法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，北风网和讲师拥有完全知识产权；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或者机构不得盗版、复制、仿造其中的创意和内容，我们保留一切通过法律手段追究违反者的权利。
- 课程详情请咨询
  - 微信公众号：北风教育
  - 官方网址：http://www.ibeifeng.com/

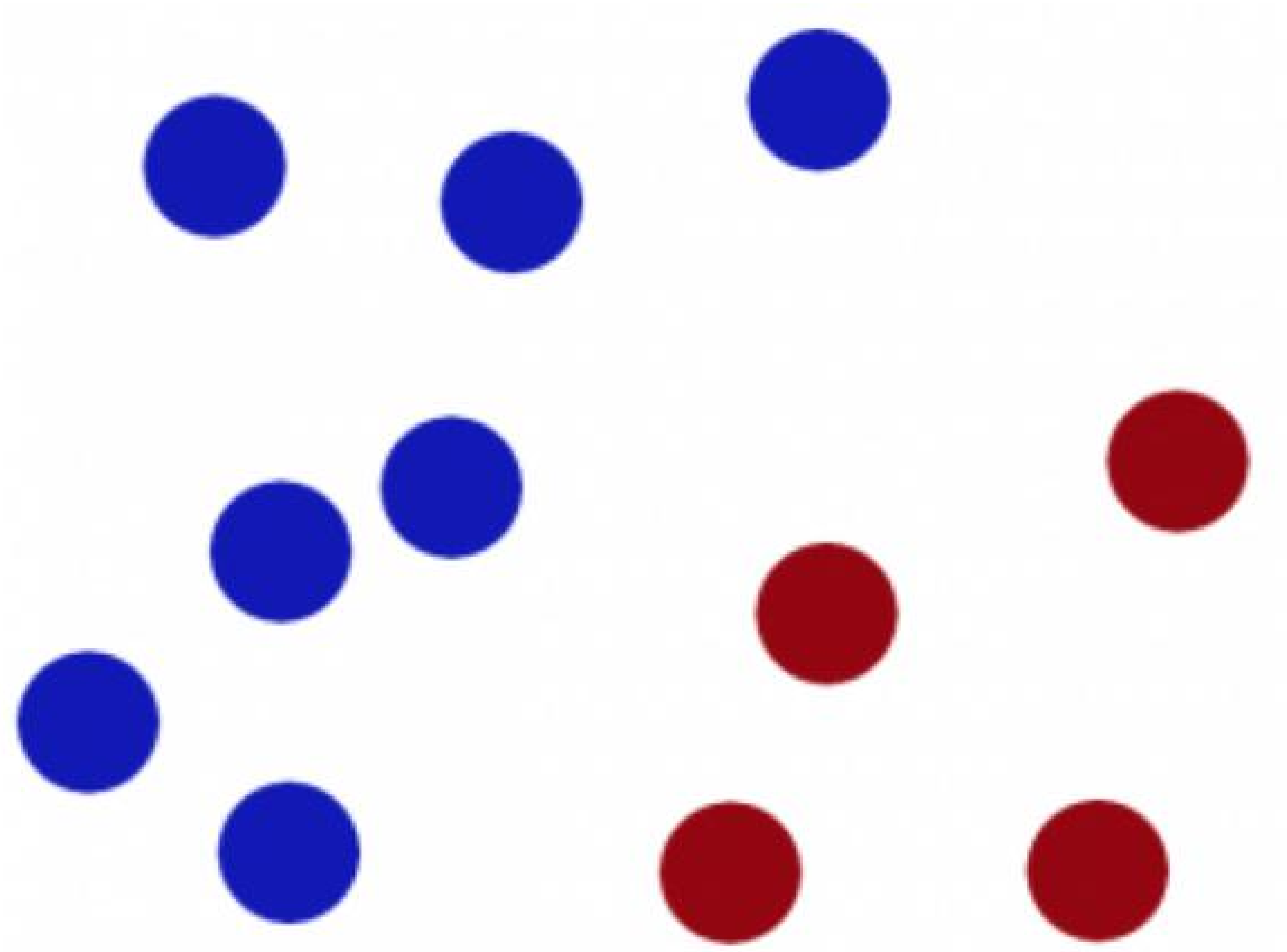
人工智能之机器学习

支持向量机（SVM）

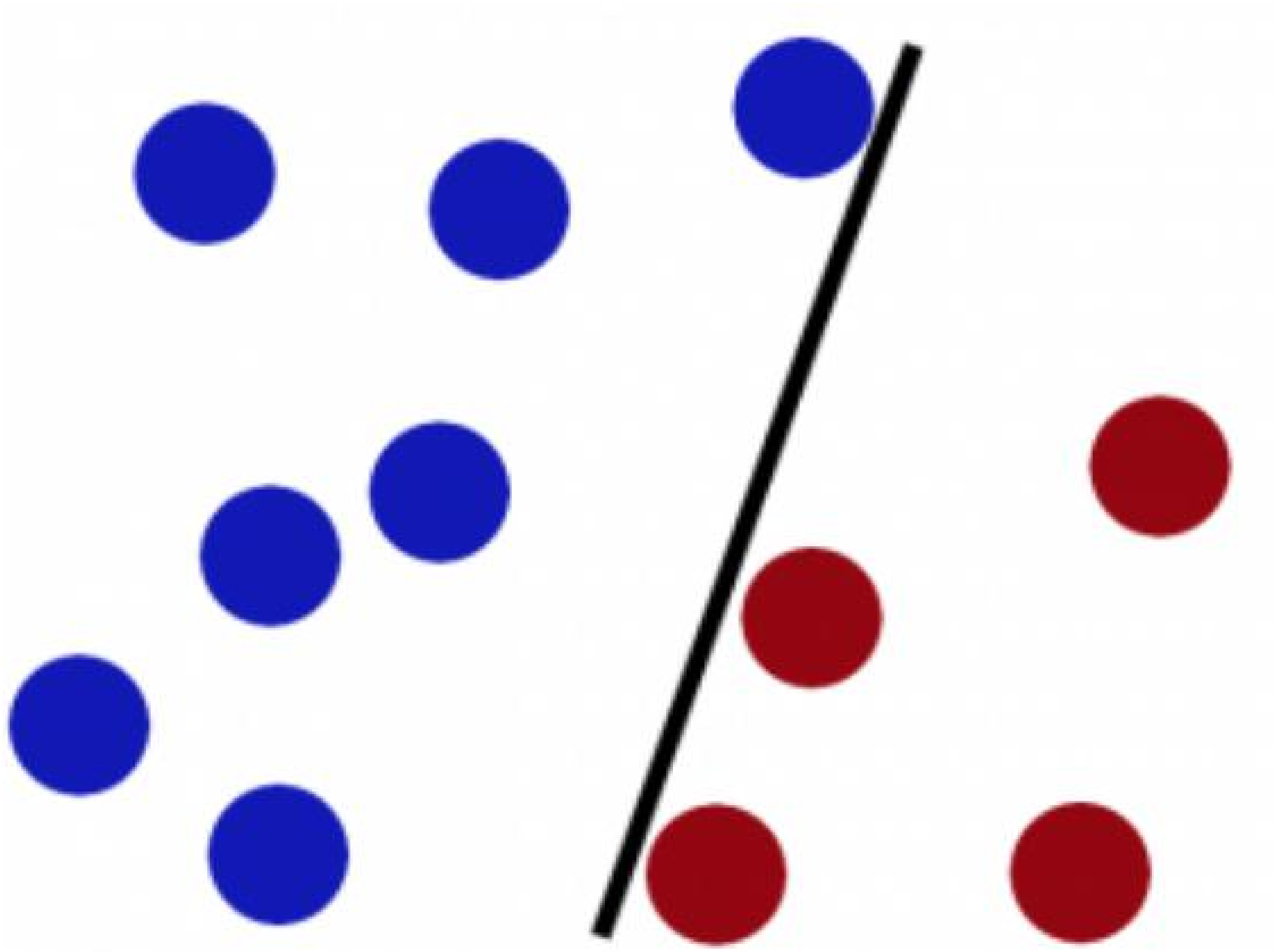
主讲人：赵翌臣

上海育创网络科技有限公司

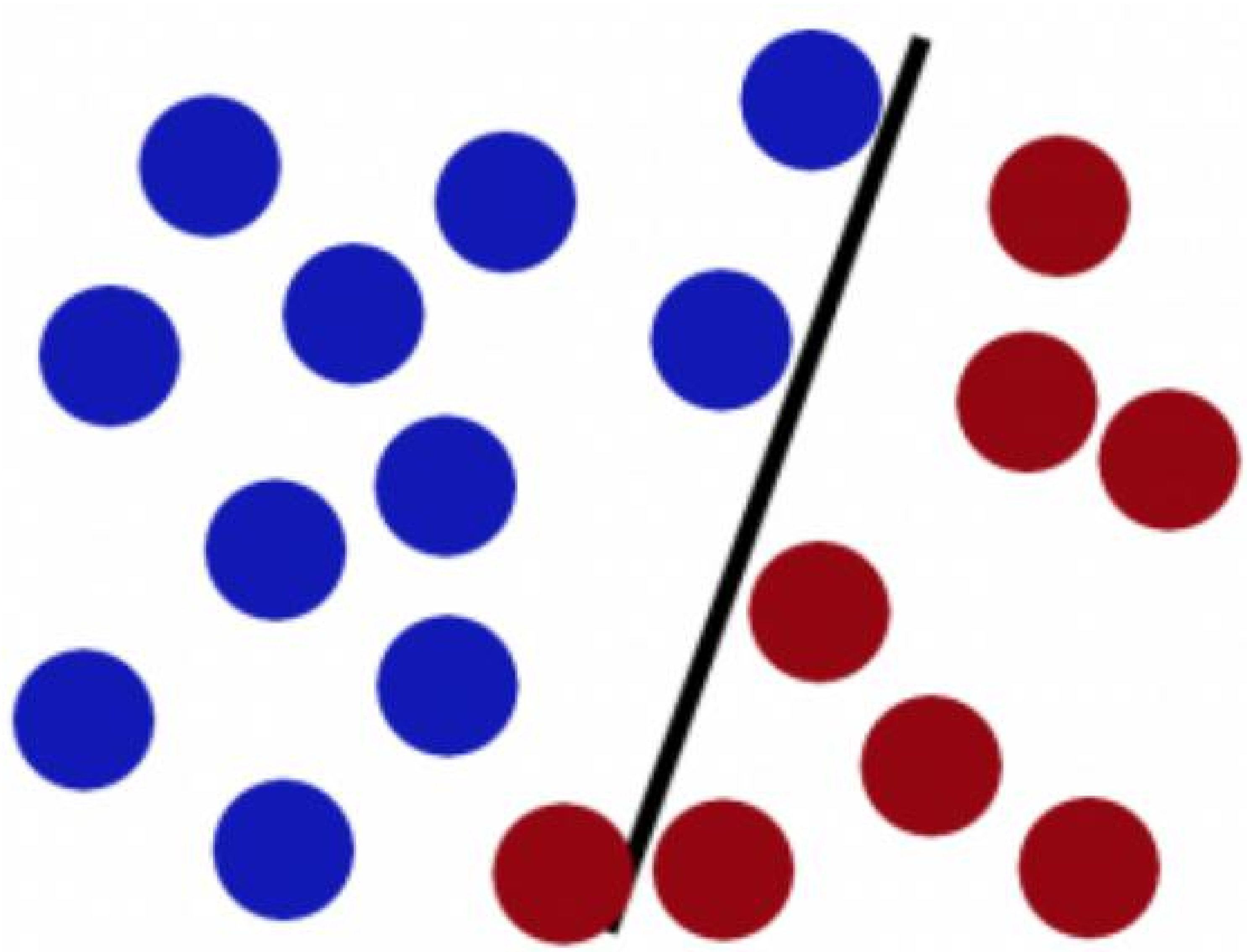
从一个故事出发



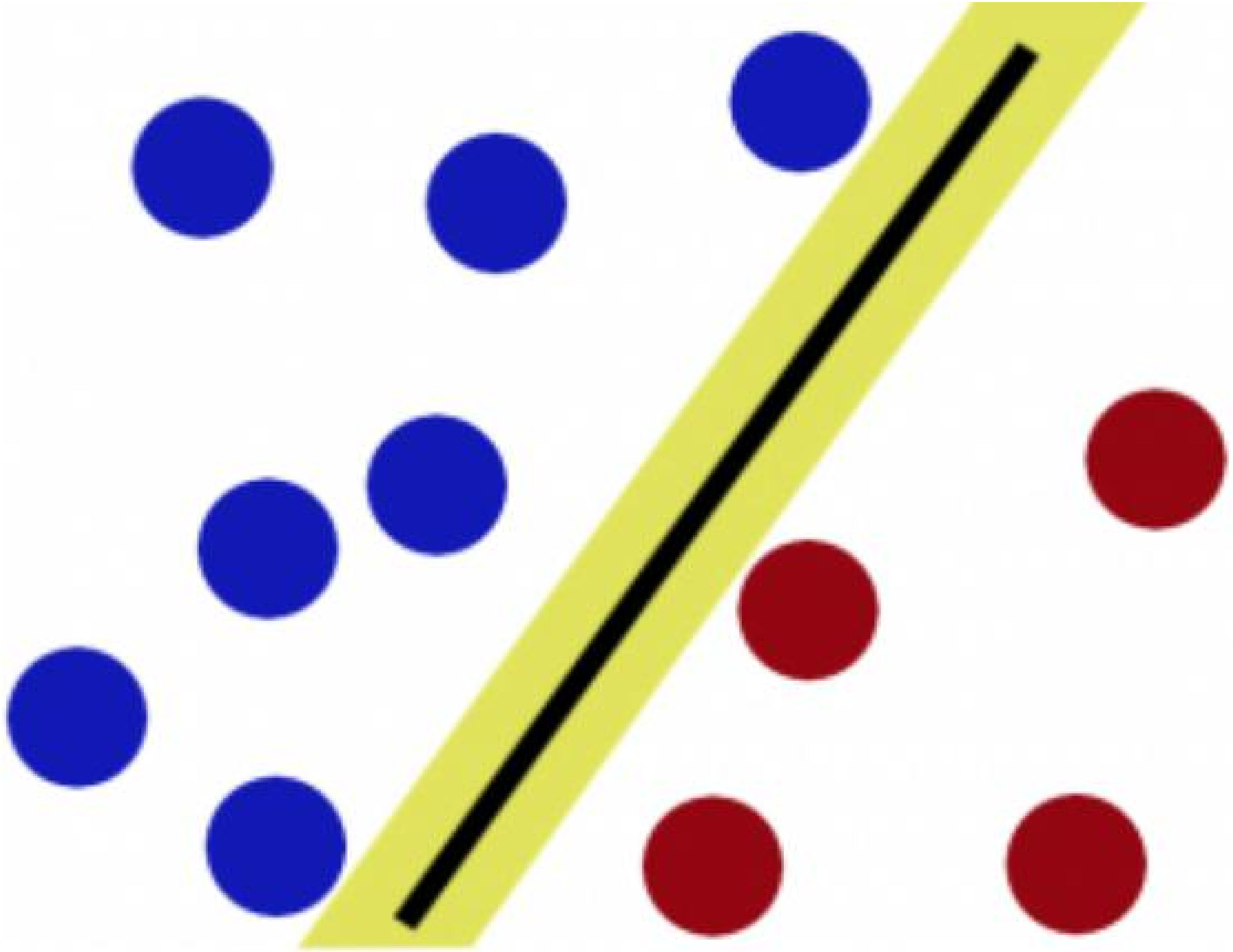
从一个故事出发



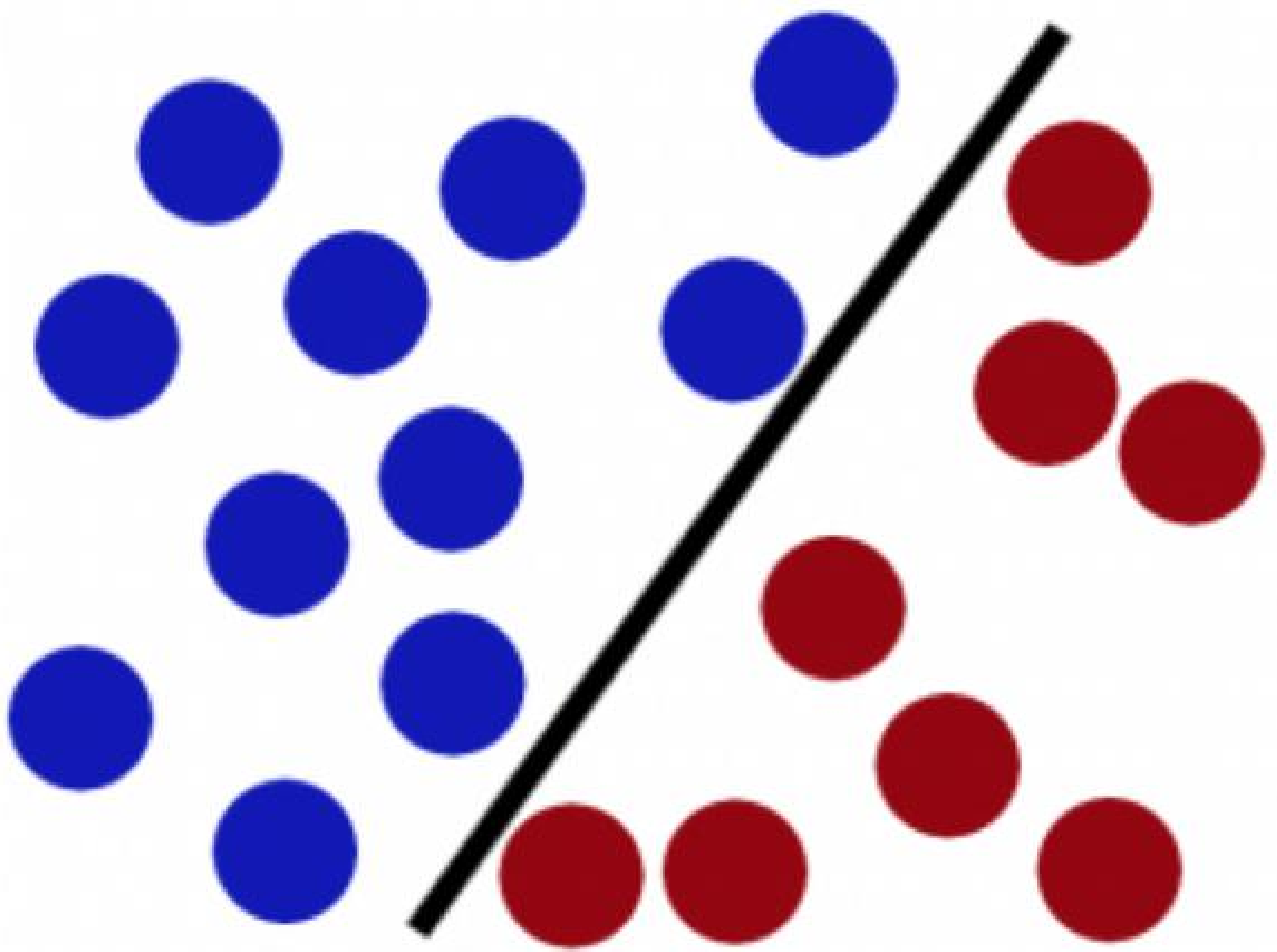
从一个故事出发



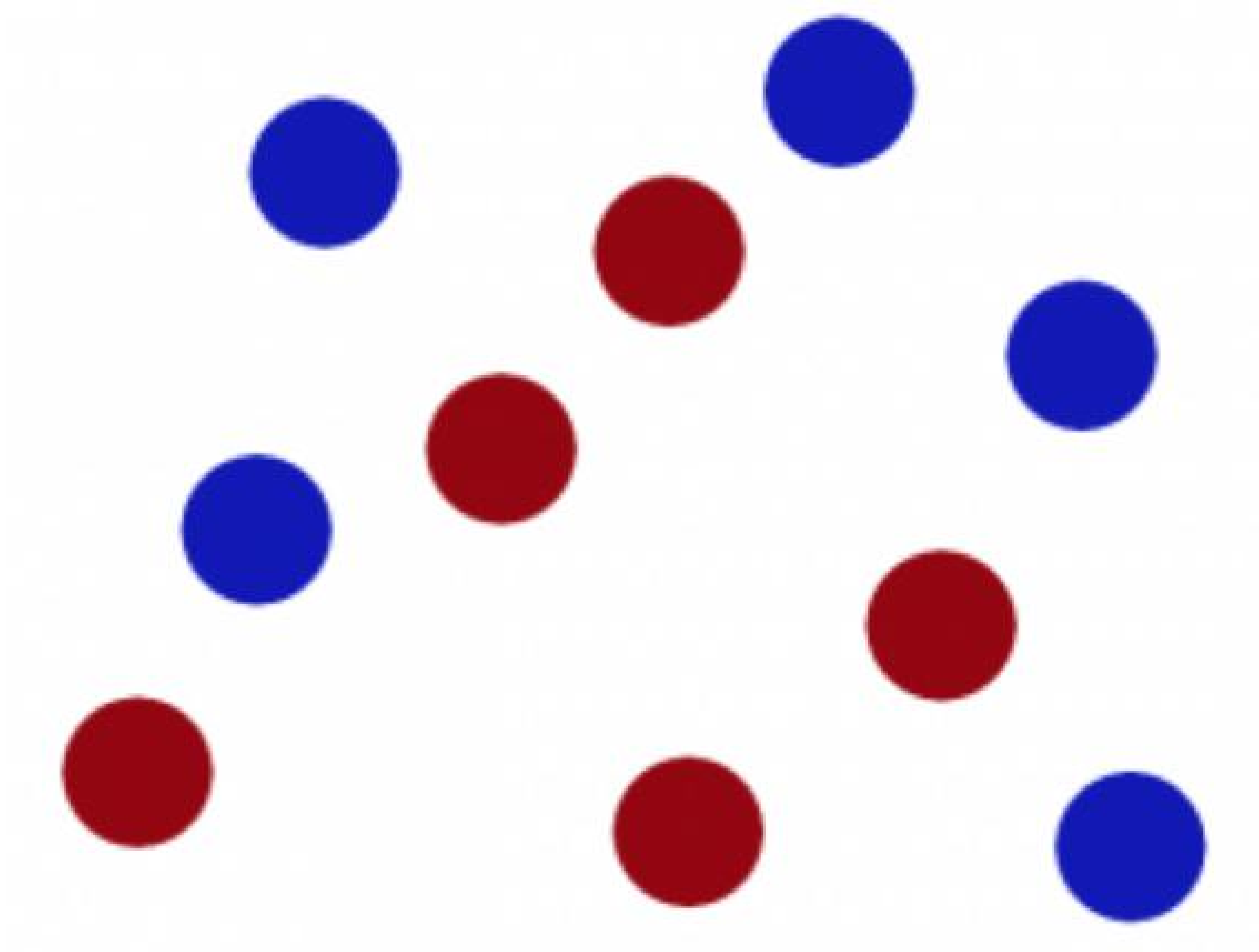
从一个故事出发



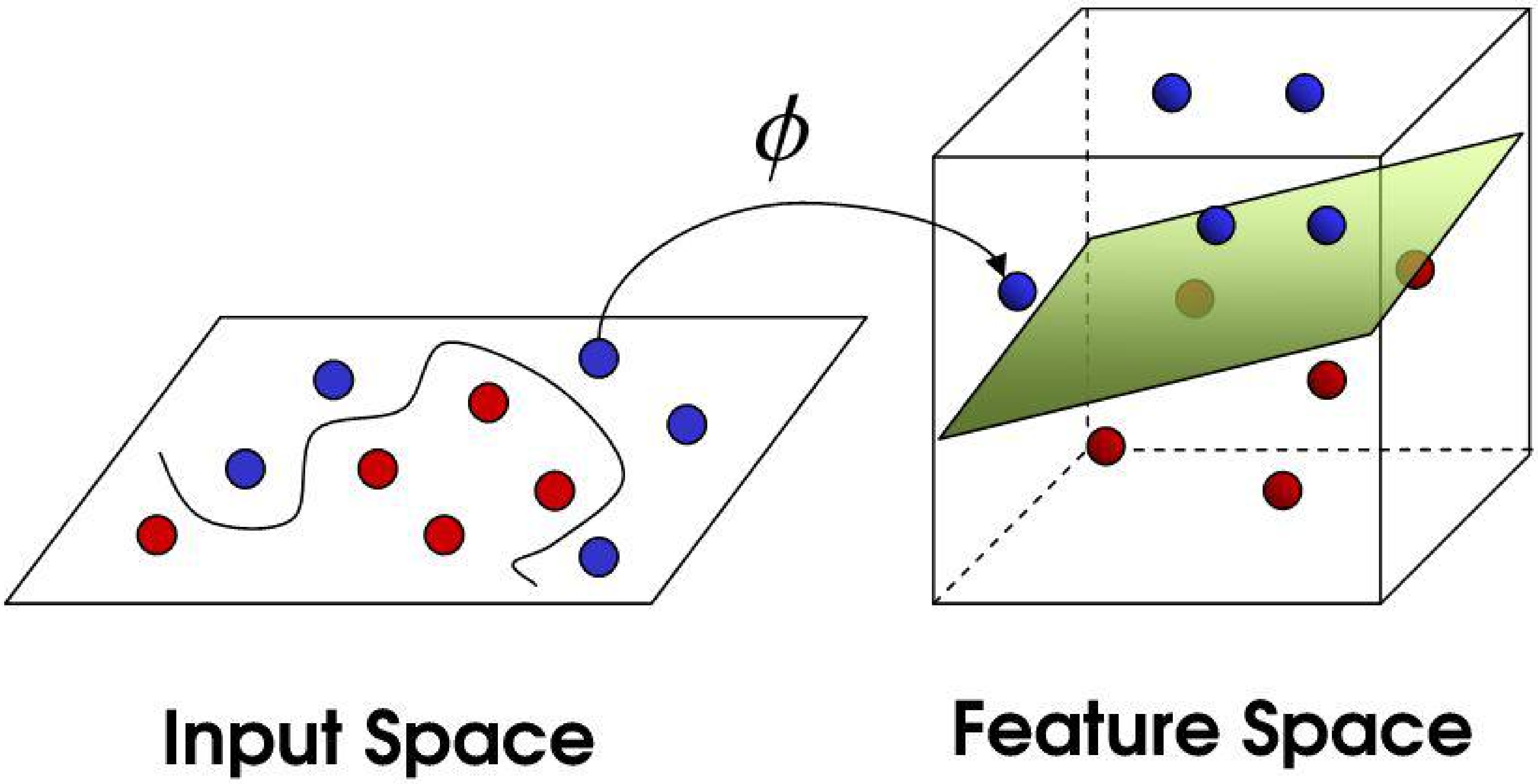
从一个故事出发



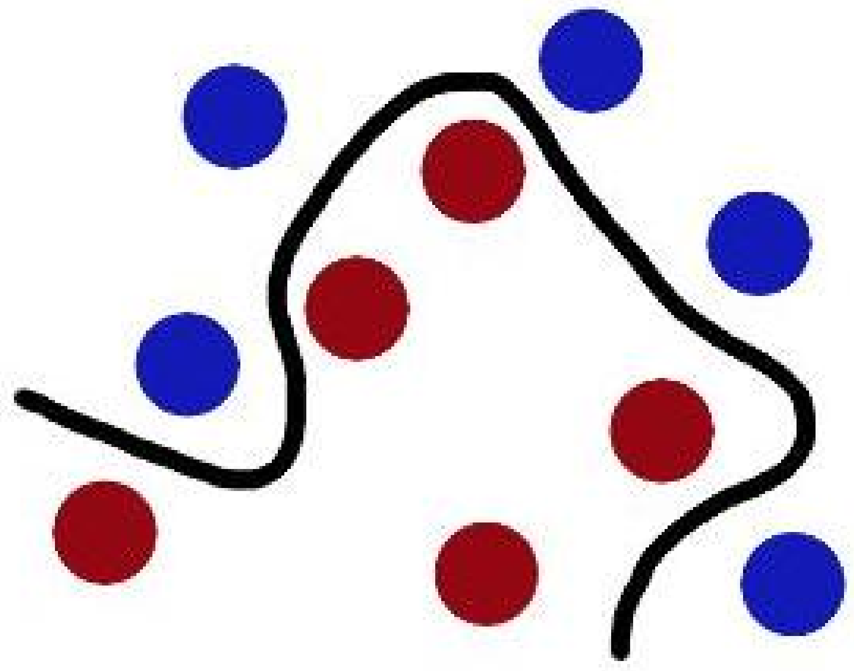
从一个故事出发



从一个故事出发



从一个故事出发



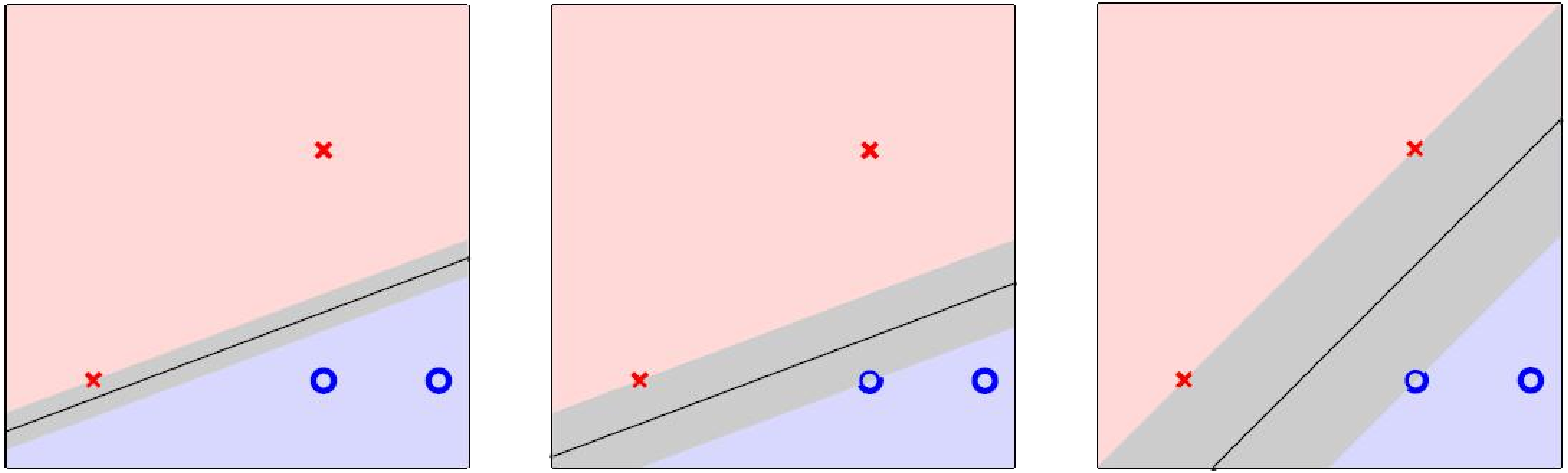
再之后，无聊的大人们把这些球叫做data，把棍子叫做classifier, 最大间隙trick叫做optimization，拍桌子叫做kernelling, 那张纸叫做hyperplane

http://bytesizebio.net/2014/02/05/support-vector-machines-explained-well/

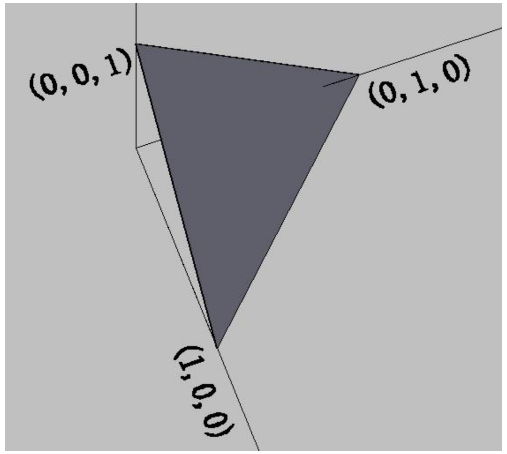
支持向量机（Support Vector Machine，1995）

- 介绍
  - 支持向量机是一种二类分类模型。
- 支持向量机学习方法包含构建由简至繁的模型
  - 线性可分支持向量机(linear support vector machine in linearly separable case)
  - 线性支持向量机(linear support vector machine)
  - 非线性支持向量机(non- linear support vector machine)

线性可分支持向量机——硬间隔最大化模型



- 目标函数：得到最“胖”的分割平面 $w^{*}\cdot x+\mathrm{b}^{*}=0$
- 分类决策函数：$f(x)=sign(w\cdot x+\mathrm{b})$

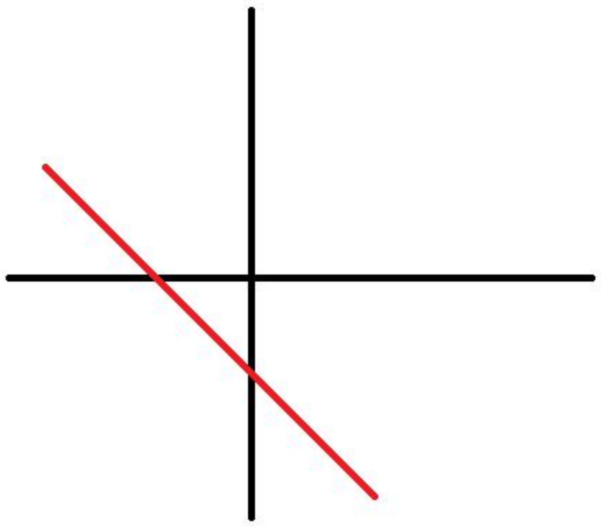


平面方程：

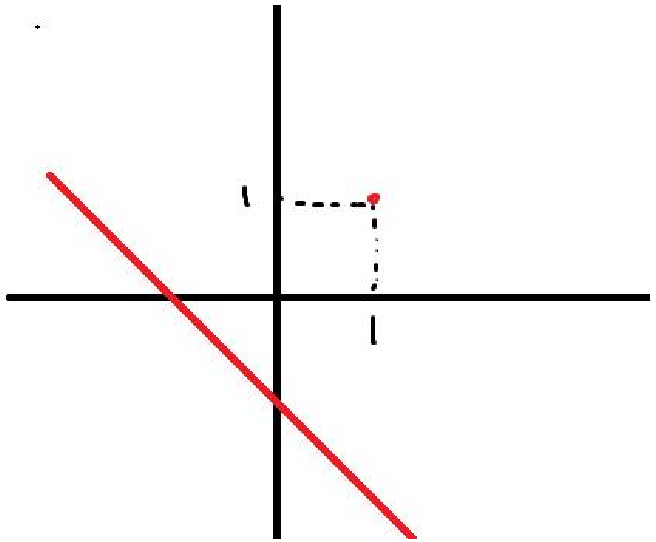
$\mathrm{Ax}+\mathrm{By}+\mathrm{Cz}+\mathrm{D}=0$

x+y+z=1

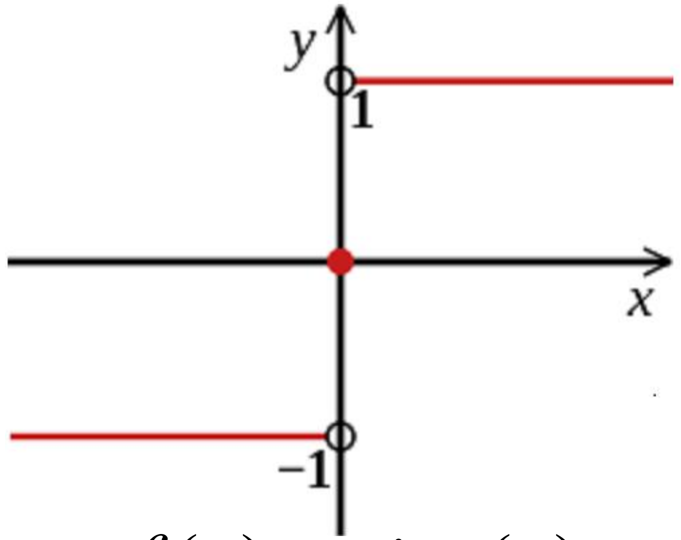
线性可分支持向量机——硬间隔最大化模型



$$x + y + 1 = 0$$



$$f(x) = sign(x + y + 1)$$



$$f(x) = sign(x)$$

平面方程：

$$\mathrm{Ax + By + Cz + D = 0}$$

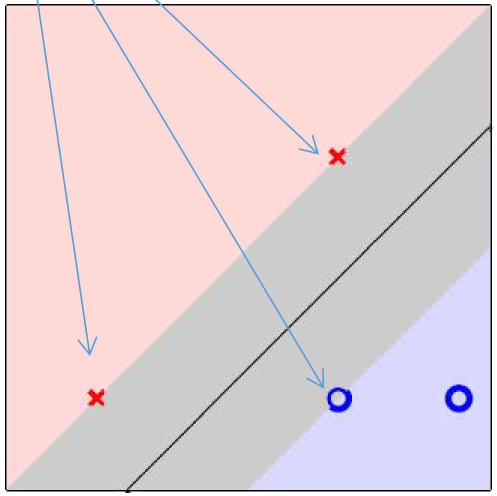
硬间隔最大化

支持向量候选人

$$\max_{b,w} \quad [\mathrm{margin}(b,\mathrm{w}) = \min_{\mathrm{i}=1,\dots,N} \frac{1}{\|w\|} y_i(w\cdot x_i + b)]$$

$$s.t. \quad every \quad y_i(w\cdot x_i + b) > 0$$

这里需要考虑会不会出现所有y(w·x+b)都>1的情况，如果能证明不会出现这种情况，那放松条件也就没有关系了。假设所有的y(w·x+b)都>1时，1/||w||为会取得最大值；如果让w和b等比缩小一些，并且仍满足y(w·x+b)都>1是完全可以做到的，且不违背条件，但此时的w已经缩小，之前的1/||w||也就不可能是最大值，矛盾！因此假设不成立，所以还是会有y(w·x+b)=1的情况出现的，这种放松条件是没有问题的



$$w\cdot x + \mathrm{b} = 0$$

$$\max_{b,w} \frac{1}{\|w\|}$$

$$s.t. \quad \min_{\mathrm{i}=1,\dots,N} y_{\mathrm{i}}(w\cdot x_{\mathrm{i}} + b) = 1$$

$$\max_{b,w} \frac{1}{\|w\|}$$

$$s.t. \quad \mathrm{every} \quad y_{\mathrm{i}}(w\cdot x_i + b) \geq 1$$

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

$$s.t. \quad y_i(w\cdot x_i + b) - 1 \geq 0 \quad , i = 1,2,\cdots,N$$

这是最终的目标函数

这里放松了条件，
但是能证明：是没有影响的

点到平面距离：

平面方程：

$$\mathrm{Ax} + \mathrm{By} + \mathrm{Cz} + \mathrm{D} = 0$$

$$\mathrm{d} = \left|\frac{\mathrm{Ax}_0 + \mathrm{By}_0 + \mathrm{Cz}_0 + \mathrm{D}}{\sqrt{\mathrm{A}^2 + \mathrm{B}^2 + \mathrm{C}^2}}\right|$$

编程——SVM求解

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

$$s.t. \quad y_i(w \cdot x_i + b) - 1 \geq 0 \quad ,i = 1,2,\cdots,N$$

例 7.1 数据与例 2.1 相同．已知一个如图 7.4 所示的训练数据集，其正例点是 $x_1=(3,3)^T$， $x_2=(4,3)^T$，负例点是 $x_3=(1,1)^T$，试求最大间隔分离超平面．

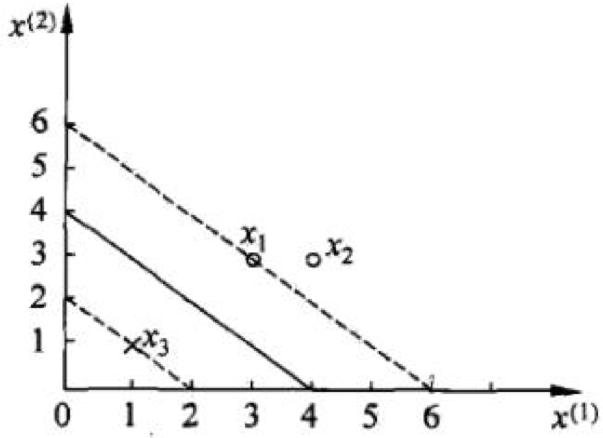


图 7.4 间隔最大分离超平面示例

转为对偶问题

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

$$s.t. \quad y_i(w \cdot x_i + b) - 1 \geq 0 \quad, i = 1,2,\cdots,N$$

- 目标函数本身是一个凸二次规划问题，能直接用现成的优化计算包求解，但我们可以有更高效的做法——求解对偶问题
- 在带有约束的最优化问题中，常常利用拉格朗日对偶性将原始问题转换为对偶问题，通过求解对偶问题得到原始问题的解，优点：
  - 一、对偶问题更容易求解
  - 二、自然引入核函数，进而推广到非线性分类问题

凸二次规划问题 => 拉格朗日极小极大问题

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

$$s.t. \quad y_i(w\cdot x_i + b) - 1 \geq 0,$$

$$i = 1,2,\cdots,N$$

构造拉格朗日函数

$$L(b,w,\alpha) =$$

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{N}\alpha_i\left(1 - y_i(w\cdot x_i + b)\right)$$

广义化：原始最优化问题转化为广义拉格朗日函数的极小极大问题

$$\min_{b,w}\left(\max_{all\ \ \alpha_i\geq 0} L(b,w,\alpha)\right)$$

若（）>0：违背了 => 那么做最大化动作的时候，就会令α->∞，整体->∞，那么再做最小化的动作也无济于事，最小值都是无穷大，那么求解也就没有意义了，也就说明，如果存在违背条件的情况，是没有解的，要想有解，必须让每一个约束都不能违背，
若（）<0：没有违背 => 那么做最大化动作时，就会令α=0，整体便最大，其实这个α对应的样本就是无名小卒，肯定不是支持向量；
若（）=0：那么α生死未卜，α>0对应的点就是支持向量
所以说这样转换是没问题的。

拉格朗日极小极大问题 =>对偶问题

$$\underbrace{\min_{b,w}\left(\max_{all\ \alpha_i\ge 0} L(b,w,\alpha)\right)}_{\text{equiv.to primal } SVM} \overset{?}{=} \underbrace{\max_{all\ \alpha_i\ge 0}\left(\min_{b,w} L(b,w,\alpha)\right)}_{Lagrange\ \ dual}$$

好处：对偶问题中，第一步是对b、w的最优化问题，是没有约束条件的，易解

定理4：
≥： weak duality
=： strong duality，true for QP if
convex primal
feasible primal(true if separable)
linear constraints

C.1定理：拉格朗日函数的极小极大问题的最优值 ≥ 拉格朗日函数的极大极小问题的最优值
C.1推论：如果拉格朗日函数的极小极大问题的最优值 = 拉格朗日函数的极大极小问题的最优值，那么原始问题和对偶问题的最优解相同
C.2定理：如果满足定理4的等式条件，则满足C.1的推论的前提
C.3定理：对偶问题最优解和原始问题最优解相等 <=> 满足KKT条件

因此：

$$\underbrace{\min_{b,w}\left(\max_{all\ \alpha_i\ge 0} L(b,w,\alpha)\right)}_{\text{equiv.to primal } SVM} = \underbrace{\max_{all\ \alpha_i\ge 0}\left(\min_{b,w} L(b,w,\alpha)\right)}_{Lagrange\ \ dual}$$

在强对偶条件下，原始问题和对偶问题的最优值相等，可以用解对偶问题代替解原始问题

对偶问题最优解和原始问题最优解相等 <=> 满足KKT条件（触发了某个神秘的机关）

简化对偶问题

$$\max_{all\ \ \alpha_i \geq 0}\left(\min_{b,w}\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\alpha_i\left(1-y_i(w\cdot x_i+b)\right)\right)$$

$$\frac{\partial L(b,w,\alpha)}{\partial b}=0=-\sum_{i=1}^{N}\alpha_i y_i \qquad ① \qquad \frac{\partial L(b,w,\alpha)}{\partial w}=0=w-\sum_{i=1}^{N}\alpha_i y_i x_i \qquad ②$$

代入①

$$\rightarrow \max_{all\ \ \alpha_i\geq 0,\sum\alpha_i y_i=0}\left(\min_{w}\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\alpha_i\left(1-y_i(w\cdot x_i)\right)\right)$$

代入②

$$\rightarrow \max_{all\ \ \alpha_i\geq 0,\sum\alpha_i y_i=0,\ \ w=\sum\alpha_i y_i x_i}\left(\min_{w}\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\alpha_i-\|w\|^2\right)$$

$$=\max_{all\ \ \alpha_i\geq 0,\sum\alpha_i y_i=0,\ \ w=\sum\alpha_i y_i x_i}\left(-\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i x_i\right\|^2+\sum_{i=1}^{N}\alpha_i\right)$$

求解对偶问题

$$\max_{all\ \ \alpha_i\ge 0,\sum\alpha_i y_i=0,\ \ w=\sum\alpha_i y_i x_i}\left(-\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i x_i\right\|^2+\sum_{i=1}^{N}\alpha_i\right)$$

standard hard-margin *SVM* *dual*:

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\ \ \sum_{i=1}^{N}\alpha_i y_i=0$$

$$\alpha_i\ge 0,\ \ i=1,2,\cdots,N$$

(convex)QP of N variables & N+1 constraints $\longrightarrow \boldsymbol{\alpha}^*$

KTT条件（触发的机关）

$$\underbrace{\min_{b,w}\left(\max_{all\ \ \alpha_i\geq 0}\frac{1}{2}\|w\|^2+\sum_{i=1}^{N}\alpha_i\big(1-y_i(w\cdot x_i+b)\big)\right)}_{primal\ \ SVM} \qquad \underbrace{\max_{all\ \ \alpha_i\geq 0,\sum\alpha_i y_i=0,\ \ w=\sum\alpha_i y_i x_i}\left(-\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i x_i\right\|^2+\sum_{i=1}^{N}\alpha_i\right)}_{Lagrange\ \ dual}$$

$$if\quad primal\text{-}dual\quad optimal(b^*,w^*,\alpha^*),$$

$$primal\quad feasible: y_i(w^*\cdot x_i+b^*)\geq 1$$

$$dual\quad feasible: \alpha_i^*\geq 0$$

$$dual\text{-}inner\quad optimal: \sum\alpha_i^* y_i=0,\quad w^*=\sum\alpha_i^* y_i x_i$$

$$primal\text{-}inner\quad optimal: \alpha_i^*(1-y_i(w^*\cdot x_i+b^*))=0$$

使用α*，通过KKT条件求解w*和b*

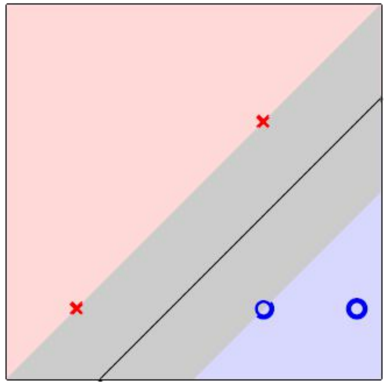
$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i=\sum_{SV}\alpha_i^* y_i x_i$$

$$b^*=y_i-w^*\cdot x_i\quad with\quad any\quad SV(x_i,y_i)$$

硬间隔最大化的最终模型

分割平面：$w^{*}\cdot x+\mathrm{b}^{*}=0$

分类决策函数：$f(x)=sign(w^{*}\cdot x+\mathrm{b}^{*})$



分割平面：$\sum_{SV}\alpha_{\mathrm{i}}^{*}y_{i}x_{i}\cdot x+\mathrm{b}^{*}=0$

分类决策函数：$f(x)=sign(\sum_{SV}\alpha_{\mathrm{i}}^{*}y_{i}x_{i}\cdot x+\mathrm{b}^{*})$

NOTE :

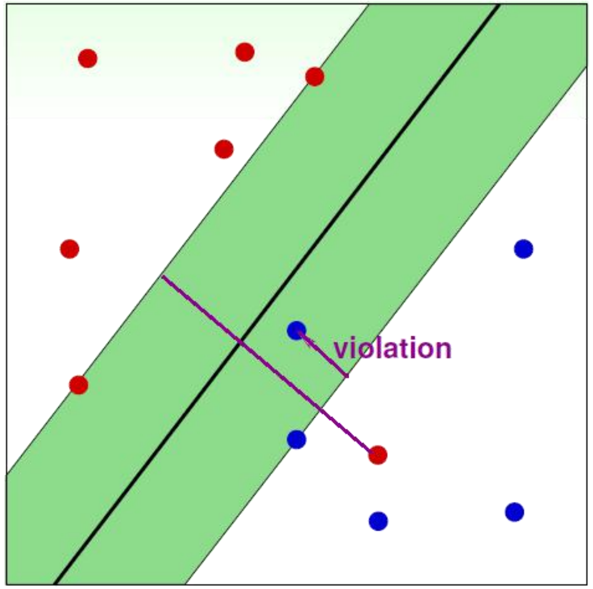
call $\alpha_{i}>0$ $examples(x_{i},y_{i})$ SV

$SV(\alpha_{i}>0)\subseteq \mathrm{point}(on\quad boundary)$

原始问题 VS 对偶问题

Primal Hard-Margin SVM	Dual Hard-Margin SVM
$\min\limits_{b,w} \quad \frac{1}{2}\lVert w\rVert^2$ $s.t. \quad y_i(w\cdot x_i+b)-1\geq 0 \quad ,i=1,2,\cdots,N$	$\min\limits_{\boldsymbol{\alpha}} \frac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i\cdot x_j)-\sum\limits_{i=1}^{N}\alpha_i$ $s.t. \quad \sum\limits_{i=1}^{N}\alpha_i y_i=0$ $\alpha_i\geq 0, \quad i=1,2,\cdots,N$
n+1个变量，N个约束，当n+1较小时适用	N个变量，N+1个约束，当N较小时适用

线性支持向量机——软间隔最大化模型



目标函数：有容错能力的最"胖"腰围的分割平面 $w^* \cdot x + b^* = 0$

分类决策函数： $f(x) = sign(w^* \cdot x + b^*)$

平面方程：

$Ax + By + Cz + D = 0$

软间隔最大化

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

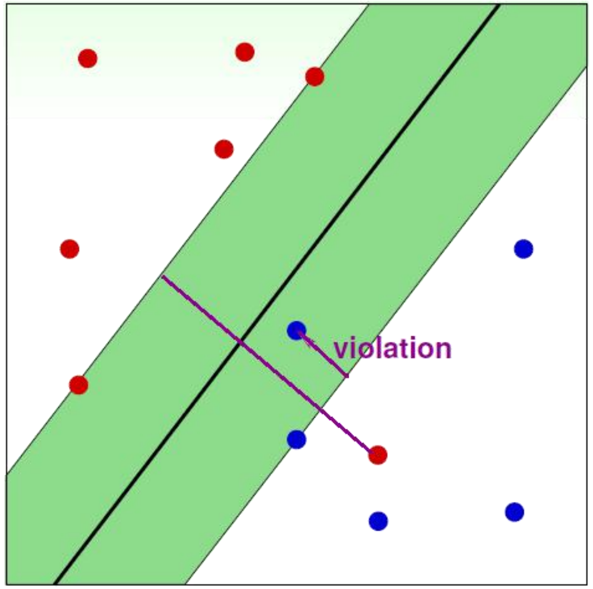
$$s.t. \quad y_i(w \cdot x_i + b) - 1 \geq 0 \quad , i = 1,2,\cdots,N$$

引入松弛变量

$$\min_{b,w,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \,, \quad i = 1,2,\cdots,N$$

$$\xi_i \geq 0, \quad i = 1,2,\cdots,N$$



C: trade-off of large margin & margin violation
large C: want less margin violation
small C: want large margin

QP of n + 1 + N variables, 2N constraints

对偶问题

$$\min_{b,w,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$s.t. \quad y_i(w\cdot x_i + b) \ge 1-\xi_i\,, \quad i=1,2,\cdots,N$$

$$\xi_i \ge 0, \quad i=1,2,\cdots,N$$

构造拉格朗日函数

$$L(b,w,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\cdot\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\big(1-\xi_i - y_i(w\cdot x_i + b)\big) + \sum_{i=1}^{N}\beta_i(-\xi_i)$$

广义化

$$\min_{b,w}\left(\max_{all\ \ \alpha_i\ge 0} L(b,w,\xi,\alpha,\beta)\right)$$

对偶问题

$$\max_{\alpha_i\ge 0,\ \ \beta_i\ge 0}\left(\min_{b,w,\xi} \quad L(b,w,\xi,\alpha,\beta)\right)$$

简化对偶问题

$$\max_{\alpha_i \ge 0,\ \beta_i \ge 0}\left(\min_{b,w,\xi}\ \frac{1}{2}\|w\|^2 + C\cdot\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\left(1-\xi_i - y_i(w\cdot x_i + b)\right) + \sum_{i=1}^{N}\beta_i(-\xi_i)\right)$$

$$\frac{\partial L}{\partial \xi_i} = 0 = \mathrm{C} - \alpha_i - \beta_i \qquad \beta_i = \mathrm{C} - \alpha_i \qquad 0 \le \alpha_i \le \mathrm{C}$$

$$\xrightarrow{\text{代入}} \max_{0\le\alpha_i\le \mathrm{C},\ \beta_i = C-\alpha_i}\left(\min_{b,w}\ \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N}\alpha_i\left(1 - y_i(w\cdot x_i + b)\right)\right)$$

$$\frac{\partial L}{\partial b} = 0 = -\sum_{i=1}^{N}\alpha_i y_i \qquad \frac{\partial L}{\partial w_i} = 0 = w_i - \sum_{i=1}^{N}\alpha_i y_i x_i$$

$$\xrightarrow{\text{代入}} \max_{0\le\alpha_i\le \mathrm{C},\ \beta_i = C-\alpha_i,\ \sum\alpha_i y_i = 0,\ w=\sum\alpha_i y_i x_i}\left(-\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i x_i\right\|^2 + \sum_{i=1}^{N}\alpha_i\right)$$

求解对偶问题

$$\max_{0\le\alpha_i\le C,\ \beta_i=C-\alpha_i,\ \sum\alpha_i y_i=0,\ w=\sum\alpha_i y_i x_i}\left(-\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i x_i\right\|^2+\sum_{i=1}^{N}\alpha_i\right)$$

standard soft-margin *SVM* *dual*:

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$0\le\alpha_i\le C,\quad i=1,2,\cdots,N$$

(convex)QP of N variables & 2N+1 constraints $\longrightarrow \boldsymbol{\alpha}^*$

谁是支持向量?

$$\text{non}\quad \text{SV}(\alpha_i = 0):$$

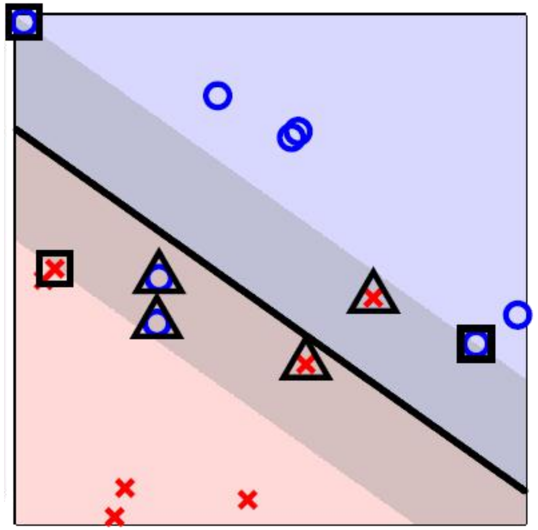
$$\xi_i = 0, away \quad from/on \quad fat \quad boundary$$

$$\square free \quad SV(0 < \alpha_i < C):$$

$$\xi_i = 0, on \quad fat \quad boundary, locates \quad b^*$$

$$\triangle bounded \quad \text{SV}(\alpha_i = C):$$

$$\xi_i = violation \quad amount, violate/on \quad fat \quad boundary$$

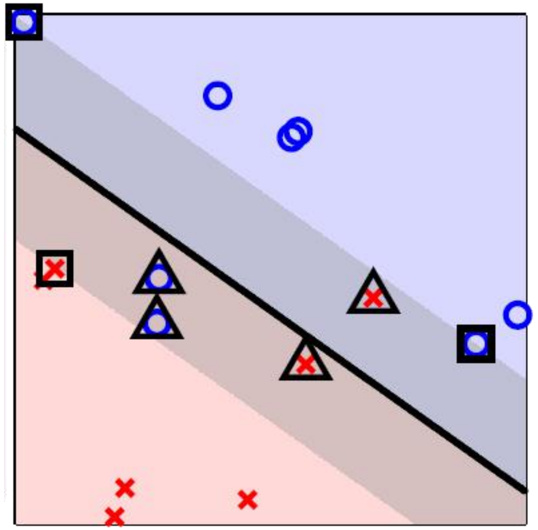


KKT条件（触发的机关）

$$\alpha_i^*(1-y_i(w^*\cdot x_i+b^*))=0$$

$$\to w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i = \sum_{SV}\alpha_i^* y_i x_i$$

$$\to b^* = y_j - w^*\cdot x_j \quad with \quad any \quad SV(x_j, y_j)$$



non $SV(\alpha_i = 0)$:

$\xi_i = 0, away \quad from/on \quad fat \quad boundary$

□ $free \quad SV(0<\alpha_i<C)$:

$\xi_i = 0, on \quad fat \quad boundary, locates \quad b^*$

△ $bounded \quad SV(\alpha_i = C)$:

$\xi_i = violation \quad amount, violate/on \quad fat \quad boundary$

从硬间隔到软间隔

$$\alpha_i^*(1-\xi_i-y_i(w^*\cdot x_i+b^*))=0$$

$$\to w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i = \sum_{SV}\alpha_i^* y_i x_i$$

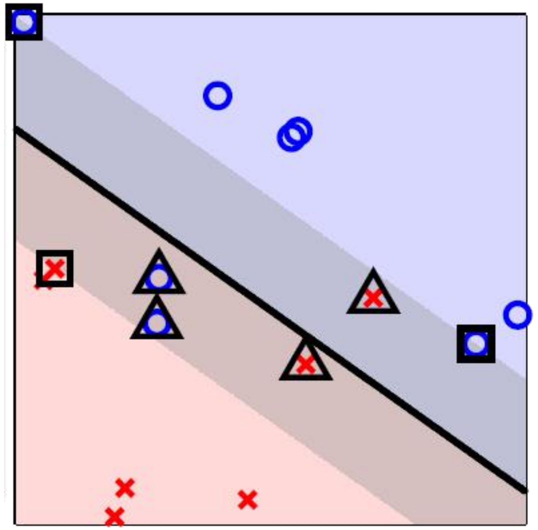
$$\to b^* = y_j - w^*\cdot x_j \quad with \quad any \quad free \quad SV(x_j, y_j)$$

b 的解可能不唯一， 所以实际计算时可以取在所有符合条件的样本点上的平均值.

软间隔最大化的最终模型

分割平面：$w^{*}\cdot x+\mathrm{b}^{*}=0$

分类决策函数：$f(x)=sign(w^{*}\cdot x+\mathrm{b}^{*})$

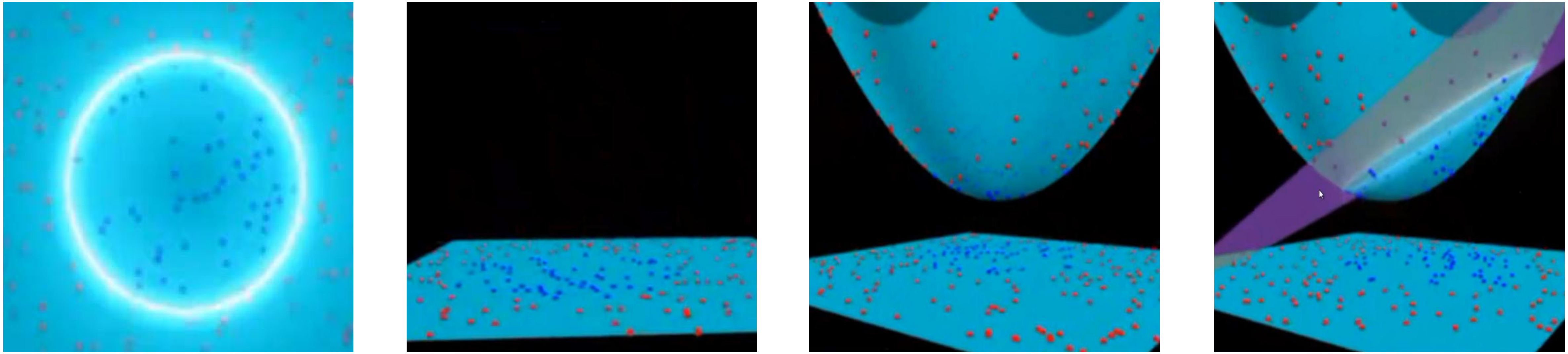


分割平面：$\sum_{SV}\alpha_{\mathrm{i}}^{*}y_{i}x_{i}\cdot x+\mathrm{b}^{*}=0$

分类决策函数：$f(x)=sign(\sum_{SV}\alpha_{\mathrm{i}}^{*}y_{i}x_{i}\cdot x+\mathrm{b}^{*})$

非线性支持向量机——使用核函数的软间隔最大化模型

问题描述：给定训练数据集$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$其中，x是向量，y=+1或-1。再假设训练数据集是线性不可分的。试学习一个SVM模型。



那么，如何求解这个间隔最大的超平面呢？

输入空间与特征空间

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C, \quad i = 1,2,\cdots,N$$

转换

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(x_i)\cdot\phi(x_j) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C, \quad i = 1,2,\cdots,N$$

比如：在输入空间上$x = (\frac{1}{2}, x_1, x_2, \cdots, x_d)$是d+1维的

映射到特征空间

$$\phi(x) = (\frac{1}{2}, x_1, x_2, \cdots, x_d, \quad x_1^2, x_1x_2, \cdots, x_1x_d, \quad x_2x_1, x_2^2, \cdots, x_2x_d, \quad \cdots, x_d^2)$$

思考：缺点是?

在高维空间使用线性支持向量机求软间隔最大的分离超平面。这样做可以成功，但是计算复杂度增加急增。如何解决呢?

多项式核函数原理

$$\phi(x) = (\frac{1}{2}, x_1, x_2, \cdots, x_d, \quad x_1{}^2, x_1x_2, \cdots, x_1x_d, \quad x_2x_1, x_2{}^2, \cdots, x_2x_d, \quad \cdots, x_d{}^2)$$

$$\phi(x') = (\frac{1}{2}, x_1', x_2', \cdots, x_d, \quad x_1{}^2, x_1x_2, \cdots, x_1x_d, \quad x_2x_1, x_2{}^2, \cdots, x_2x_d, \quad \cdots, x_d{}^2)$$

$$\phi(x) \cdot \phi(x') = \frac{1}{4} + \sum_{i=1}^{d} x_i x_i' + \sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j x_i' x_j'$$

$$= \frac{1}{4} + \sum_{i=1}^{d} x_i x_i' + \sum_{i=1}^{d} x_i x_i' \sum_{j=1}^{d} x_j x_j'$$

$$= \frac{1}{4} + x \cdot x' + (x \cdot x')(x \cdot x') \equiv \mathrm{K}(x, x')$$

PS：多项式核函数

$$(x \cdot z + \frac{1}{2})^2$$

$$= (x \cdot z)^2 + x \cdot z + \frac{1}{4}$$

$$= (x \cdot z)(x \cdot z) + (x \cdot z) + \frac{1}{4}$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j z_i z_j + \sum_{i=1}^{d} x_i z_i + \frac{1}{4} = \phi(x) \cdot \phi(z)$$

其中$\phi(x) = (\frac{1}{2}, x_1, x_2, \cdots, x_d, \quad x_1^{\ 2}, x_1 x_2, \cdots, x_1 x_d, \quad x_2 x_1, x_2^{\ 2}, \cdots, x_2 x_d, \quad \cdots, x_d^{\ 2})$

核函数+软间隔最大化的最终模型

分割平面：$w^* \cdot x + \mathrm{b}^* = 0$

分类决策函数：$f(x) = sign(w^* \cdot x + \mathrm{b}^*)$

$$w^* = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_{\mathrm{i}}^* y_i x_i = \sum_{SV} \alpha_{\mathrm{i}}^* y_i x_i$$

$$b^* = y_{\mathrm{j}} - w^* \cdot x_j \quad with \quad any \quad \text{free} \quad SV(x_j, y_j)$$

分割平面：$\sum_{SV} \alpha_{\mathrm{i}}^* y_i x_i \cdot x + \mathrm{b}^* = 0$

分类决策函数：$f(x) = sign(\sum_{SV} \alpha_{\mathrm{i}}^* y_i x_i \cdot x + \mathrm{b}^*)$

其中，$b^* = y_{\mathrm{j}} - \sum_{SV} \alpha_{\mathrm{i}}^* y_i x_i \cdot x_{\mathrm{j}}$ b 的解可能不唯一

分割平面：$w^* \cdot \phi(x) + \mathrm{b}^* = 0$

分类决策函数：$f(x) = sign(w^* \cdot \phi(x) + \mathrm{b}^*)$

$$w^* = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_{\mathrm{i}}^* y_i \phi(x_i) = \sum_{SV} \alpha_{\mathrm{i}}^* y_i \phi(x_i)$$

$$b^* = y_{\mathrm{j}} - w^* \cdot \phi(x_j) \quad with \quad any \quad \text{free} \quad SV(\phi(x_j), y_j)$$

分割平面：$\sum_{SV} \alpha_{\mathrm{i}}^* y_i \mathrm{K}(x, x_i) + \mathrm{b}^* = 0$

分类决策函数：$f(x) = sign(\sum_{SV} \alpha_{\mathrm{i}}^* y_i \mathrm{K}(x, x_i) + \mathrm{b}^*)$

其中，$b^* = y_{\mathrm{j}} - \sum_{SV} \alpha_{\mathrm{i}}^* y_i K(x_i, x_j)$ b 的解可能不唯一

常用核函数

- 线性核函数
  - 公式：$K(x,z)=x\cdot z$
  - 优点：安全、快速、容易解释；　缺点：无法处理非线性问题
- 多项式核函数
  - 公式：$K(x,z)=(\mathrm{m}x\cdot z+\mathrm{n})^p$
  - 优点：能处理非线性问题；　缺点：p较大时数值容易溢出、参数较多不宜找到好的组合
- 高斯核函数（属于rbf核函数）
  - 公式：$K(x,z)=\exp(-\frac{\|x-z\|^2}{2\sigma^2})$
  - 优点：能做出很复杂的边界、数值不容易溢出、只有一个参数；缺点：不易解释、速度较慢、易过拟合

面试题：高斯核函数是如何将数据映射到无穷维度的?

$$when \quad \mathbf{x}=(x), \phi(x)=\exp(-x^2)\left(1,\sqrt{\frac{2}{1!}}x,\sqrt{\frac{2^2}{2!}}x^2,\ldots\right)$$

$$\phi(x)\cdot\phi(x')=\sum_{i=0}^{\infty}\left(\exp(-x^2)\exp\left(-(x')^2\right)\sqrt{\frac{2^i}{i!}}\sqrt{\frac{2^i}{i!}}(x)^i(x')^i\right)$$

$$=\exp(-x^2)\exp\left(-(x')^2\right)\left(\sum_{i=0}^{\infty}\frac{(2xx')^i}{i!}\right)$$

$$=\exp(-x^2)\exp\left(-(x')^2\right)\exp(2xx')$$

$$=\exp\left(-(x-x')^2\right)$$

$$more \quad generally, Gaussian \quad Kernel \quad K(\mathbf{x},\mathbf{x}')=\exp\left(-\frac{\left\|\mathbf{x}-\mathbf{x}'\right\|^2}{2\sigma^2}\right)$$

PS：泰勒公式在高斯核函数中的使用

$$f(x)=\frac{f(x_0)}{0!}+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\ldots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x)$$

$$\exp(x)=\frac{1}{0!}+\frac{1}{1!}x+\frac{1}{2!}x^2+\ldots=\sum_{i=0}^{\infty}\frac{x^i}{i!}$$

SVM的优缺点

优点

- 能将数据映射到高维空间，善于处理非线性问题
- 样本量不是海量数据的时候，分类准确率高，泛化能力强。

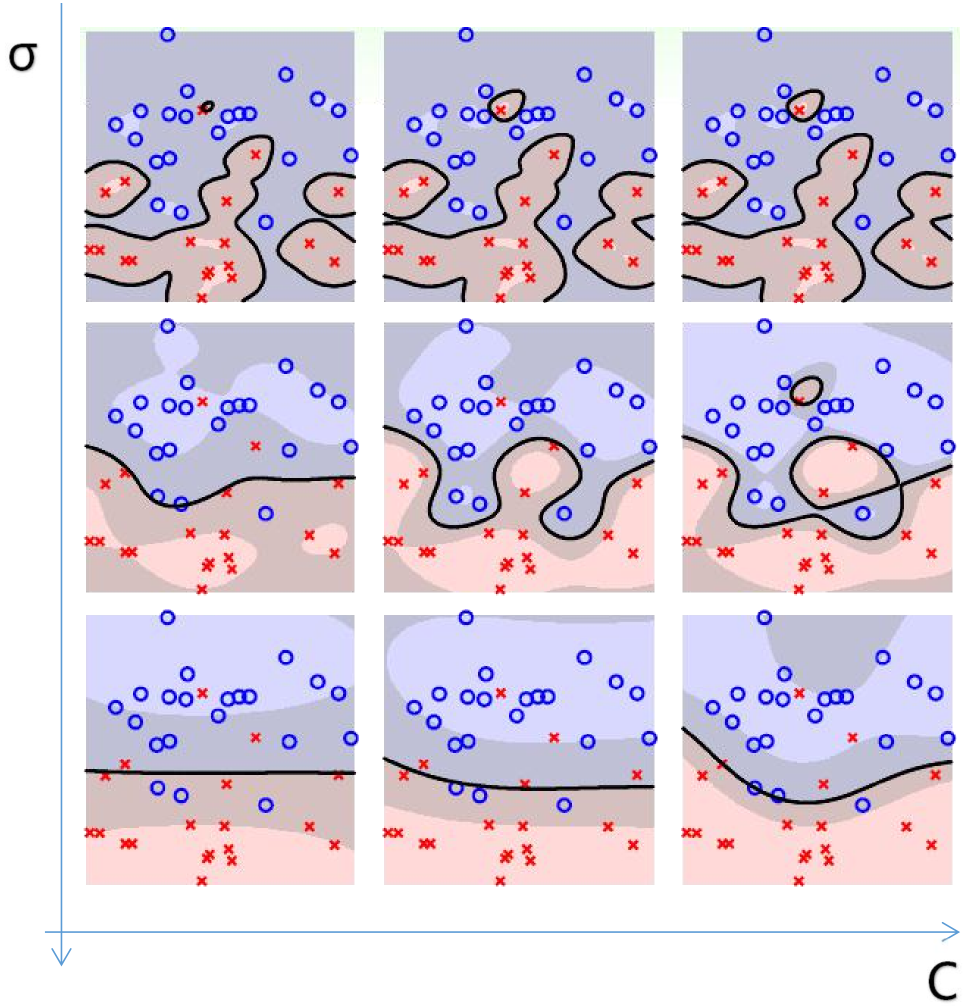
缺点

- 如果特征维度远远大于样本数，则SVM表现一般
- SVM在样本量非常大，核函数映射维度非常高时，计算量过大，不太适合使用。
- 非线性问题的核函数的选择没有通用标准，难以选择一个合适的核函数
- 算法很难并行化，spark中只实现了线性核函数，不过面试中SVM核函数原理属于高频面试题

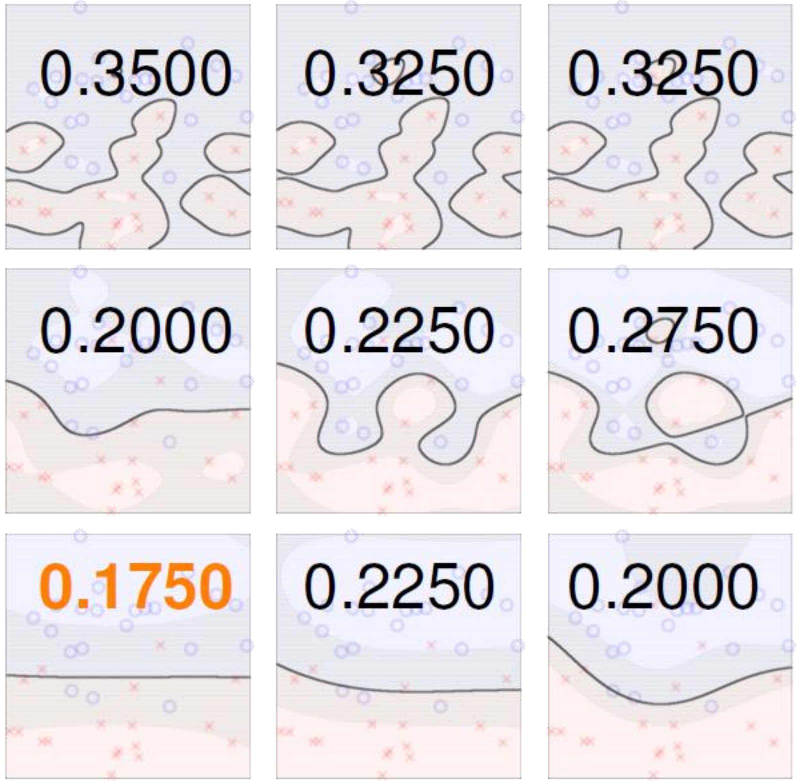
SVM的其他知识

- 支持向量机虽然诞生只有短短的二十多年，但是自一诞生便由于它良好的分类性能席卷了机器学习领域，并牢牢压制了神经网络领域好多年。如果不考虑集成学习的算法，不考虑特定的训练数据集，在分类算法中的表现SVM说是排第一估计是没有什么异议的。
- SVM是一个二元分类算法，线性分类和非线性分类都支持。经过演进，现在也可以支持多元分类，同时经过扩展，也能应用于回归问题。
- 在Spark MLlib中只实现了SVM的线性核函数（从合页损失函数角度做优化，可以参考李航《统计学习方法》7.2.4节）

高斯核函数SVM模型选择



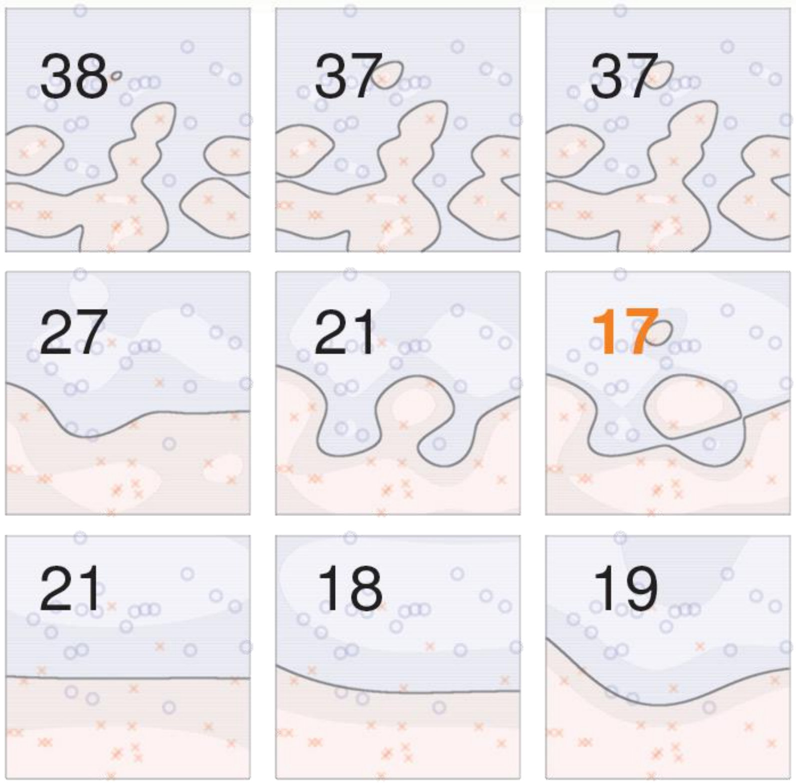
交叉验证



non-smooth——difficult to optimize

proper models can be chosen by V-fold cross validation on a few grid values of (C, σ)

根据SV数量进行模型预选择



dangerous models can be ruled out by nSV

编程——SVM综合案例之森林植被类型预测

- 数据集：
  - https://archive.ics.uci.edu/ml/datasets/covertype
- 解释：
  - 该数据集记录了美国科罗拉多州不同地块的森林植被类型。每个样本包含了描述每块土地的若干特征，包括海拔、坡度、到水源的距离、遮阳情况和土壤类型，并且随同给出了地块的已知森林植被类型。我们需要总共54 个特征中的其余各项来预测森林植被类型

Data Set Characteristics:	Multivariate	Number of Instances:	581012	Area:	Life
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	54	Date Donated	1998-08-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	185453

THANK YOU

上海育创网络科技有限公司

序列最小最优化方法（SMO）*

- 背景
  - 支持向量机的学习问题可以形式化为求解凸二次规划问题。这样的凸二次规划问题具有全局最优解，并且有多种算法可以用于这一问题的求解。当训练样本容量非常大时，这些算法往往非常低效。而序列最小最优化(sequential minimal optimization:SMO）算法可以高效求解。
- SMO算法的思路：
  - 若所有变量都满足条件，则最优化问题的解就得到了。
  - 否则，选择两个变量的同时固定其他所有变量，针对这两个变量构建一个二次规划子问题。
    - 这个二次规划子问题关于这两个变量的解应该更接近原始二次规划问题的解，因为这会使得原始二次规划问题的目标函数值变得更小。
    - 更重要的是，这个二次规划子问题可以通过解析的方法求解。