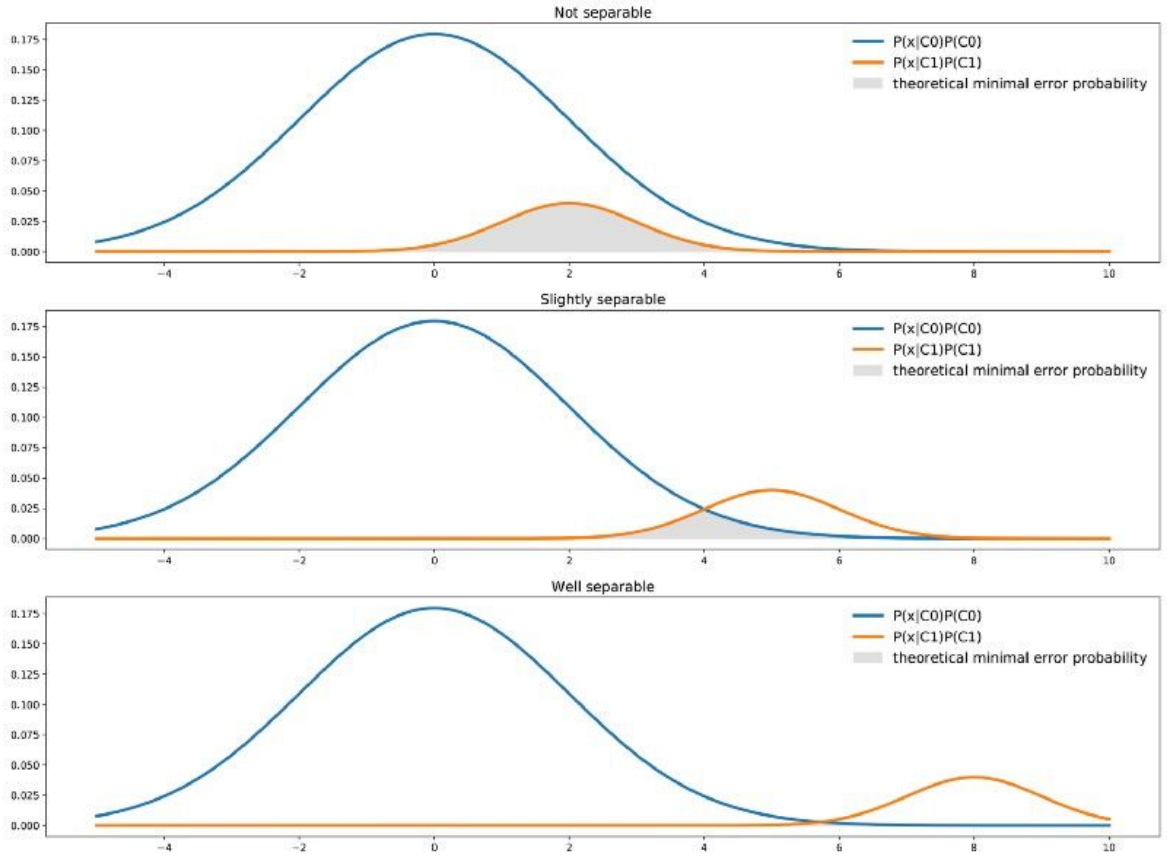


非参数估计

$t(\mathbf{x}) = f(\mathbf{x}) + \epsilon$ vs. $p(\mathbf{x}|C)$

- 分类器具有理论意义上的最小误差概率，即由两条类别样本分布曲线最小值下的面积给出。
- 如果能从已知数据推断出每个分类的分布曲线（属于无监督学习），就可据此进行预测。
- 参数法
  - 假定曲线的公式已知（带参数），如高斯分布。应用最大似然法。
- 非参数法：

  不需假设曲线形状。
  - 直方图方法与核密度估计法。
  - 近邻法。



直方图方法

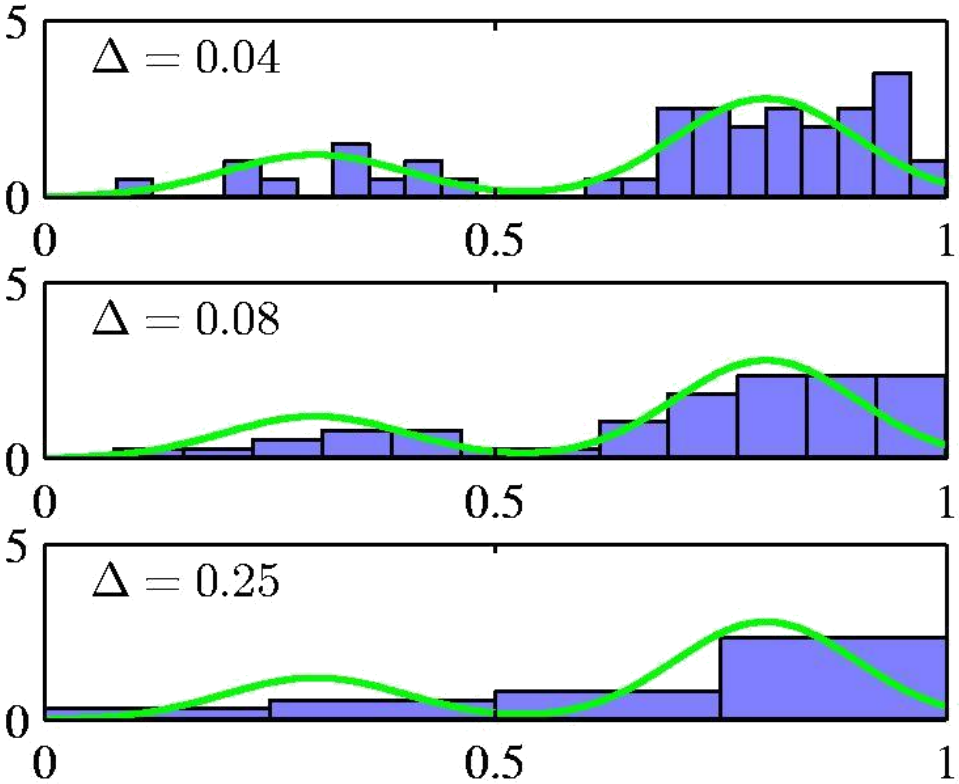
- 将 x 轴分成一些隔间（bin），宽度Δ_i，并计算落在其中的数据个数n_i。概率密度估计为

$$p_i = \frac{n_i}{N\Delta_i}$$

- 通常取$\Delta_i = \Delta$（模型的唯一（超）参数）。
- 缺点：
  - 高维空间下面临维度灾难，

$$\propto M^D$$

  - 结果曲线不光滑。



- 把$\mathbf{x}$理解成随机变量，一个小区域$\mathcal{R}$内的概率：

$$P = \int_{\mathcal{R}} p(\mathbf{x})d\mathbf{x}$$

- 则N个点中有K个点落入区域$\mathcal{R}$的概率服从二项式分布：

$$p(K|N,P) = C_N^K P^K (1-P)^{N-K}$$

- 如果N与K都很大，则分布很窄（大数定律），$K \approx NP$
- 如果区域$\mathcal{R}$体积（V）很小，则其内部$p(\mathbf{x})$近似是个常数，$P \approx p(\mathbf{x})V$
- 因此

$$p(\mathbf{x}) \approx \frac{K}{NV}$$

- 区间$\mathcal{R}$选择：需较小以便其中的$p(\mathbf{x})$基本是个常数，又需较大以使K较大以便获得好的统计特性。

核密度估计法

$$p(\mathbf{x}) \approx \frac{K}{NV}$$

- 固定区域$\mathcal{R}$的体积V，从已知数据中估计K
- $\mathcal{R}$为中心处于$\mathbf{x}$、边长为h的小立方块，定义核函数（Parzen window）

$$k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) = \begin{cases} 1, & \text{if } \left|\frac{x_i - x_{n,i}}{h}\right| < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

则 $K = \sum_n k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$

$$p(\mathbf{x}) = \frac{1}{N}\sum_n \frac{1}{h^D} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$

可以不把公式理解成落到以$\mathbf{x}$为中心的cube有多少数据点，而是$\mathbf{x}$落到多少个以数据点$\mathbf{x}_n$为中心的cubes。

- 为了避免得到不连续的$p(\mathbf{x})$，可采用光滑的核函数$k(\mathbf{x}, \mathbf{x}_n)$
  - 如高斯函数

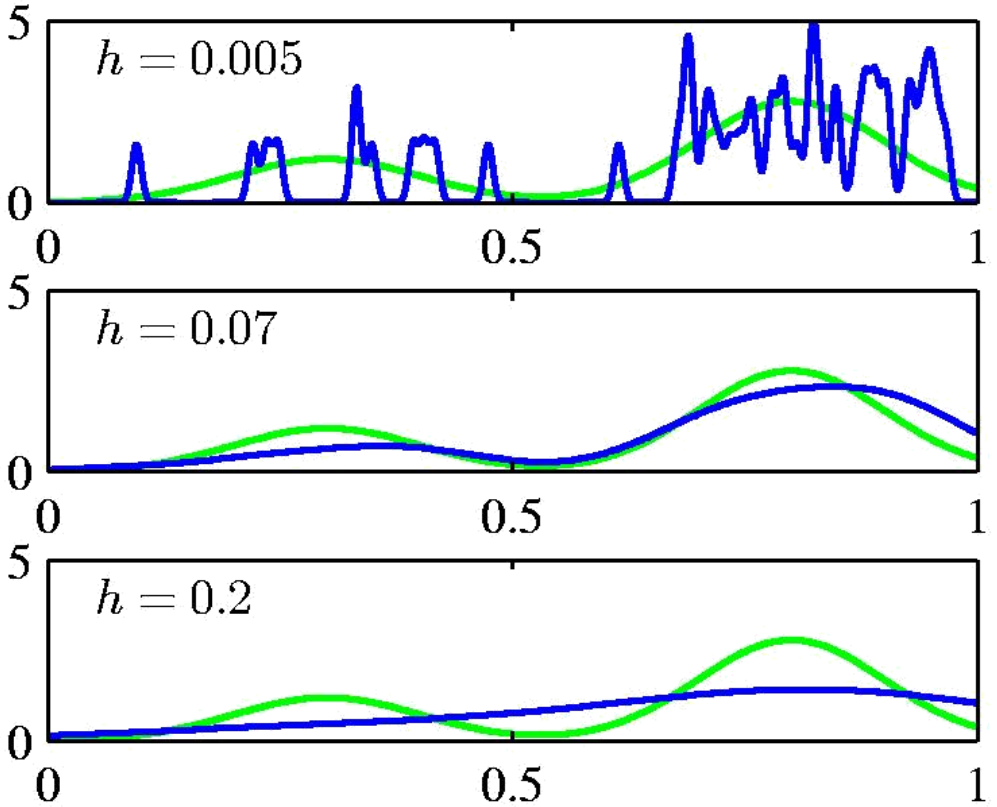
$$p(\mathbf{x}) = \frac{1}{N}\sum_n \frac{1}{(2\pi h^2)^{D/2}} \exp\left(-\frac{|\mathbf{x}-\mathbf{x}_n|^2}{2h^2}\right)$$

- 任何满足

$$k(\mathbf{u}) \geq 0, \quad \int k(\mathbf{u})d\mathbf{u} = 1$$

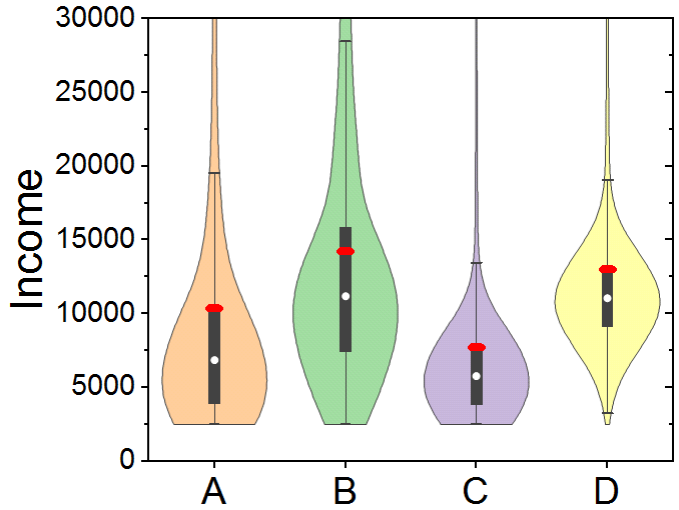
的$k(\mathbf{u})$都可以作为核函数。

h acts as a smoother.



- 题外：

画图软件给出的小提琴图



近邻法（Nearest Neighbour）

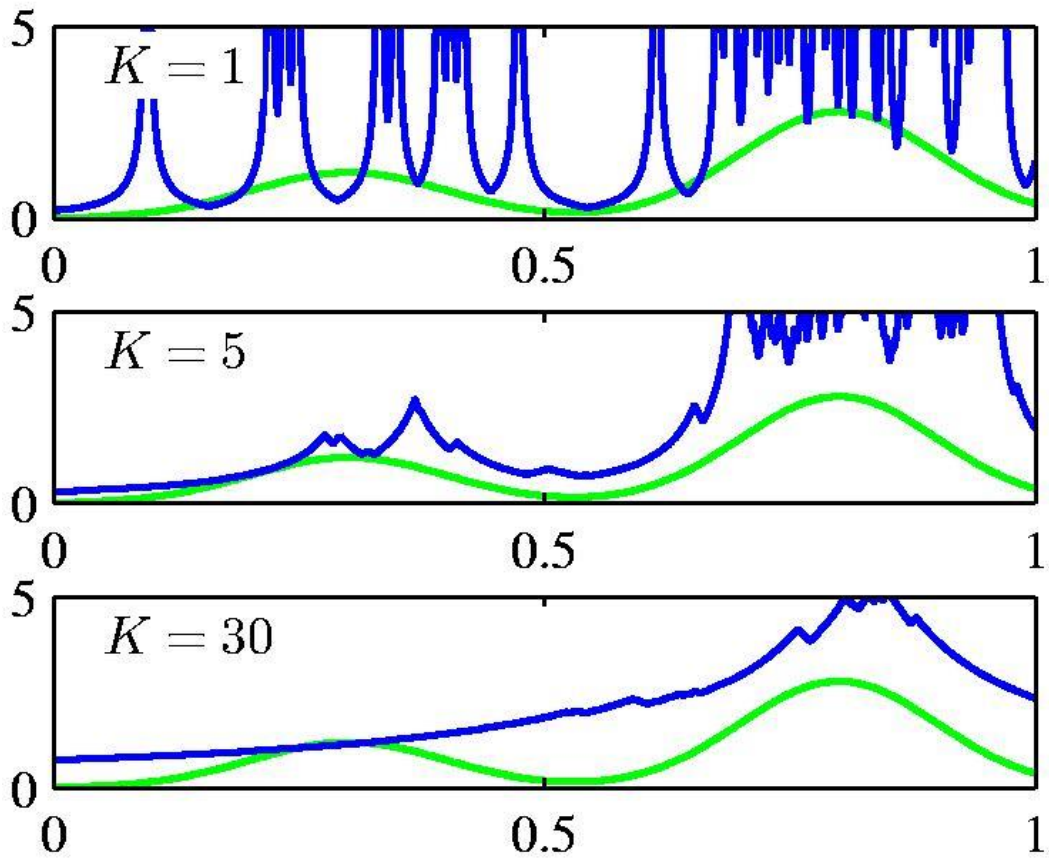
- 核方法的缺点是h与V固定，不容易兼顾密度大与密度小的地方。
- 近邻法：固定K，从已知数据中估计V

$$p(\mathbf{x}) \approx \frac{K}{NV}$$

  - 例如：不断膨胀的球，
    直到包含进K个数据点。

- K近邻算法（KNN）

- $K = 1$：最近邻算法。



在分类问题中的应用…

- 第k类的概率

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}$$

- 先验概率

$$p(\mathcal{C}_k) = \frac{N_k}{N}$$

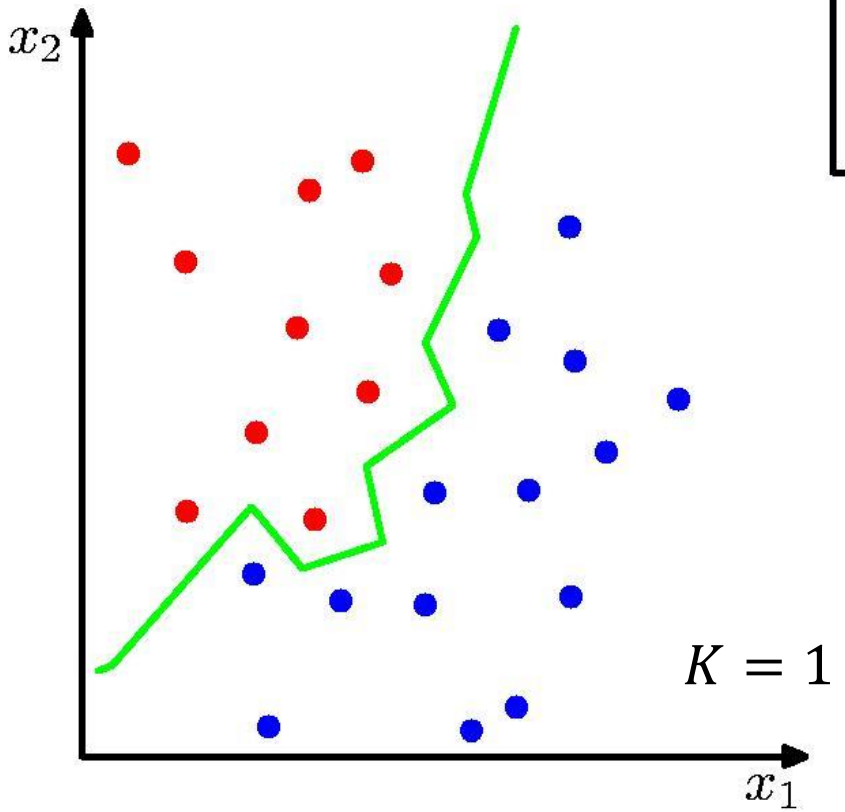
- 因此后验概率

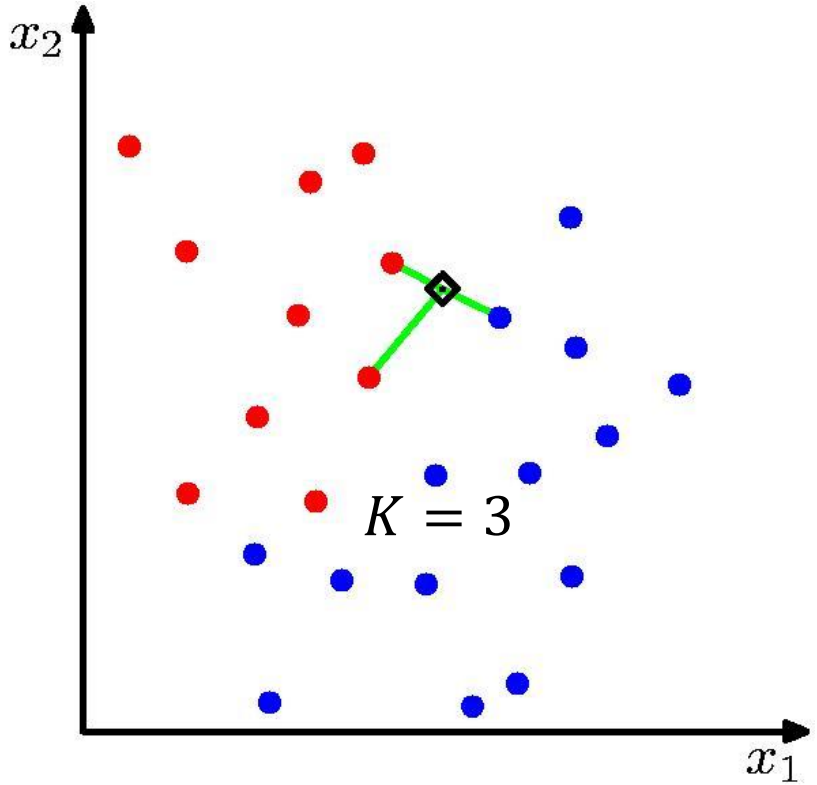
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

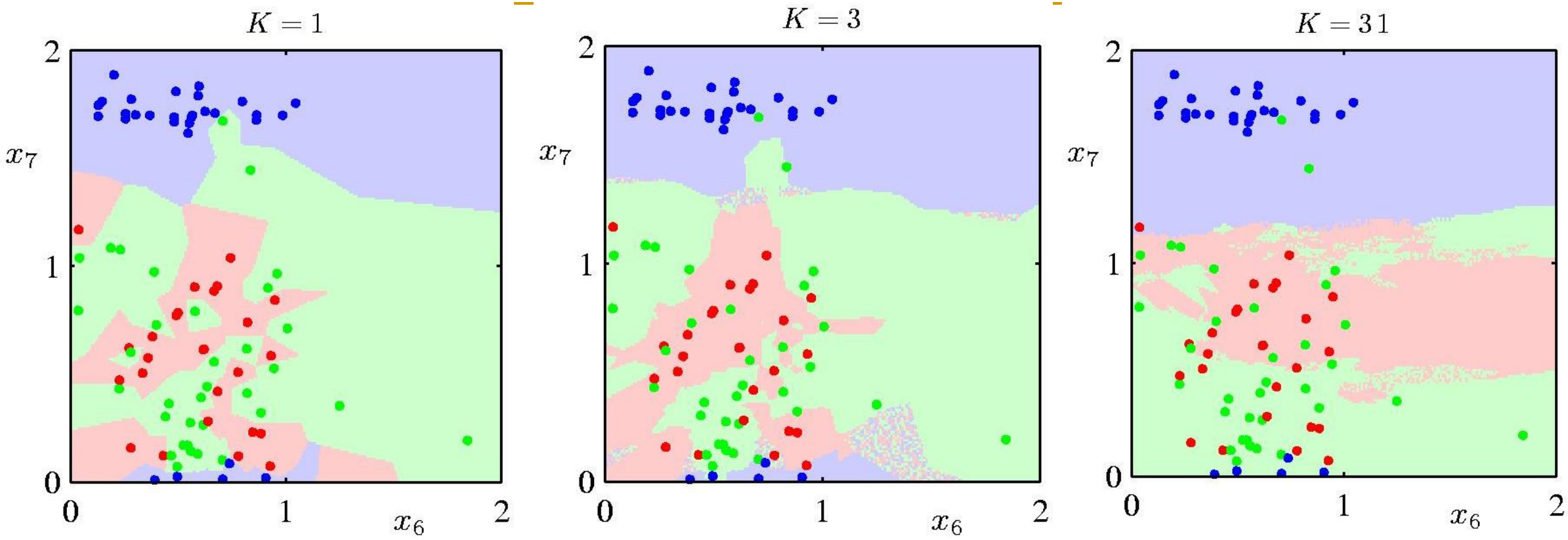
$$= \frac{\frac{K_k}{N_k V}\cdot\frac{N_k}{N}}{\frac{K}{NV}} = \frac{K_k}{K}$$

N_k：第k类的数据个数。

K_k：最近的K个数据中属于第k类的个数。







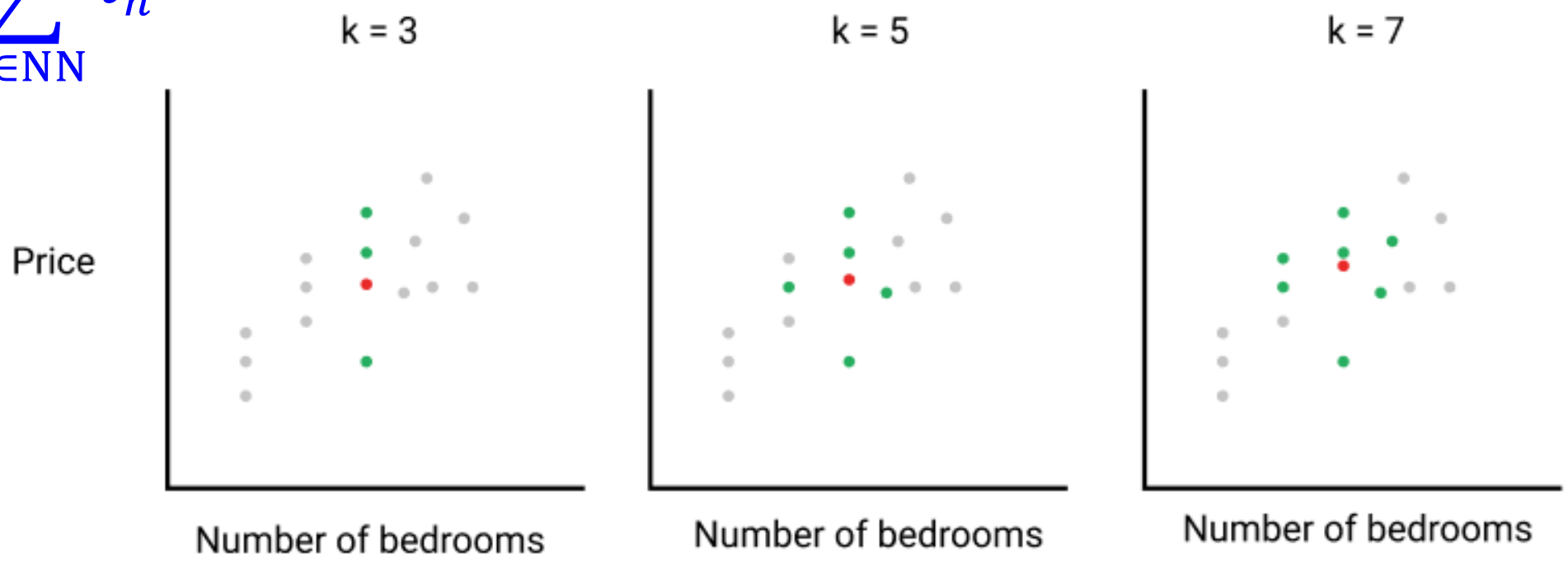
- 核心思想：像什么就是什么！
- 最近邻算法是人类有史以来发明的最简单、最快速的学习算法。
- 最近邻算法是史上第一个能够利用不限数量的数据来掌握任意复杂概念的算法。

--《终极算法》

在回归问题中的应用…

- 预测值

$$f(\mathbf{x}) = \frac{1}{K}\sum_{n\in \mathrm{NN}} t_n$$



- 例子：我的房子有3个卧室，租多少钱呢？

非参数法的优缺点

- （K是超参数，可通过交叉验证来确定。）
- 优点：不需假设曲线性质
  - 在$N \to \infty$时能拟合任何函数曲线。
- 缺点1：需保存所有数据用于预测
  - 而参数法则只需要保存训练得到的参数。
- 缺点2：需要大量数据
  - 估计的无偏性（偏差-方差分解中的方差为零），只有在局部邻域中的点很多时才能得到保障。

小结

- 非参数估计：并不假设曲线形状，从已知数据推断每个分类的分布曲线，并据此进行预测。
- 核密度估计法：固定分块区域体积，

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$

- 近邻法：固定分块区域内数据个数，$p(\mathcal{C}_k|\mathbf{x}) = \frac{N_k}{K}$

2. 核方法：主要想法
(Kernel Methods)

- 普适意义下的核方法

机器学习的工作机制基本上是相同的：

将临近的样例归类到同一个类别中。

《机器学习那些事》

核方法的主要想法

- 前述方法的核心概念是两个点之间的距离（相似性）。
  - 本质上所有的学习器都是将临近的样本归类到同一个类别中；关键的不同之处在于“临近”的意义。
- 当引入基函数$\{\phi_j(\mathbf{x})\}$后，可以基于$\phi_j(\mathbf{x})$重新定义两个$\mathbf{x}$点之间的距离。
  - 相当于先变换坐标，再在新坐标系下讨论相似性（距离）。
- 既然重要的是距离，也可以不显式讨论$\{\phi_j(\mathbf{x})\}$，而是直接引入某种形式的距离（内积）的定义，它的信息由核函数给出。
  - 很多方法都可以结合核方法，如线性回归（Bishop 6.1-6.2）

以线性回归为例…

[涉及到对偶理论，数学复杂，详见Bishop 6.1，过程略]

- 定义核函数： $k(\mathbf{x},\mathbf{x}')=\sum_{j=0}^{M}\phi_j(\mathbf{x})\phi_j(\mathbf{x}')\equiv\phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}')$
  - 在最简单的情况下，$\phi(\mathbf{x})=\mathbf{x}$，上式变成内积，称为linear kernel。
- 定义Gram矩阵，$\mathbf{K}=\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$，是$N\times N$对称矩阵，矩阵元

$$K_{mn}=\phi(\mathbf{x}_m)^{\mathrm{T}}\phi(\mathbf{x}_n)\equiv k(\mathbf{x}_m,\mathbf{x}_n)$$

- 则在正则化条件下的预测结果可写成

$$y(\mathbf{x})=\mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\mathbf{t}$$

  - 其中$\mathbf{k}(\mathbf{x})$是N维矢量：$k_n(\mathbf{x})=k(\mathbf{x},\mathbf{x}_n)$
  - 因此，结果是由表征距离性质的核函数所决定的，而且是已知函数值$\mathbf{t}$的某种（线性）组合。

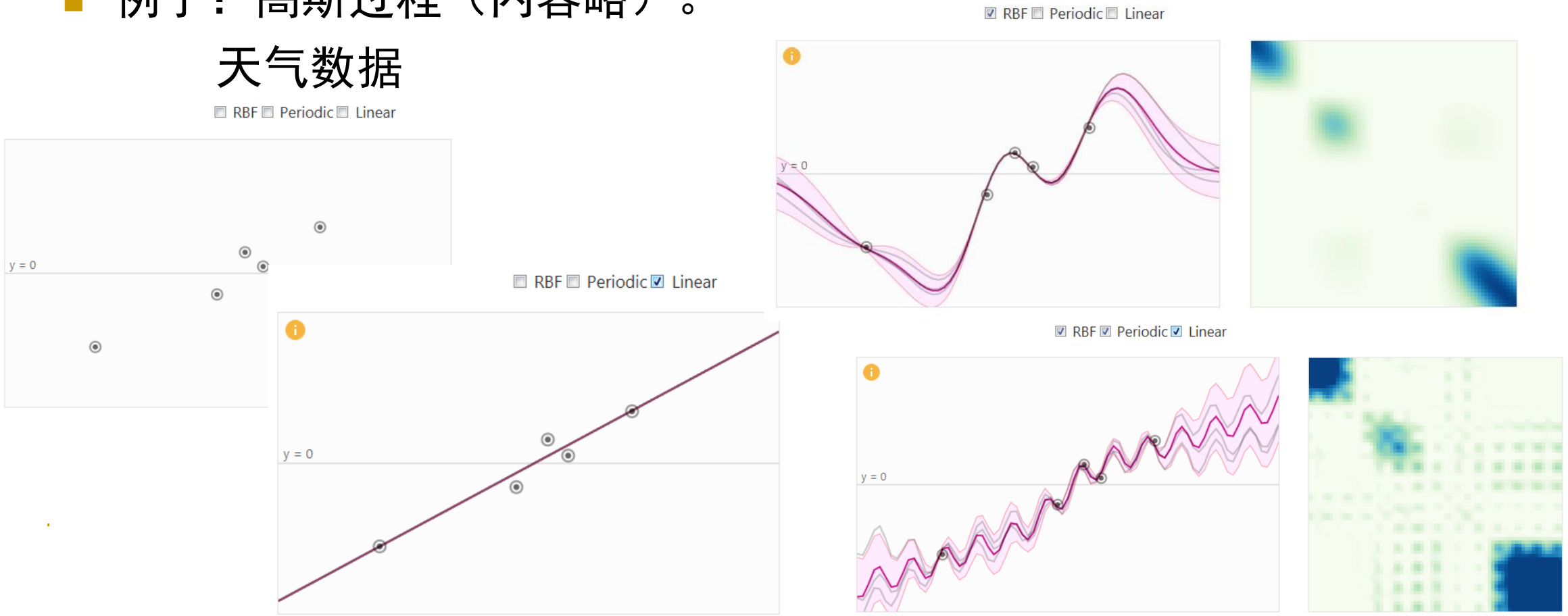
核方法的灵活性

- 核函数使得专家可以把某个领域的知识引入到模型中，以捕捉训练数据中的趋势。

https://www.jiqizhixin.com/articles/2019-02-12-3

- 例子：高斯过程（内容略）。

天气数据

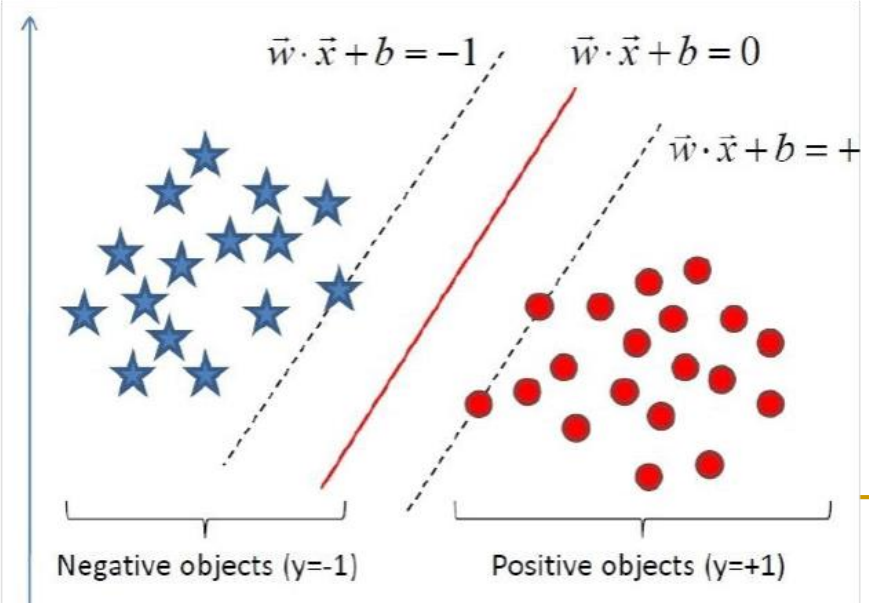


小结

- 核方法
  - 直接以核函数的形式给出不同点之间的“距离”（相关性）；
  - 核函数具有很大的灵活性，可将领域知识引入；

3. 支持向量机

（support vector machine, SVM）



一种清晰、强大的机器学习方法。

历史

- 1963年，Vapnik在解决模式识别问题时提出了支持向量方法。
  - 起决定性作用的样本为支持向量。
- 1971年，Kimeldorf基于支持向量构建核空间
- 1995年，Vapnik等人正式提出支持向量机。

Corinna Cortes, Vladimir Vapnik. Support-Vector Networks. *Machine Leaming* **20**, 273-297 (1995)（引用36433次）

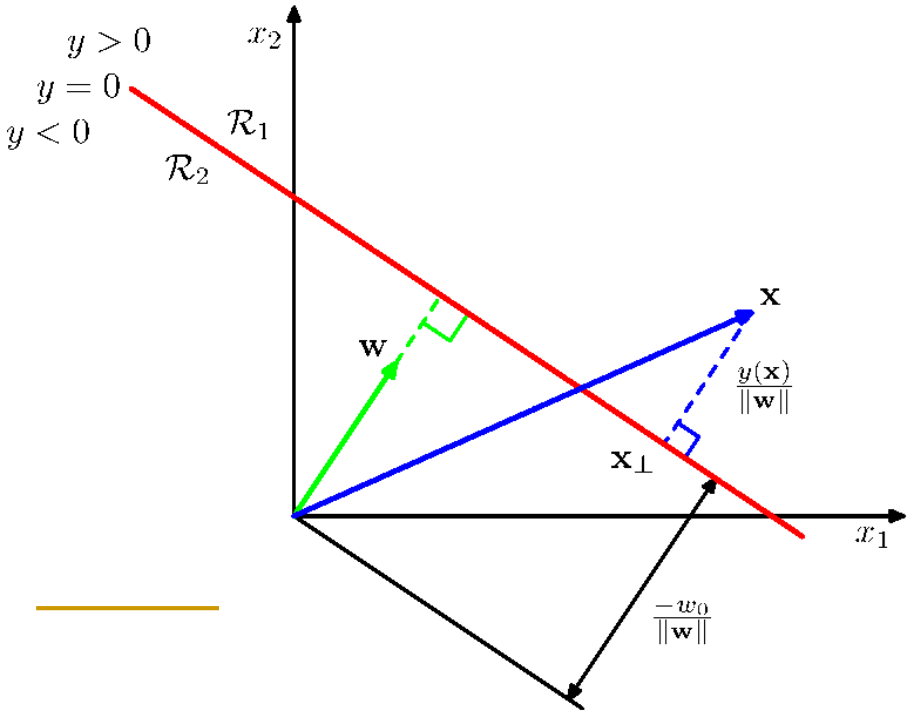
  - 提出了软边距的非线性SVM并将其应用于手写字符识别问题。
- 在手写数字识别、文本分类等众多领域获得了成功。

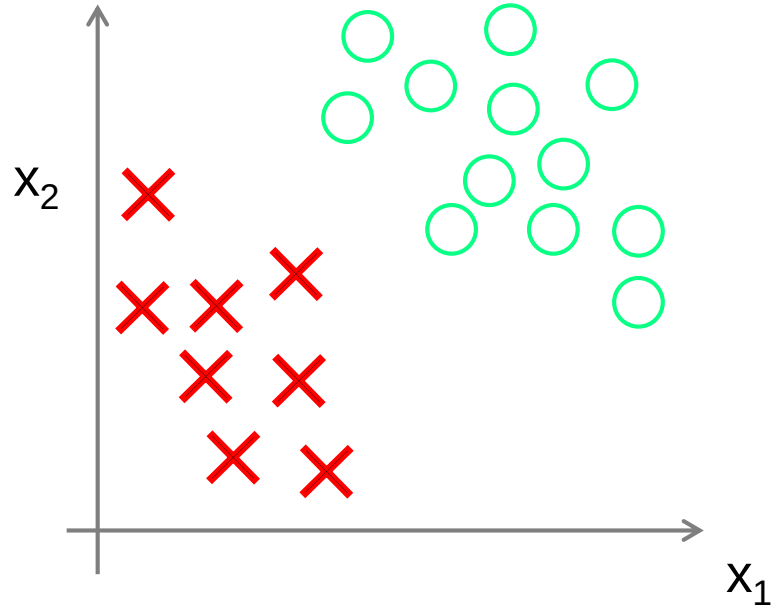
想法：类间间隔的最大化

$$y(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

$$\Longrightarrow y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + w_0$$

$$(w_0 = b)$$

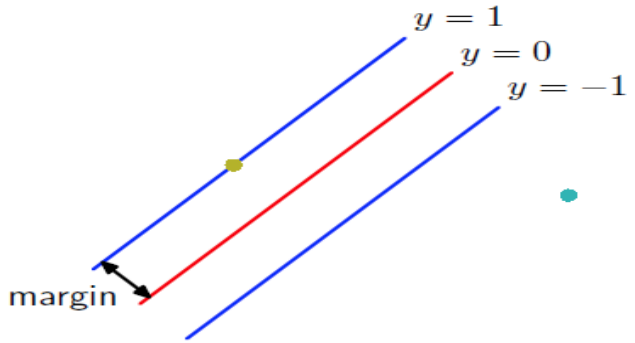


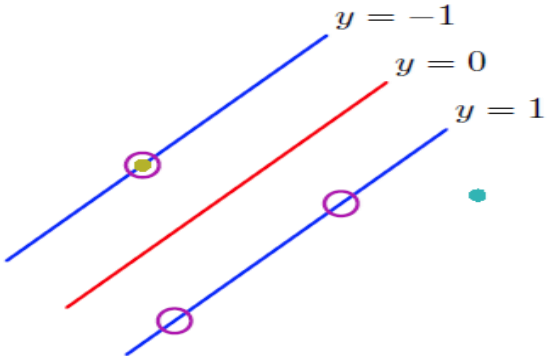


- $\mathbf{w}$代表了决策面$(y(\mathbf{x}) = \mathbf{0})$的垂直方向。因此，$\mathbf{x}$与决策面的（有向）距离为

$$d(\mathbf{x}) = \mathbf{x} \cdot \hat{\mathbf{w}} - d_0 = \frac{\mathbf{w} \cdot \mathbf{x}}{\|\mathbf{w}\|} - d_0 = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$$

线性可分体系





- $\mathbf{x}_n$与决策面的距离可写成

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n[\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b]}{\|\mathbf{w}\|}$$

$t_n = \pm 1$

- 最小距离，即间隔/边距（margin）：

$$\frac{1}{\|\mathbf{w}\|}\min_n\{t_n[\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b]\}$$

- 间隔最大化：

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_n\{t_n[\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b]\}\right\}$$

求解方法…

- 当$\mathbf{w}$与b同时增大κ倍时，$t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]$也增大κ倍，但$\frac{t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]}{\|\mathbf{w}\|}$不变。
- 利用这个性质，可以令离决策面最近的数据点n（边界点）满足

$$t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]=1$$

则对任何数据点都有

$$t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]\geq 1$$

称为决策平面的规范表示（canonical representation）。

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_n\{t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]\}\right\}\Longrightarrow\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\right\}\Longrightarrow\arg\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2$$

数学背景：拉格朗日方法与KKT条件

- 等式约束

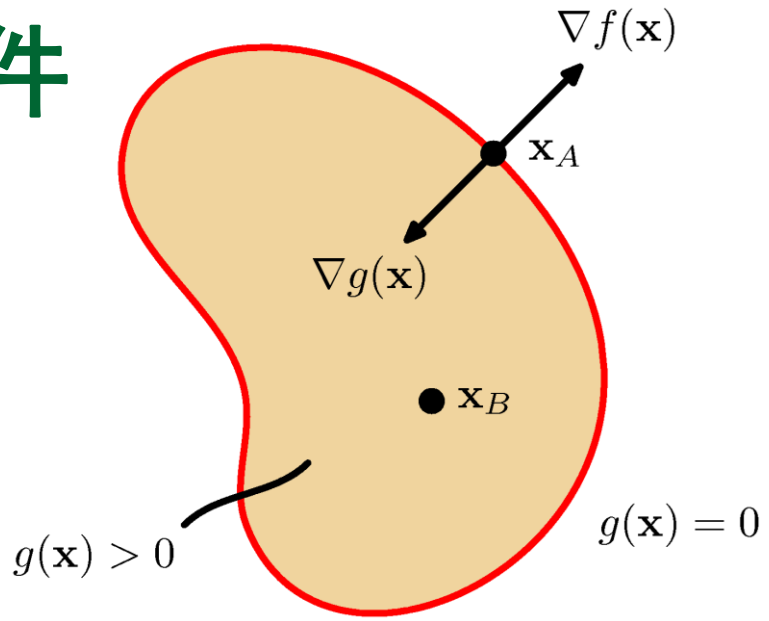
$f(\mathbf{x})$极值；
约束：$g(\mathbf{x}) = 0$

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})\text{极值}$$

- 不等式约束

$f(\mathbf{x})$极值；
约束：$g(\mathbf{x}) \geq 0$

引入：
$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$



结果有：

$$\lambda g(\mathbf{x}) = 0$$

$$\frac{\partial}{\partial \mathbf{x}} L(\mathbf{x}) = 0$$

$$g(\mathbf{x}) \geq 0$$

可能解1：（内部）$\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = 0,\ g(\mathbf{x}) > 0$

可能解2：（边界）$\frac{\partial}{\partial (\mathbf{x}, \lambda)} L(\mathbf{x}) = 0$ → （极大值）$g(\mathbf{x}) = 0, \lambda > 0$；（极小值）$g(\mathbf{x}) = 0, \lambda < 0$

间隔最大化

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$，约束条件$t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b] \geq 1$

- 引入拉格朗日因子λ_n：

$$L(\mathbf{w},b,\lambda) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_n \lambda_n\{t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]-1\}$$

约束条件：$\lambda_n\{t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]-1\}=0$

- 其极值满足

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w},b,\lambda) = 0 \Rightarrow \mathbf{w} = \sum_n \lambda_n t_n \phi(\mathbf{x}_n)$$

- 代入$y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$，得

$$y(\mathbf{x}) = \sum_{n,i} \lambda_n t_n \phi_i(\mathbf{x}_n)\phi_i(\mathbf{x}) + b = \sum_n \lambda_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$$

与核函数$k(\mathbf{x}, \mathbf{x}_n) = \sum_i \phi_i(\mathbf{x}_n)\phi_i(\mathbf{x})$有关。

- λ_n的值需进一步求解。但由于条件

$$\lambda_n\{t_n[\mathbf{w}^T\phi(\mathbf{x}_n) + b] - 1\} = 0$$

的存在，不在间隔边缘$t_n[\mathbf{w}^T\phi(\mathbf{x}_n) + b] = 1$上的点都有$\lambda_n = 0$，不出现在求和中。

- 只有间隔边缘的数据点（支持向量）才对$y(\mathbf{x})$的计算有贡献！

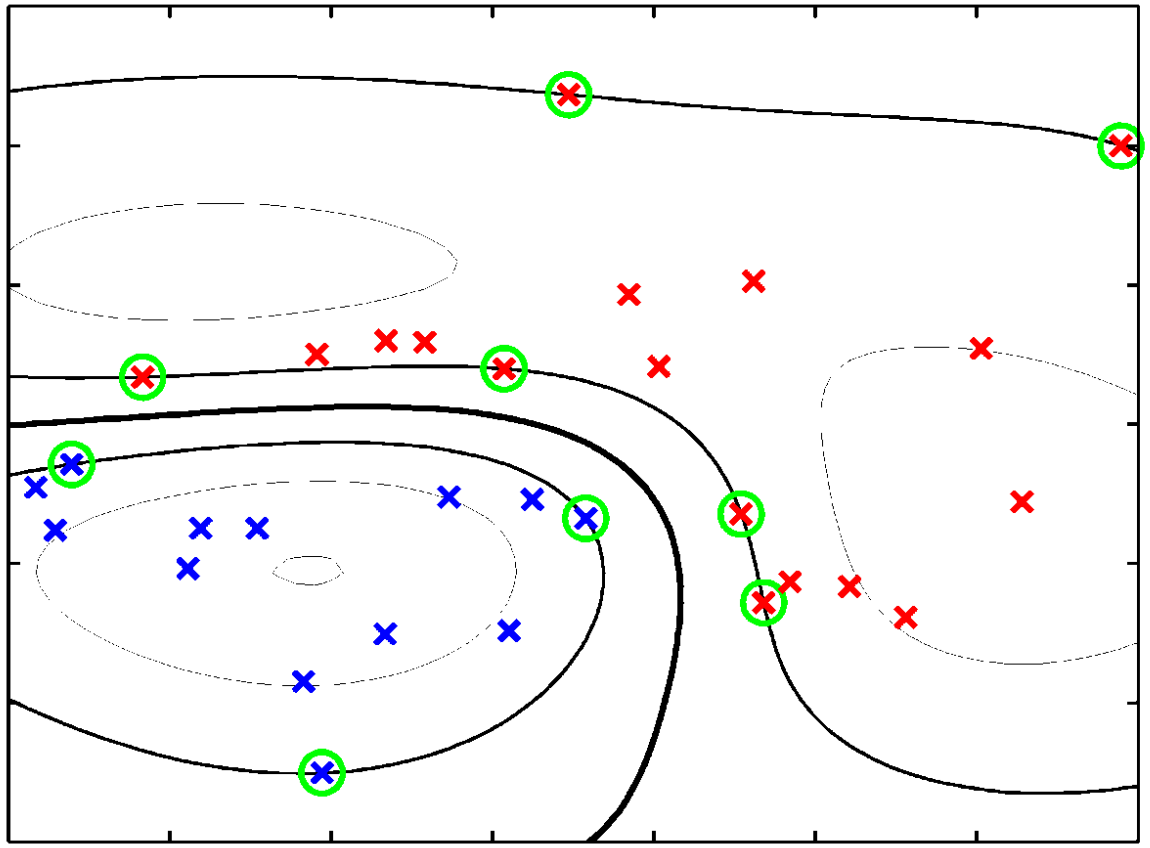
实际解法…

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$，约束条件$t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b] \geq 1$

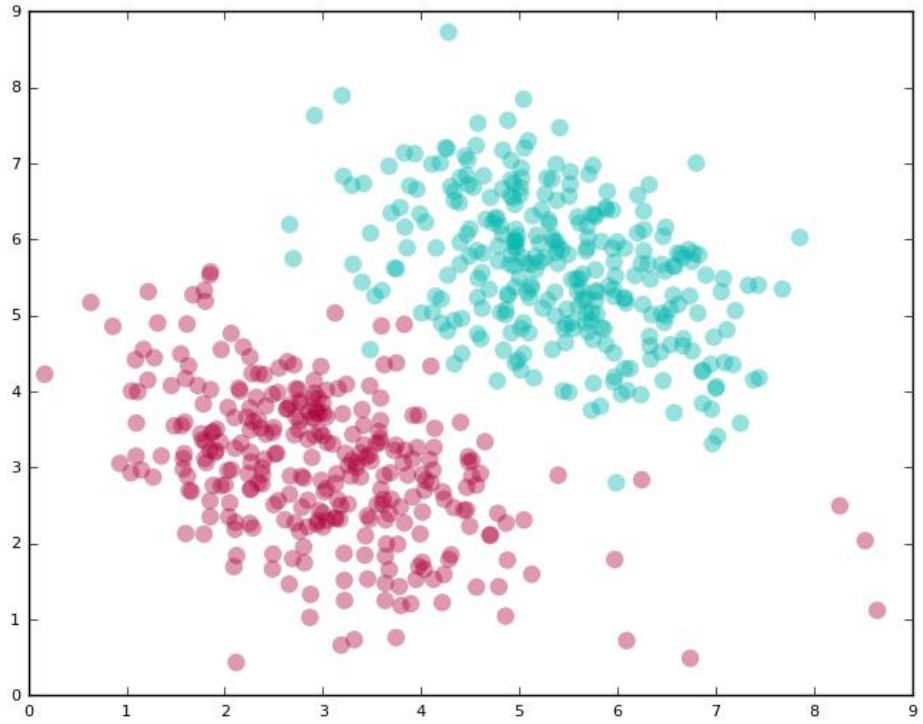
- 这是一个凸优化问题，目标函数是二次的，约束条件是线性的。可以用现成的优化包进行求解。
- 更加方便的是将其化成对偶问题，并用SMO算法进行求解（复杂，略）。

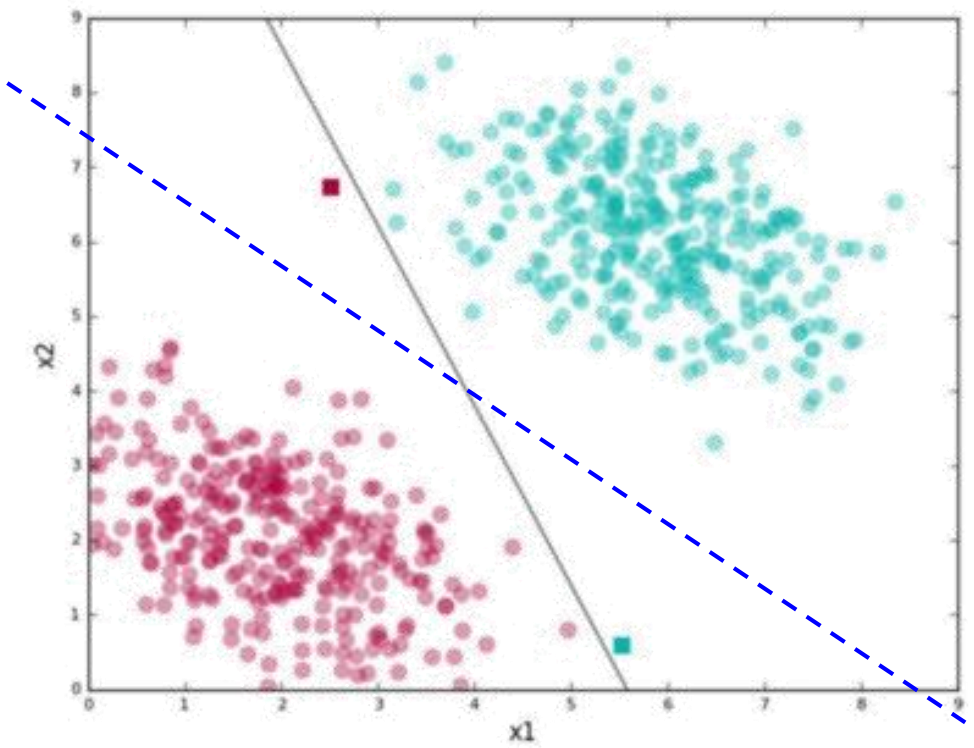
例子

- 高斯核下的支持向量机结果。
- 支持向量机具有鲁棒性的原因：努力用一个最大间距来分离样本。



线性不可分体系：软间隔的使用





- 线性不可分体系：不存在符合要求的硬间隔
- 右边：如果允许分类错误，则可增加间隔，效果更好。

软间隔…

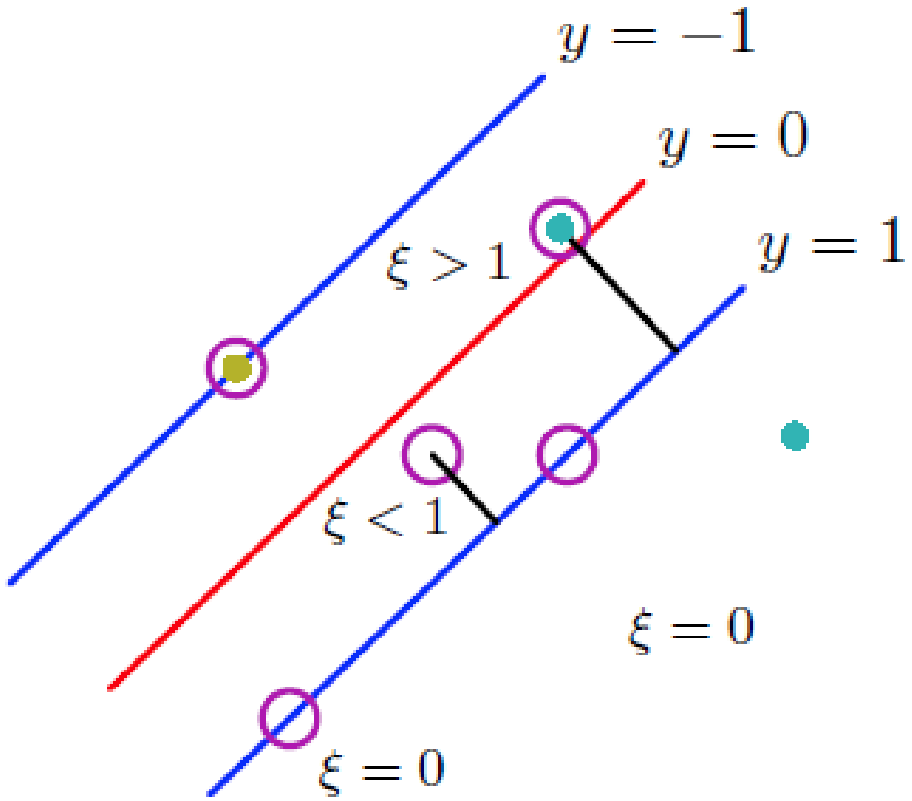
- 为了允许误差的存在，引入松弛变量（slack variables）$\xi_n \geq 0$，使约束条件放宽为

$$t_n y(\mathbf{x}_n) \geq 1 - \xi_n$$

ξ_n可看做是数据点越过margin边界的距离，即某种误差。求解时可将其视为独立变量。

- 代价函数写成

$$C \sum_n \xi_n + \frac{1}{2} \|\mathbf{w}\|^2$$



解法…

$$\arg\min_{\mathbf{w},b}\left[C\sum_{n}\xi_n+\frac{1}{2}\|\mathbf{w}\|^2\right]$$

约束条件：$t_n[\mathbf{w}^T\phi(\mathbf{x}_n)+b]\geq 1-\xi_n,\ \xi_n\geq 0$

- 通过引入引入拉格朗日函数

$$L=C\sum_n\xi_n+\frac{1}{2}\|\mathbf{w}\|^2-\sum_n a_n[t_n y(\mathbf{x}_n)-1+\xi_n]-\sum_n\mu_n\xi_n$$

将问题变成（复杂，略）

$$\arg\max_{\{a_n\}}\left[\sum_n a_n-\sum_{m,n}a_m a_n t_m t_n k(\mathbf{x}_m,\mathbf{x}_n)\right]$$

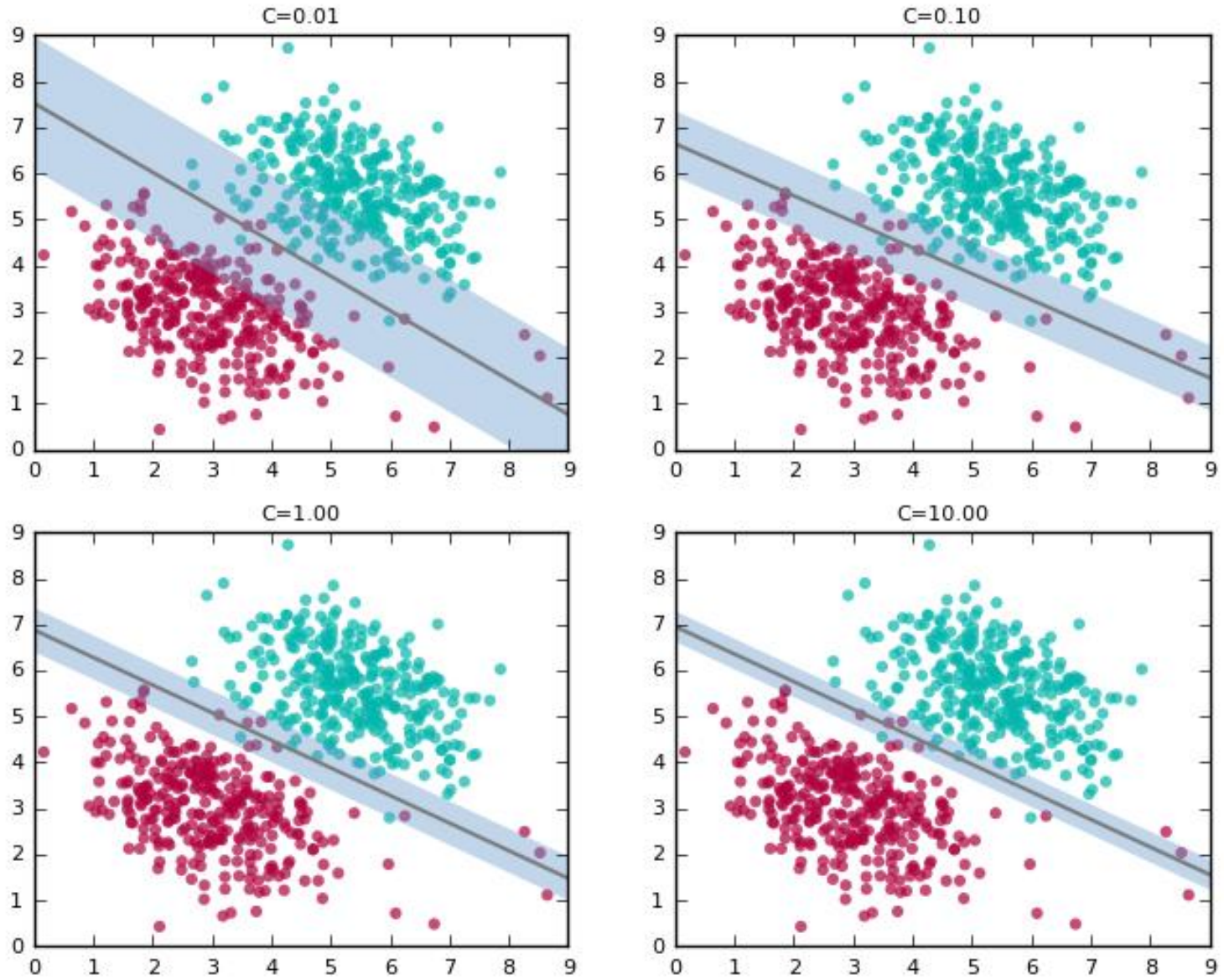
约束条件：$0\leq a_n\leq C,\ \sum_n a_n t_n=0$。一般采用SMO算法求解。

例子：

- 线性基

$$y(\mathbf{x}) = \sum_n a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$$

- C是超参数，可通过交叉验证等方法优化。



与逻辑回归的联系

- 代价函数可重新写成

$$\sum_n E_{\mathrm{SV}}(y_n t_n) + \lambda \|\mathbf{w}\|^2$$

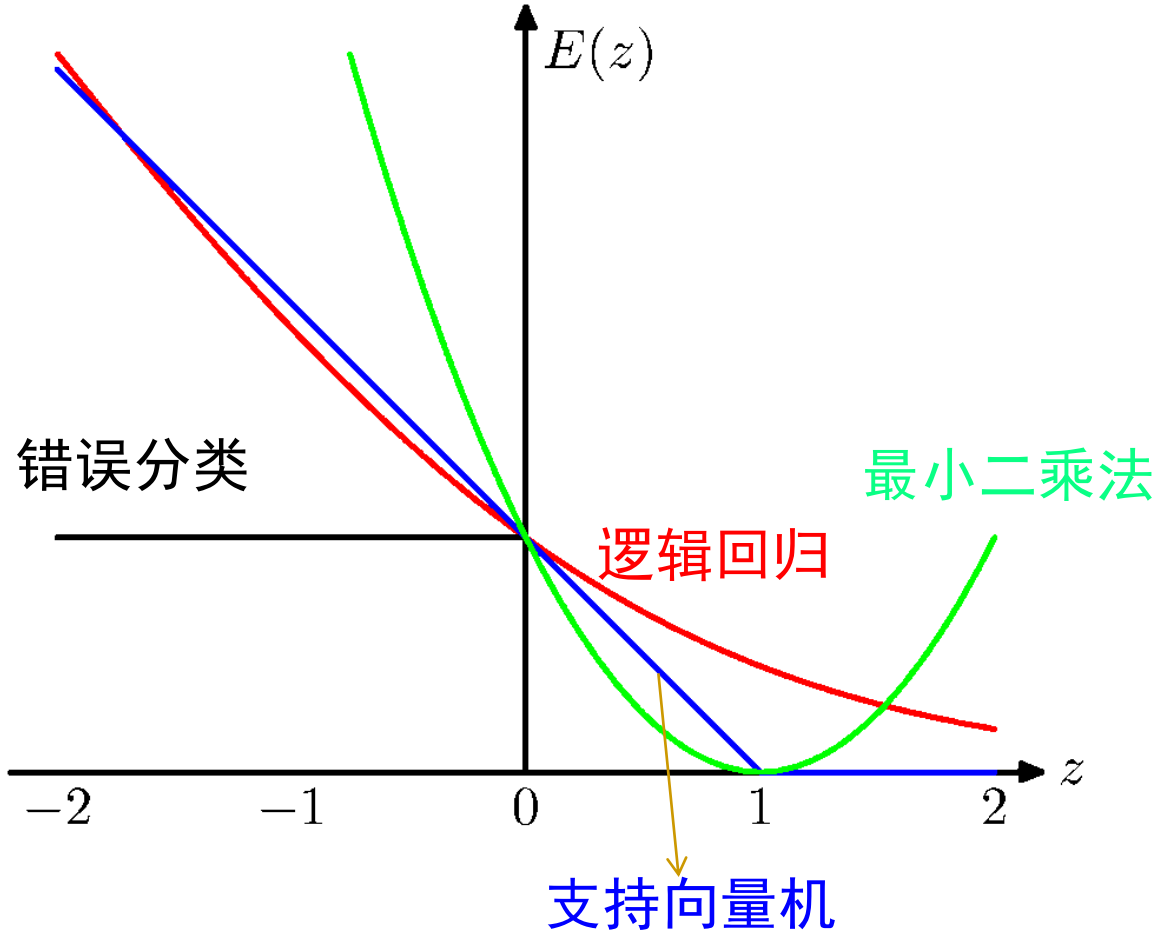
其中

$$E_{\mathrm{SV}}(y_n t_n) = [1 - y_n t_n]_+$$

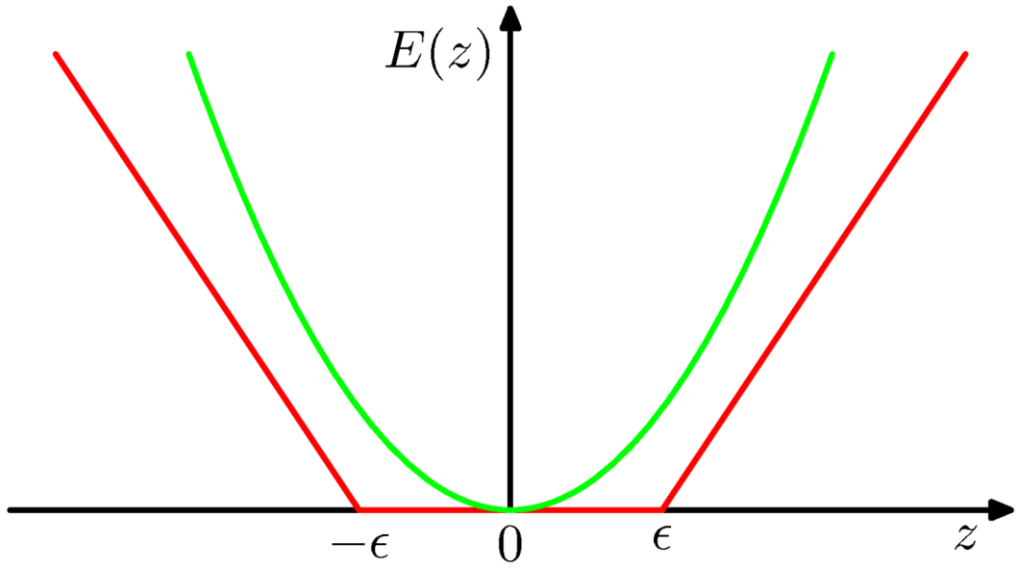
只取其正值，否则为零

- 对于逻辑回归，有（分析略）

$$E_{\mathrm{SV}}(y_n t_n) = E_{\mathrm{SV}}(z) = \ln(1 + e^{-z})$$



SVM用于回归问题

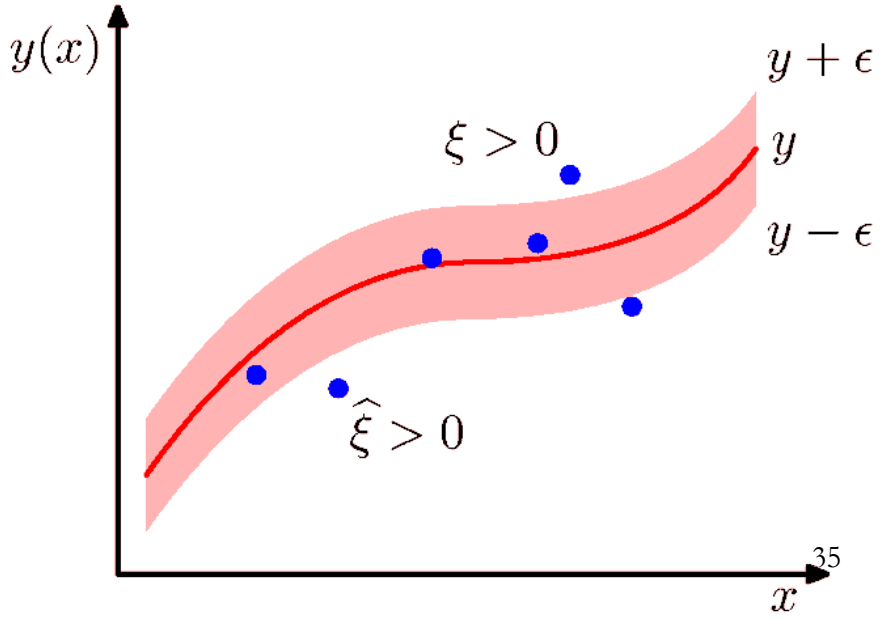


- 回归问题的二乘误差为

$$\frac{1}{2}\sum_n (y_n - t_n)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- 借鉴SVM分类中的思想，将误差定义为

$$E_\varepsilon(y(\mathbf{x}) - t) = \begin{cases} 0, & \text{if } |y(\mathbf{x}) - t| < \varepsilon \\ |y(\mathbf{x}) - t| - \varepsilon, & \text{otherwise} \end{cases}$$

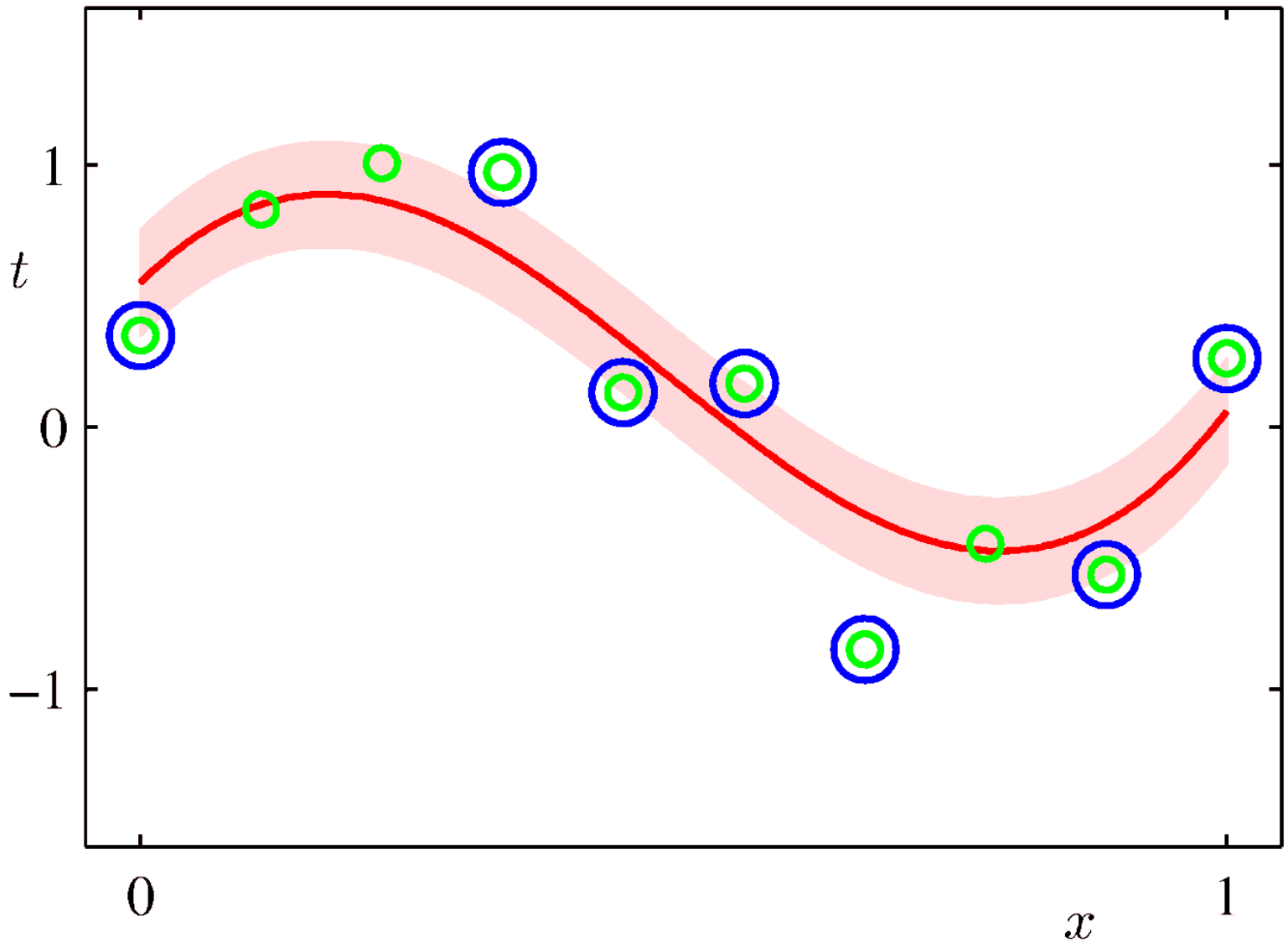


- 或者引入松弛变量$\xi_n, \hat{\xi}_n \geq 0$

$$\begin{cases} t_n \leq y(\mathbf{x}_n) + \varepsilon + \xi_n \\ t_n \geq y(\mathbf{x}_n) - \varepsilon - \hat{\xi}_n \end{cases}$$

代价函数变成：$C\sum_n (\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2$

例子



与逻辑回归的比较/选择

- 同为凸优化，只有一个优化解。
- 特征数M，样本数N，则
  - N较小（允许比M小），训练集数据量不够支持我们训练一个复杂的非线性模型，可选用逻辑回归模型或者不带核函数的支持向量机。
  - 如果M较小（例如在 1-1000 之间），而且N大小中等（10-10000之间），可使用高斯核函数的支持向量机。
  - 如果M较小，而N较大（例如大于50000），则使用支持向量机会非常慢，解决方案是创造、增加更多的特征，然后使用逻辑回归或不带核函数的支持向量机。

（吴恩达建议）

与神经网络的比较

- 神经网络适用性很广，但它的优化问题是非凸的，训练可能非常慢。
- 训练结束后不需要保存训练数据点，预测阶段计算快，在高频交易事务（如股票）中有优势。

与近邻法的比较

- 表面上看，支持向量机很像加权近邻算法。
- 但它只需记住那些用于确定边界的关键例子。
- 近邻算法的边界是锯齿状的，但支持向量机的边界是光滑的。

小结

- 支持向量机
  - 通过最大化不同类之间的间隔（margin）来取得好的效果。
  - 与其它的核方法相比，预测时只需要保存间隔边缘的数据点（支持向量）。

应用例子1：

Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier & Alexander J. Norquist.

Machine-learning-assisted materials discovery using failed experiments.

Nature **533**, 73 (2016).

学习资料：na73.pdf, na73s1.pdf

MOFs合成（水热合成法）

- 从实验记录（laboratory notebooks）录入，约50条/小时。
  - 组分信息；
  - 反应条件（pH值，加热曲线）；
  - 产物结果
    - 晶体（1：无固态物；2：无定形物；3：多晶；4：单晶（> 0.01 mm）
    - 纯度（1：多相；2：单相）
- 复查（1.89%错误）
- 筛选：含有机组分、无机组分、溶剂，反应条件，产物
- 数据个数：3955

特征选择（总数：273）

- 有机反应物
  - 物化性质（分子量，pH值下的氢键受体与给体数目，极化表面积）
  - 19个直接性质+6个摩尔比
- 无机反应物
  - 12个原子性质（离子化能、电子亲和力、电负性、硬度、原子半径）
  - 22个逻辑值（各种金属）
  - 28个逻辑值（周期表位置）
  - 8个逻辑值（金属价态）
- 反应条件
  - 5个（温度、时间、pH值等）

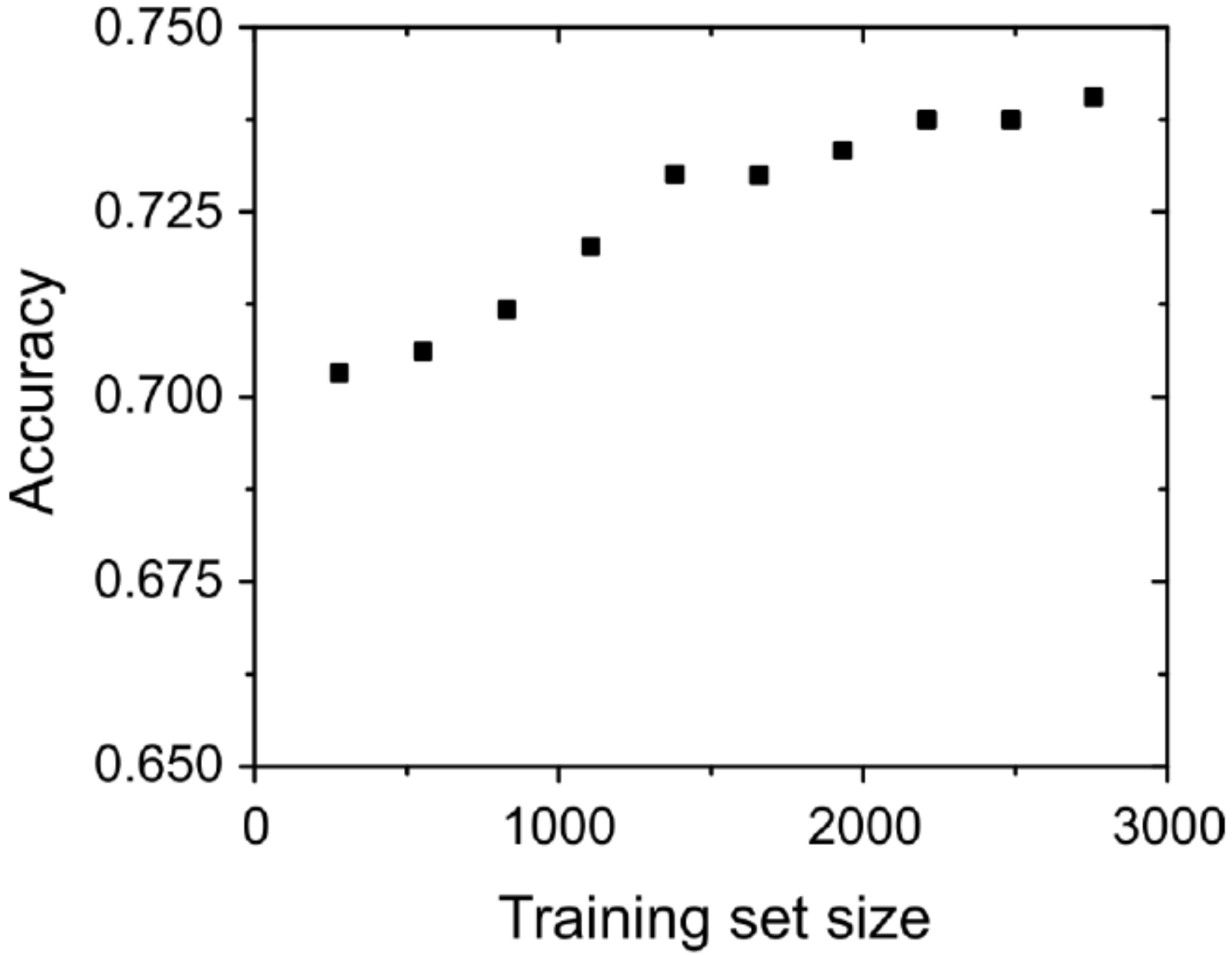
一些细节…

- Data normalization
- a standard 1/3-test and 2/3-training data split.
- 测试集不是随机抽取：all of the reactions containing a particular set of inorganic and organic reactants were placed into either the test or training set.
- 二分类：outcomes of ‘3’ or ‘4’ were considered successes and ‘1’ and ‘2’ were grouped together as failed reactions.
  - 2783 positive，1228 negative
- 支持向量机！

结果

Technique	Accuracy (%)
Decision tree (J48)	67.5
Random forest (size 100 and 1000)	69.8, 70.5
Logistic regression	69.2
k-Nearest neighbors (K = 1, 2, and 3)	69.1, 66.9, 68.4
SMO SVM	74.1

学习曲线

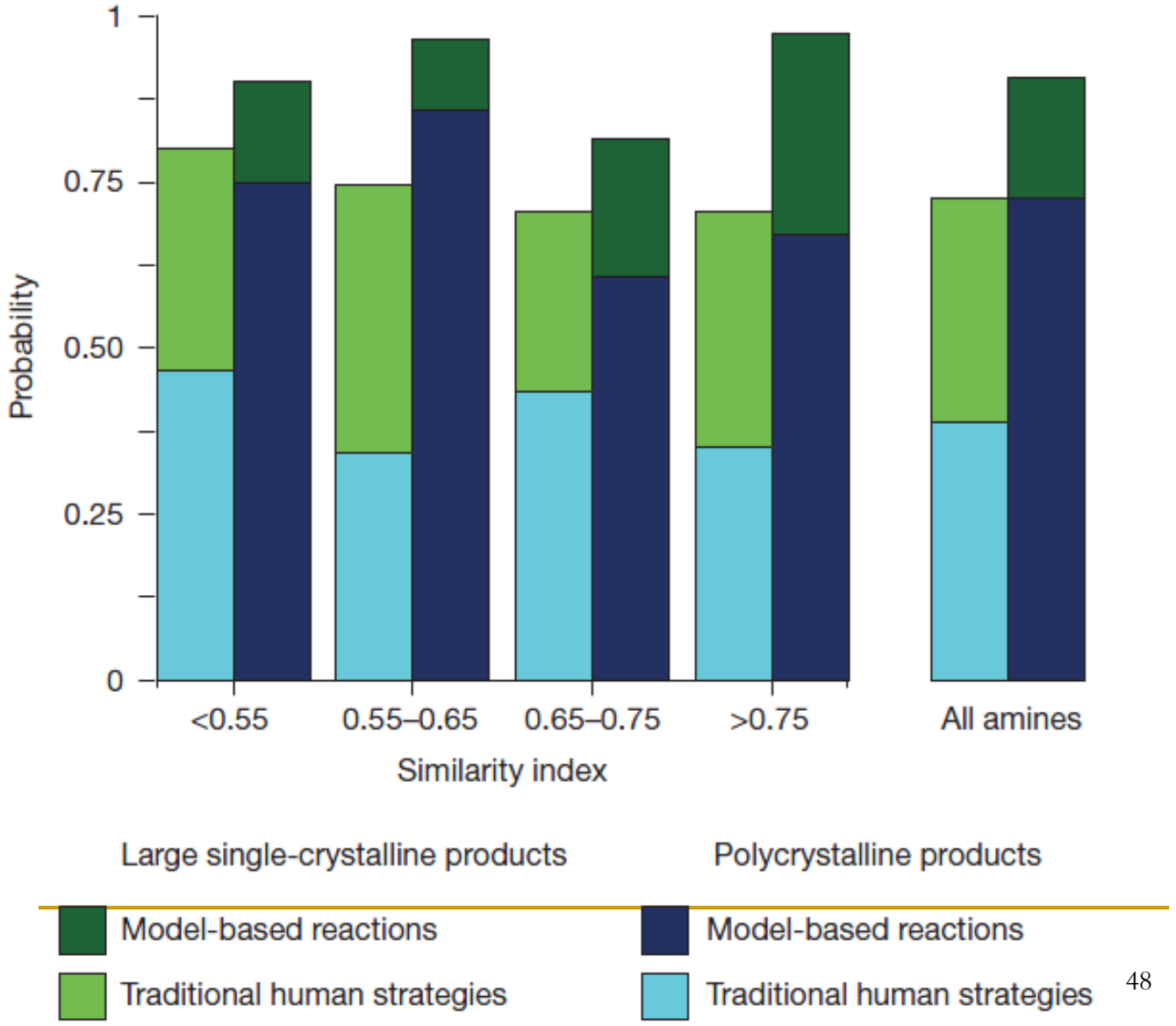

0.750
0.725
0.700
0.675
0.650
Accuracy
0
1000
2000
3000
Training set size

一些分析…

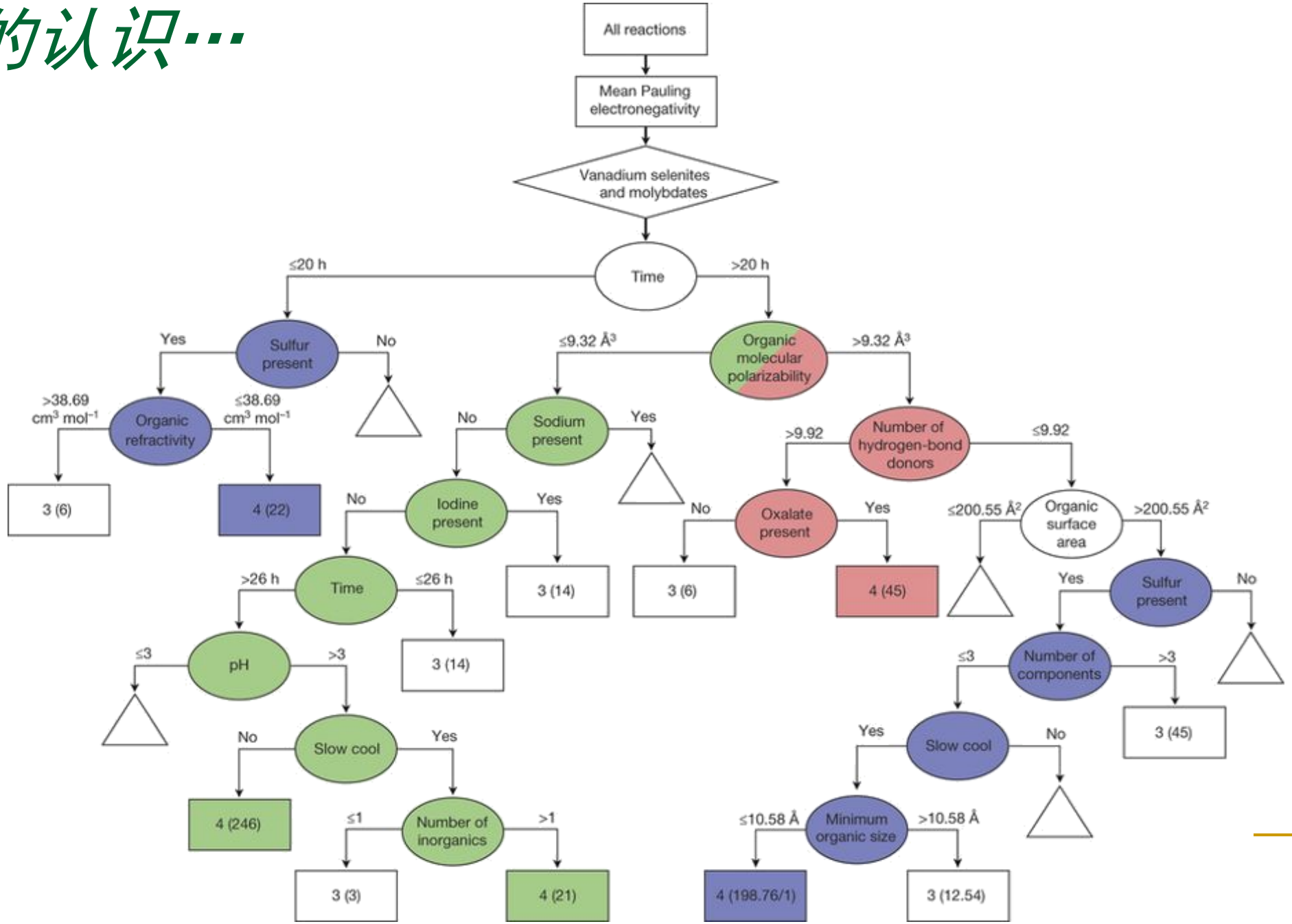
- High-dimensional feature spaces are not problematic for SVMs
- Feature selection was performed on the model to identify the properties with the most influence on classification success.
  - organic amines (van der Waals surface area, solvent-accessible surface area of positively charged atoms and the number of hydrogen-bond donors) and inorganic components (mean of the Pauling electronegativites of the metals, their mole-weighted hardness and mean mole-weighted atomic radii).
  - Using only these six features lowers the model accuracy to 70.7%;
  - the six selected features listed above appear in the decision-tree description of the model

新反应物的预测

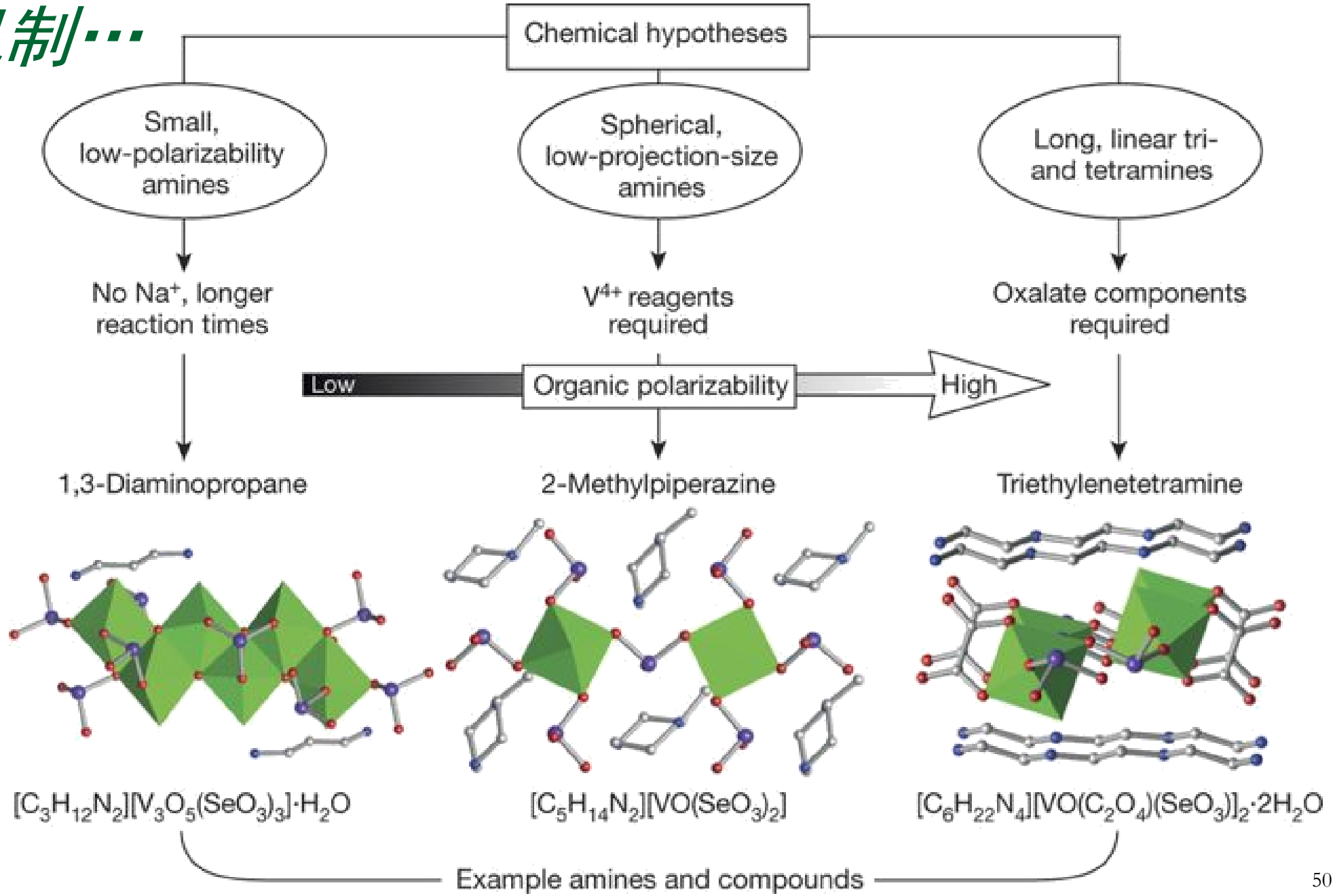
- 34 new diamines（二胺）from 1680。
- 新做276个实验。
- 成功率：89%
- 人类（规则）：78%



机器的认识…


All reactions
Mean Pauling electronegativity
Vanadium selenites and molybdates
Time
≤20 h
>20 h
Sulfur present
Yes
No
Organic refractivity
>38.69 cm³ mol⁻¹
≤38.69 cm³ mol⁻¹
3 (6)
4 (22)
Organic molecular polarizability
≤9.32 Å³
>9.32 Å³
Sodium present
No
Yes
Iodine present
No
Yes
3 (14)
Time
>26 h
≤26 h
3 (14)
pH
≤3
>3
Slow cool
No
Yes
4 (246)
Number of inorganics
≤1
>1
3 (3)
4 (21)
Number of hydrogen-bond donors
>9.92
≤9.92
Oxalate present
No
Yes
3 (6)
4 (45)
Organic surface area
≤200.55 Å²
>200.55 Å²
Sulfur present
Yes
No
Number of components
≤3
>3
3 (45)
Slow cool
Yes
No
Minimum organic size
≤10.58 Å
>10.58 Å
4 (198.76/1)
3 (12.54)

三种机制…


Chemical hypotheses
Small, low-polarizability amines
Spherical, low-projection-size amines
Long, linear tri- and tetramines
No Na+, longer reaction times
V4+ reagents required
Oxalate components required
Low
Organic polarizability
High
1,3-Diaminopropane
2-Methylpiperazine
Triethylenetetramine
[C3H12N2][V3O5(SeO3)3]·H2O
[C5H14N2][VO(SeO3)2]
[C6H22N4][VO(C2O4)(SeO3)]2·2H2O
Example amines and compounds

数据可及性

- **Code availability.** All code for this project is available at https://github.com/darkreactions. The code is licensed under the GPL version 3.

应用例子2：

Jarosław M. Granda, Liva Donina, Vincenza Dragone, De-Liang Long & Leroy Cronin.

Controlling an organic synthesis robot with machine learning to search for new reactivity.

Nature **559**, 377 (2018).

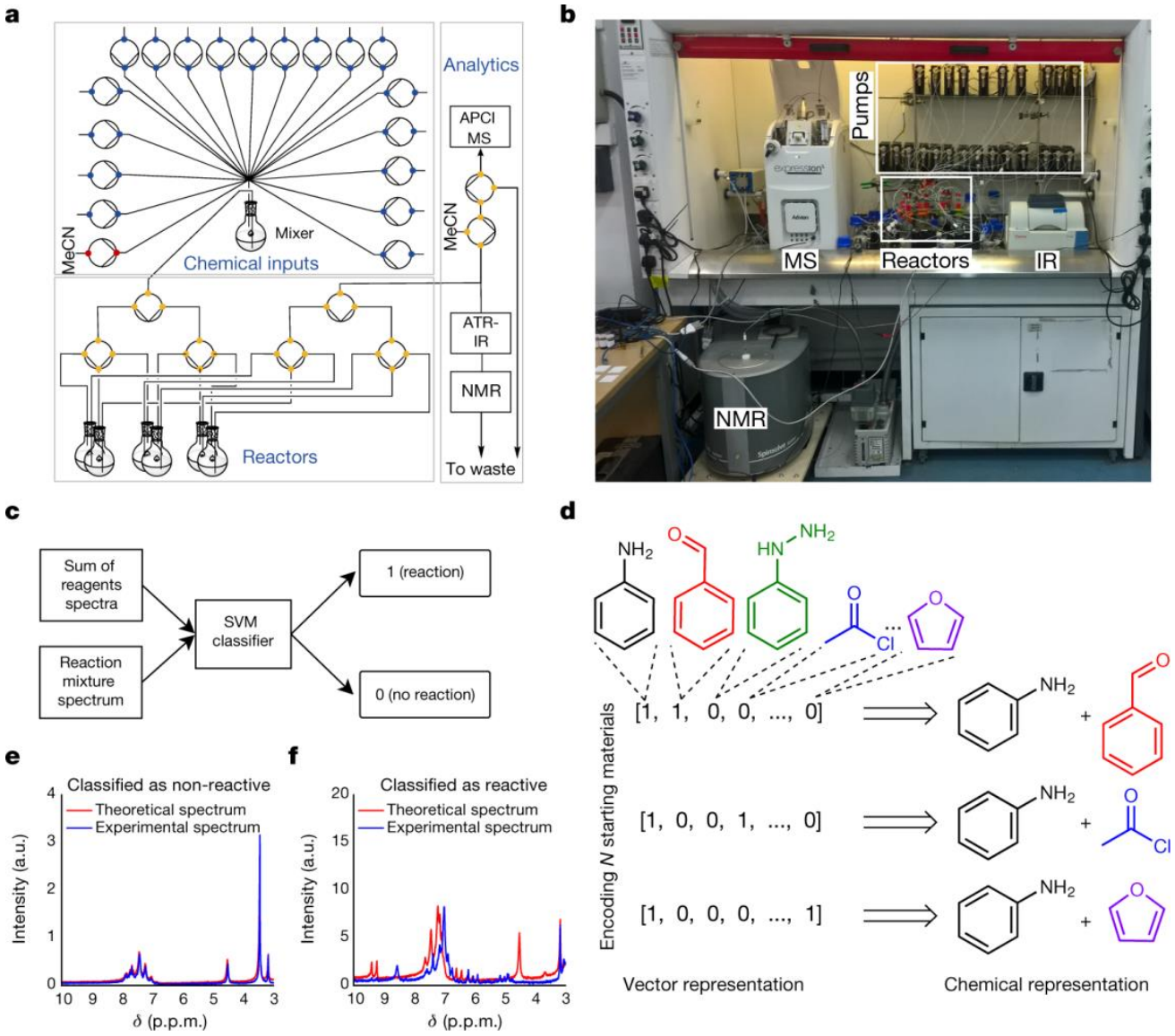
学习资料：na377.pdf

- http://m.sohu.com/a/244701109_610519

  三思而后行！Nature报道像化学家一样探索新反应的“AI”机器人

有机反应能否进行？

- 有机合成机器人
  - 同时做6个实验，一天36个。
- 根据实验前后谱学数据预测（检验）是否发生反应。
  - SVM with a linear kernel
  - leave-one-out cross-validation.
  - 72个数据，预测精度86%
  - One-hot encoding
  - using the sci-kit learn package

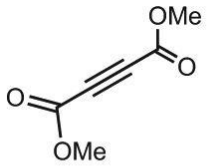


指导实验

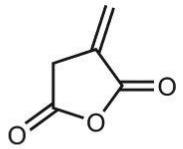
- 用已知反应性数据，训练线性分类器，并预测未知（未做）反应的反应性作为新实验选取的依据。
  - a linear discriminant analysis (LDA)
  - 18种成分，

    选2或3种，共

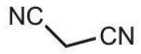
$C_{18}^{2} + C_{18}^{3} = 969$



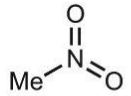
1, activated alkyne



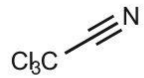
2, anhydride



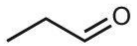
3, activated cyanide



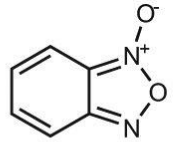
4, nitrocompound



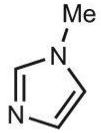
5, activated cyanide



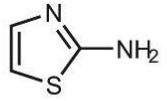
6, aldehyde



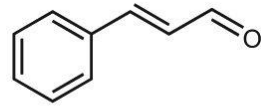
7, electronpoor heterocycle



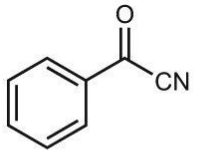
8, imidazole



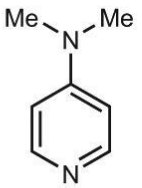
9, aminothiazole



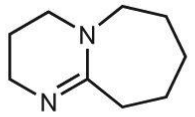
10, a,b-usuturated aldehyde



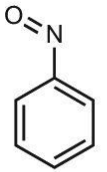
11, activated carbonyl group



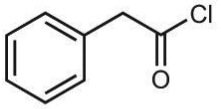
12, DMAP



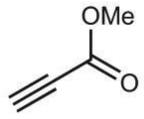
13, non-nucleophilic base



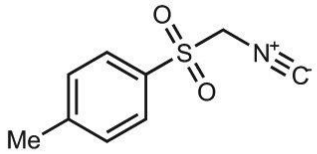
14, nitroso compound



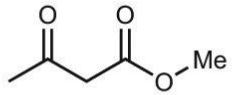
15, acyl chloride



16, activated alkyne

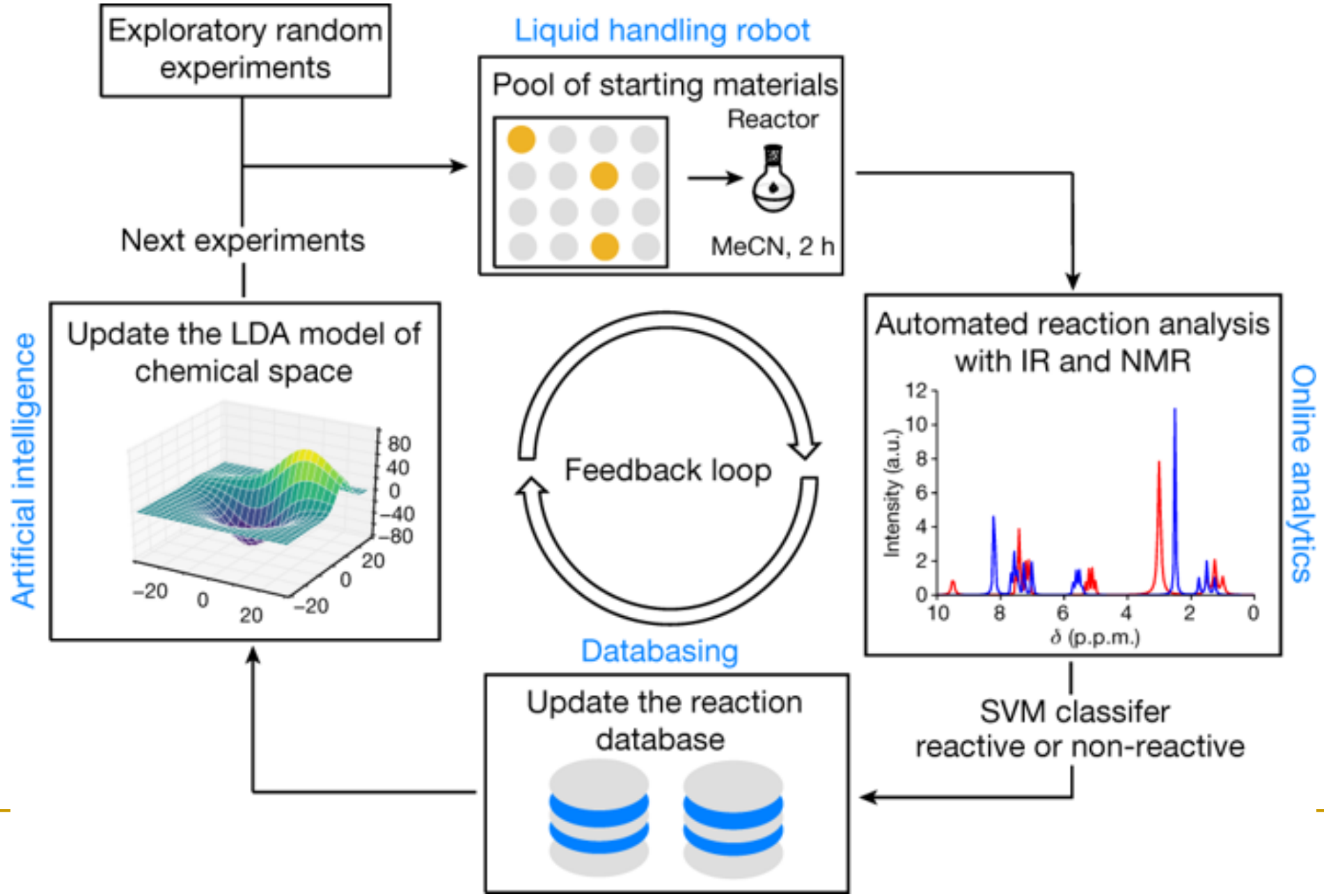


17, isocyanide

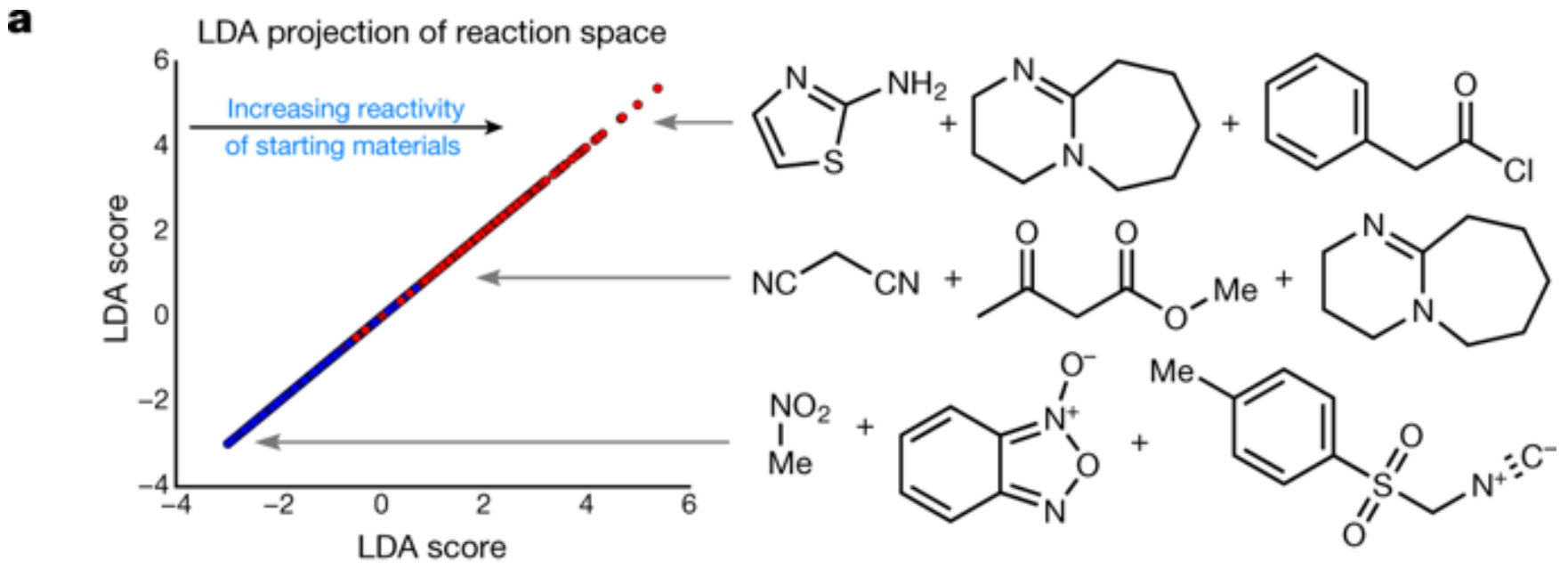


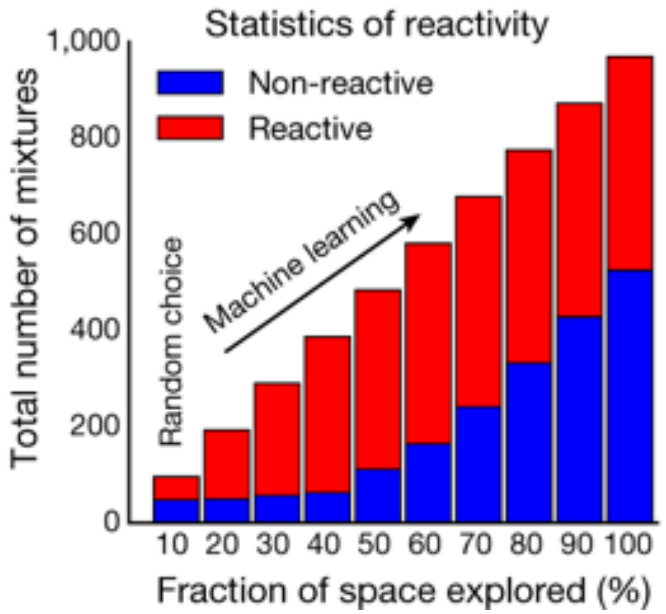
18, enolizable molecule

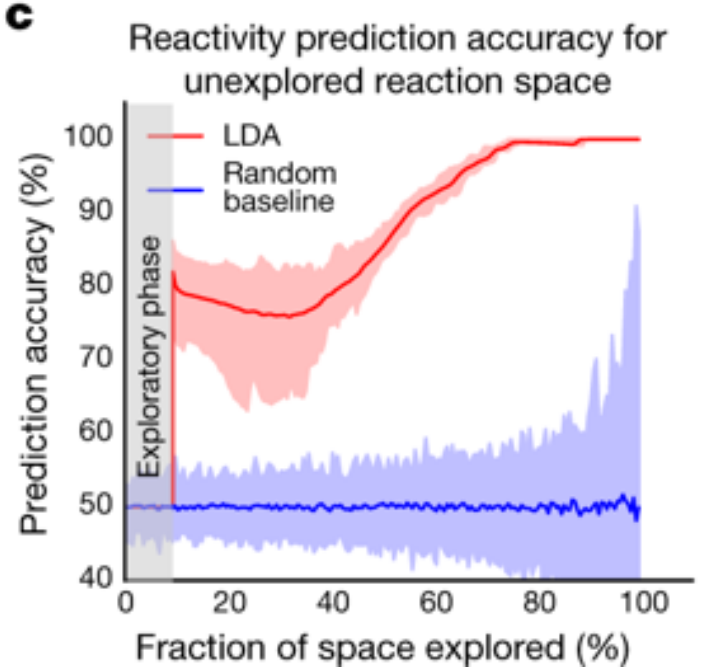
流程


Exploratory random experiments
Liquid handling robot
Pool of starting materials
Reactor
MeCN, 2 h
Next experiments
Automated reaction analysis with IR and NMR
Intensity (a.u.)
δ (p.p.m.)
Online analytics
Artificial intelligence
Update the LDA model of chemical space
Feedback loop
Databasing
Update the reaction database
SVM classifer reactive or non-reactive

结果

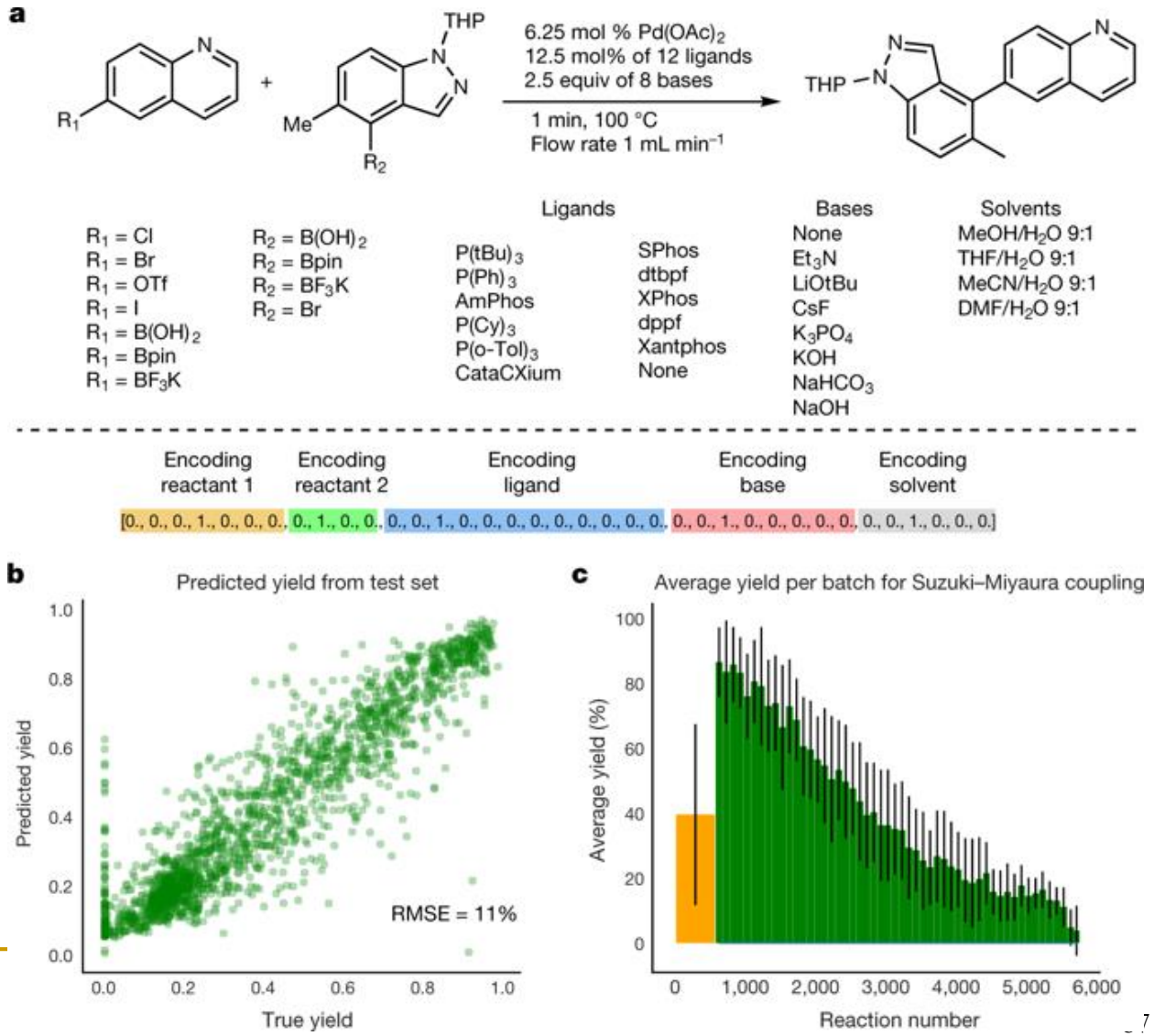

a
LDA projection of reaction space
Increasing reactivity of starting materials
LDA score
LDA score
6
4
2
0
−2
−4
−4
−2
0
2
4
6
NH2
N
S
N
N
O
Cl
NC
CN
O
O
Me
O
N
N
NO2
Me
O⁻
N⁺
O
N
Me
O
S
O
N⁺
C⁻


b
Statistics of reactivity
Non-reactive
Reactive
Random choice
Machine learning
Total number of mixtures
1,000
800
600
400
200
0
10 20 30 40 50 60 70 80 90 100
Fraction of space explored (%)

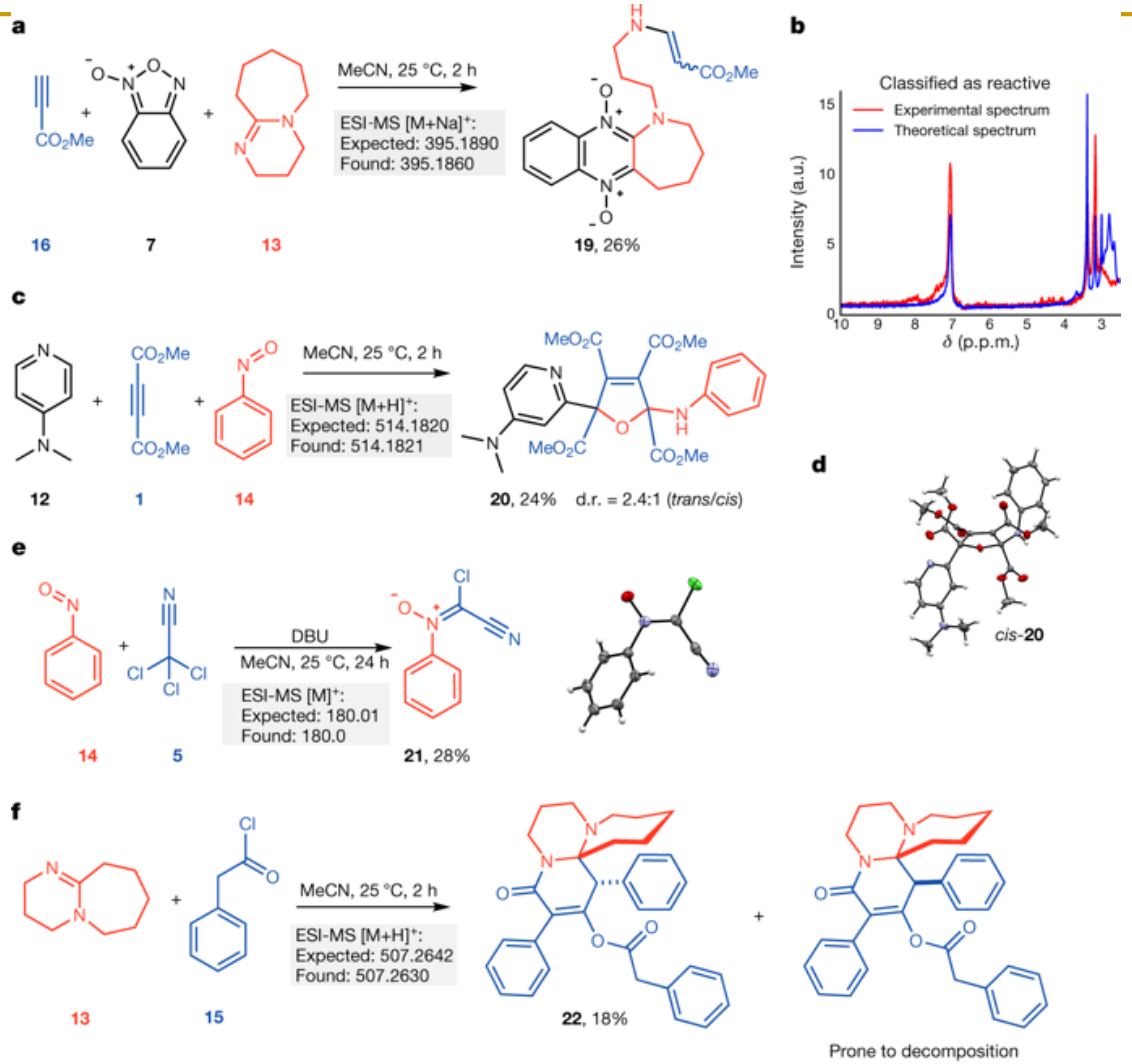

c
Reactivity prediction accuracy for unexplored reaction space
LDA
Random baseline
Exploratory phase
Prediction accuracy (%)
100
90
80
70
60
50
40
0
20
40
60
80
100
Fraction of space explored (%)

Suzuki–Miyaura反应

- 预测产率
- 文献数据：5760
- 神经网络
  - 误差0.11
- This approach is valuable because it shows that by realizing only 10% of the total number of reactions, we can predict the outcomes of the remaining 90% without needing to carry out the experiments.



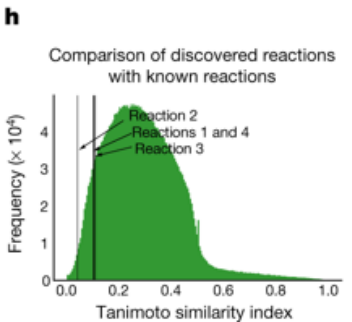
新发现



g

Comparison of discovered reactions with known reactions

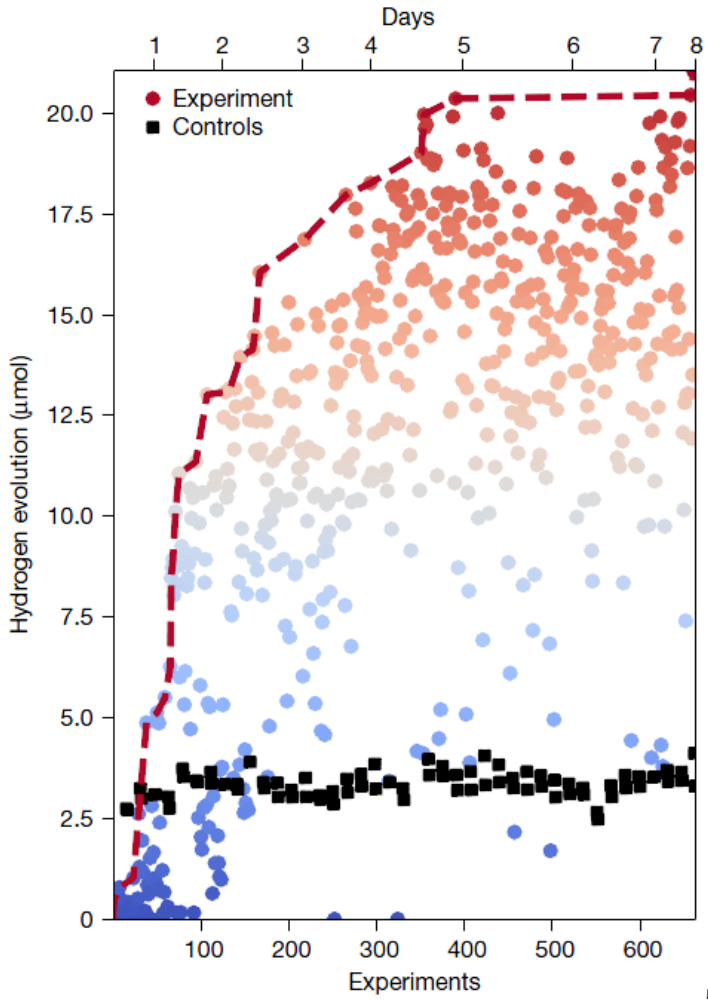
Reaction	Synthesis of compound	Percentile of the data
1	**19**	7.6
2	**20**	0.78
3	**21**	9.0
4	**22**	8.0



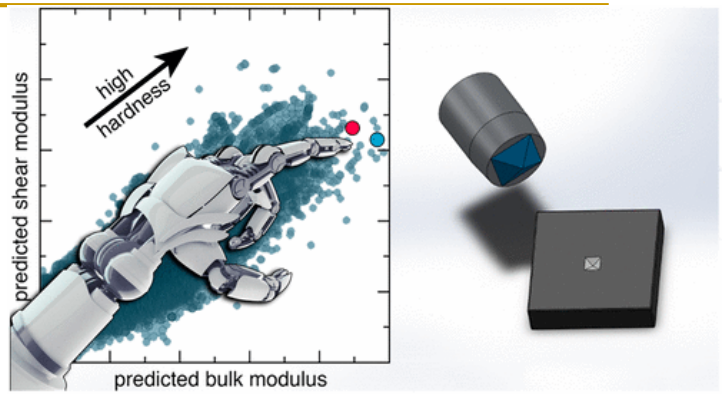
相关工作…

Benjamin Burger, …, Andrew I. Cooper, *Nature* **583**, 237 (2020). na237.pdf

- 水光解产氢的催化剂及配方优化。
- 贝叶斯优化及高斯过程



应用例子3：



Aria Mansouri Tehrani, Anton O. Oliynyk, Marcus Parry, Zeshan Rizvi, Samantha Couper, Feng Lin, Lowell Miyagi, Taylor D. Sparks, and Jakoah Brgoch.

Machine learning directed search for ultraincompressible,

Superhard materials.

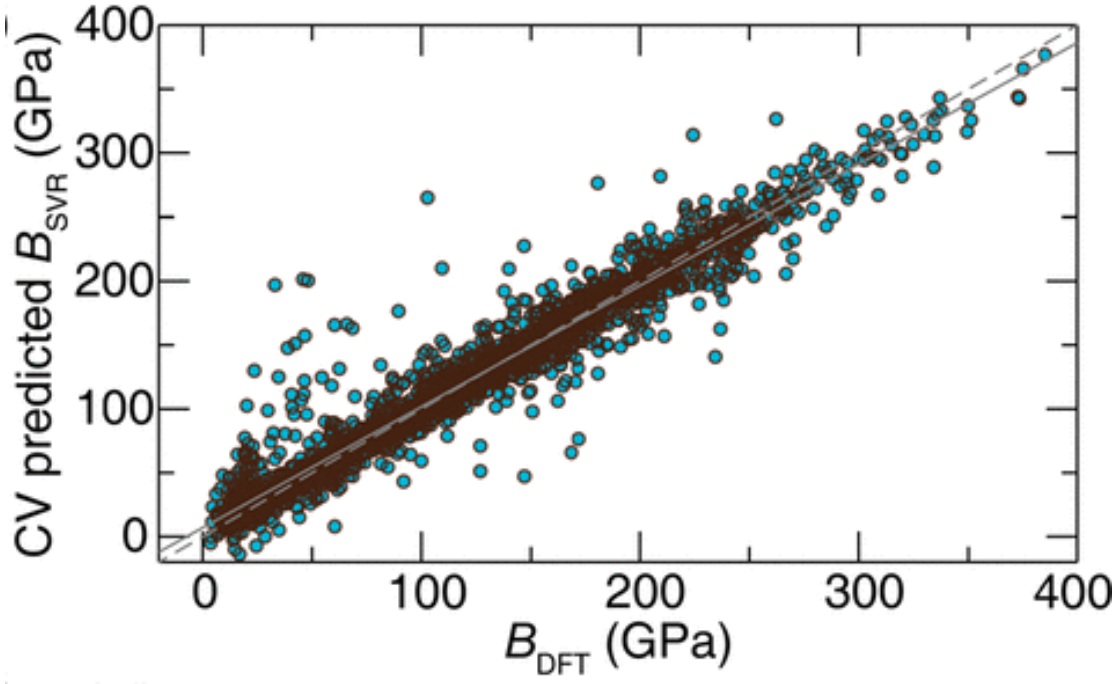
J. Am. Chem. Soc **140**, 9844 (2018).

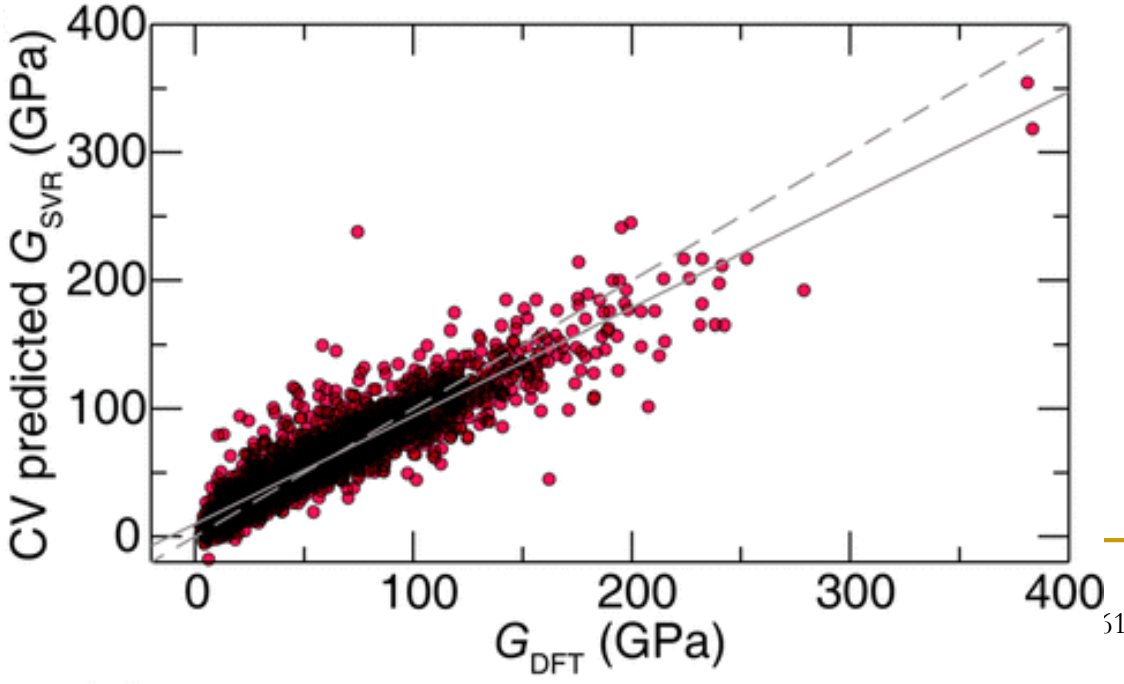
学习资料：jacs9844.pdf

模型训练

- 超硬材料，弹性模量
- 2572个数据（DFT）
- 150个特征

The descriptors included 34 distinct compositional variables describing the elemental properties such as position on the periodic table, electronic structure, and physical properties as well as their associated math expressions (difference, average, largest value, and smallest value). Additionally, 14 structure descriptors related to variables including crystal system, space group, and unit cell volume.

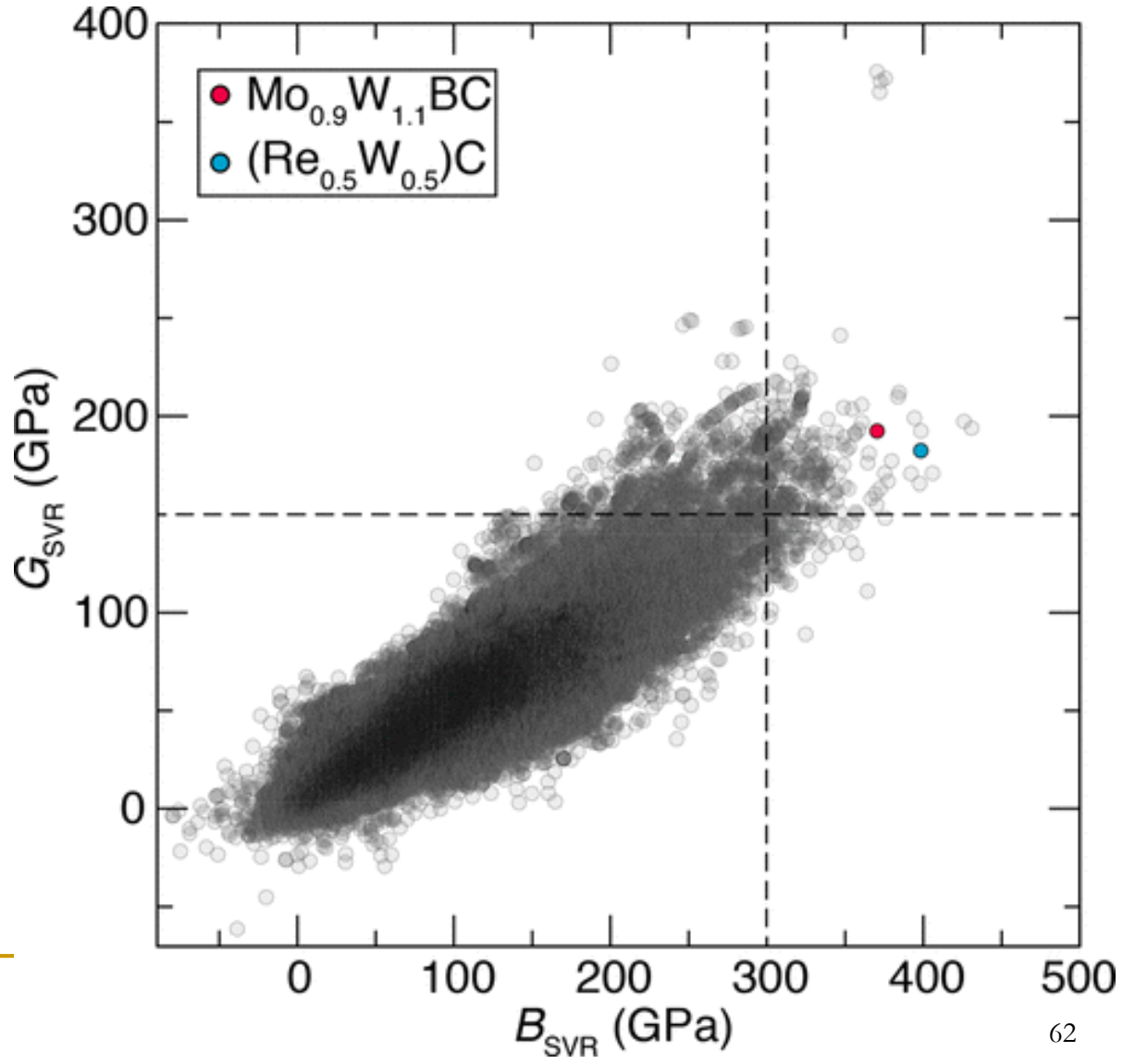




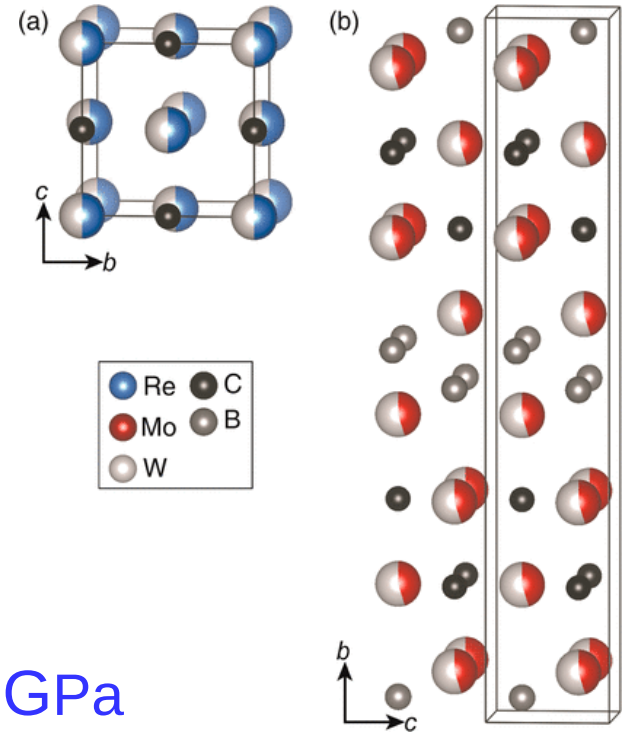
预测：寻找新的超硬材料

- The SVR employed a radial basis function (RBF) as the kernel function and was trained with a 10-fold cross-validation scheme.

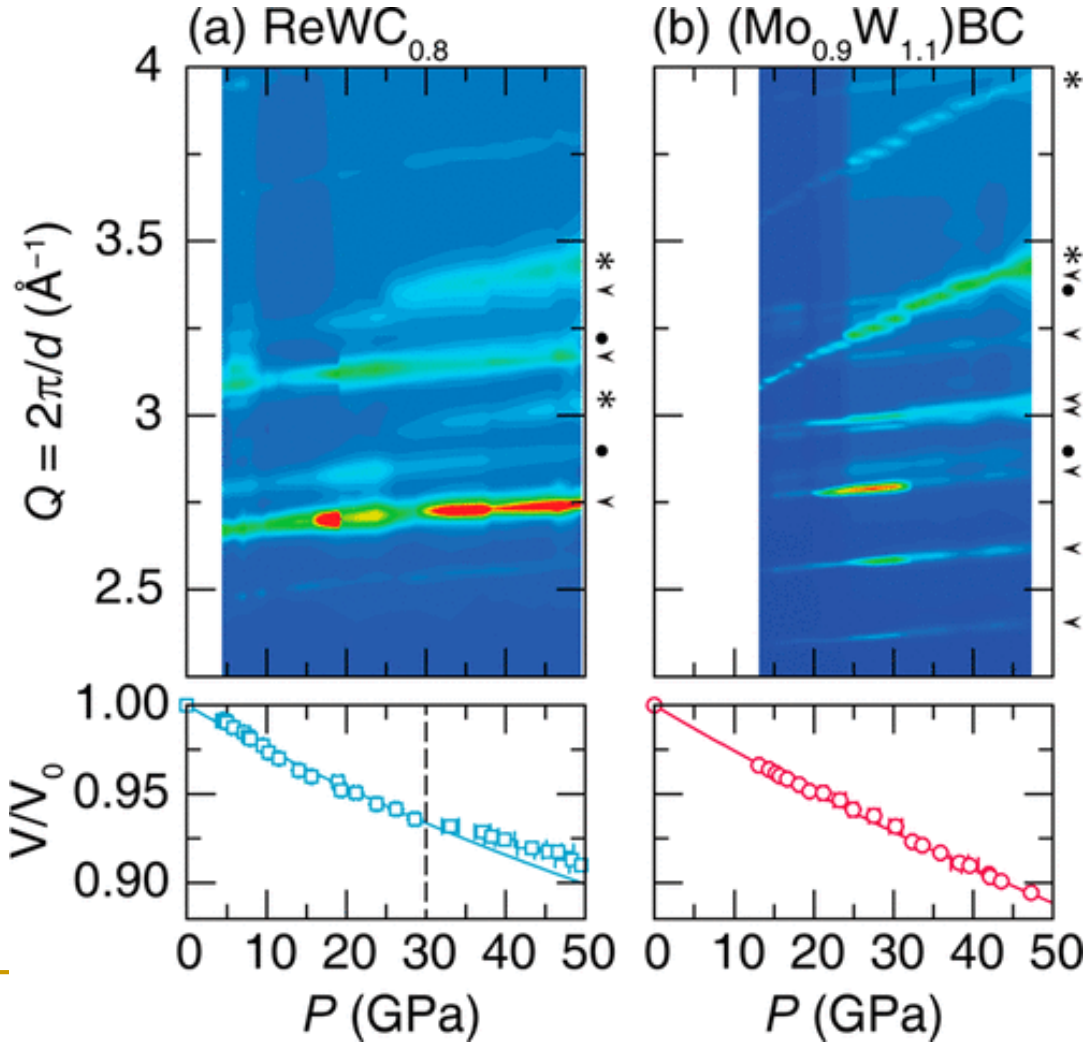
- 118287 compounds available in the PCD were assembled in a database. Their elastic moduli, bulk and shear, were predicted using ML.



实验验证



- $ReWC_{0.8}$
  - B_0 = 380 Gpa
  - Predicted: 398 GPa
- $Mo_{0.9}W_{1.1}BC$
  - B_0 = 373 GPa.
  - Predicted: 370 GPa



Scikit-Learn相关内容

https://scikit-learn.org/

https://sklearn.apachecn.org/

- **1.7. Gaussian Processes**
  - **gaussian_process.GaussianProcessRegressor**
- **1.4. Support Vector Machines**
  - **svm.SVC，NuSVC，LinearSVC**
  - **svm.SVR，NuSVR，LinearSVR**
  - **svm.OneClassSVM**

Reference:

- Bishop 2.3, 2.5, 6.1, 6.2, 6.4.1, 6.4.2, 6.4.5, 14.E；
- Elements 5.8。
- 深度学习 5.2.1。
- 实战 2, 6。
- 吴恩达 12。

扩展阅读：

- https://www.jiqizhixin.com/articles/2019-02-12-3
- https://www.jgoertler.com/visual-exploration-gaussian-processes

看得见的高斯过程：这是一份直观的入门解读.mht

- ieee67.pdf：No Free Lunch Theorems for Optimization
- https://cloud.tencent.com/developer/article/1033696

支持向量机入门简介.mht

- https://blog.csdn.net/v_july_v/article/details/7624837

支持向量机通俗导论（理解SVM的三层境界）.mht

- http://m.sohu.com/a/244701109_610519

三思而后行！Nature报道像化学家一样探索新反应的“AI”机器人.mht

- https://baijiahao.baidu.com/s?id=1594514807888322634

机器学习萌新必学的Top10算法.mht

- https://www.yantuo.com.cn/8010.html

《Nature》封面：化学家失业在即？不需要休息！无情的科研机器人横空出世！

谢谢大家！