监督学习2

主讲：郭春乐、刘夏雷
南开大学计算机学院

致谢：本课件主要内容来自浙江大学吴飞教授、
南开大学程明明教授

单选题 1分

设置

1. 请判断下面说法是否正确:
线性判别分析是在最大化类间方差和类内方差的比值(　)

A 正确

B 错误

提交

单选题 1分

设置

2. 在决策树建立过程中,使用一个属性对某个节点对应的数据集合进行划分后,结果具有高信息熵(high entropy),对于结果的描述,最贴切的是(　)

A 纯度高

B 纯度低

C 有用

D 无用

E 以上描述都不贴切

提交

单选题 1分

设置

3. 在一个监督学习任务中,每个数据样本有4个属性和一个类别标签,每种属性分别有3、2、2和2种可能的取值,类别标签有3种不同的取值。请问可能有多少种不同的样本?(注意,并不是在某个数据集中最多有多少种不同的样本,而是考虑所有可能的样本)( )

A 3

B 6

C 12

D 24

E 48

F 72

提交

设置

4. 加入L_2标准化（normalization）后，对于包含参数w的线性回归损失函数的标准形式为：

$$L = (Y - Xw)^{\mathrm{T}}(Y - Xw) + \lambda w^T w \quad 其中\lambda > 0$$

（1）假设L_2标准化项被误写为$\lambda Y^T Y$，请解释为什么该项起不到标准化的作用。

（2）在上述L_2标准化中，如果λ小于 0，请解释为什么起不到标准化的作用。

（1）

该标准化项与参数w无关，该项对w的导数永远为 0，对w的优化求解没有作用。

（2）

L2 标准化项通过惩罚过大的参数w来避免过拟合，λ小于 0 意味着该损失函数倾向于更大的w，从而激励过拟合，失去了标准化的作用。

正常使用主观题需2.0以上版本雨课堂

作答

主观题 10分

设置

5.某城镇人均耐用消费品指出、人均年可支配收入以及耐用消费品价格指数如下表所示。请建立多元回归模型，即Y与X1,X2的回归关系（可利用电脑矩阵计算）。（多元回归分析的计算不在考核范围）

年份	人均耐用消费品支出 Y（元）	人均年可支配收入 X1（元）	耐用消费品价格指数 X2（1990 年=100）
1991	137.16	1181.4	115.96
1992	124.56	1375.7	133.35
1993	107.91	1501.2	128.21
1994	102.96	1700.6	124.85
1995	125.24	2026.6	122.49
1996	162.45	2577.4	129.86
1997	217.43	3496.2	139.52
1998	253.42	4283.0	140.44
1999	251.07	4838.9	139.12
2000	285.85	5160.3	133.35
2001	327.26	5425.1	126.39

正常使用主观题需2.0以上版本雨课堂

作答

提纲

一、机器学习基本概念

二、回归分析

三、决策树

四、线性区别分析

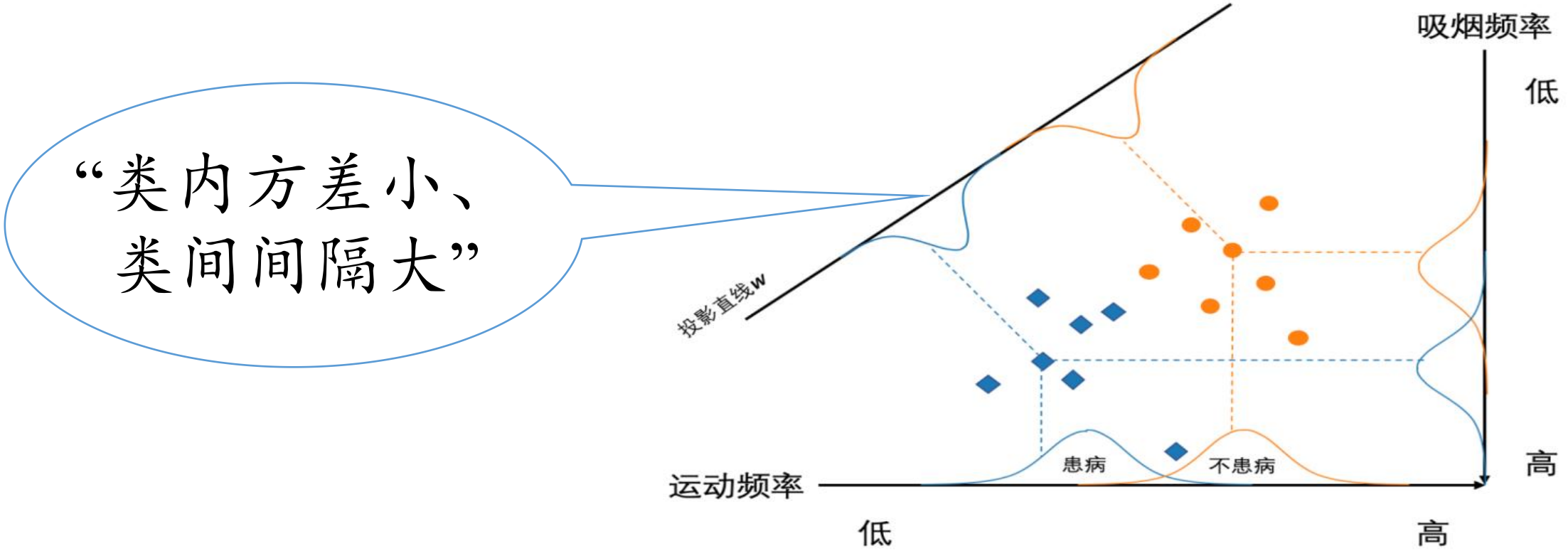
五、Ada Boosting

六、支持向量机

七、生成学习模型

线性区别分析 (linear discriminant analysis, LDA)

- 一种基于监督学习的降维方法
  - 也称为Fisher线性判别分析 (FDA) [Fisher 1936]
  - LDA利用类别信息，将高维数据样本线性投影到一个低维空间



线性区别分析：符号定义

- **假设样本集为$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)\}$，$\boldsymbol{x}_i \in \mathbb{R}^d$**
  - 其中，y_i的取值范围是$\{C_1, C_2, ..., C_K\}$，即一共有K类样本
  - 定义K为所有样本构成集合、X_i为第i类样本的集合
  - N_i为第i个类别所包含样本个数
  - $\boldsymbol{m}$为所有样本的均值向量、$\boldsymbol{m}_i$为第i类样本的均值向量
- **Σ_i为第i类样本的协方差矩阵，定义为:**

$$\Sigma_i = \sum_{\boldsymbol{x} \in \boldsymbol{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T$$

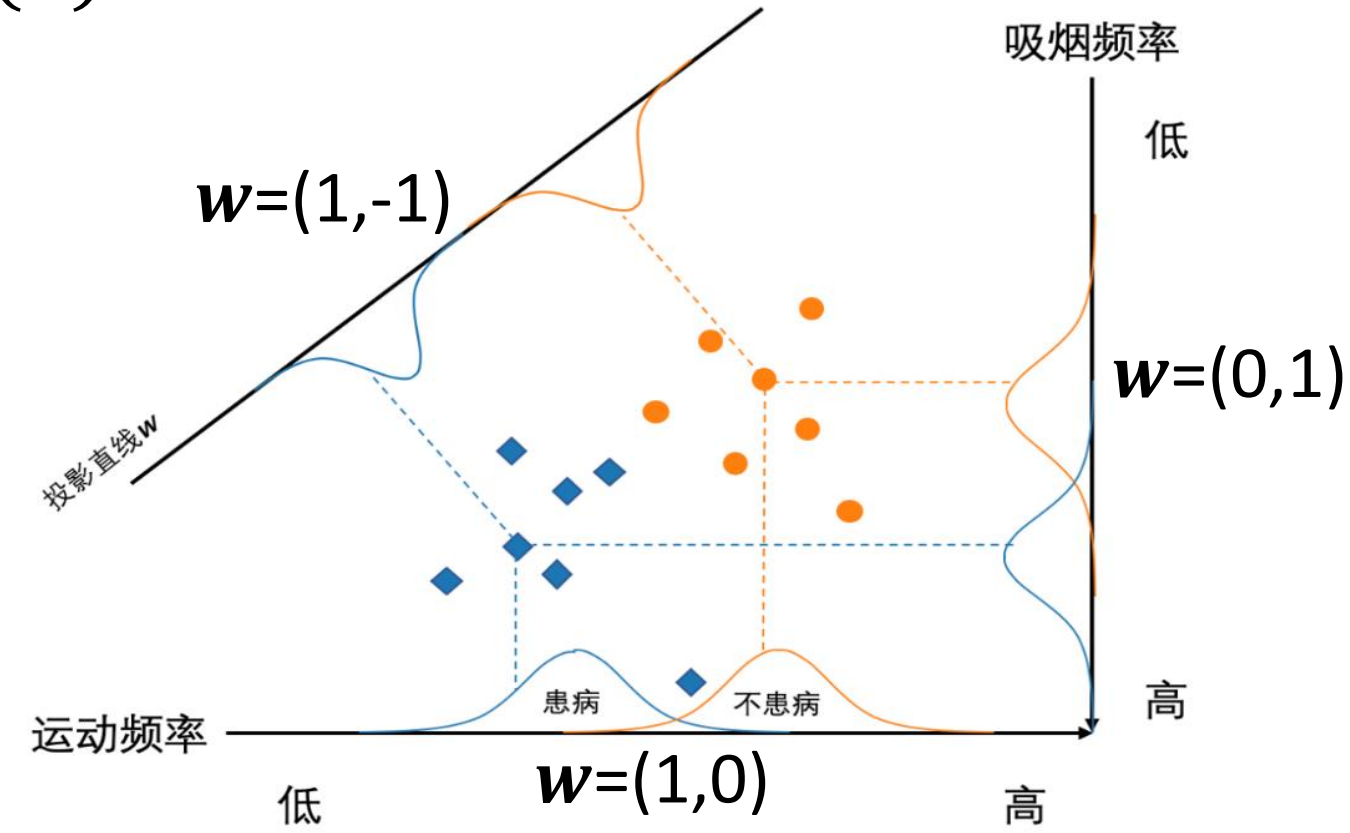
线性区别分析：二分类问题

- **先来看$K = 2$的情况：训练样本归属于C_1或C_2两个类别**
  - 过如下的线性函数投影到一维空间上（其中$\boldsymbol{w} \in \mathbb{R}^n$）

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

节点(1,1), (2,2), (3,3), (4,4)

都会投影到同一个点。



线性区别分析：二分类问题

- 先来看$K = 2$的情况：训练样本归属于$\mathcal{C}_1$或$\mathcal{C}_2$两个类别
  - 过如下的线性函数投影到一维空间上（其中$\boldsymbol{w} \in \mathbb{R}^n$）

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$$

- 投影之后类别$\mathcal{C}_1$的协方差矩阵s_1为：

$$s_1 = \sum_{\boldsymbol{x}\in\mathcal{C}_1} (\boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{m}_1)^2 = \boldsymbol{w}^T \sum_{\boldsymbol{x}\in\mathcal{C}_1} [(\boldsymbol{x} - \boldsymbol{m}_1)(\boldsymbol{x} - \boldsymbol{m}_1)^T]\boldsymbol{w}$$

- 同理可得到投影之后类别$\mathcal{C}_2$的协方差矩阵s_2

线性区别分析：二分类问题

- 投影后两个协方差矩阵为$s_1 = \boldsymbol{w}^T \Sigma_1 \boldsymbol{w}$和$s_2 = \boldsymbol{w}^T \Sigma_2 \boldsymbol{w}$
  - 为了使同类本尽可能靠近(分散程度低)，需要最小化$s_1 + s_2$
- 投影后，归属于两个类别的数据样本中心为：

$$m_1 = \boldsymbol{w}^T \boldsymbol{m}_1, \quad m_2 = \boldsymbol{w}^T \boldsymbol{m}_2$$

- 为使不同类样本尽可能彼此远离，需要最大化

$$\|m_2 - m_1\|_2^2$$

- 总体需要最大化的目标( )定义为

$$J(\boldsymbol{w}) = \frac{\|m_2 - m_1\|_2^2}{s_1 + s_2}$$

线性区别分析：二分类问题

$$J(\boldsymbol{w}) = \frac{\|\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)\|_2^2}{\boldsymbol{w}^T\Sigma_1\boldsymbol{w} + \boldsymbol{w}^T\Sigma_2\boldsymbol{w}} = \frac{\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\boldsymbol{w}}{\boldsymbol{w}^T(\Sigma_1 + \Sigma_2)\boldsymbol{w}} = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$$

- **$\boldsymbol{S}_b$称为类间散度矩阵(between-class scatter matrix)**
  - 衡量两个类别均值点之间的“分离”程度:

$$\boldsymbol{S}_b = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T$$

- **$\boldsymbol{S}_w$称为类内散度矩阵(within-class scatter matrix)**
  - 衡量每个类别中数据点的“分离”程度:

$$\boldsymbol{S}_w = \Sigma_1 + \Sigma_2$$

- **由于$J(\boldsymbol{w})$的分子和分母都是关于$\boldsymbol{w}$的二项式，因此解只与$\boldsymbol{w}$的方向有关，与$\boldsymbol{w}$的长度无关，因此可令分母$\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w} = 1$，用拉格朗日乘子法来求解**

线性区别分析：二分类问题

- 对应拉格朗日函数为：

$$L(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - 1)$$

- 对$\boldsymbol{w}$求偏导并使其求导结果为零，可得$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w} = \lambda\boldsymbol{w}$
  - λ和$\boldsymbol{w}$分别是$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$的特征根和特征向量
  - $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w} = \lambda\boldsymbol{w}$也被称为Fisher线性判别

$$\boldsymbol{S}_w{}^{-1}\boldsymbol{S}_b\boldsymbol{w} = \boldsymbol{S}_w{}^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \times \lambda_w = \lambda\boldsymbol{w}$$

- 由于对$\boldsymbol{w}$的放大缩小不影响结果，可约去未知数λ和λ_w：

$$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$$

线性区别分析：多分类问题

假设n个原始高维数据所构成的类别种类为K、每个原始数据被投影映射到低维空间中的维度为r。

令投影矩阵$\boldsymbol{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_r)$，可知$\boldsymbol{W}$是一个$n\times r$矩阵。于是，$\boldsymbol{W}^T\boldsymbol{m}_i$为第$i$类样本数据中心在低维空间的投影结果，$\boldsymbol{W}^T\Sigma_i\boldsymbol{W}$为第$i$类样本数据协方差在低维空间的投影结果。

类内散度矩阵$\boldsymbol{S}_w$重新定义如下：

$$\boldsymbol{S}_w=\sum_{i=1}^{K}\Sigma_i,\ \text{其中}\Sigma_i=\sum_{\boldsymbol{x}\in class\ i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T$$

在上式中，$\boldsymbol{m}_i$是第i个类别中所包含样本数据的均值。

类间散度矩阵$\boldsymbol{S}_b$重新定义如下：

$$\boldsymbol{S}_b=\sum_{i=1}^{K}\frac{N_i}{N}(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T$$

线性区别分析：多分类问题

将多类LDA映射投影方向的优化目标$J(\boldsymbol{W})$改为：

$$J(\boldsymbol{W}) = \frac{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}}$$

其中，$\prod_{diag} \mathbb{A}$为矩阵$\mathbb{A}$主对角元素的乘积。

继续对$J(\boldsymbol{W})$进行变形：

$$J(\boldsymbol{W}) = \frac{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\prod_{diag} \boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}} = \frac{\prod_{i=1}^{r} \boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\prod_{i=1}^{r} \boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i} = \prod_{i=1}^{r} \frac{\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i}$$

显然需要使乘积式子中每个$\frac{\boldsymbol{w}_i^T \boldsymbol{S}_b \boldsymbol{w}_i}{\boldsymbol{w}_i^T \boldsymbol{S}_w \boldsymbol{w}_i}$取值最大，这就是二分类问题的求解目标，即每一个$\boldsymbol{w}_i$都是$\boldsymbol{S}_w^{-1} \boldsymbol{S}_b \boldsymbol{W} = \lambda \boldsymbol{W}$的一个解。

线性区别分析：线性判别分析的降维步骤

对线性判别分析的降维步骤描述如下：

1. 计算数据样本集中每个类别样本的均值
2. 计算类内散度矩阵S_w和类间散度矩阵S_b
3. 根据${S_w}^{-1}S_bW = \lambda W$来求解${S_w}^{-1}S_b$所对应前$r$个最大特征值所对应特征向量$(w_1, w_2, ..., w_r)$，构成矩阵$W$
4. 通过矩阵W将每个样本映射到低维空间，实现特征降维。

线性区别分析：与主成分分析法的异同

	线性判别分析	主成分分析
是否需要样本标签	监督学习	无监督学习
降维方法	优化寻找特征向量$\boldsymbol{w}$	优化寻找特征向量$\boldsymbol{w}$
目标	类内方差小、类间距离大	寻找投影后数据之间方差最大的投影方向
维度	LDA降维后所得到维度是与数据样本的类别个数K有关	PCA对高维数据降维后的维数是与原始数据特征维度相关

提纲

一、机器学习基本概念

二、回归分析

三、决策树

四、线性区别分析

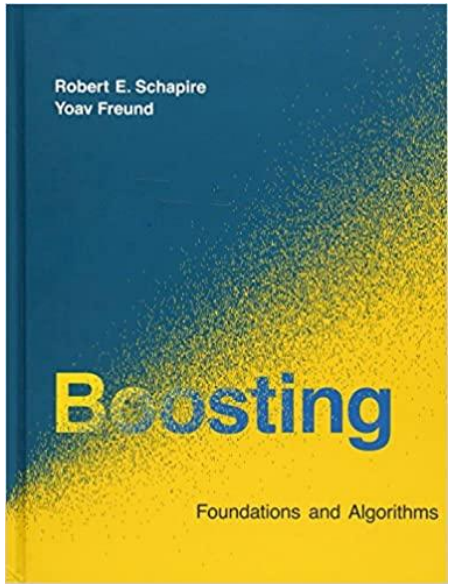
五、Ada Boosting

六、支持向量机

七、生成学习模型

Boosting（adaptive boosting, 自适应提升)

- **Boosting: a machine learning approach**
  - Creating a highly accurate predictor by combining many weak and inaccurate “rules of thumb.”
  - A remarkably rich theory has evolved around boosting
    - With connections to statistics, game theory, convex optimization, and information geometry.
  - Enjoyed practical success in
    - Biology, vision, and speech processing.



Boosting: Foundations and Algorithms by Robert E. Schapire and Yoav Freund.

Adaptive boosting

- 对于一个复杂的分类任务，可以将其分解为若干子任务，然后将若干子任务完成方法综合，最终完成该复杂任务。
- 将若干个弱分类器(weak classifiers)组合起来，形成一个强分类器(strong classifier)。

能用众力，则无敌于天下矣；能用众智，则无畏于圣人矣

《三国志·吴志·孙权传》

计算学习理论 (Computational Learning Theory)

- 可计算：什么任务是可以计算的？图灵可停机
- 可学习：什么任务是可以被学习的、从而被学习模型来完成？
- Leslie Valiant (2010年图灵奖获得者)和其学生Michael Kearns两位学者提出了这个问题并进行了有益探索，逐渐完善了计算学习理论。

计算学习理论：霍夫丁不等式(Hoeffding's inequality)

- **学习任务：统计某个电视节目在全国的收视率。**
  - 方法：不可能去统计整个国家中每个人是否观看电视节目、进而算出收视率。只能抽样一部分人口，然后将抽样人口中观看该电视节目的比例作为该电视节目的全国收视率。
- **霍夫丁不等式：全国人口收视率x与抽样人口中收视率y满足**

$$P(|x-y| \geq \epsilon) \leq 2e^{-2N\epsilon^2}$$

  - 其中，N是采样人口总数、$\epsilon \in (0,1)$是可容忍误差范围

当N足够大时，“全国人口收视率”与“样本人口收视率”差值超过误差范围ϵ的概率非常小。

计算学习理论：概率近似正确 (PAC)

- **对于统计收视率这样的任务，可以用不同的采样方法来计算**
  - 即用不同模型，每个模型会产生不同的误差。
- **这就是概率近似正确（probably approximately correct, PAC）要回答的问题**
  - 如果得到完成任务的若干“弱模型”，是否可以将这些弱模型组合起来，形成一个“强模型”，使其误差很小呢?

计算学习理论：概率近似正确 (PAC)

- 在PAC背景下，有“强可学习模型”和“弱可学习模型”

强可学习 (strongly learnable)	学习模型能够以较高精度对绝大多数样本完成识别分类任务
弱可学习 (weakly learnable)	学习模型仅能完成若干部分样本识别与分类，其精度略高于随机猜测。
强可学习和弱可学习是等价的，也就是说，如果已经发现了“弱学习算法”，可将其提升（boosting）为“强学习算法”。Ada Boosting算法就是这样的方法。具体而言，Ada Boosting将一系列弱分类器组合起来，构成一个强分类器。	

Ada Boosting：思路描述

- **Ada Boosting算法中两个核心问题：**
  - 在每个弱分类器学习过程中，如何改变训练数据的权重：提高在上一轮中分类错误样本的权重。
  - 如何将一系列弱分类器组合成强分类器：通过加权多数表决方法来提高分类误差小的弱分类器的权重，让其在最终分类中起到更大作用。同时减少分类误差大的弱分类器的权重，让其在最终分类中仅起到较小作用。

Ada Boosting：算法描述---数据样本权重初始化

- 给定包含N个标注数据的训练集合$\Gamma = \{(x_1, y_1), \ldots, (x_N, y_N)\}$，$x_i (1 \le i \le N) \in X \subseteq R^n, y_i \in Y = \{-1, 1\}$
- Ada Boosting算法将从这些标注数据出发，训练得到一系列弱分类器，并将这些弱分类器线性组合得到一个强分类器。

1. 初始化每个训练样本的权重

$$D_1 = (w_{11}, \ldots, w_{1i}, \ldots, w_{1N}), \quad 其中\ w_{1i} = \frac{1}{N} (1 \le i \le N)$$

2. 迭代地利用加权样本训练弱分类器并增加错分类样本权重
3. 以线性加权形式来组合弱分类器

Ada Boosting：算法描述---第𝑚个弱分类器训练

- 迭代地利用加权样本训练弱分类器并增加错分类样本权重
- 对$m = 1,2,\ldots,M$
  - 使用具有分布权重D_m的训练数据来学习得到第m个弱分类

$$G_m(x): X \to \{-1,1\}$$

  - 计算$G_m(x)$在训练数据集上的分类误差，其中$I(\cdot)$为示性函数

$$err_m = \sum_{i=1}^{N} w_{mi} I(G_m(x_i) \neq y_i)$$

  - 计算弱分类器$G_m(x)$的权重：$\alpha_m = \frac{1}{2}\ln\frac{(1-err_m)}{err_m}$
  - 更新训练样本数据的分布权重D_{m+1}为$w_{m+1,i} = \frac{w_{m,i}}{Z_m} e^{-\alpha_m y_i G_m(x_i)}$
    - 其中归一化因子$Z_m = \sum_{i=1}^{N} w_{m,i} e^{-\alpha_m y_i G_m(x_i)}$使得$D_{m+1}$为概率分布

Ada Boosting：算法描述---弱分类器组合成强分类器

- 以线性加权形式来组合弱分类器$f(x)$

$$f(x) = \sum_{i=1}^{M} \alpha_m G_m(x)$$

- 得到强分类器$G(x)$

$$G(x) = sign\big(f(x)\big) = sign\big(\sum_{i=1}^{M} \alpha_m G_m(x)\big)$$

Ada Boosting：算法解释

- **第m个弱分类器$G_m(x)$在训练数据集上产生的分类误差**
  - 该误差为被错误分类的样本所具有权重的累加

$$err_m = \sum_{i=1}^{N} w_{m,i} \, I(G_m(x_i) \neq y_i)$$

  - 这里$I(\cdot)$为示性函数

Ada Boosting：算法解释

- **计算第m个弱分类器$G_m(x)$的权重$\alpha_m = \frac{1}{2}\ln\frac{1-err_m}{err_m}$**
  - 当$G_m(x)$错误率$err_m = \sum_{i=1}^{N} w_{mi}\, I(G_m(x_i) \neq y_i) = 1$，意味每个样本分类出错，则$\alpha_m = \frac{1}{2}\ln\frac{1-err_m}{err_m} \to -\infty$。
  - 当$G_m(x)$错误率为1/2，$\alpha_m = \frac{1}{2}\ln\frac{1-err_m}{err_m} = 0$。如果错误率$err_m$小于1/2，权重$\alpha_m$为正($err_m < 1/2$、$\alpha_m > 0$)。可知权重$\alpha_m$随$err_m$减少而增大，即错误率越小的弱分类器会赋予更大权重。
  - 如果错误率为1/2，可视为该弱分类器仅相当于随机分类效果

Ada Boosting：算法解释

- **在训练第$m+1$个弱分类器$G_{m+1}(x)$前调整训练数据权重**
  - 如果某个样本无法被第m个弱分类器$G_m(x)$分类成功，则增大该样本权重，否则减少该样本权重。被错误分类样本在训练第$m+1$个弱分类器$G_{m+1}(x)$时会被“重点关注”
  - 在每一轮学习过程中，Ada Boosting算法均在划重点（重视当前尚未被正确分类的样本）

$$w_{m+1,i}=\begin{cases}\dfrac{w_{m,i}}{Z_m}e^{-\alpha_m}, & G_m(x_i)=y_i\\ \dfrac{w_{m,i}}{Z_m}e^{\alpha_m}, & G_m(x_i)\neq y_i\end{cases}$$

Ada Boosting：算法解释

- 弱分类器构造强分类器
  - $f(x)$是M个弱分类器的加权线性累加。分类能力越强的弱分类器具有更大权重。
  - α_m累加之和并不等于1。
  - $f(x)$符号决定样本x分类为1或-1。如果$\sum_{i=1}^{M}\alpha_m G_m(x)$为正，则强分类器$G(x)$将样本$x$分类为1；否则为-1。

$$G(x) = sign\big(f(x)\big) = sign\left(\sum_{i=1}^{M}\alpha_m G_m(x)\right)$$

Ada Boosting：回看霍夫丁不等式

- 个弱分类器G_m的线性组合所产生误差满足

$$P\left(\sum_{i=1}^{M} G_m(x) \neq \zeta(x)\right) \leq e^{-\frac{1}{2}M(1-2\epsilon)^2}$$

  - $\zeta(x)$是真实分类函数、$\epsilon \in (0,1)$
  - 学习分类误差随弱分类器数增长呈指数级下降，直至为零
  - 两个前提条件：每个弱分类器1) 误差相互独立； 2) 误差率小于50%
  - 每个弱分类器均在同一个训练集上产生，条件1) 难以满足。因此，分类结果的“准确性”和分类器的“差异性”难以同时满足。
  - Ada Boosting 采取了序列化学习机制。

Ada Boosting：优化目标

- AdaBoost实际在最小化指数损失函数(exponential loss)

$$\sum_i e^{-y_i f(x_i)} = \sum_i e^{-y_i \sum_{m=1}^{M} \alpha_m G_m(x_i)}$$

- AdaBoost的分类误差上界为：

$$\frac{1}{N}\sum_{i=1}^{N} I(G(x_i) \neq y_i) \leq \frac{1}{N}\sum_i e^{-y_i f(x_i)} = \prod_m Z_m$$

  - 在第m次迭代中，AdaBoost趋向于将具有最小误差的模型选做本轮生成的弱分类器G_m，使得累积误差快速下降。

Ada Boosting：算法例子

Pattern Recognition and Machine Learning, C.M. Bishop, 2006.

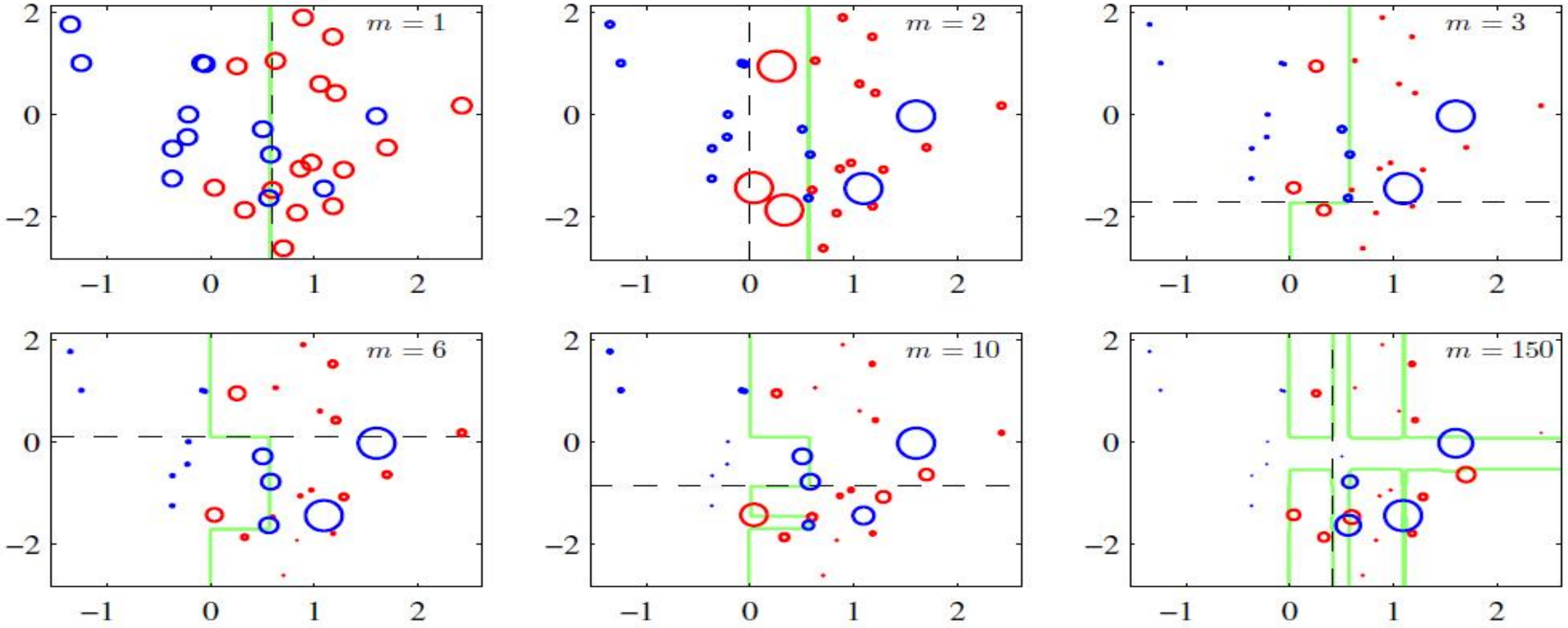


Figure 14.2 Illustration of boosting in which the base learners consist of simple thresholds applied to one or other of the axes. Each figure shows the number m of base learners trained so far, along with the decision boundary of the most recent base learner (dashed black line) and the combined decision boundary of the ensemble (solid green line). Each data point is depicted by a circle whose radius indicates the weight assigned to that data point when training the most recently added base learner. Thus, for instance, we see that points that are misclassified by the $m = 1$ base learner are given greater weight when training the $m = 2$ base learner.

回归与分类的区别

- 均是学习将输入变量映射到输出变量的潜在关系模型

$$f: \quad x \rightarrow y, \; x \in A, \; y \in B$$

映射函数（学习模型）　输入变量　输出变量　变量域　值域

- 在回归分析中，学习一个函数将输入变量映射到连续输出空间
  - 如价格和温度等，即值域是连续空间
- 在分类模型中，学习一个函数将输入变量映射到离散输出空间
  - 如人脸和汽车等，即值域是离散空间

提纲

一、机器学习基本概念

二、回归分析

三、决策树

四、线性区别分析

五、Ada Boosting

六、支持向量机

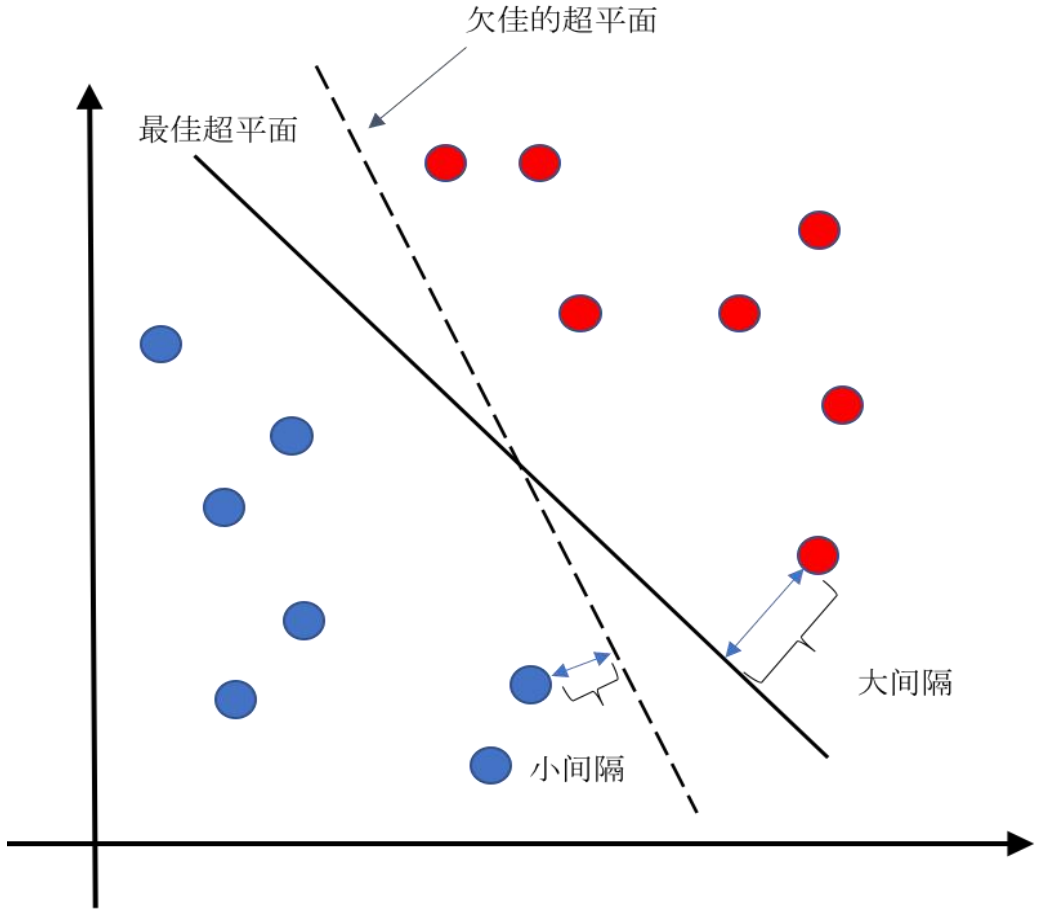
七、生成学习模型

支持向量机：VC维与结构风险最小化

- **传统识别理论认为，算法模型性能可用从训练样本集所得经验风险（empirical risk）来衡量。**
  - 经验风险指模型在训练样本集中所有数据上所得误差的累加
  - 显然，经验风险越小，算法模型对训练数据拟合程度越好
  - 实际中，一味降低经验风险容易造成过学习问题(over-fitting)

支持向量机：VC维与结构风险最小化

- **支持向量机（support vector machine, SVM）**
  - 通过结构风险(structural risk)最小化来解决过学习问题



一个两类分类问题的最佳分类平面。图中存在多个可将样本分开的超平面。支持向量机学习算法会去寻找一个最佳超平面，使得每个类别中距离超平面最近的样本点到超平面的最小距离最大。

支持向量机：VC维与结构风险最小化

- **支持向量机认为：分类器对未知数据（即测试数据）进行分类时所产生的期望风险（即真实风险）不是由经验风险单独决定的，而是由两部分组成：**
  - 1）从训练集合数据所得经验风险（如果经验风险小、期望风险很大，则是过学习）；
  - 2）置信风险（confidence risk），它与分类器的[illegible]维及训练样本数目有关。

支持向量机：VC维与结构风险最小化

- **Vapnik推导出期望风险$\Re$和经验风险$\Re_{emp}$以$1-\eta$的概率满足:**

$$\Re \leq \Re_{emp} + \sqrt{\frac{h(log\,(2n/h)+1)-log\,(\eta/4)}{n}}$$

  - 其中，$\sqrt{\frac{h(log\,(2n/h)+1)-log\,(\eta/4)}{n}}$叫做“□□置信值”
  - $0 \leq \eta \leq 1, n$是训练样本个数，h是反映学习机复杂程度的□□维

支持向量机：VC维与结构风险最小化

- **Vapnik推导出期望风险$\Re$和经验风险$\Re_{emp}$以$1-\eta$的概率满足:**

$$\Re \leq \Re_{emp} + \sqrt{\frac{h(log\,(2n/h)+1) - log\,(\eta/4)}{n}}$$

  - 期望风险$\Re$代表了分类器对未知数据分类推广能力，$\Re$越小越好
  - VC置信值是h的增函数，$\Re_{emp}$是h的减函数，于是选择一个折中的h值可以使期望风险$\Re$达到最小。支持向量机使用结构风险最小化准则来选取VC维h，使每一类别数据之间的分类间隔(Margin)最大，最终使实期望风险$\Re$最小。

支持向量机：VC维与结构风险最小化

- 假设空间的VC维
  - 将d个数据分为两类，可以有2^n种分法，即可理解成有2^n个学习问题。若存在一个假设$\mathcal{H}$，能准确无误地将2^n种问题进行分类，那么d就是$\mathcal{H}$的VC维。
  - 在d维空间中，线性决策面的VC维为$d+1$。VC维就是对假设空间$\mathcal{H}$复杂度的一种度量。当样本数d固定时，如果VC维越高，则算法模型的复杂性越高。VC维越大，通常推广能力越差，置信风险会变大。如果算法模型一味降低经验风险，则会提高模型复杂度，导致VC维很高，置信风险大，使得期望风险就大。

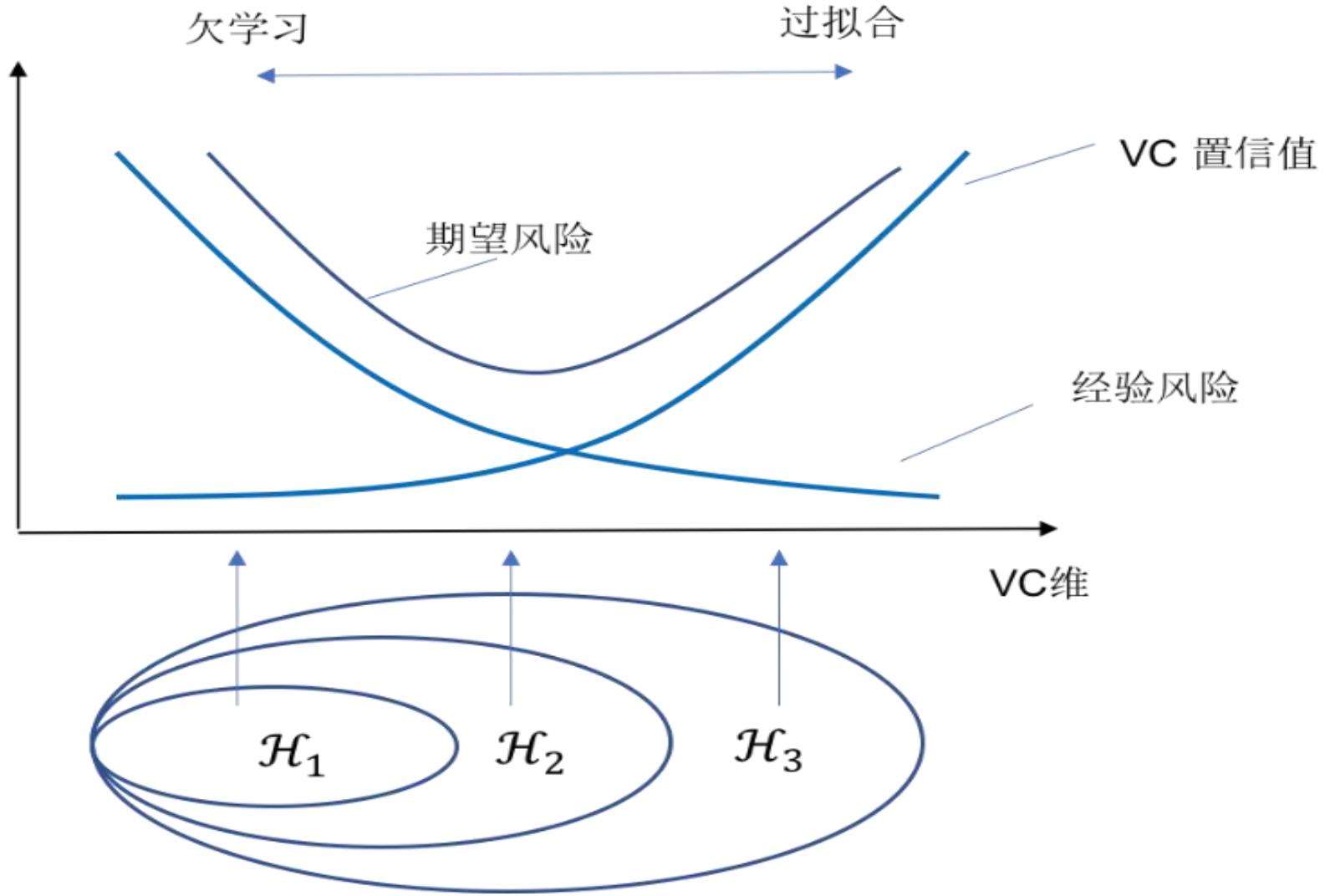
支持向量机：VC维与结构风险最小化

- **通过VC理论，可认识到期望风险（即真实风险）$\Re$与经验风险$\Re_{emp}$之间是有差别的，这个差别项被称为置信风险**
  - 它与训练样本个数和模型复杂度都有密切的关系
  - 用复杂度高的模型去拟合小样本，往往会导致过拟合
  - 因此需要给经验风险$\Re_{emp}$加上一个惩罚项或者正则化项，以同时考虑经验风险与置信风险。这一思路被称为结构风险最小化。
  - 这样，在小样本情况下可取得较好性能。
  - 在保证分类精度高（经验风险小）同时，有效降低算法模型VC维，可使算法模型在整个样本集上的期望风险得到控制。

支持向量机：VC维与结构风险最小化

- 经验风险、期望风险、VC置信度、VC维、过学习和欠学习的关系

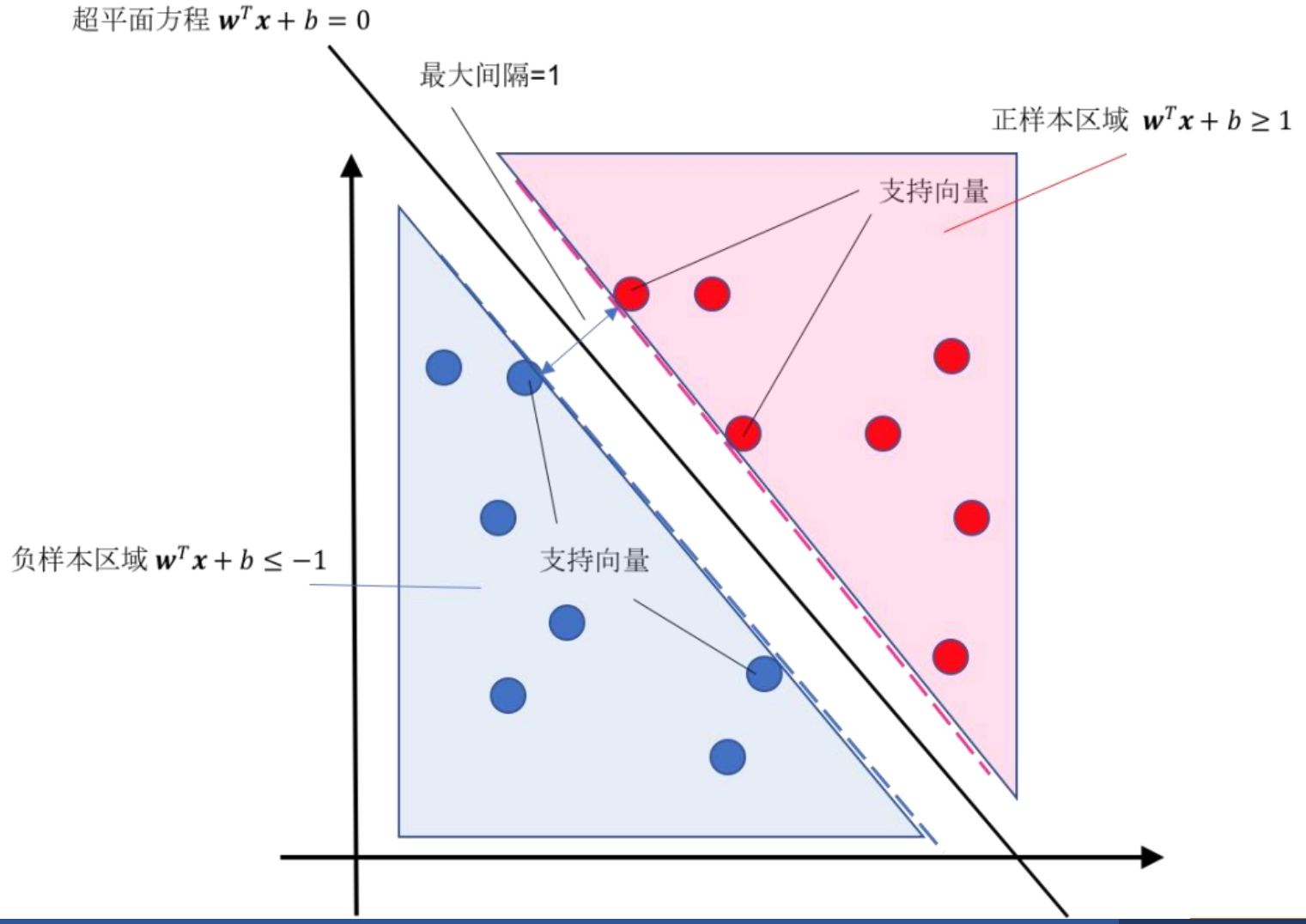
训练样本给定，
分类间隔越大，
则分类超平面集
合的 VC 维就越小



支持向量机：线性可分支持向量机

- 经验风险、期望风险、VC置信度、VC维、过学习和欠学习的关系

寻找一个最优的超平面，其方程为$\boldsymbol{w^T x} + b = 0$。这里$\boldsymbol{w} = (w_1, w_2, \ldots, w_d)$为超平面的法向量，与超平面的方向有关；b为偏置项，是一个标量，其决定了超平面与原点之间的距离。



支持向量机：线性可分支持向量机

- 由于法向量w中的值可按比例任意缩放而不改变法向量方向，使得分类平面不唯一。为此，对w和b添加如下约束：

$$r_{min_i}|w^T x_i + b| = 1$$

- 即离超平面最近的正负样本代入超平面方程后其绝对值为1。
- 于是对超平面的约束变为：$y_i(w^T x_i + b) \geq 1$

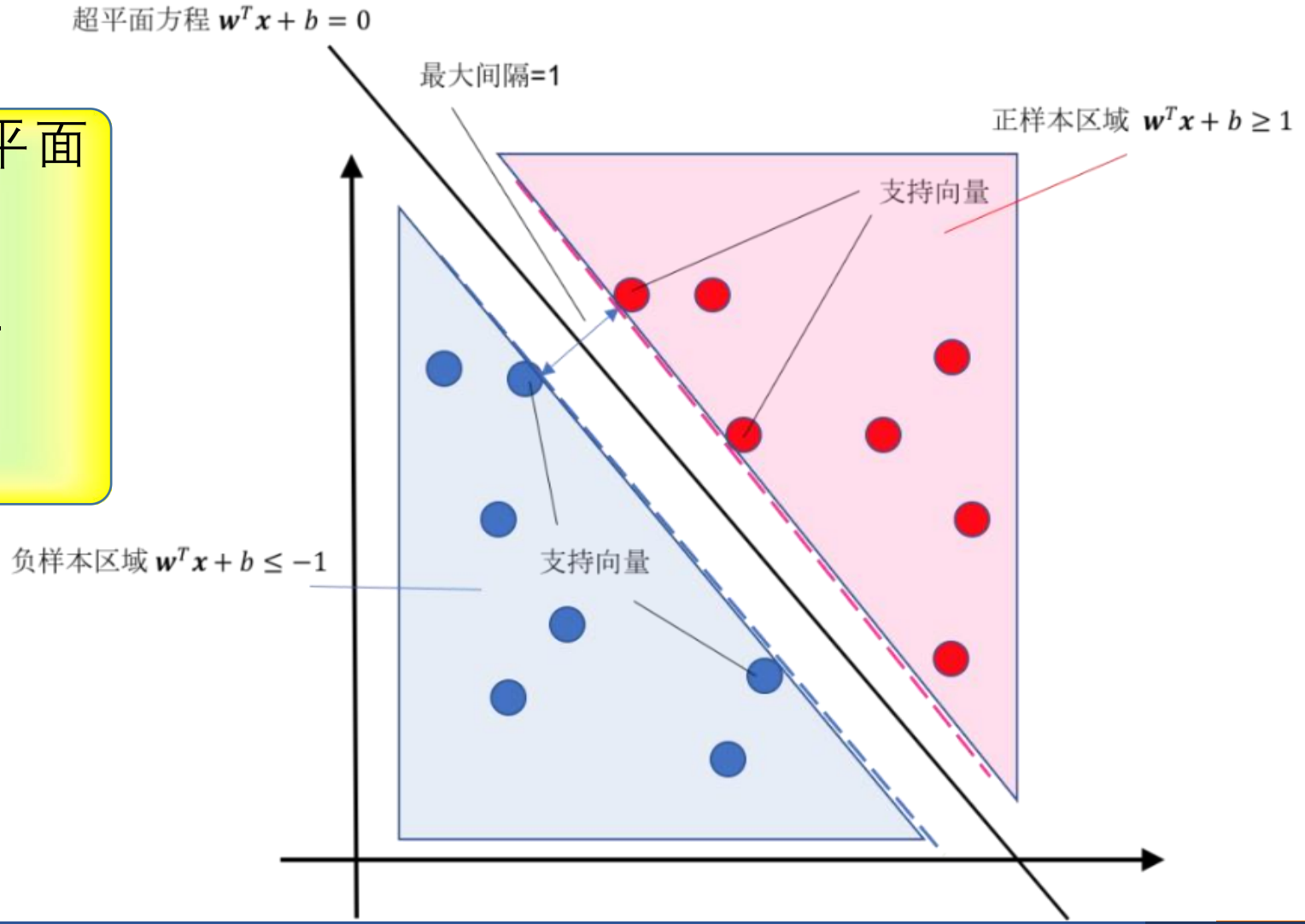
支持向量机：线性可分支持向量机

- 对超平面的约束为：$y_i(\boldsymbol{w}^T\boldsymbol{x_i}+b)\geq 1$

样本空间中任意样本$\boldsymbol{x}$到该平面距离可表示为:

$$r=d(\boldsymbol{w},b,x)=\frac{|\boldsymbol{w}^T\boldsymbol{x}+b|}{||\boldsymbol{w}||_2}$$

其中$||\boldsymbol{w}||_2=\sqrt{\boldsymbol{w}^T\boldsymbol{w}}$



支持向量机：线性可分支持向量机

- 两类样本中离分类超平面最近的数据之间的距离$d(\boldsymbol{w}, b)$为：

$$min_{(\boldsymbol{x}_k, y_k=1)} d(\boldsymbol{w}, b, \boldsymbol{x}_k) + min_{(\boldsymbol{x}_m, y_m=-1)} d(\boldsymbol{w}, b, \boldsymbol{x}_m)$$

$$= min_{(\boldsymbol{x}_k, y_k=1)} \frac{|\boldsymbol{w}^T \boldsymbol{x}_k + b|}{||\boldsymbol{w}||_2} + min_{(\boldsymbol{x}_m, y_m=-1)} \frac{|\boldsymbol{w}^T \boldsymbol{x}_m + b|}{||\boldsymbol{w}||_2}$$

$$= \frac{1}{||\boldsymbol{w}||_2} \left(min_{(\boldsymbol{x}_k, y_k=1)} |\boldsymbol{w}^T \boldsymbol{x}_k + b| + min_{(\boldsymbol{x}_m, y_m=-1)} |\boldsymbol{w}^T \boldsymbol{x}_m + b| \right)$$

$$= \frac{2}{||\boldsymbol{w}||_2}$$

支持向量机：线性可分支持向量机

- 两类样本中离分类超平面最近的数据间的距离$d(\boldsymbol{w}, b) = \frac{2}{||\boldsymbol{w}||_2}$
- 支持向量机的基本形式就是最大化分类间隔，即在满足约束的条件下找到参数$\boldsymbol{w}$和b使得间隔γ最大，等价于：

$$\min_{\boldsymbol{w},b} \frac{||\boldsymbol{w}||^2}{2} = \min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x} + \text{b}) \geq 1, i = 1,2,\ldots,n$$

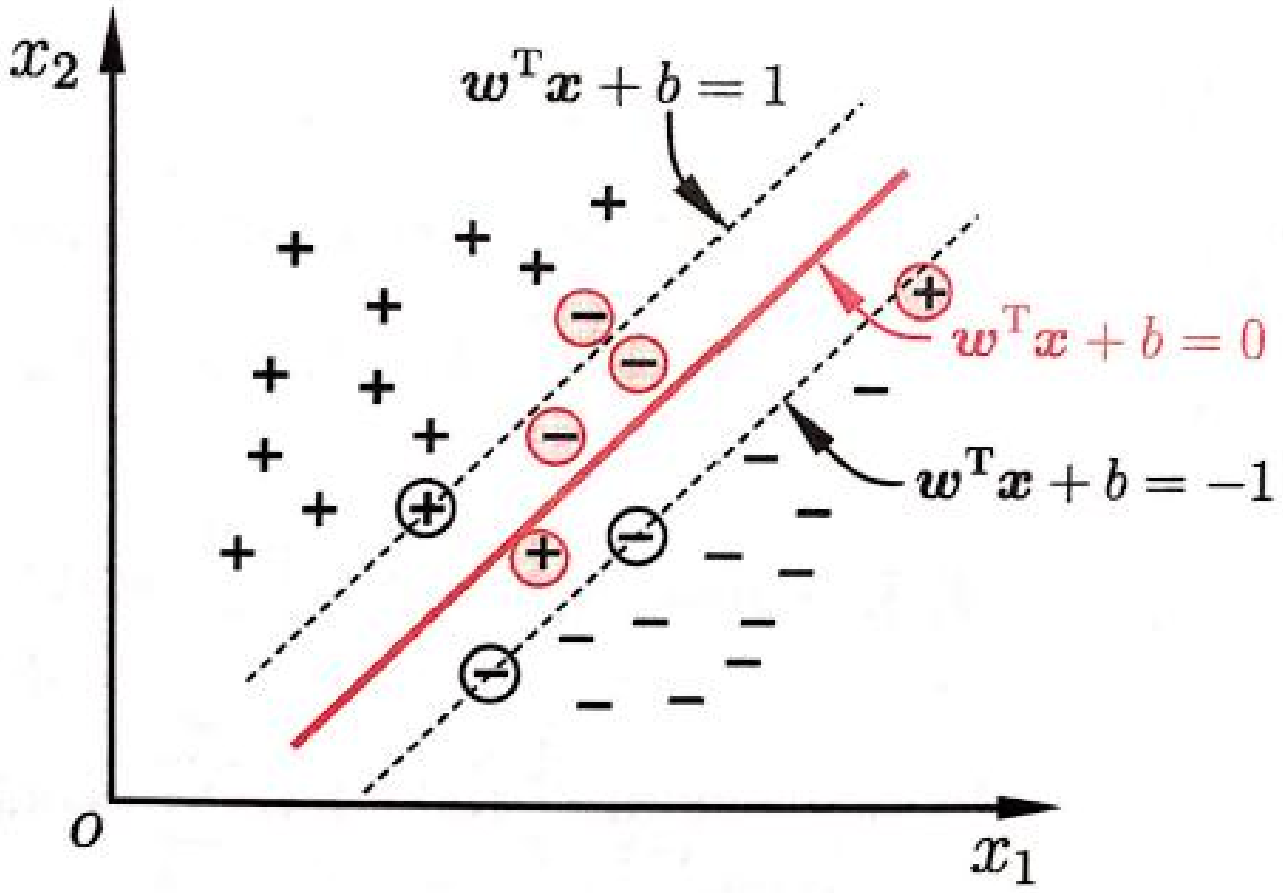
凸二次规划

支持向量机：松弛变量，软间隔与hinge损失函数

- **“硬间隔”(hard margin)**
  - 假设所有训练样本数据是线性可分，即存在一个线性超平面能将不同类别样本完全隔开
- **与硬间隔相对的是“软间隔”(soft margin)**
  - 软间隔指允许部分错分给定的训练样本。

支持向量机：松弛变量，软间隔与hinge损失函数

- “硬间隔”(hard margin) 与 “软间隔”(soft margin)



$$\min_{w,b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+c\times\sum_{i=1}^{n}[\![y_i\neq sign(\boldsymbol{w}^T\boldsymbol{x}_i+b)]\!]$$

$$s.t.\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1\ for\ correct\ \boldsymbol{x}_i$$

难以直接求解

支持向量机：松弛变量，软间隔与hinge损失函数

- hinge损失函数：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c \times \sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

  - 正确分类数据的hinge损失$\max\left(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\right) = 0$

- 记$\xi_i = \max\left(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\right)$为第$i$个变量的“松弛变量”
  - 显然$\xi_i \geq 0$。每一个样本对应一个松弛变量(slack variables)，用来表示该样本被分类错误所产生的损失。

支持向量机：软间隔支持向量机

- hinge损失函数：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c \times \sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

- 记$\xi_i = \max\left(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\right)$为第$i$个变量的“松弛变量”，则

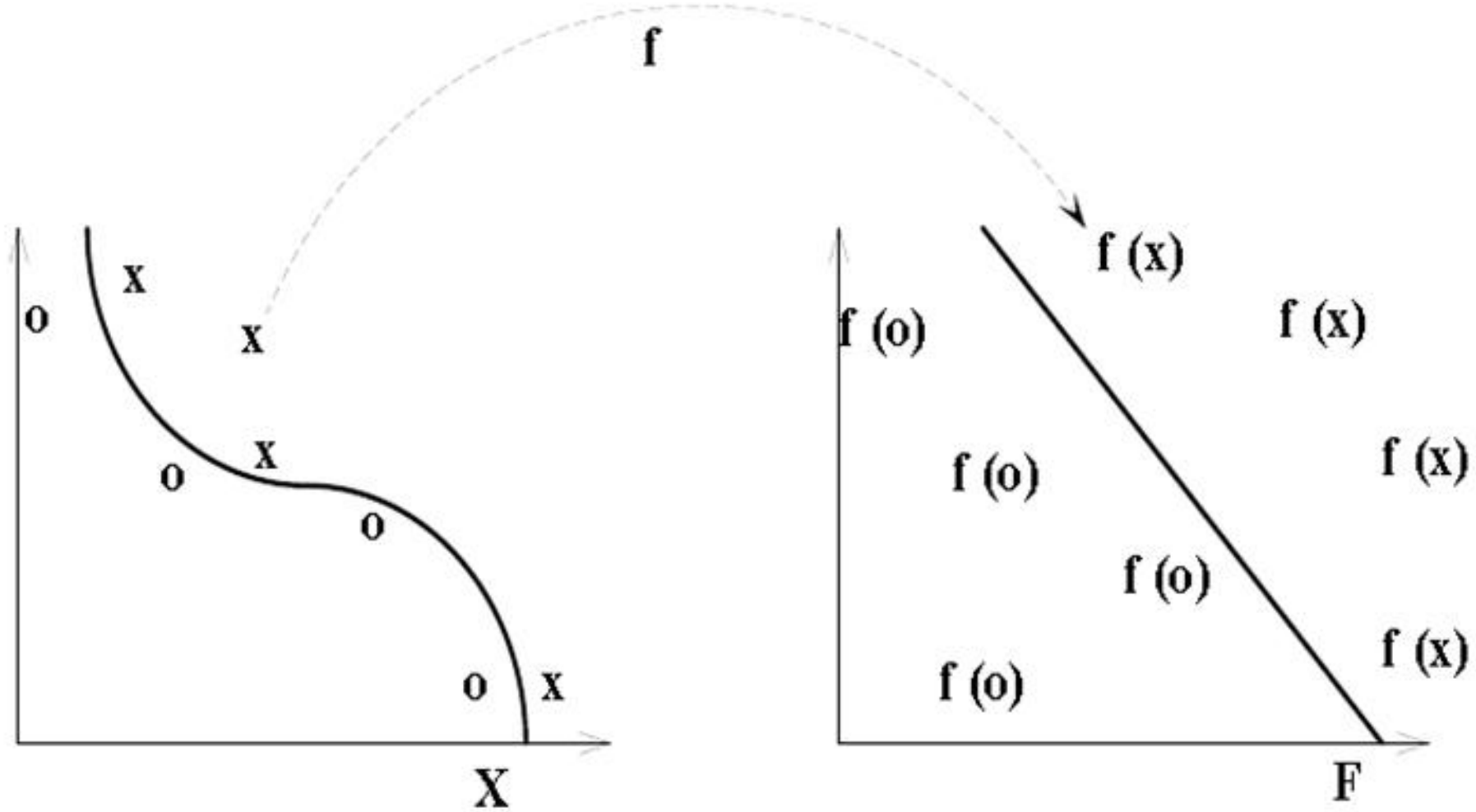
$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c \times \sum_{i=1}^{n} \xi_i$$

$$\text{s. t.} \; y_i(\boldsymbol{w}^T\boldsymbol{x} + \mathrm{b}) \geq 1 - \xi_i$$

$$\xi_i \geq 0, i = 1, 2, \ldots, n$$

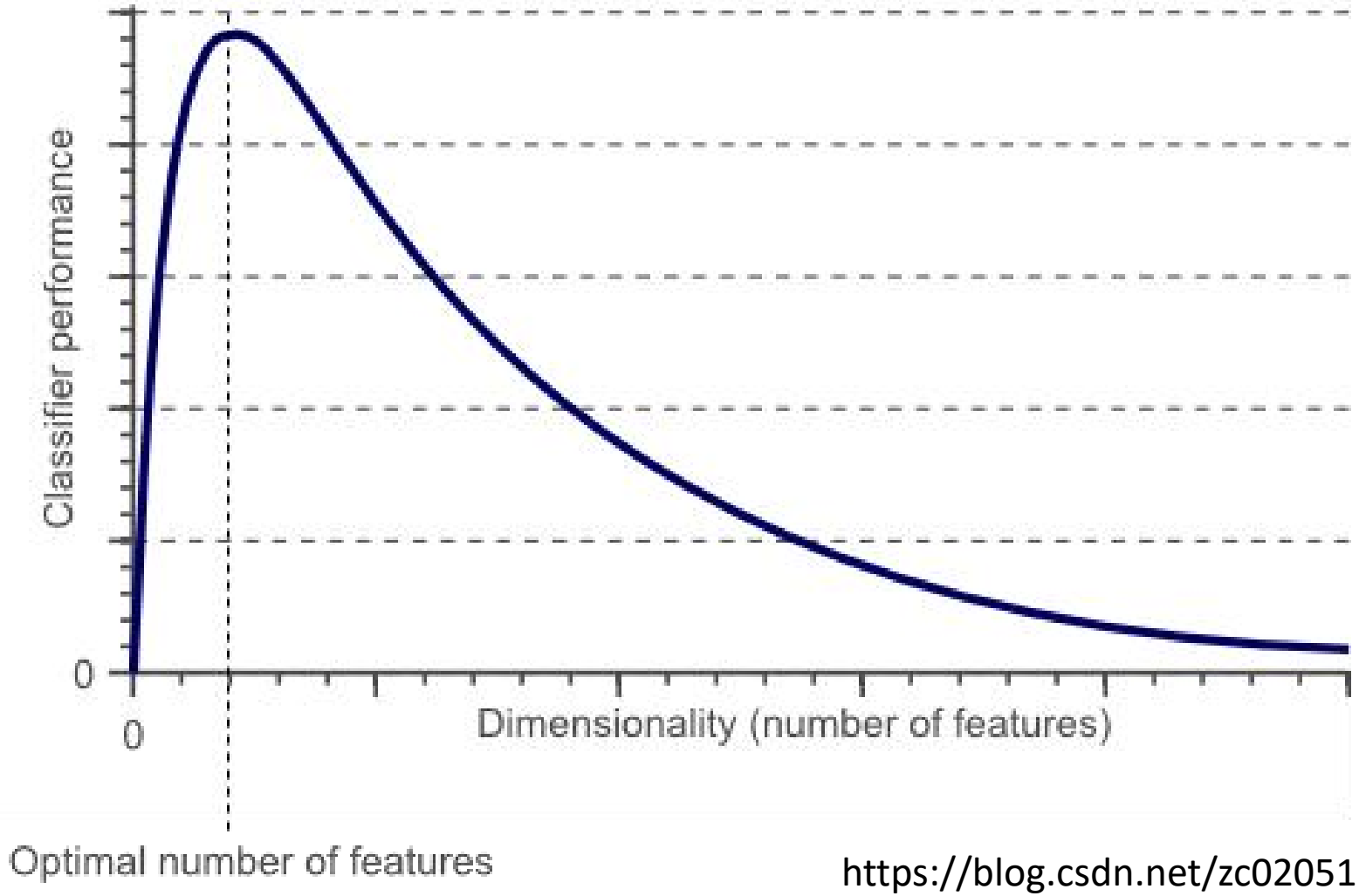
支持向量机：用高维空间映射解决线性不可分

- 把数据映射到线性可分的高维空间



维数灾难 (The Curse of Dimensionality)

- 以一个简单分类问题为例，用颜色、纹理等特征区分猫狗

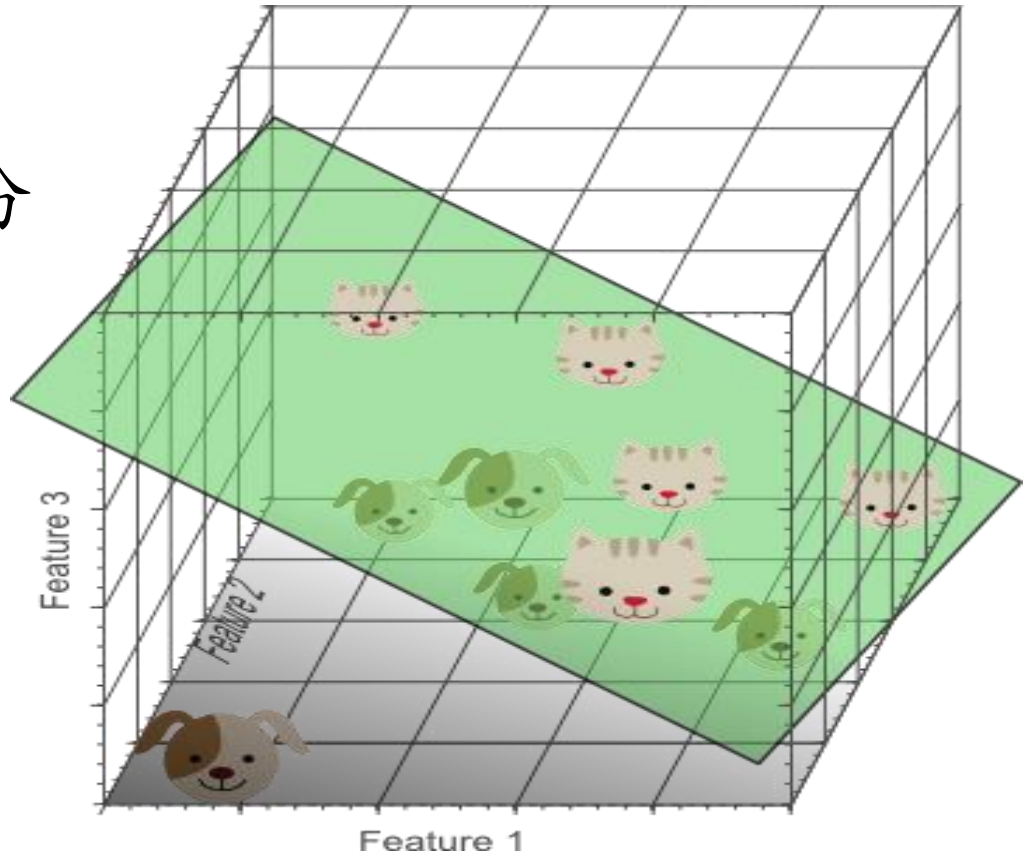


https://blog.csdn.net/zc02051126/article/details/49618633

维数灾难 (The Curse of Dimensionality)

- 维数灾难与过拟合
  - 例子：10个样本，每个维度分5份
  - 1D和3D对应样本密度为2和$\frac{10}{125}$

单一特征的分类器，
表现并不好

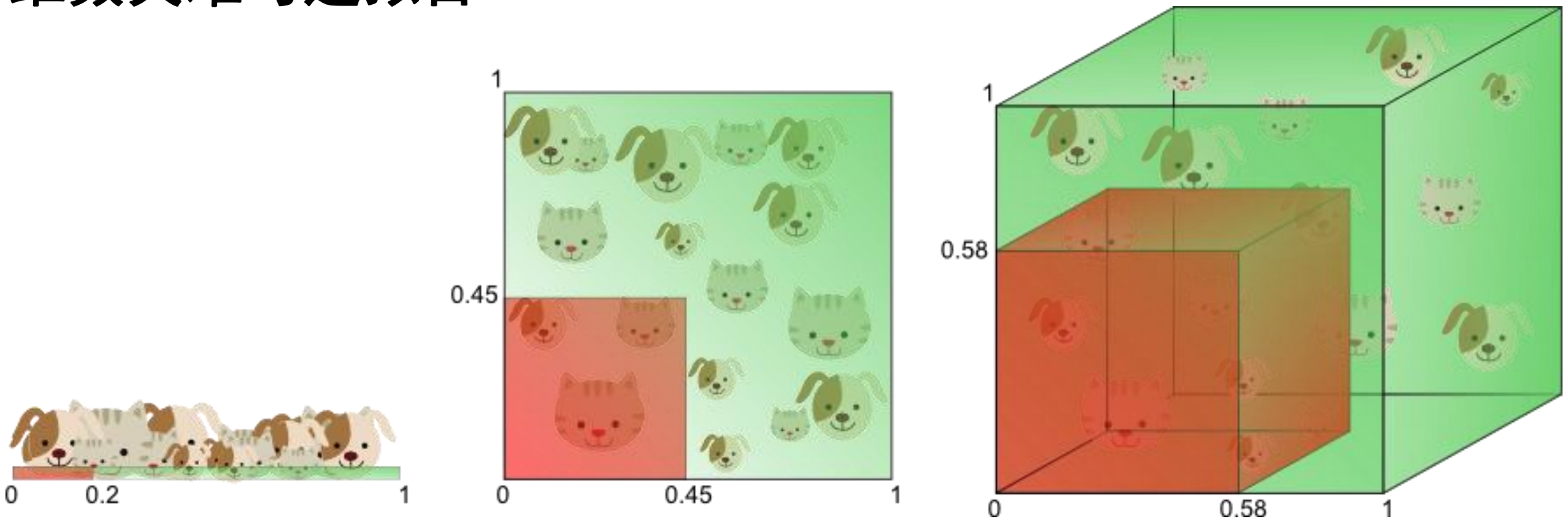


高维空间中，似乎能得到
更优的分类器

https://blog.csdn.net/zc02051126/article/details/49618633

维数灾难 (The Curse of Dimensionality)

- 维数灾难与过拟合

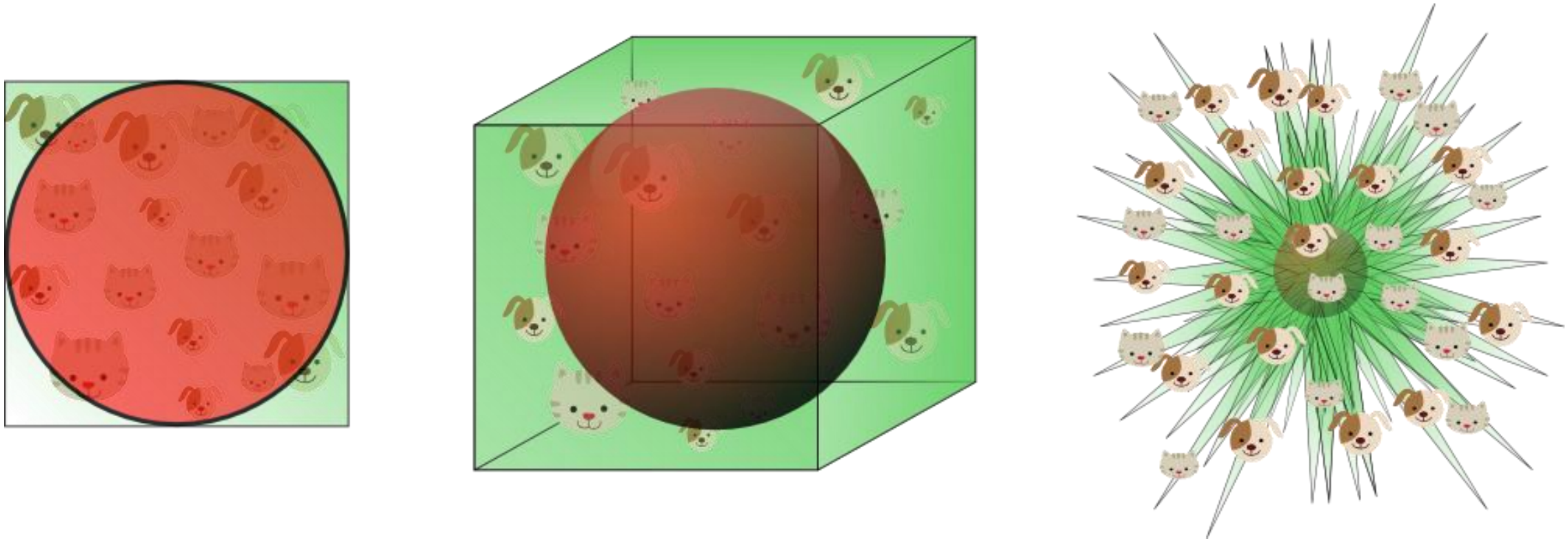


取得相同数量样本(例如20%)需要的空间大小

https://blog.csdn.net/zc02051126/article/details/49618633

维数灾难 (The Curse of Dimensionality)

- 维数灾难与过拟合



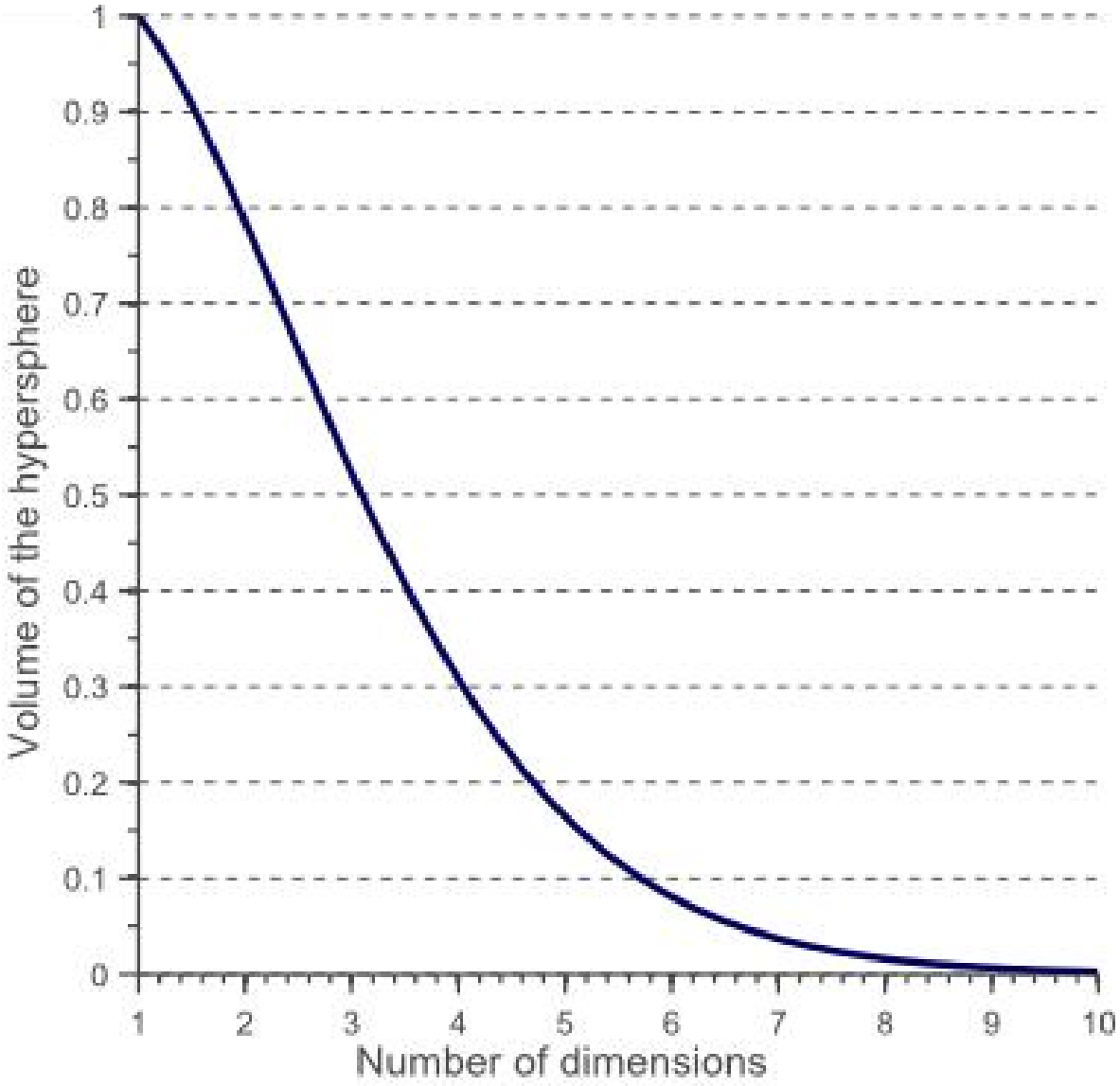
随着特征维度增加样本在样本空间中的分布情况

https://blog.csdn.net/zc02051126/article/details/49618633

维数灾难 (The Curse of Dimensionality)

- 维数灾难与过拟合

$$V(d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} 0.5^d.$$



https://blog.csdn.net/zc02051126/article/details/49618633

维数灾难 (The Curse of Dimensionality)

- **由理查德·贝尔曼（Richard Bellman）在考虑优化问题时提出**
  - 描述当空间维度增加时，分析和组织高维空间中的数据，因体积指数增加而遇到各种问题场景
  - 维度增加，空间的体积增加得很快，使得可用的数据变得稀疏
  - 稀疏性对于任何要求有统计学意义的方法而言都是一个问题
  - 为了获得统计学上可靠的结果，需要的数据量呈指数级增长
  - 高维空间中，所有的数据都很稀疏，从很多角度看都不相似，因而平常使用的数据组织策略变得极其低效

“维数灾难”通常是用来作为不要处理高维数据的借口

支持向量机：核函数$\mathcal{K}(x_i, x_j) = \phi(x_i)^T \phi(x_j)$

- **将线性不可分样本从原始空间映射到一个高维的特征空间**
  - 使得样本在这个特征空间中高概率线性可分
  - 如果原始空间是有限维，那么一定存在一个高维特征空间使样本可分[Shawe-Taylor, J. 2004]。

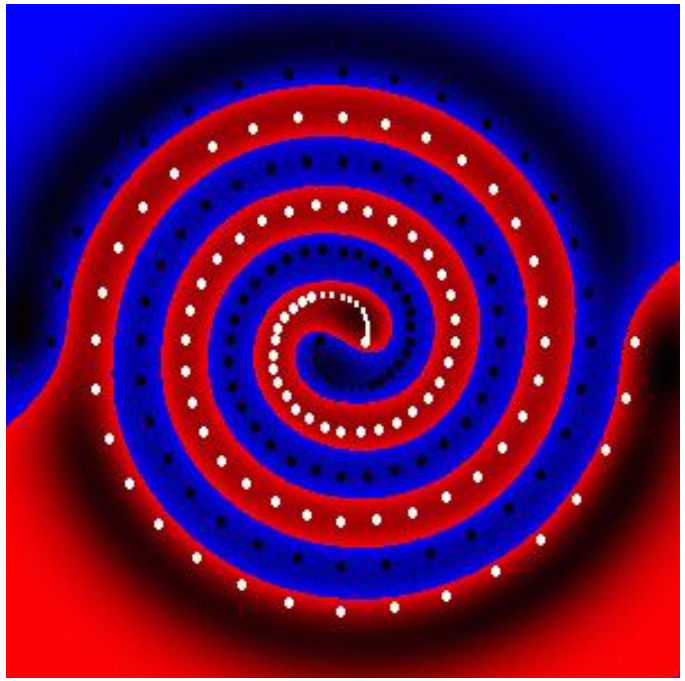
常见核函数

线性	$x_i^T x_j$
多项式	$(\gamma x_i^T x_j + c)^n$
径向基函数	$e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$
Sigmoid	$\tanh(x_i, x_j - \gamma)$

单选题 1分

设置

图中的特征映射采用了那种核函数?



A 线性

B 多项式

C 径向基

D Sigmoid

提交

线性	$\boldsymbol{x}_i^T \boldsymbol{x}_j$
多项式	$\left(\gamma \boldsymbol{x}_i^T \boldsymbol{x}_j + c\right)^n$
径向基	$e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$
Sigmoid	$\tanh\left(\boldsymbol{x}_i, \boldsymbol{x}_j - \boldsymbol{\gamma}\right)$

提纲

一、机器学习基本概念

二、回归分析

三、决策树

四、线性区别分析

五、Ada Boosting

六、支持向量机

七、生成学习模型

生成学习模型

- 生成学习方法从数据中学习联合概率分布$P(X,C)$，然后求出条件概率分布$P(C|X)$作为预测模型，即$P(c_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x},c_i)}{P(\boldsymbol{x})}$。

$$P(\boldsymbol{x},c_i) = \overbrace{P(\boldsymbol{x}|c_i)}^{\text{似然概率}} \times \overbrace{P(c_i)}^{\text{先验概率}}$$

$$\overbrace{P(c_i|\boldsymbol{x})}^{\text{后验概率}} = \frac{\overbrace{P(\boldsymbol{x},c_i)}^{\text{联合概率}}}{P(\boldsymbol{x})} = \frac{\overbrace{P(\boldsymbol{x}|c_i)}^{\text{似然概率}} \times \overbrace{P(c_i)}^{\text{先验概率}}}{P(\boldsymbol{x})}$$

生成学习模型：一个简单的例子

输入样本和类别标签的联合概率分布为：$P(0,\text{阳性})=\frac{6}{12}=\frac{1}{2}$、$P(0,\text{阴性})=\frac{2}{12}=\frac{1}{6}$、$P(1,\text{阳性})=0$、$P(1,\text{阴性})=\frac{4}{12}=\frac{1}{3}$。一旦给出输入数据，假定输入数据的概率为某个常数，就可以通过计算$\frac{P(0,\text{阳性})}{P(0)}$或者$\frac{P(0,\text{阴性})}{P(0)}$以及$\frac{P(1,\text{阳性})}{P(1)}$或者$\frac{P(1,\text{阴性})}{P(1)}$，将输入数据归属到所得结果最大所对应的类别。这里要注意，样本-标签数据的联合概率分布累加之和为1。

$c=$	阳性	阴性
$x=0$	6	2
$x=1$	0	4

生成学习模型

判别式式学习

输入样本和类别标签的条件概率分布为：$P(\text{阳性}|0)=\frac{6}{8}=\frac{3}{4}$、$P(\text{阳性}|1)=\frac{2}{8}=\frac{1}{4}$；$P(\text{阳性}|1)=\frac{0}{4}=0$、$P(\text{阳性}|1)=\frac{4}{4}=1$。这里要注意，输入样本为 0 或为 1 前提下，其所对应的类别概率累加之和为 1。

常见的生成学习模型有隐马尔可夫模型、隐狄利克雷分布(latent dirichlet allocation, LDA)等。

谢谢!