第10章 支持向量机

支持向量机（SVM）

弗拉基米尔·万普尼克

英文名：Vladimir Naumovich Vapnik

俄罗斯统计学家、数学家

统计学习理论（Statistical Learning Theory）的主要创建人之一，该理论也被称作VC理论（Vapnik Chervonenkis theory）。

支持向量机（SVM）

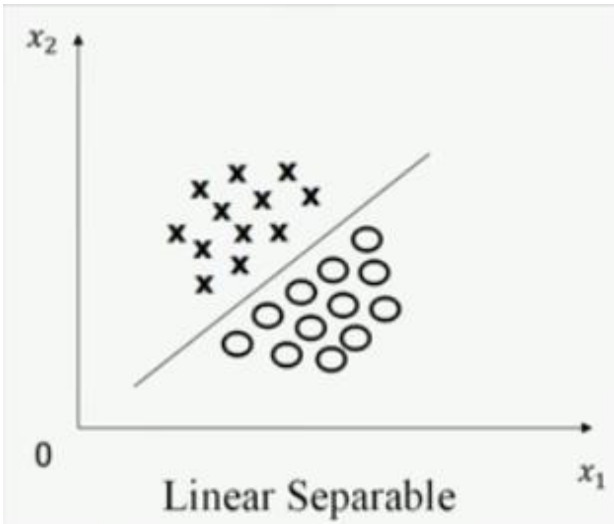
20世纪70年代
创建了支持向量机的主要理论框架

冷战时期，前苏联和西方世界对立

90年代初，前苏联解体，来到美国

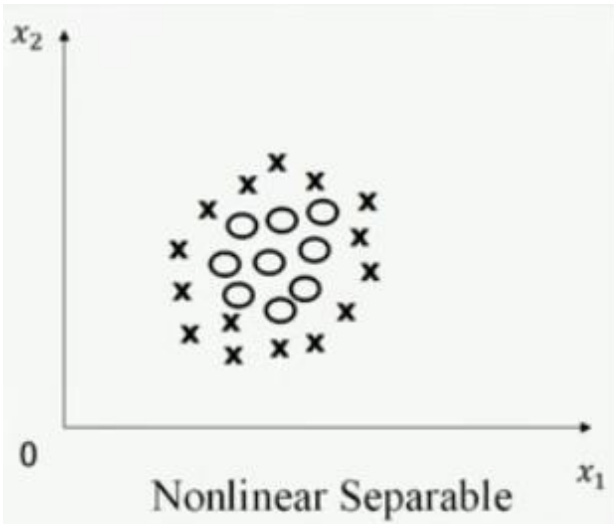
发表到欧美主流的期刊，获得认可

线性可分与线性不可分



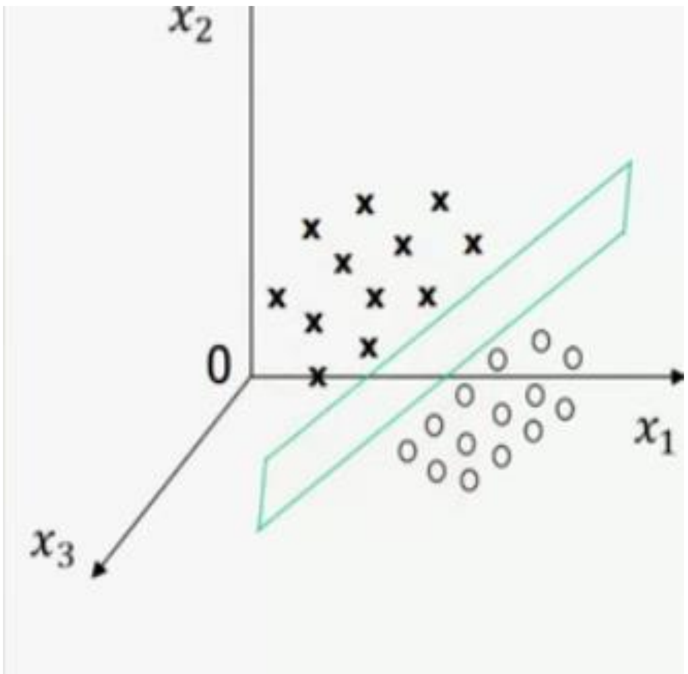
存在一条直线，可以将两类样本分开。

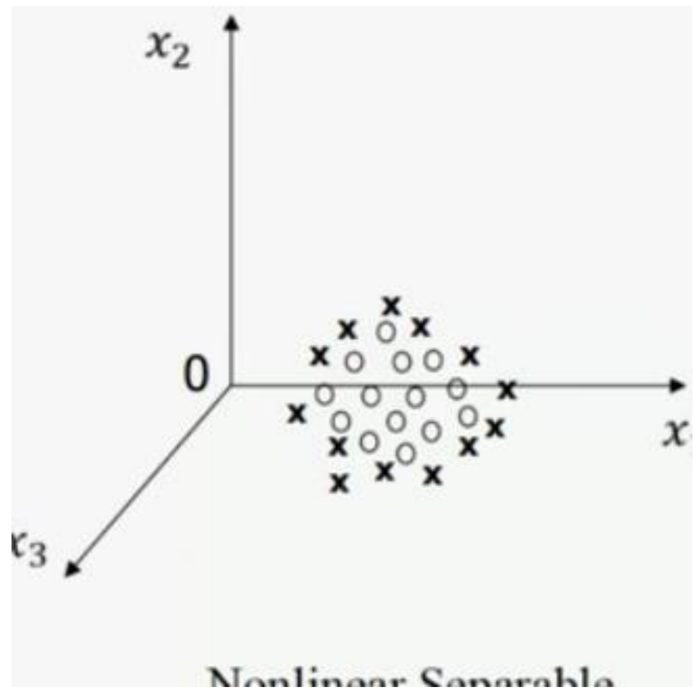
线性可分与线性不可分



不存在一条直线，可以将两类样本分开。

线性可分与线性不可分





线性可分与线性不可分

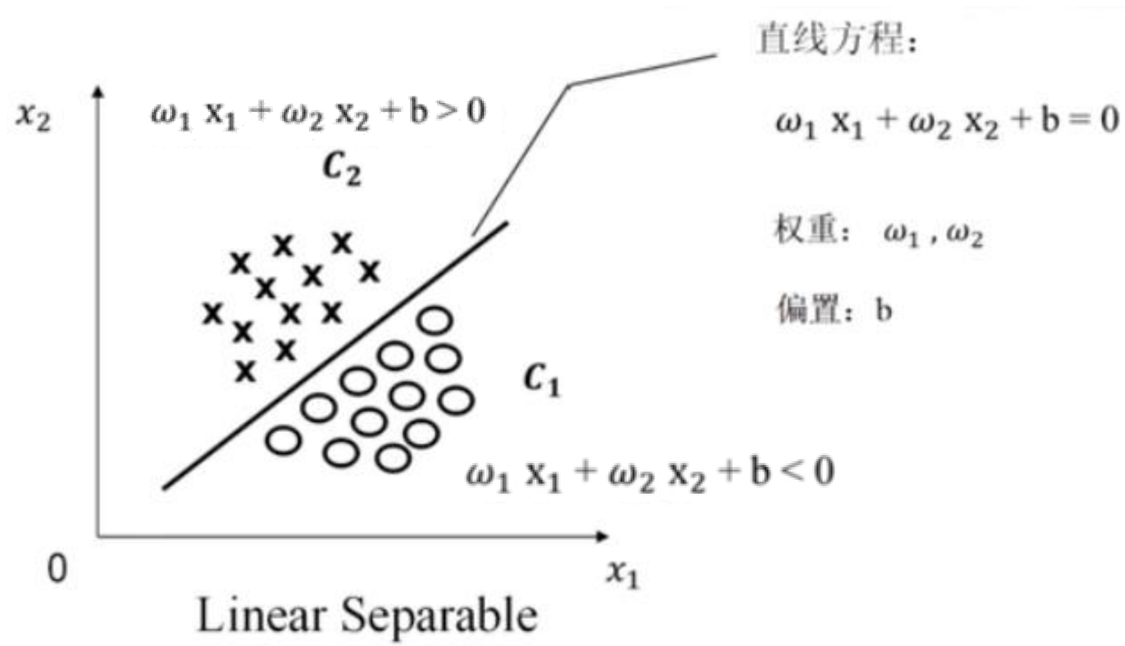
特征空间维度=2维：直线

特征空间维度=3维：平面

特征空间维度≥4维：超平面

如何表示：？ → **数学**

线性可分与线性不可分



假设： $\omega_1' = -\omega_1$, $\omega_2' = -\omega_2$, $b' = -b$

线性可分与线性不可分

用数学定义训练样本及其标签

假设：有N个训练样本及其标签$\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$

其中

$$x_i = [x_{i1}, x_{i2}]$$

$$y_i = \{+1, -1\}$$

$x_i \in C_1$ $\qquad$ $x_i \in C_2$

线性可分与线性不可分

用数学严格地定义线性可分

线性可分的严格定义：一个训练样本集$\{(x_1, y_1), (x_2, y_2),$ …,$(x_i, y_i)\}$，在i=1~N线性可分，是指存在(ω_1, ω_2, b)，使得对i=1~N，有：

(1) 若$y_i = +1$，则 $\omega_1 \mathrm{x}_{i1} + \omega_2 \mathrm{x}_{i2} + \mathrm{b} > 0$

(2) 若$y_i = -1$，则 $\omega_1 \mathrm{x}_{i1} + \omega_2 \mathrm{x}_{i2} + \mathrm{b} < 0$

线性可分与线性不可分

用向量形式来定义线性可分

假设:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix} \qquad \omega = \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}$$

(1) 若$y_i = +1$, 则 $\omega^{\mathrm{T}} x_i + b > 0$

(2) 若$y_i = -1$, 则 $\omega^{\mathrm{T}} x_i + b < 0$

线性可分与线性不可分

线性可分定义的最简化形式

(1) 若$y_i = +1$, 则 $\omega^{\mathrm{T}} x_i + b > 0$

(2) 若$y_i = -1$, 则 $\omega^{\mathrm{T}} x_i + b < 0$

若 $y_i = +1$或-1，则线性可分定义变为

一个训练样本集$\{(x_1, y_1), (x_2, y_2),\ldots,(x_i, y_i)\}$，在i=1~N线性可分，是指存在$(\omega, b)$，使得对i=1~N，有：

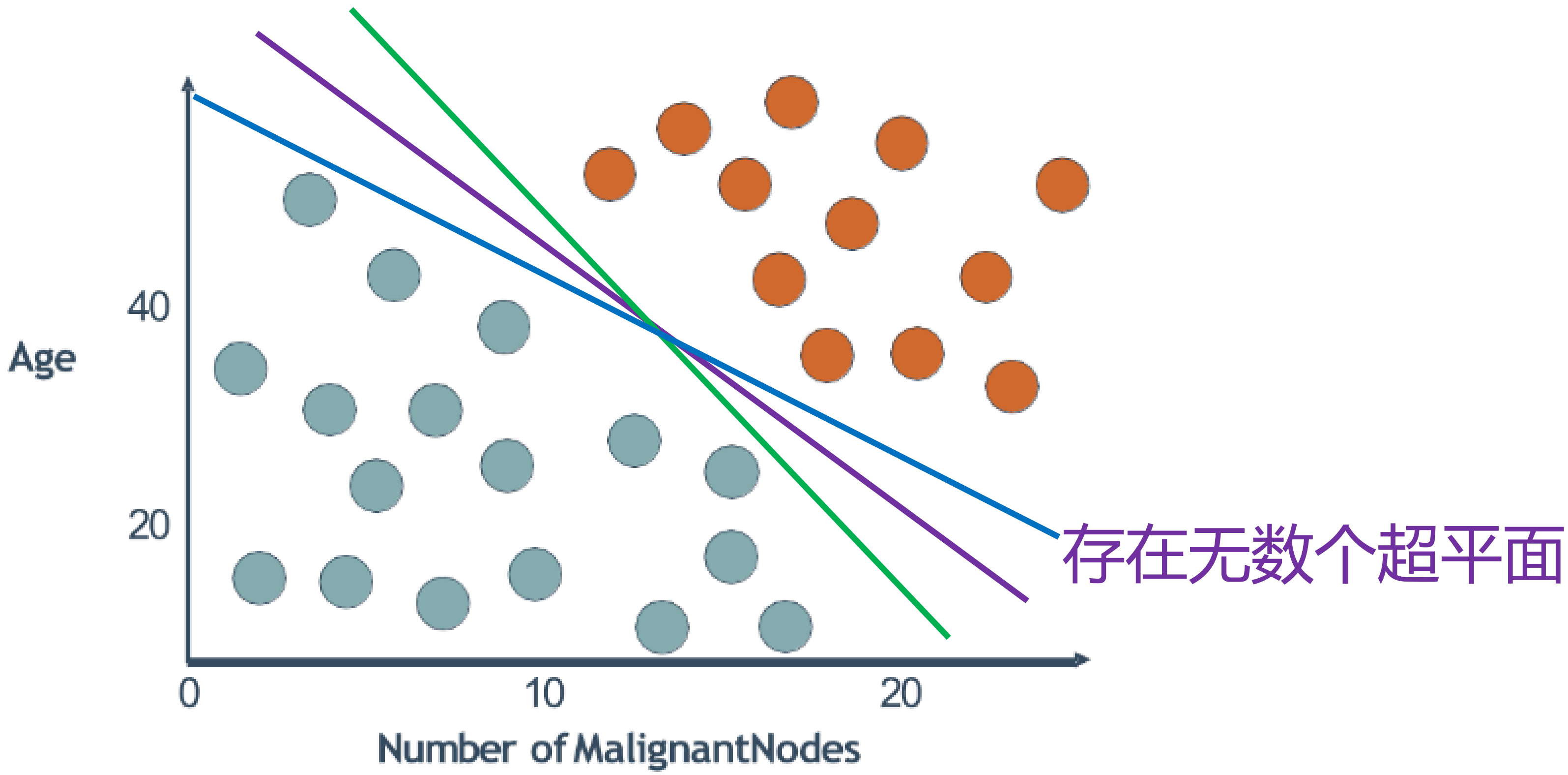
$$y_i(\omega^{\mathrm{T}} x_i + b) > 0$$

SVM算法

- 解决线性可分问题
- 再将线性可分问题中获得的结论推广到线性不可分情况

SVM算法

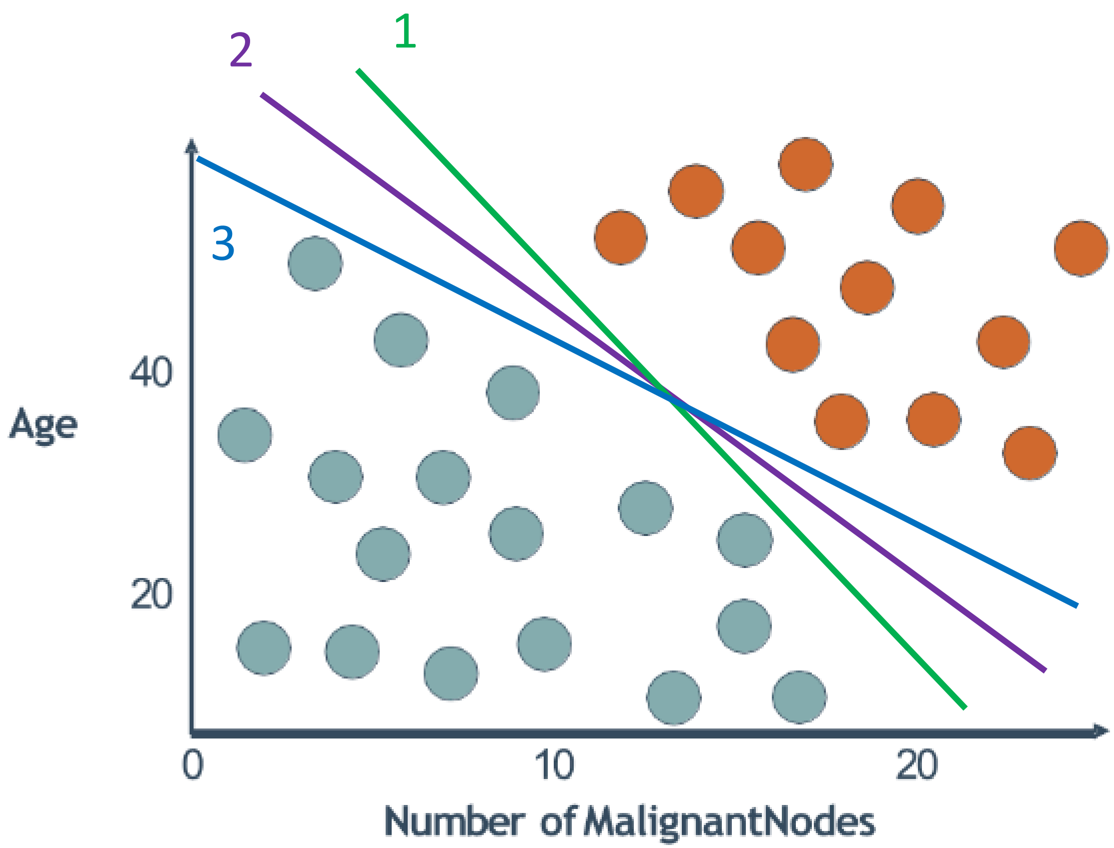
线性可分问题



SVM算法

- 在这无数个分开各个类别的超平面中，哪一个最好?

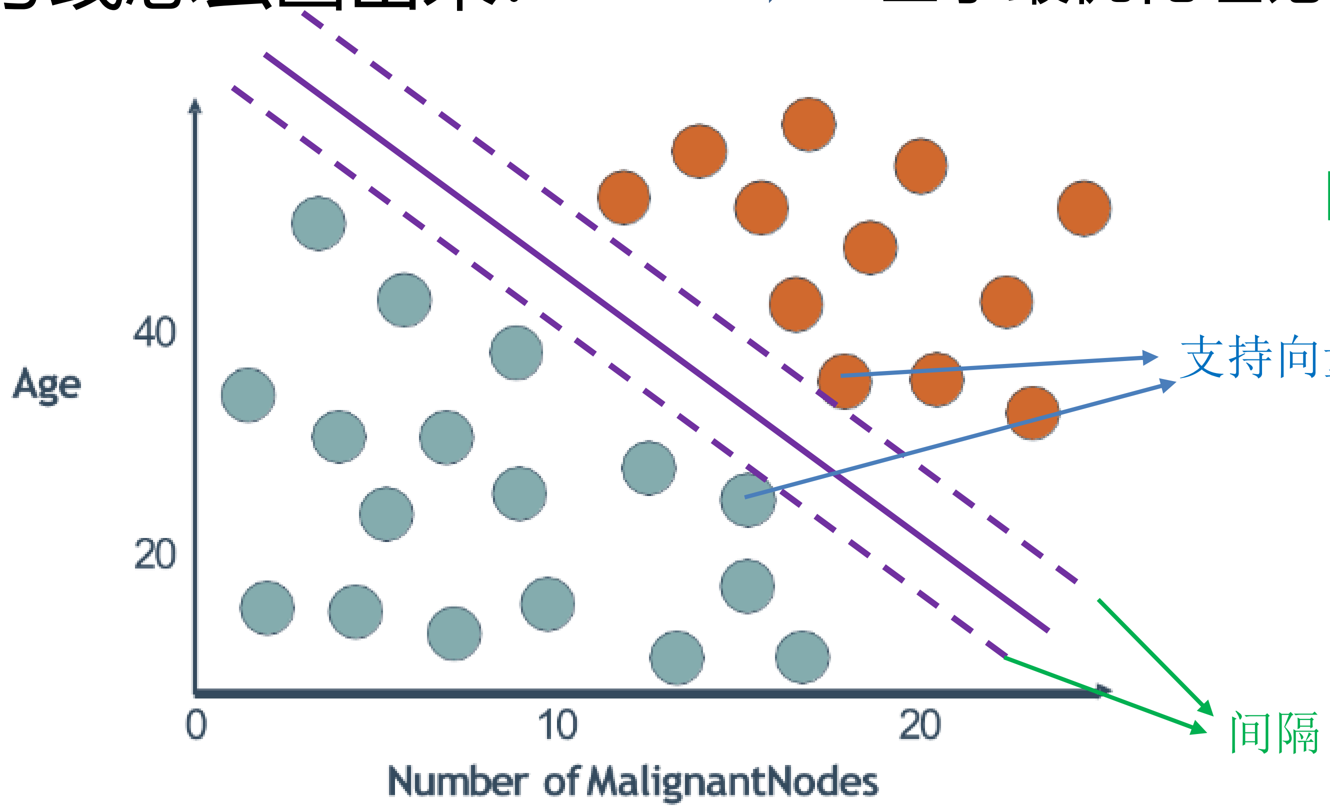
SVM算法



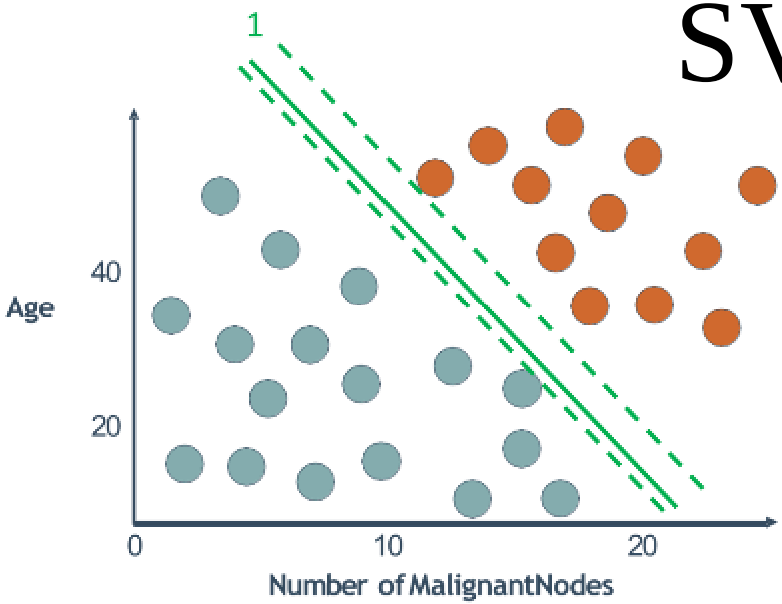
哪一个最好?

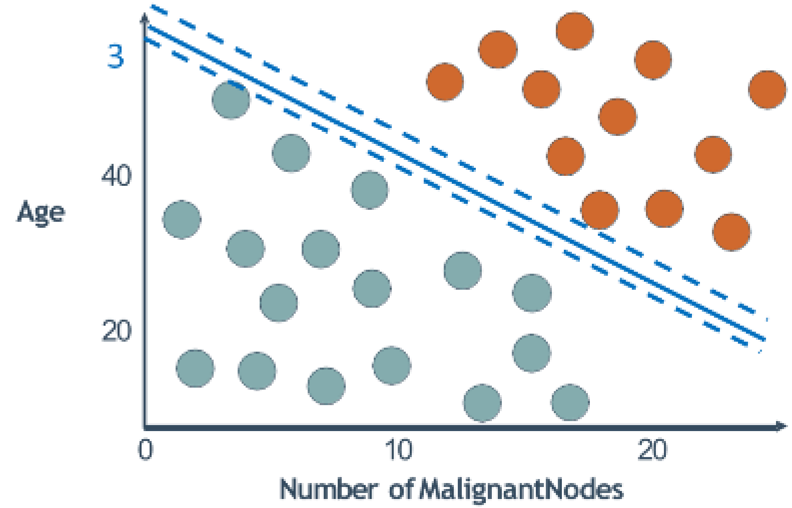
SVM算法

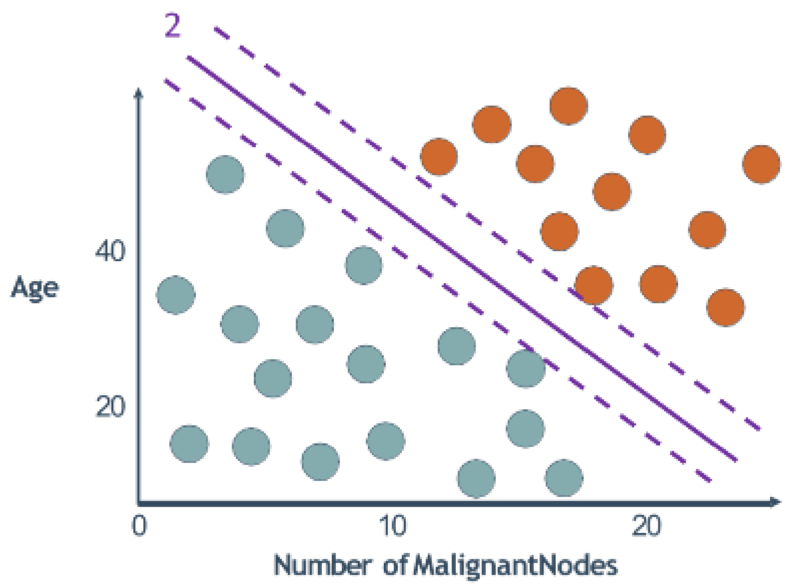
2号线怎么画出来？ → 基于最优化理论 → 间隔最大



SVM算法


1
Age
40
20
0
10
20
Number of MalignantNodes


3
Age
40
20
0
10
20
Number of MalignantNodes

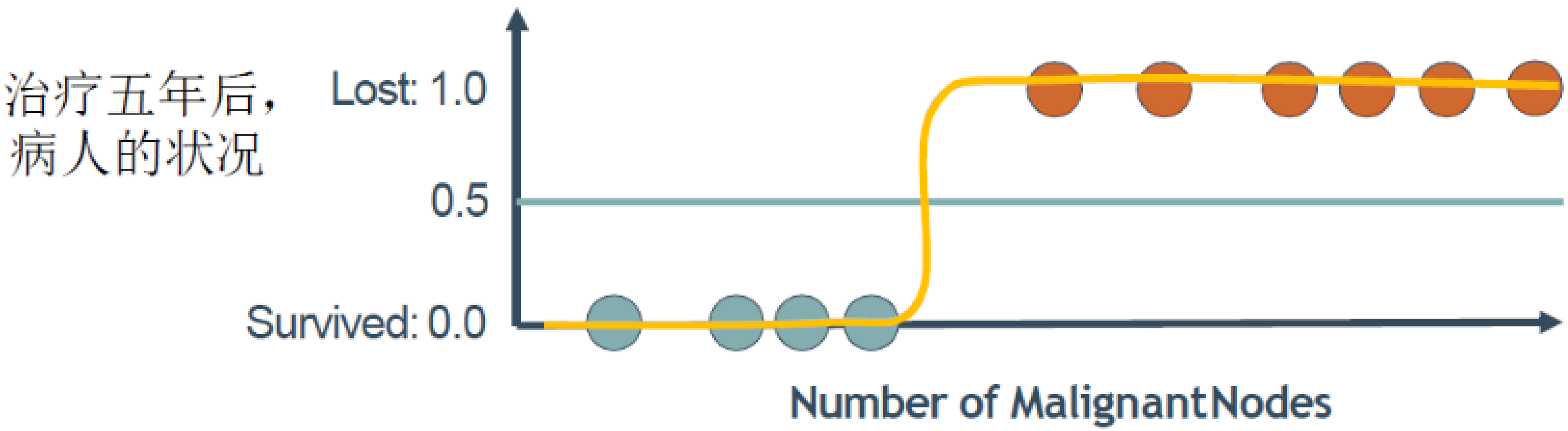

2
Age
40
20
0
10
20
Number of MalignantNodes

SVM算法

SVM寻找的最优分类直线应满足：

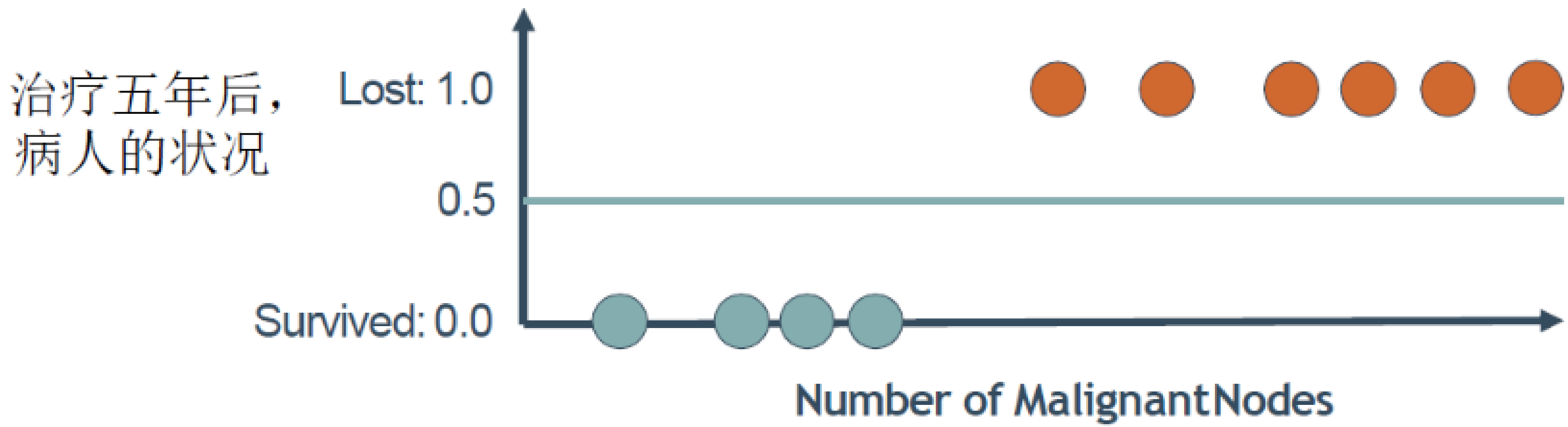
- 该直线分开了两类
- 该直线最大化了间隔
- 该直线处于间隔的中间，到所有支持向量的距离相等

SVM与逻辑回归



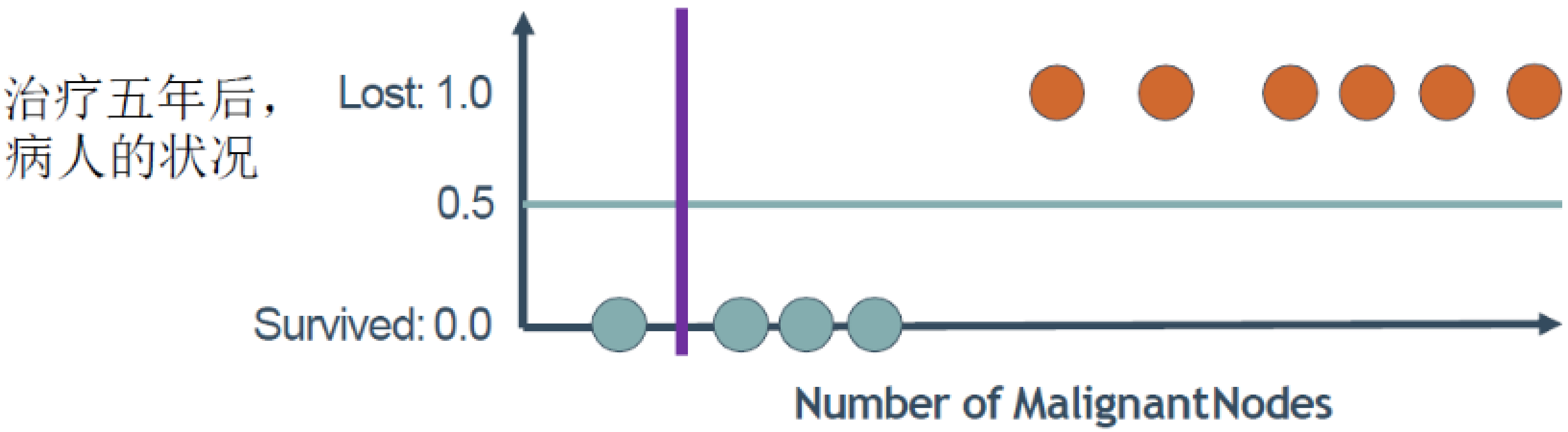
$$y(x) = \frac{1}{1 + e^{-(\omega^{\mathrm{T}} x + b)}}$$

SVM与逻辑回归



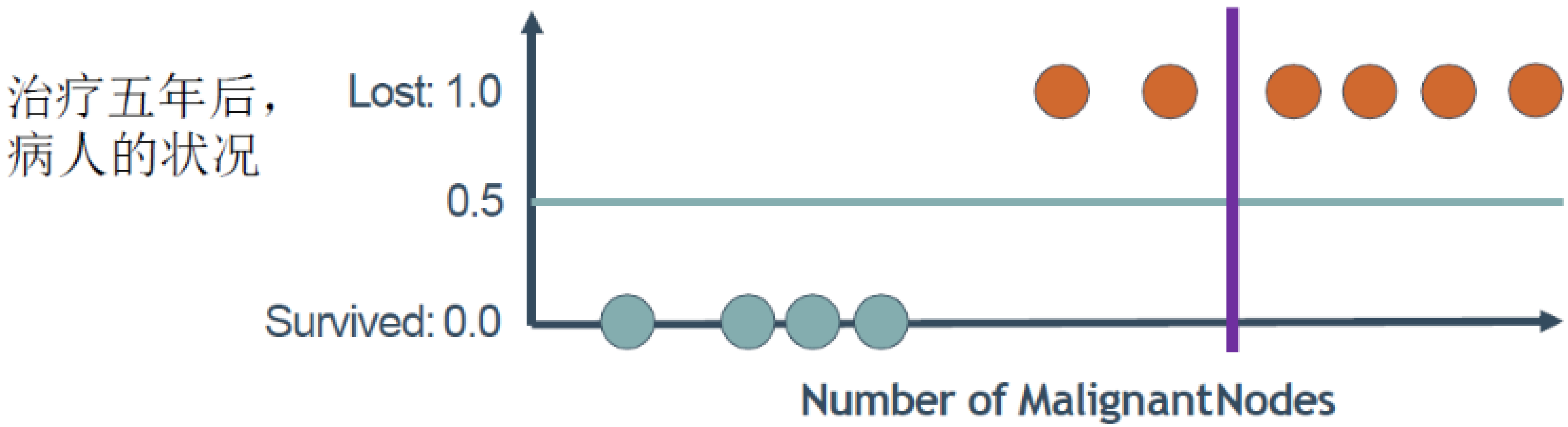
治疗五年后，
病人的状况
Lost: 1.0
0.5
Survived: 0.0
Number of MalignantNodes

SVM与逻辑回归



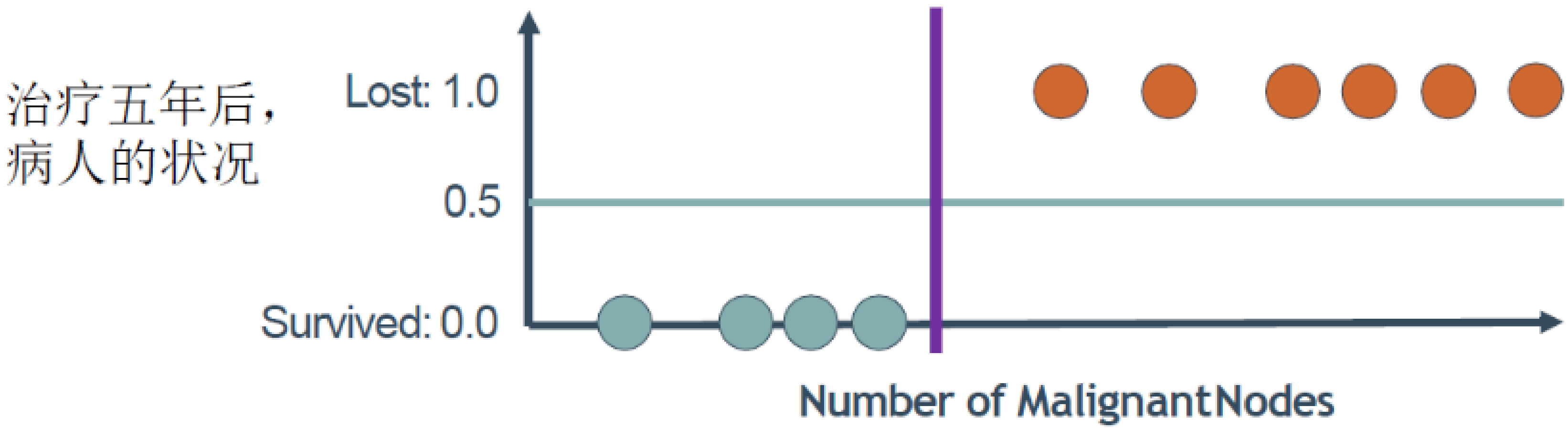
三个分类错误

SVM与逻辑回归



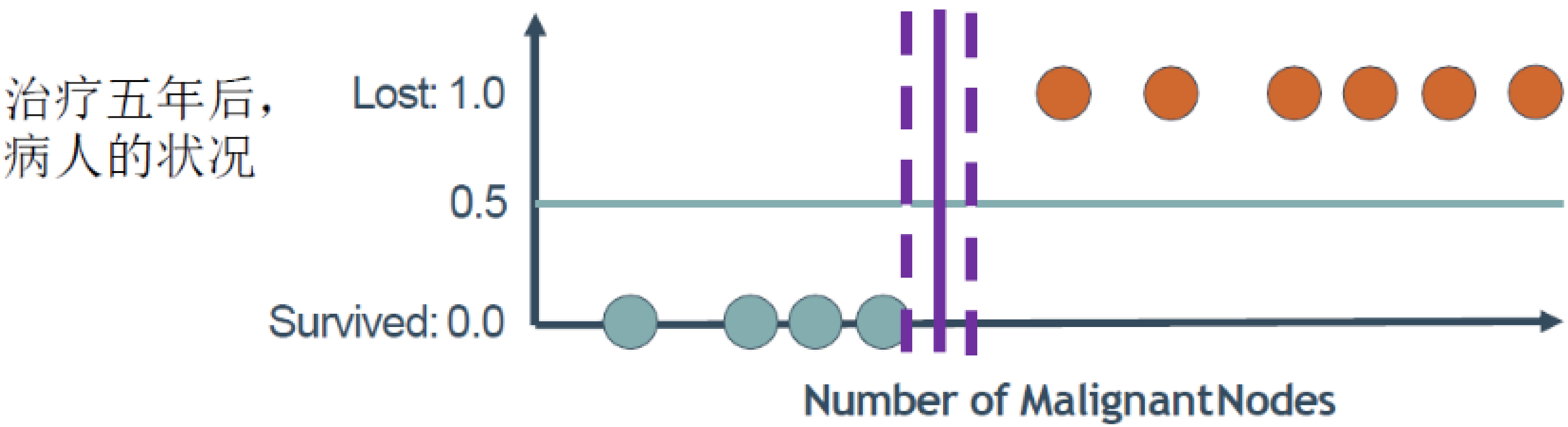
两个分类错误

SVM与逻辑回归



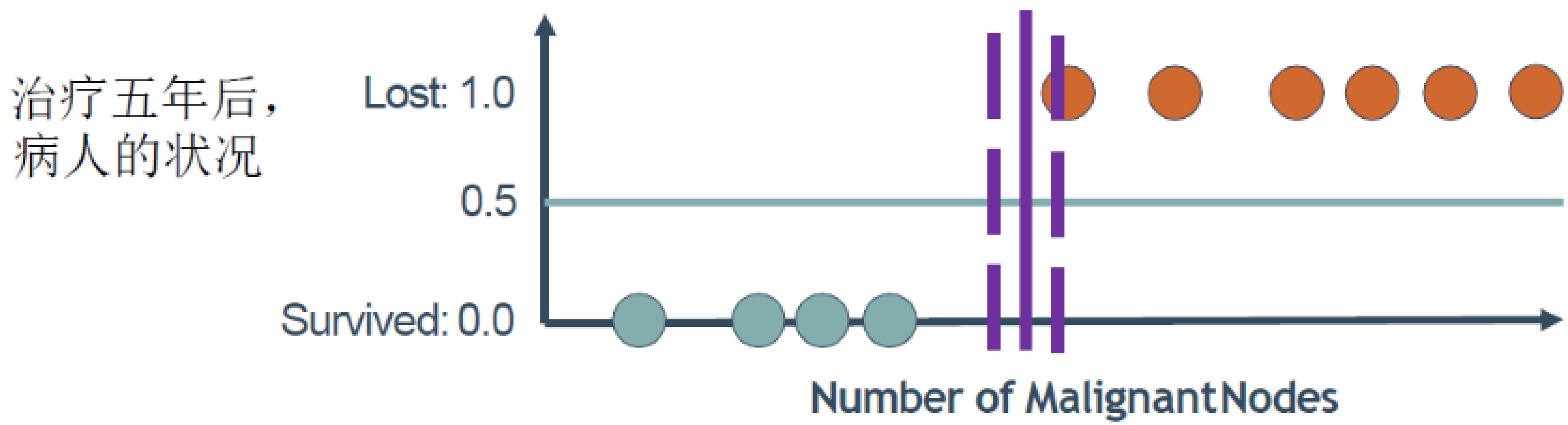
无分类错误

SVM与逻辑回归



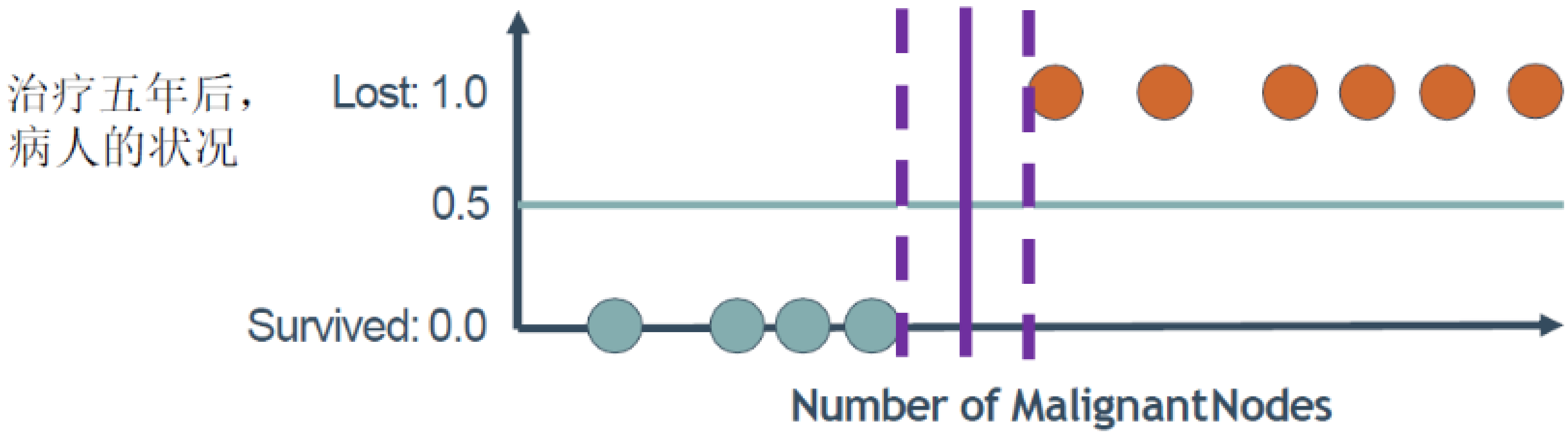
无分类错误，但是否是最佳的分类位置？

SVM与逻辑回归



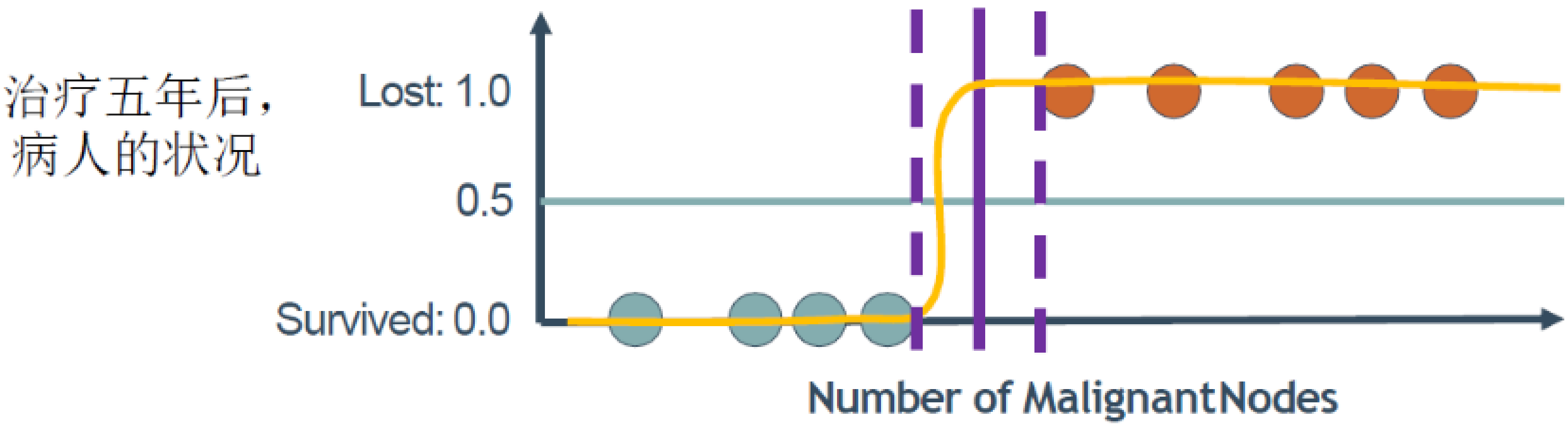
无分类错误，但是否是最佳的分类位置？

SVM与逻辑回归



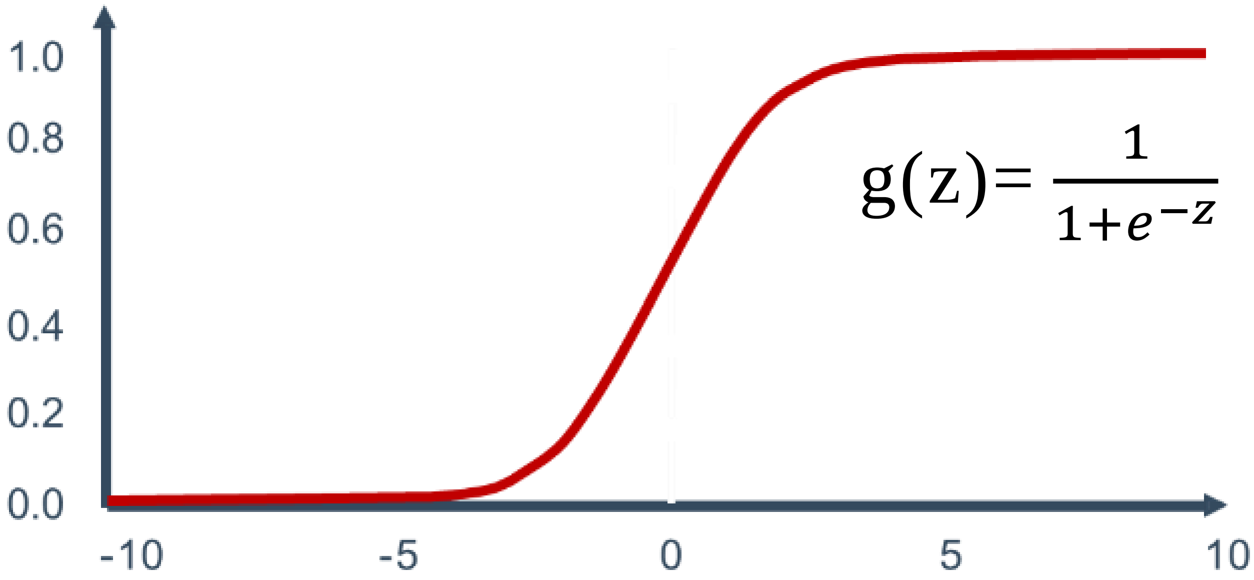
最大化类别之间的区域

SVM与逻辑回归



治疗五年后，
病人的状况
Lost: 1.0
0.5
Survived: 0.0
Number of MalignantNodes

SVM与逻辑回归

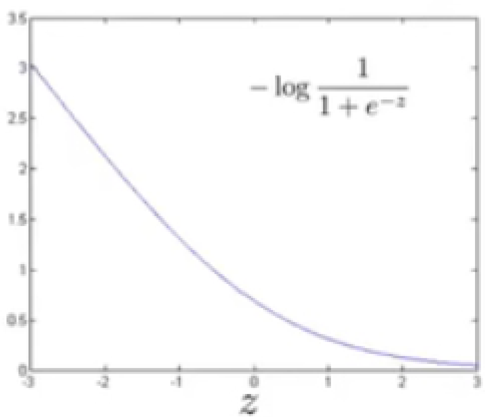


SVM与逻辑回归

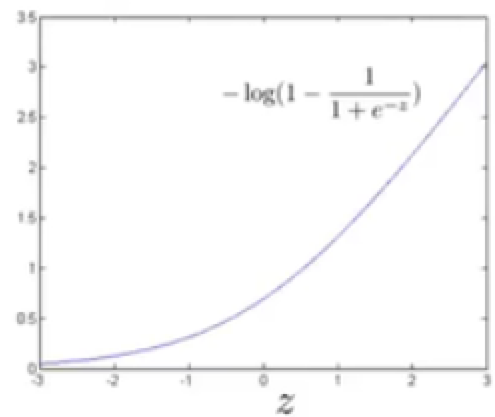
$$J(\omega) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} \times \log\left(h_\omega\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \times \log\left(1 - h_\omega\left(x^{(i)}\right)\right)\right]$$

$$-y^{(i)} \times \log \frac{1}{1 + e^{-(\omega^{\mathrm{T}} x + b)}} - \left(1 - y^{(i)}\right) \times \log\left(1 - \frac{1}{1 + e^{-(\omega^{\mathrm{T}} x + b)}}\right)$$

$y = 1$



$y = 0$



SVM优化问题

寻找最优分类超平面

最优化问题

SVM优化问题

线性可分定义：

一个训练样本集{(x_1, y_1), (x_2, y_2),…,(x_i, y_i)}，在i=1~N线性可分，是指存在(ω, b)，使得对i=1~N，有：

(1) 若$y_i = +1$, 则 $\omega^{\mathrm{T}} x_i + b > 0$

(2) 若$y_i = -1$, 则 $\omega^{\mathrm{T}} x_i + b < 0$

或 $y_i(\omega^{\mathrm{T}} x_i + b) > 0$

超平面： $\omega^{\mathrm{T}} x + b = 0$

ω为超平面的法向量，b为位移

SVM优化问题

样本空间任一点x 到超平面(ω, b)的距离：

$$r = \frac{\left|\omega^{\mathrm{T}} x + b\right|}{\|\omega\|}$$

对于支持向量(x_0, y_0), 有

$$\omega^{\mathrm{T}} x_0 + b = \lambda, \text{ 若} y_i = +1, \lambda > 0$$

$$y_i = -1, \lambda < 0$$

SVM优化问题

用a对(ω,b)进行缩放：

$\omega^{\mathrm{T}}x + b = 0$与$a\omega^{\mathrm{T}}x + ab = 0$是一个平面

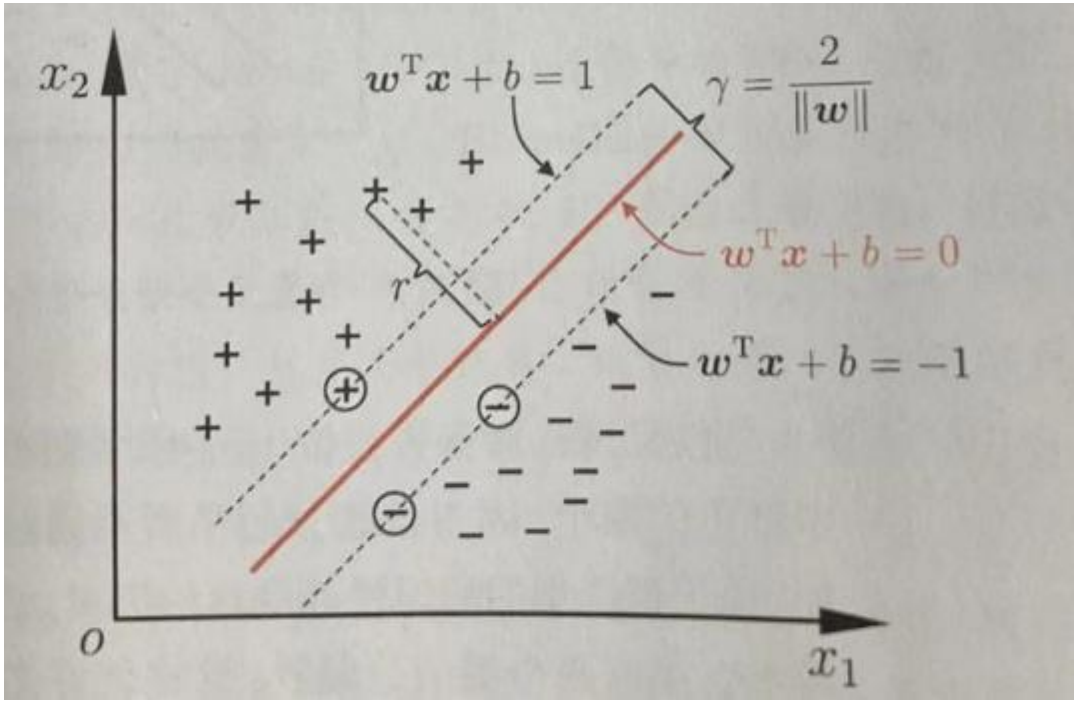
(ω,b)→ ($a\omega$,ab)

使在支持向量上$\left|\omega^{\mathrm{T}}x_i + b\right| = 1$，在非支持向量上

$\left|\omega^{\mathrm{T}}x_i + b\right| > 1$，则

$$\begin{aligned} \omega^{\mathrm{T}}x_i + b &\geq +1, y_i = +1 \\ \omega^{\mathrm{T}}x_i + b &\leq -1, y_i = -1 \end{aligned}$$ 或 $y_i(\omega^{\mathrm{T}}x_i + b) \geq +1$

SVM优化问题



间隔 $$\gamma=\frac{2}{\|\omega\|}$$

SVM优化问题

最大化间隔：

$$\max_{\omega,b} \frac{2}{\|\omega\|}$$

限制条件：$y_i(\omega^{\mathrm{T}} x_i + b) \geq +1, (i = 1 \sim N)$

最优化问题为：

目标函数：$\min_{\omega,b} \frac{1}{2}\|\omega\|^2$

限制条件：$y_i(\omega^{\mathrm{T}} x_i + b) \geq +1, (i = 1 \sim N)$

SVM对偶问题

拉格朗日乘子法

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2+\sum_{i=1}^{N}\alpha_i(1-y_i(\omega^{\mathrm{T}}x_i+b))$$

其中，$\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_N),\alpha_i\geq 0$

令，$L(\omega,b,\alpha)$对ω和b的偏导为0，可得

$$\omega=\sum_{i=1}^{N}\alpha_i y_i x_i \qquad 0=\sum_{i=1}^{N}\alpha_i y_i$$

SVM对偶问题

最优化问题的对偶问题：

$$\max_{\alpha}\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j\right)$$

限制条件：
$$\sum_{i=1}^{N}\alpha_i y_i=0,$$

$$\alpha_i \geq 0, i=1,2,\ldots,N$$

SVM对偶问题

SVM模型：

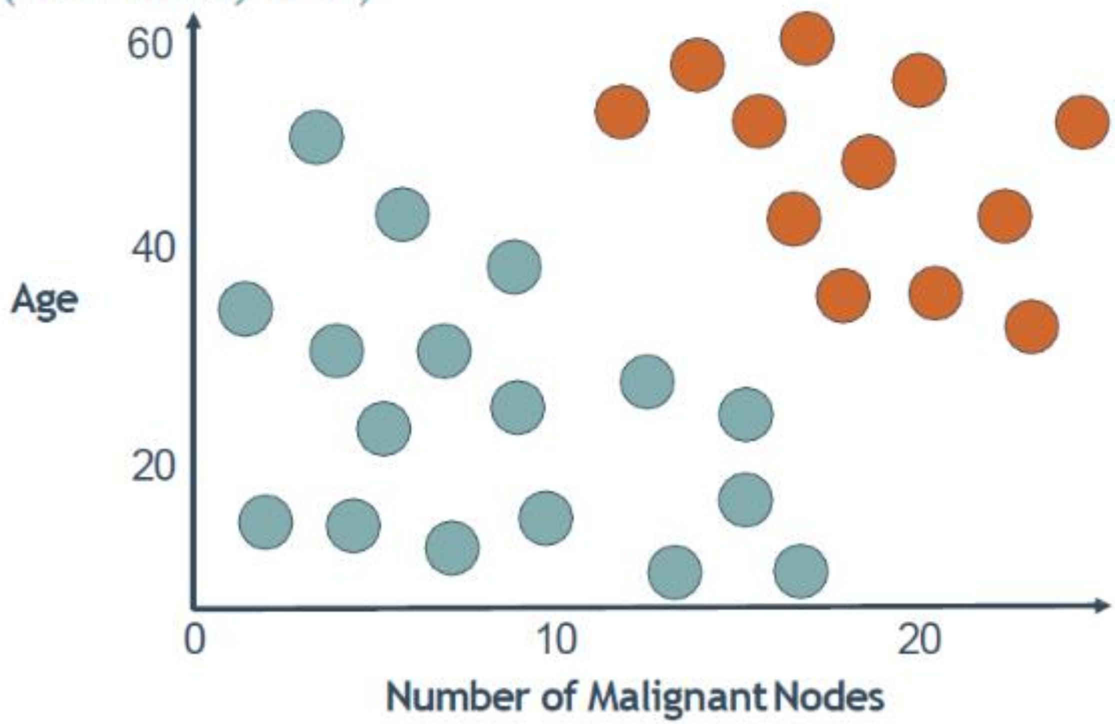
$$f(x) = \omega^{\mathrm{T}} x + b = \sum_{i=1}^{N} \alpha_i y_i x_i^T x + b$$

需满足KKT(Karush-Kuhn-Tucker)条件：

$$\begin{cases} \alpha_i \geq 0; \\ y_i f(x_i) - 1 \geq 0; \\ \alpha_i (y_i f(x_i) - 1) = 0. \end{cases}$$

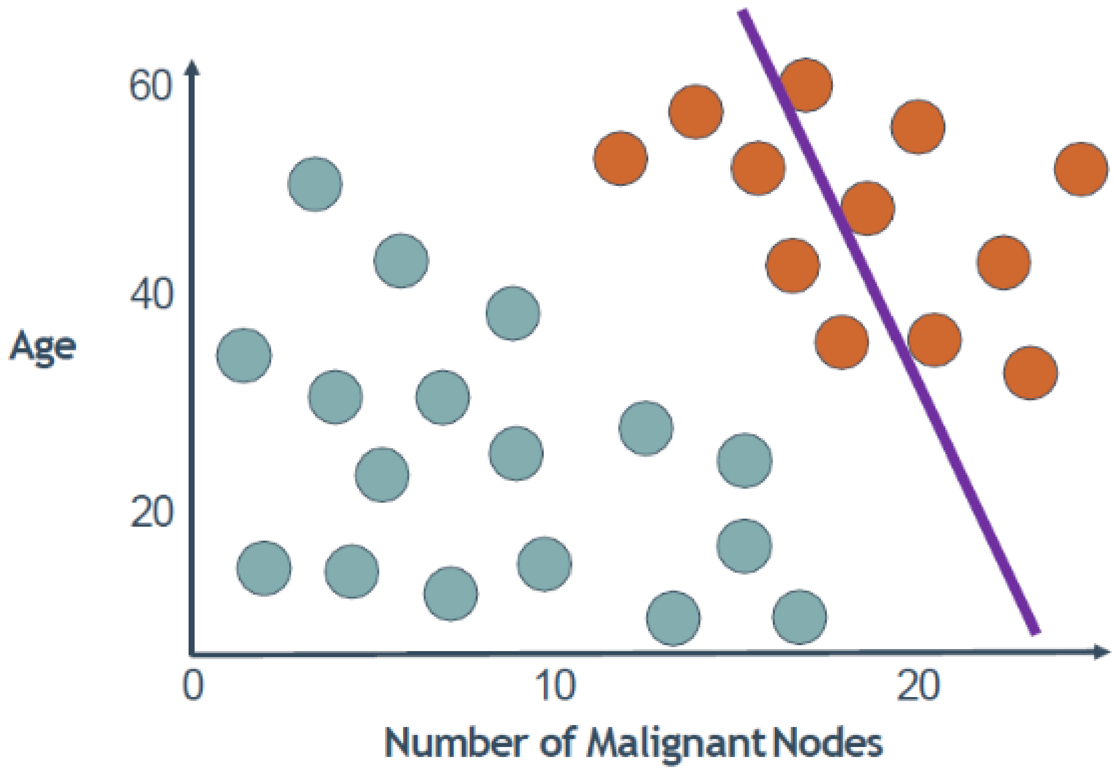
SVM分类

两个特征(nodes, age)
两类标签(survived, lost)



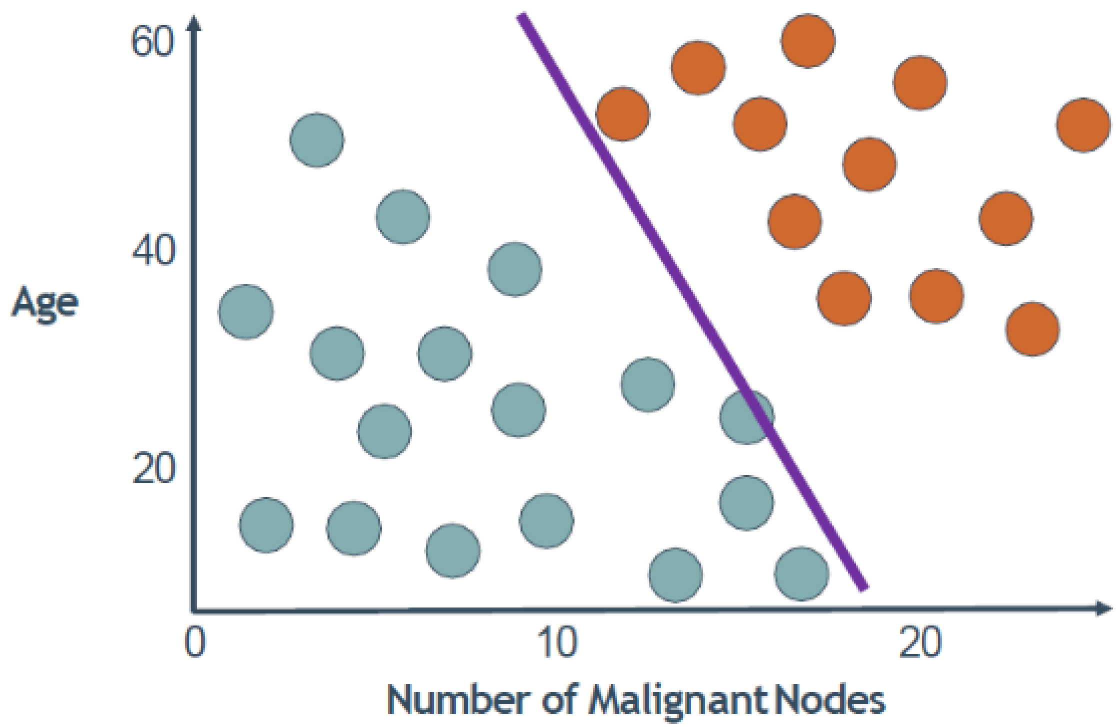
SVM分类

找出能最佳划分两类的线



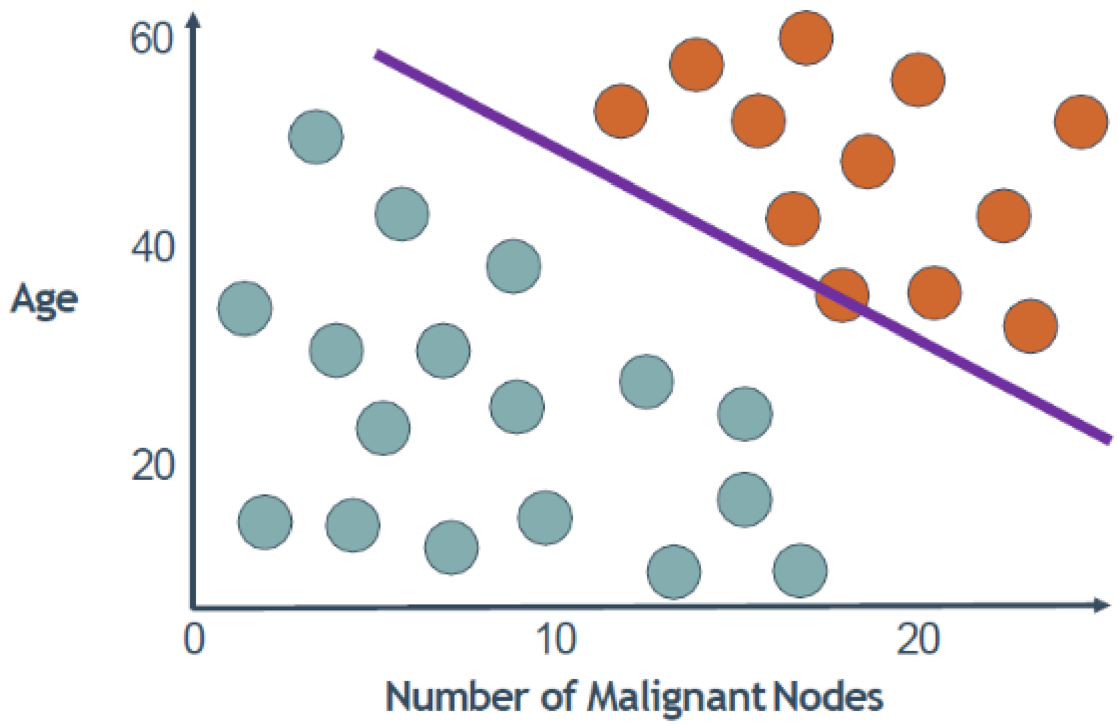
SVM分类

找出能最佳划分两类的线



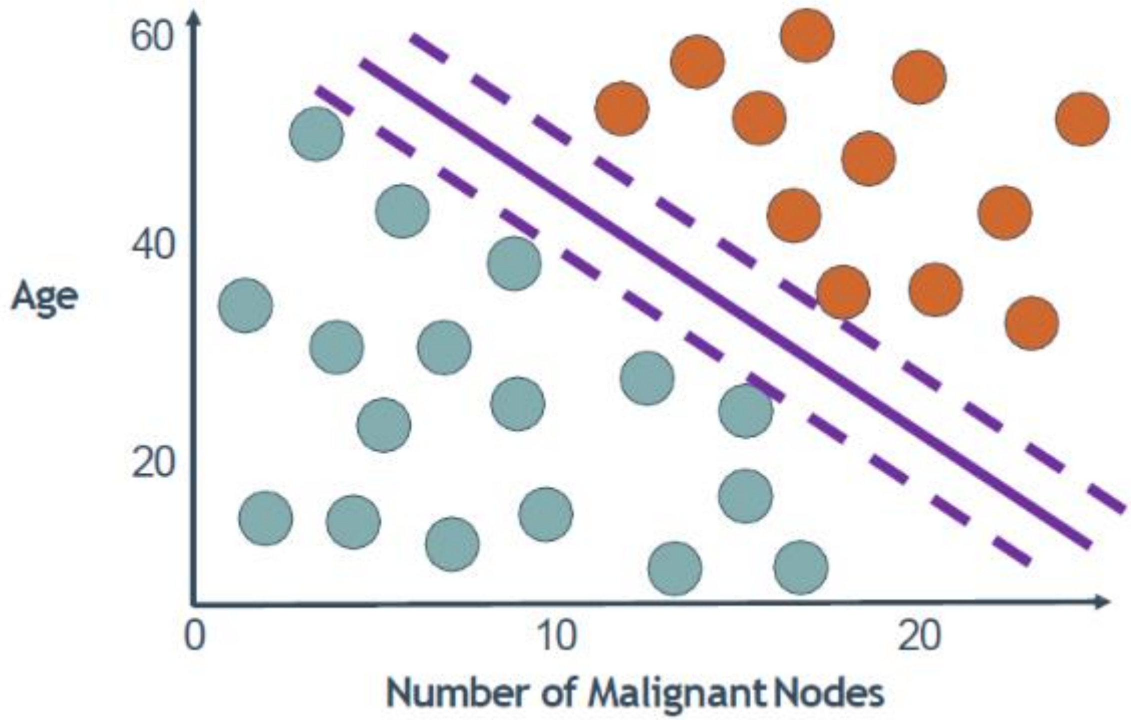
SVM分类

找出能最佳划分两类的线

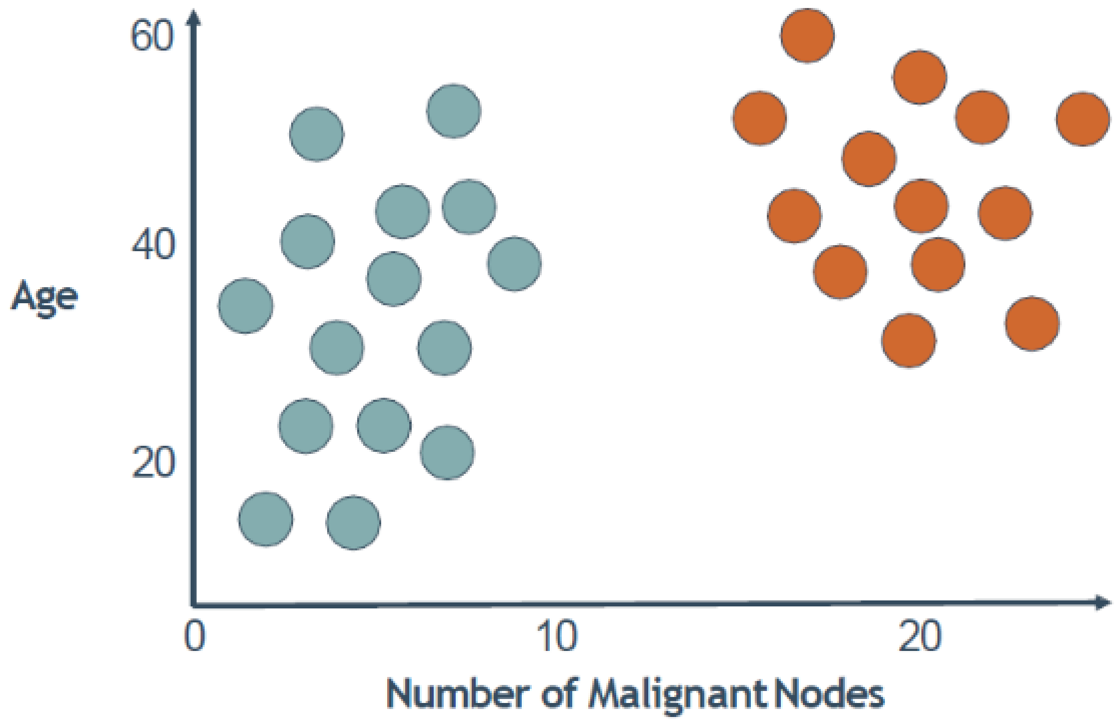


SVM分类

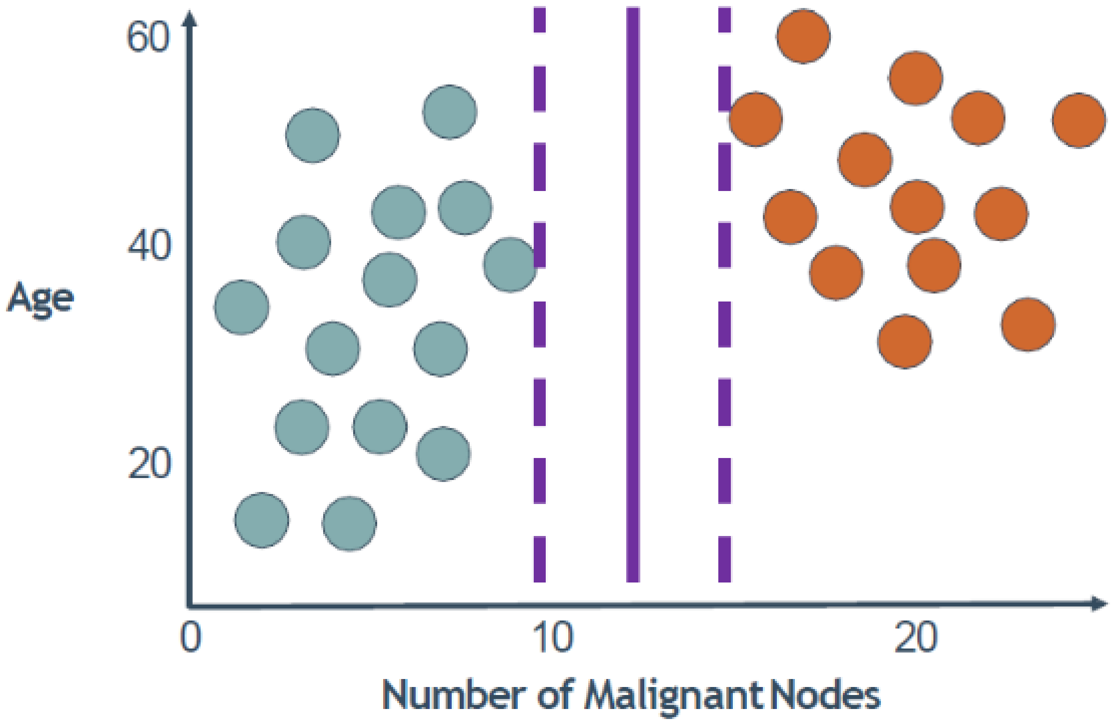
并且具有最大可能的间隔（margin）



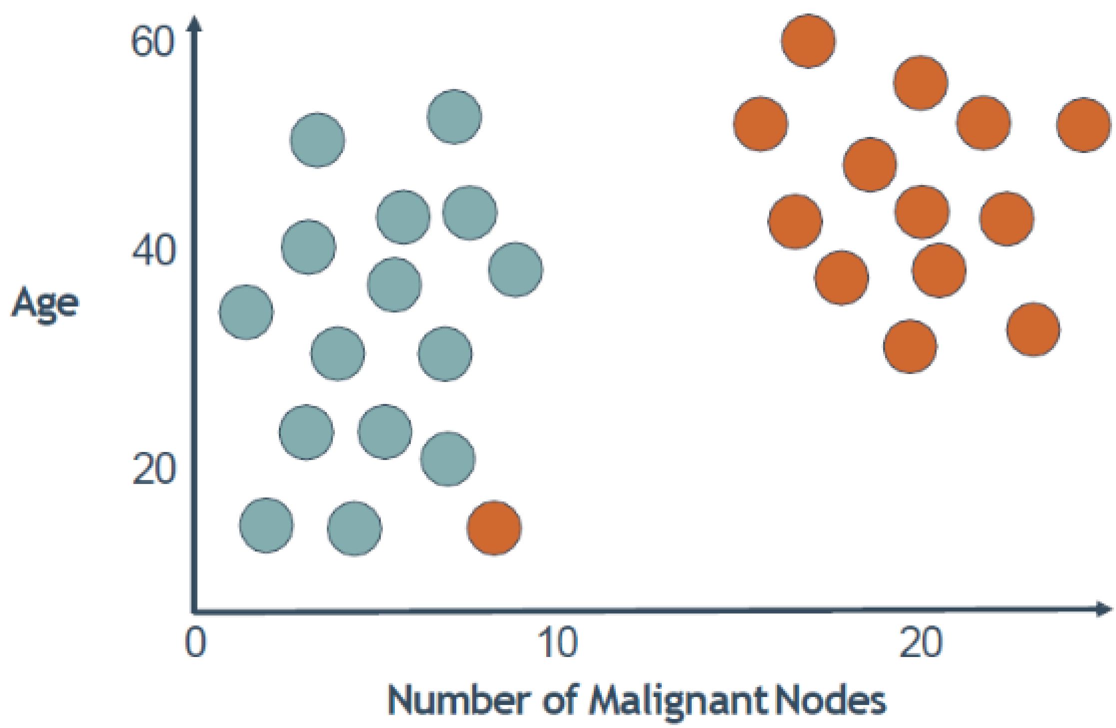
SVM对离群值的敏感性



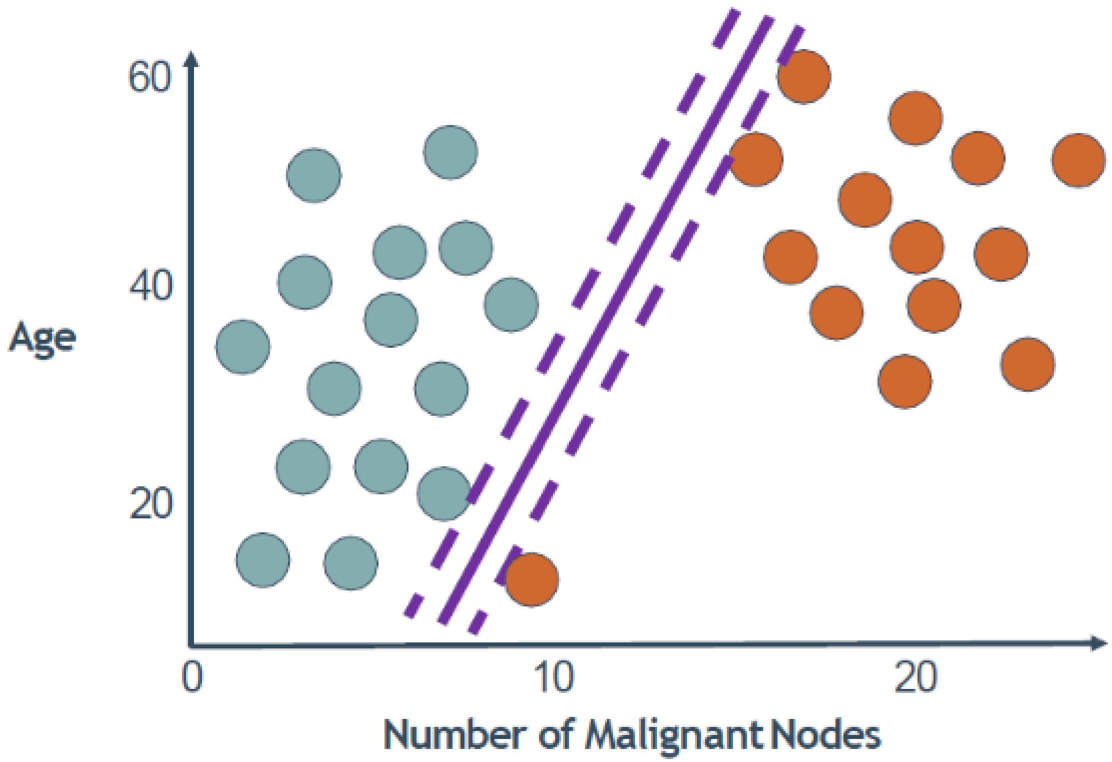
SVM对离群值的敏感性



SVM对离群值的敏感性

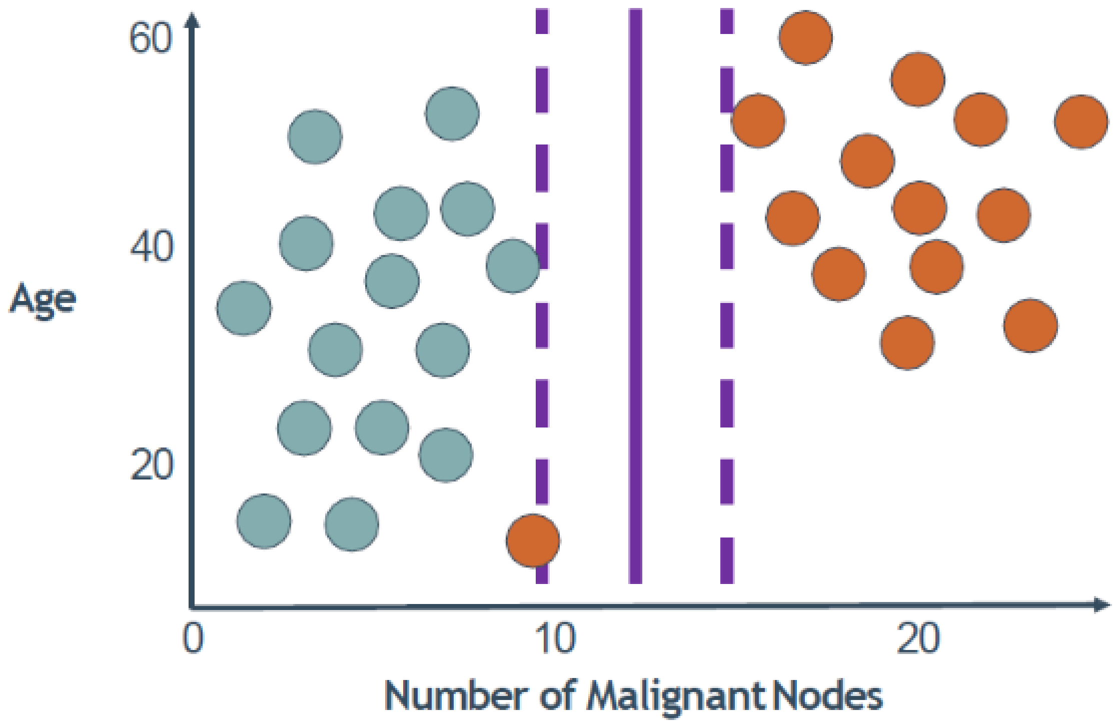


SVM对离群值的敏感性



SVM对离群值的敏感性

这可能仍是最佳的边界线



软间隔与正则化

设置松弛变量(slack variable) δ_i，适当放松限制条件

限制条件：$y_i(\omega^{\mathrm{T}}x_i + b) \geq 1 - \delta_i, (i = 1 \sim N)$

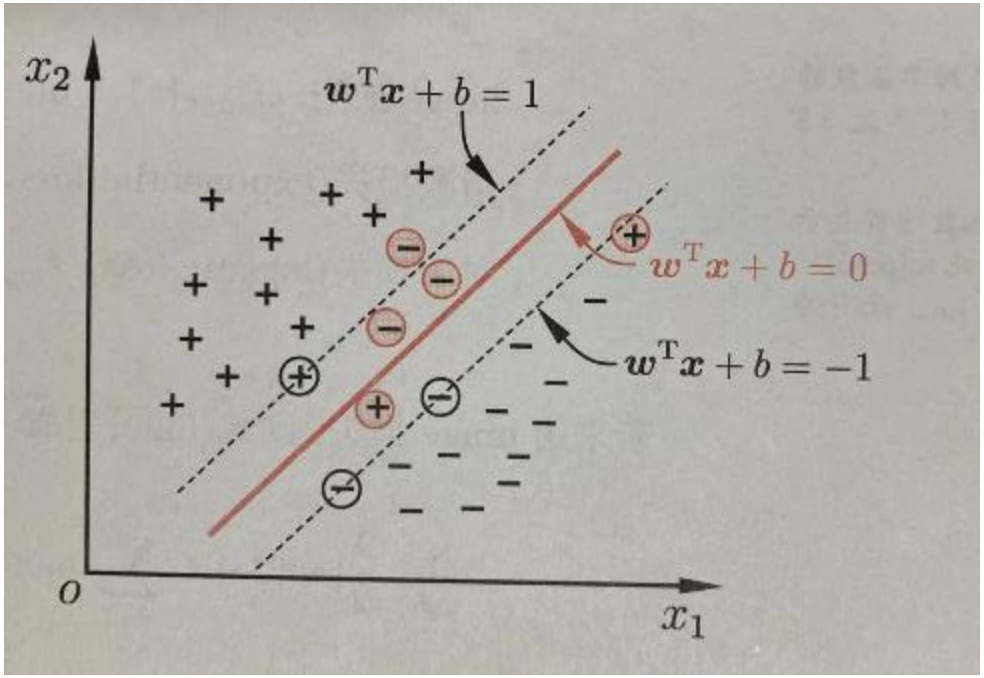
改造后的SVM优化版本：

最小化：$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}\delta_i$或$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}{\delta_i}^2$

限制条件：$\delta_i \geq 0, (i = 1 \sim N)$

$y_i(\omega^{\mathrm{T}}x_i + b) \geq 1 - \delta_i, (i = 1 \sim N)$

软间隔与正则化


x_2
$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=1$
$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$
$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=-1$
o
x_1

SVM线性不可分情况

线性可分最优化问题为：

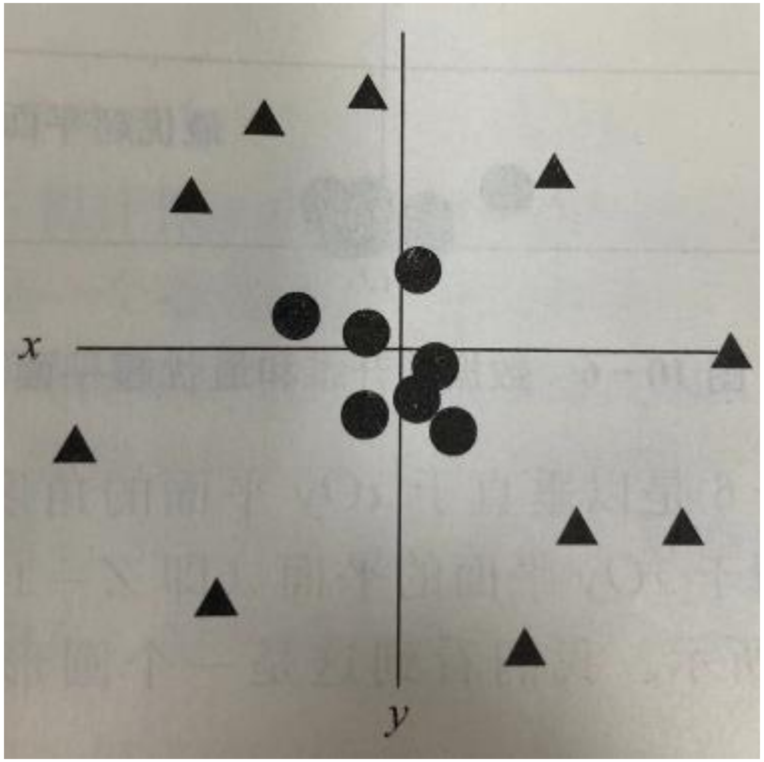
目标函数： $\min\limits_{\omega,b} \frac{1}{2}\|\omega\|^2$

限制条件： $y_i(\omega^{\mathrm{T}} x_i + b) \geq +1, (i = 1 \sim N)$

线性不可分情况，解?

无解

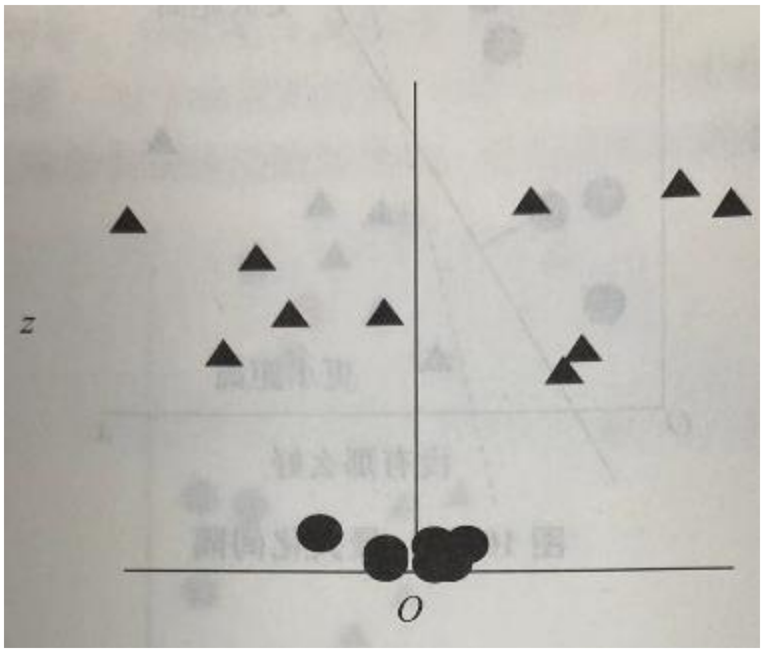
SVM线性不可分情况

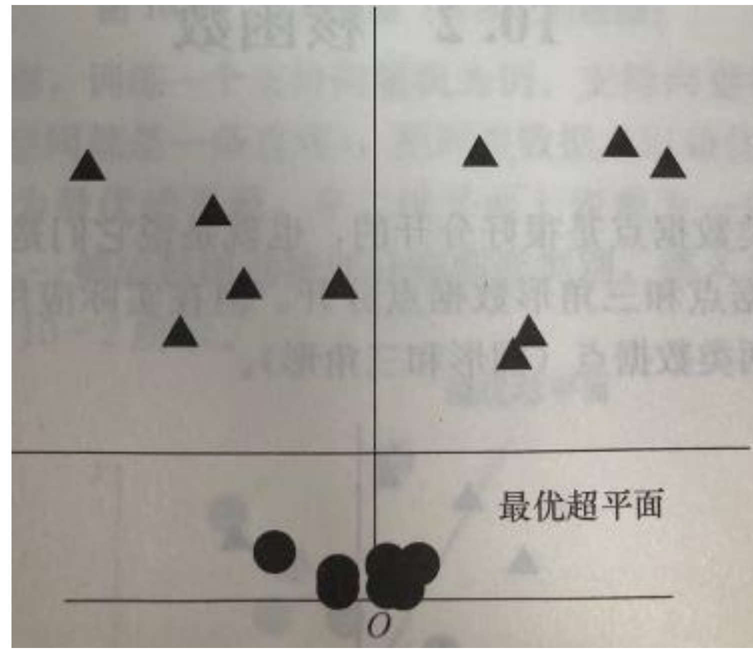


扩大可选函数范围

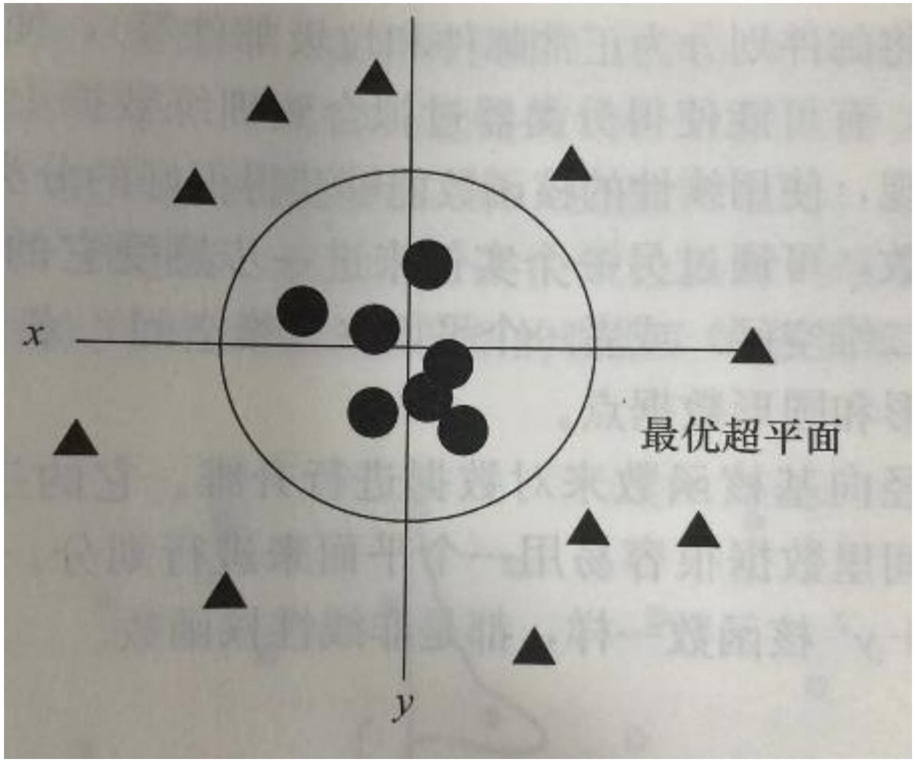
如何扩大?

SVM线性不可分情况





SVM线性不可分情况



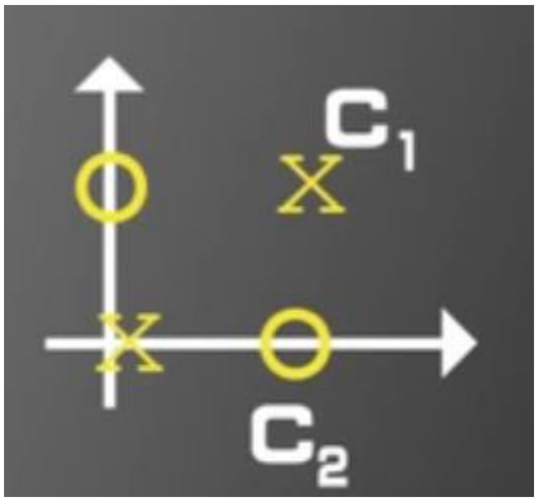
SVM低维到高维的映射

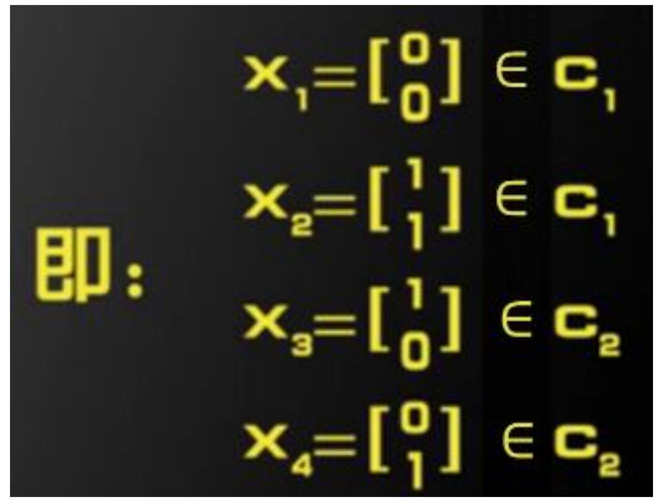
扩大可选函数范围

特征空间由低维映射到高维

用线性超平面对数据进行分类

SVM低维到高维的映射





构造一个二维到五维的映射$\varphi(x)$

$$\varphi(x):\ x=\begin{bmatrix}a\\b\end{bmatrix}\ \rightarrow\ \varphi(x)=\begin{bmatrix}a^2\\b^2\\a\\b\\ab\end{bmatrix}$$

SVM低维到高维的映射

设：

$$\omega = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 6 \end{bmatrix} \qquad b = 1$$

SVM低维到高维的映射

定理：

假设在一个M维空间上随机取N个训练样本，随机地对每个训练样本赋予标签+1或-1

假设这些训练样本线性可分的概率为P(M)，则

当M趋于无穷大时，P(M)=1

SVM低维到高维的映射

将训练样本由低维映射到高维

增大线性可分的概率

$\varphi(x)$

SVM低维到高维的映射

假设$\varphi(x)$已经确定，划分超平面：

$$f(x) = \omega^{\mathrm{T}}\varphi(x) + b$$

SVM优化问题为：

目标函数：$\min\limits_{\omega,b} \frac{1}{2}\|\omega\|^2$

限制条件：$y_i(\omega^{\mathrm{T}}\varphi(x_i) + b) \geq 1, (i = 1 \sim N)$

注意：$\omega, \varphi(x_i)$维度

SVM核函数

其对偶问题：

$$\max_{\alpha}\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boxed{\varphi(x_i)^T\varphi(x_j)}\right)$$

限制条件：

$$\sum_{i=1}^{N}\alpha_i y_i=0,$$

$$\alpha_i\geq 0, i=1,2,\ldots,N$$

SVM核函数

构造核函数（kernel function）：

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$$

对偶问题变为：

$$\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$$

限制条件：

$$\sum_{i=1}^{N} \alpha_i y_i = 0,$$

$$\alpha_i \geq 0, i = 1,2, \ldots, N$$

SVM核函数

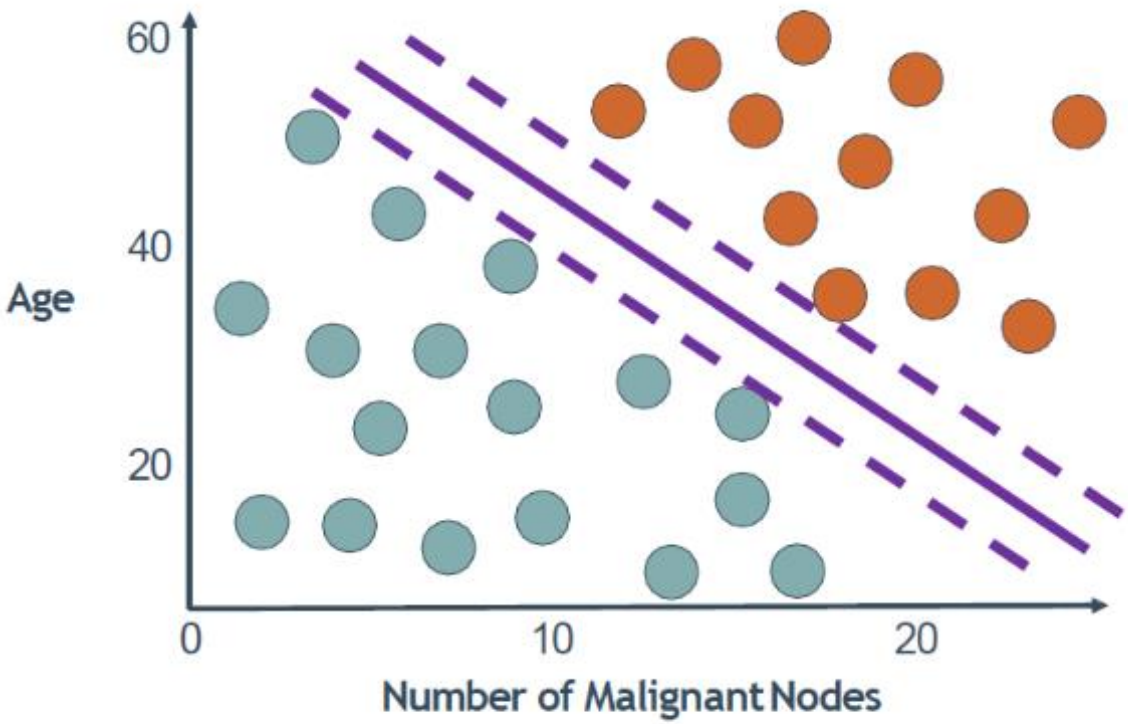
求解后得到：

$$f(x) = \omega^{\mathrm{T}}\varphi(x) + b$$

$$= \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i)^T \varphi(x) + b$$

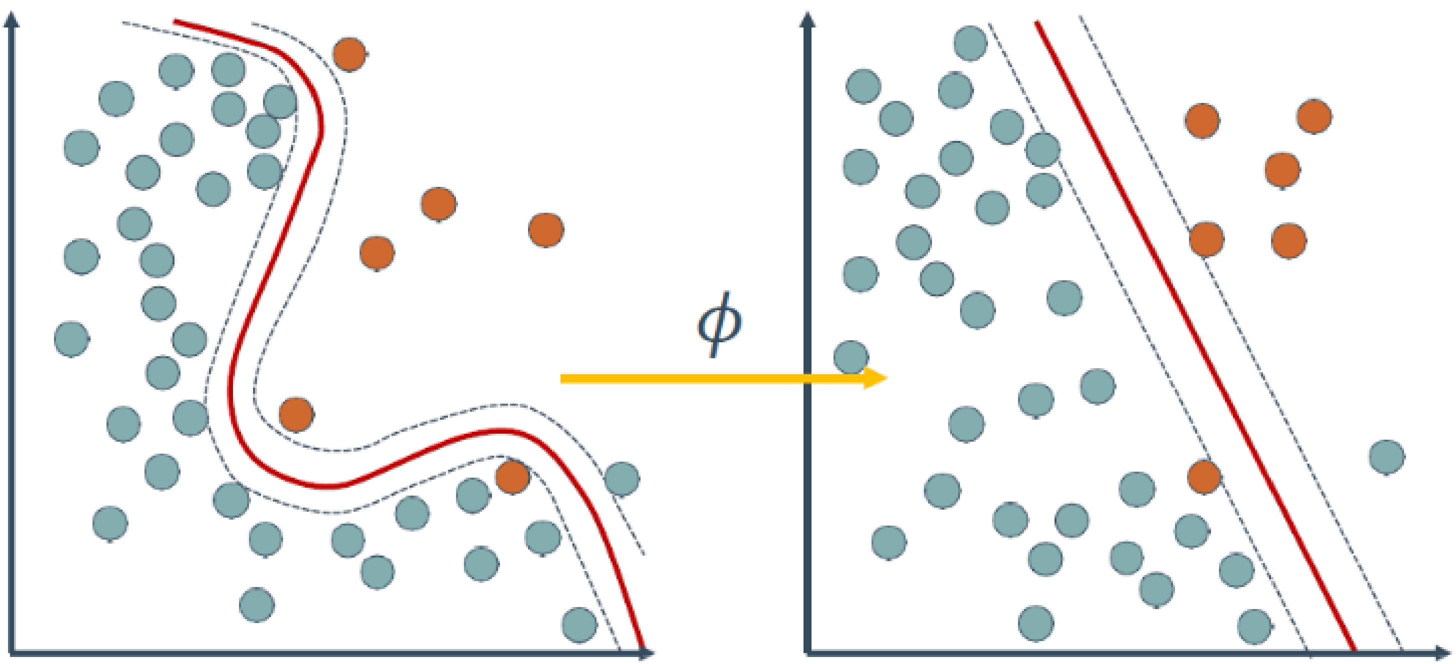
$$= \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b$$

用SVM分类



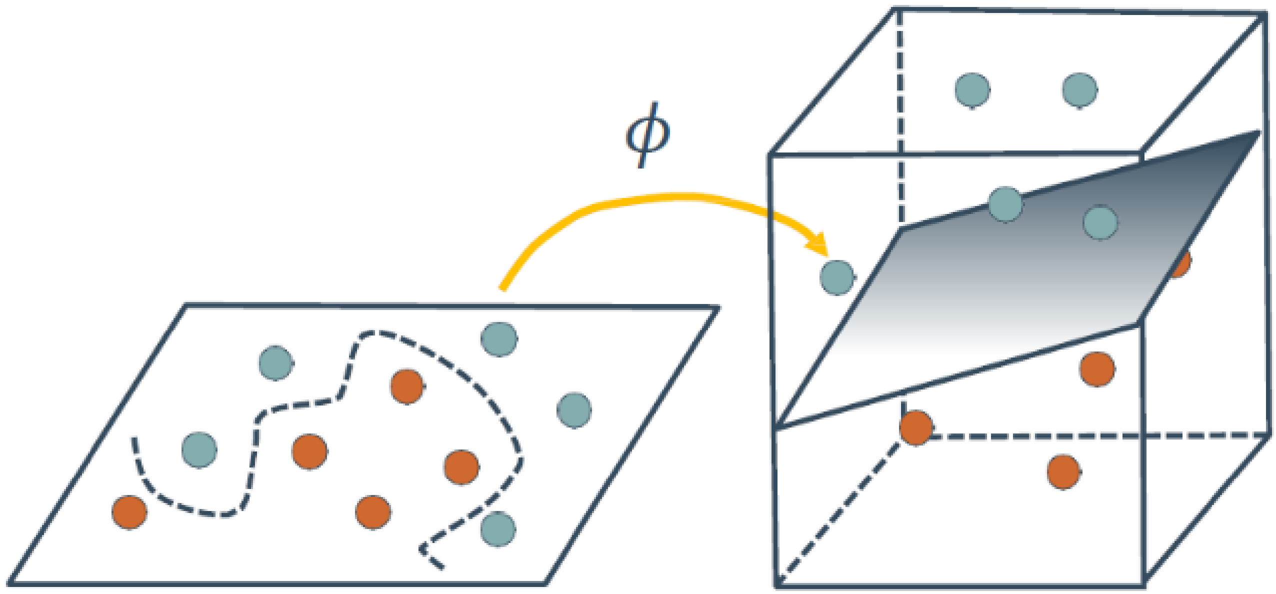
非线性判定边界

非线性数据在高维空间可能被转换为线性的

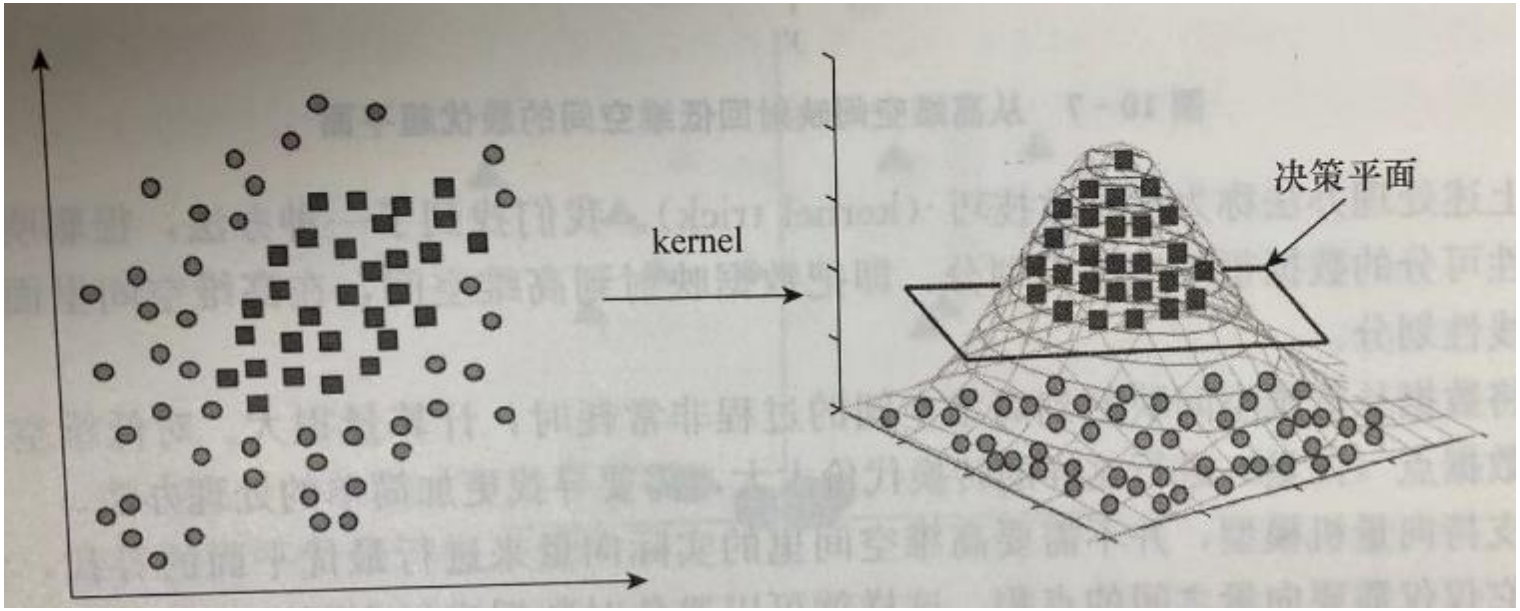


核函数

把数据转换为线性可分的



SVM高斯核函数

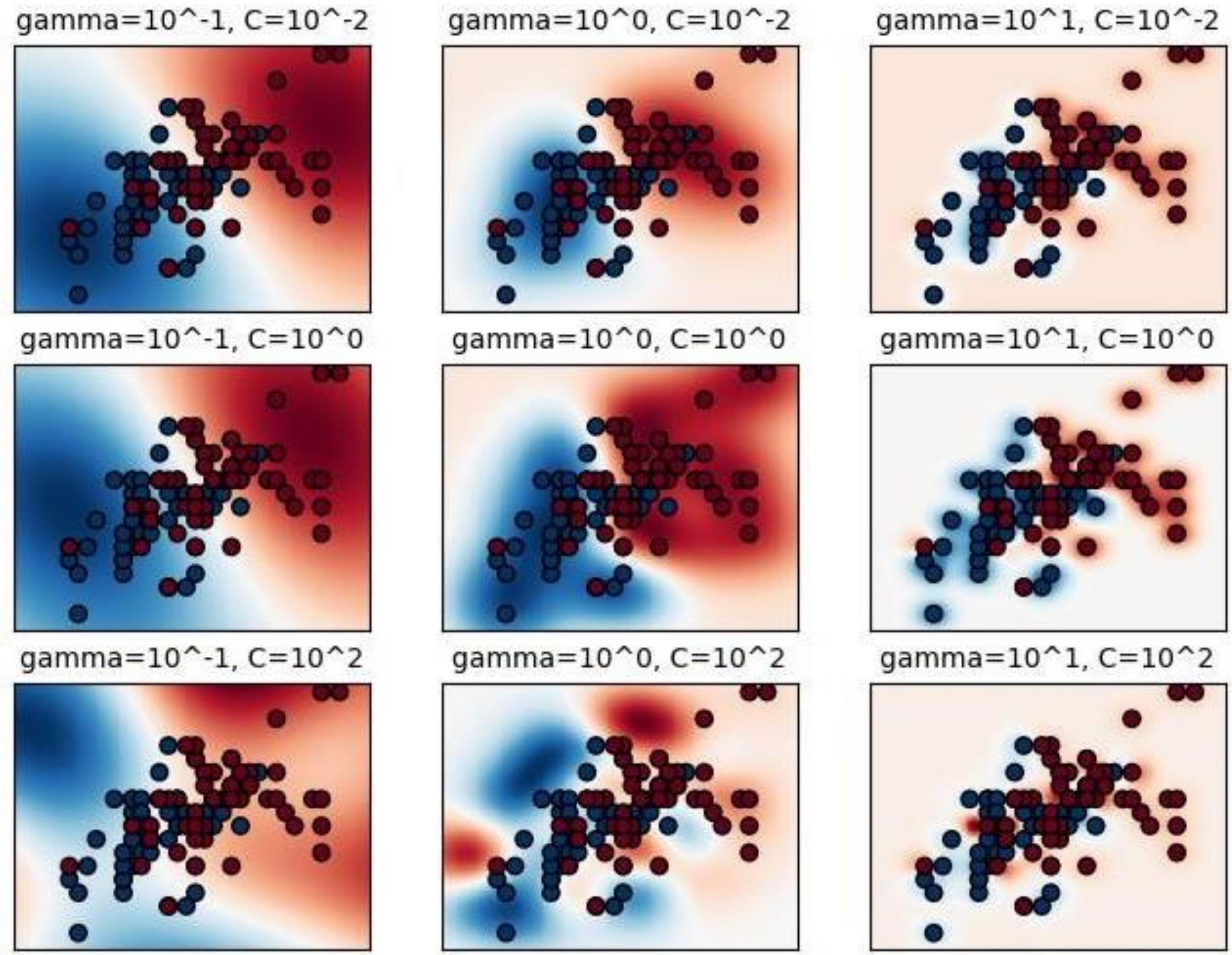


高斯径向基核函数（RBF)

各种核函数

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)^d$	$d \geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert}{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \theta)$	tanh 为双曲正切函数, $\beta > 0$, $\theta < 0$

径向基核函数参数gamma和C



逻辑回归vs. 支持向量机

联系：

- 都是监督的分类算法。
- 都是线性分类方法(不考虑核函数时）。
- 都是判别模型。

区别：

- 损失函数的不同，LR是**对数损失函数**，SVM是**hinge损失函数**。
- SVM不能产生概率，LR可以产生概率。
- SVM自带**结构风险**最小化，LR则是**经验风险**最小化。
- SVM可以用核函数，而LR一般不用核函数。

相关概念

- **判别模型**：由数据直接学习决策函数Y=f(X)，或者由条件概率分布P(Y|X)作为预测模型。判别方法关心的是给定输入X，应该预测出什么样的输出Y。SVM、LR、KNN、决策树都是判别模型。
- **生成模型**：由数据学习联合概率密度分布P(X,Y)，然后求出条件概率分布P(Y|X)。生成方法关心的是给定输入X产生输出Y的生成关系。朴素贝叶斯、隐马尔可夫模型等是生成模型。

相关概念

- **经验风险**：对所有训练样本都求一次损失函数，再累加求平均。即：模型对训练样本中所有样本的预测能力。
- **期望风险**：对所有样本（包含未知样本和已知的训练样本）的预测能力，是全局概念。（经验风险则是局部概念，仅表示决策函数对训练数据集里样本的预测能力。）
- **结构风险**：对经验风险和期望风险的折中。结构风险在经验风险的基础上加上表示模型复杂度的正则化项或惩罚项。

什么时候使用逻辑回归或SVM

特征	数据	选择模型
大量 (~10K特征)	少量 (1K 行)	简单，逻辑回归或LinearSVC
少量(<100特征)	中等 (~10k 行)	带RBF核函数的SVC
少量(<100特征)	大量 (>100K 行)	增加特征, 逻辑回归, LinearSVC 或者核近似