监督学习2

主讲：刘夏雷、郭春乐、王亚星

南开大学计算机学院

https://mmcheng.net/xliu/

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

作业回顾

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

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

作业回顾

- **正则化项/惩罚项**

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。

机器学习的分类

监督学习(supervised learning)
数据有标签、一般为回归或分类等任务

无监督学习(un-supervised learning)
数据无标签、一般为聚类或若干降维任务

半监督学习
(semi-supervised learning)

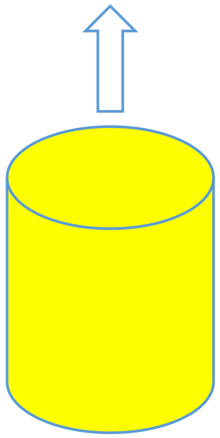
强化学习(reinforcement learning)
序列数据决策学习，一般为与从环境交互中学习

监督学习：损失函数

- **训练集共有n个标注数据，第i个记为(x_i, y_i)**
- **从训练数据中学习映射函数$f(x_i)$**
  - 损失函数就是真值y_i与预测值$f(x_i)$之间差值的函数。
- **在训练过程中希望映射函数在训练数据集上得到“损失”最小**
  - 即$\min \sum_{i=1}^{n} Loss(f(x_i), y_i)$。

训练映射函数f

使得$f(x_i)$尽量等于y_i

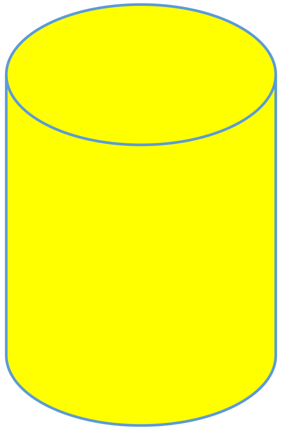


训练数据集

$\boldsymbol{(x_i, y_i), i = 1, \ldots, n}$

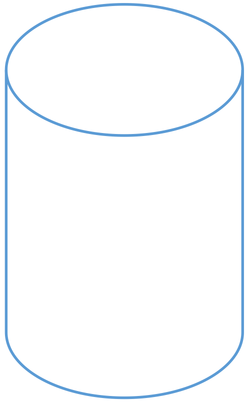
监督学习：训练数据和测试数据

从训练数据集学习
得到映射函数f



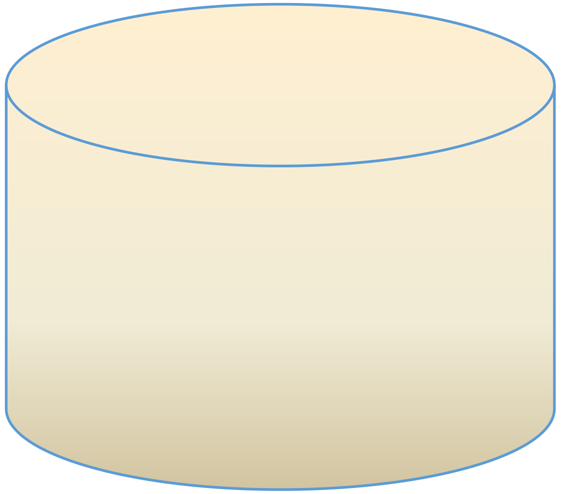
训练数据集
$(x_i, y_i), i = 1, \ldots, n$

在测试数据集
测试映射函数f



测试数据集
$(x_i{}', y_i{}'), i = 1, \ldots, m$

未知数据集
上测试映射函数f



监督学习：经验风险和期望风险

- 模型**泛化能力**与经验风险、期望风险的关系

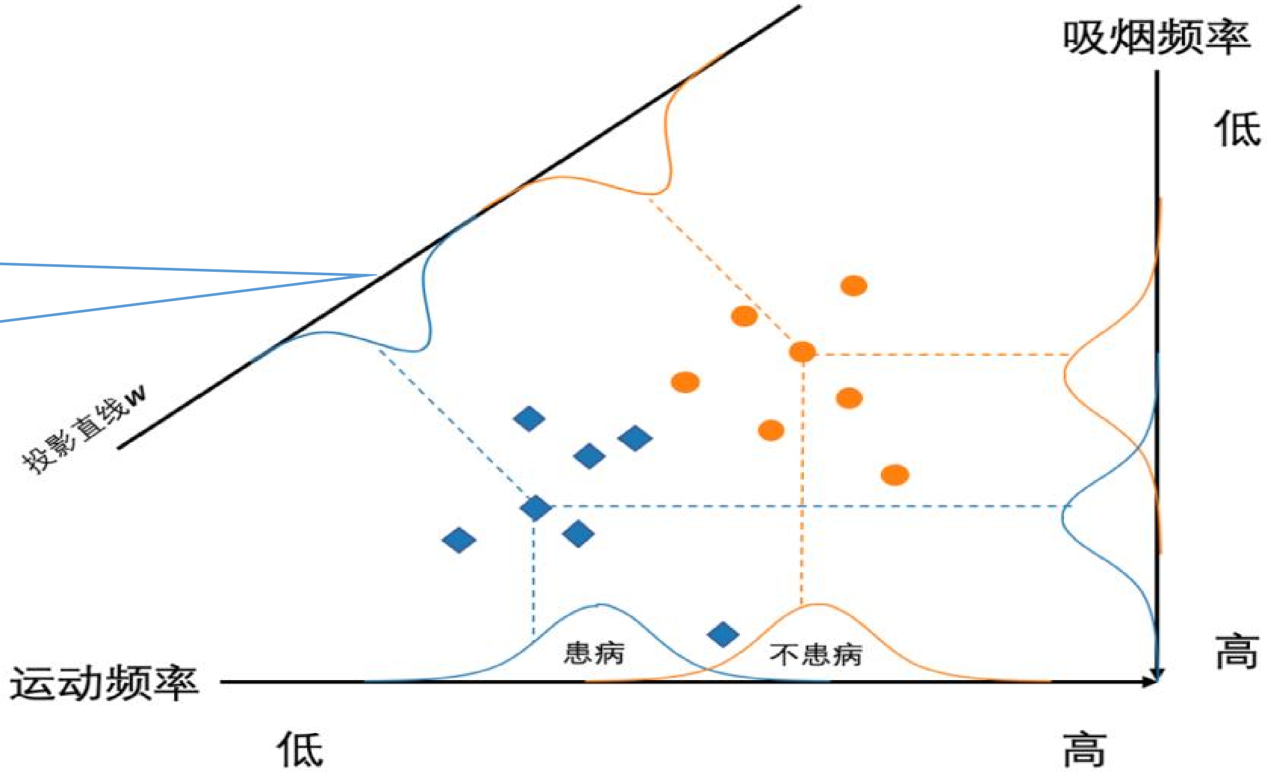
训练集上表现	测试集上表现	
经验风险小	期望风险小	泛化能力强
经验风险小	期望风险大	过学习 (模型过于复杂)
经验风险大	期望风险大	欠学习
经验风险大	期望风险小	“神仙算法”或“黄粱美梦”

提纲

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

“类内方差小、类间间隔大”



线性区别分析：符号定义

- **假设样本集为**$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_N,y_N)\}$**，**$\boldsymbol{x}_i\in\mathbb{R}^d$
  - 其中，y_i的取值范围是$\{\mathcal{C}_1,\mathcal{C}_2,\ldots,\mathcal{C}_K\}$，即一共有$K$ 类样本
  - 定义$\boldsymbol{X}$ 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:**

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

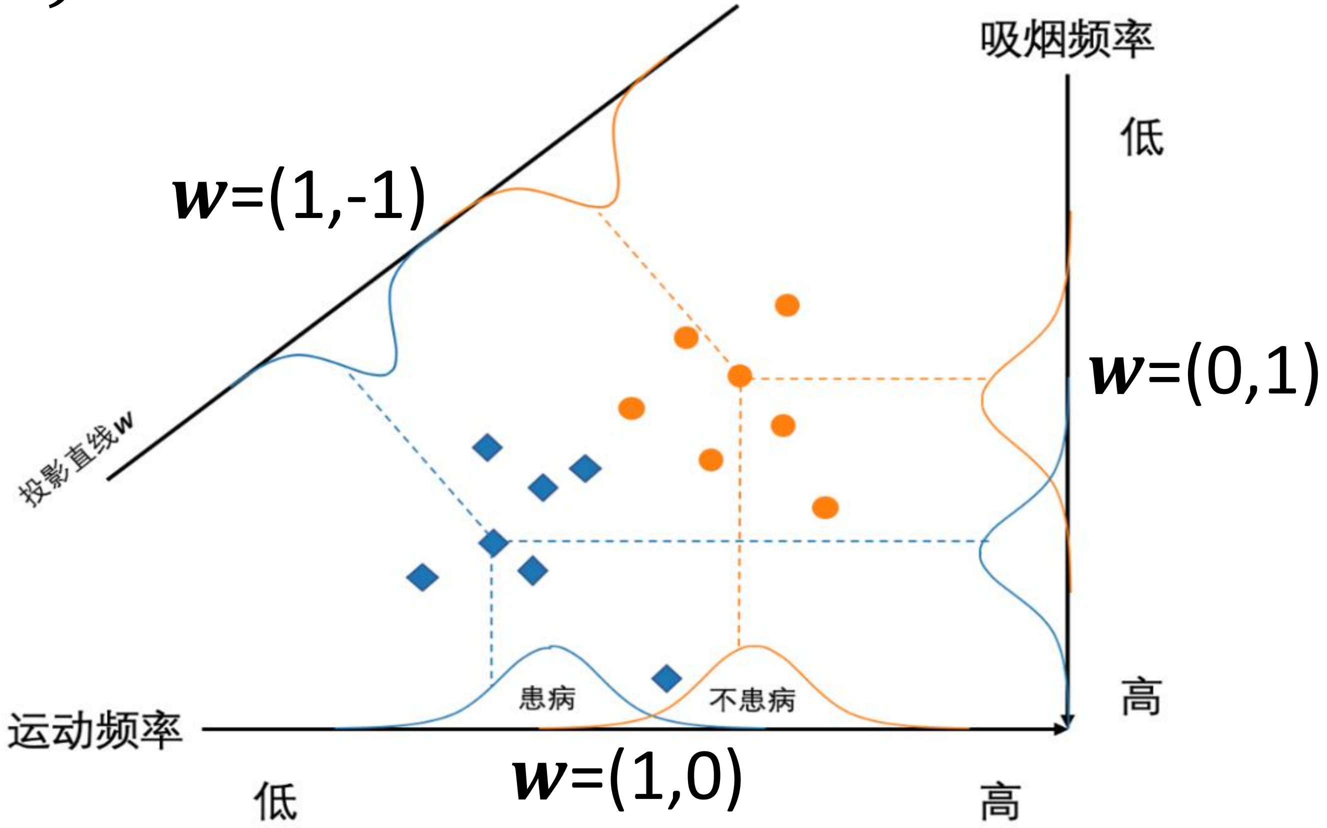
线性区别分析：二分类问题

- **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J ($\boldsymbol{w}$)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:

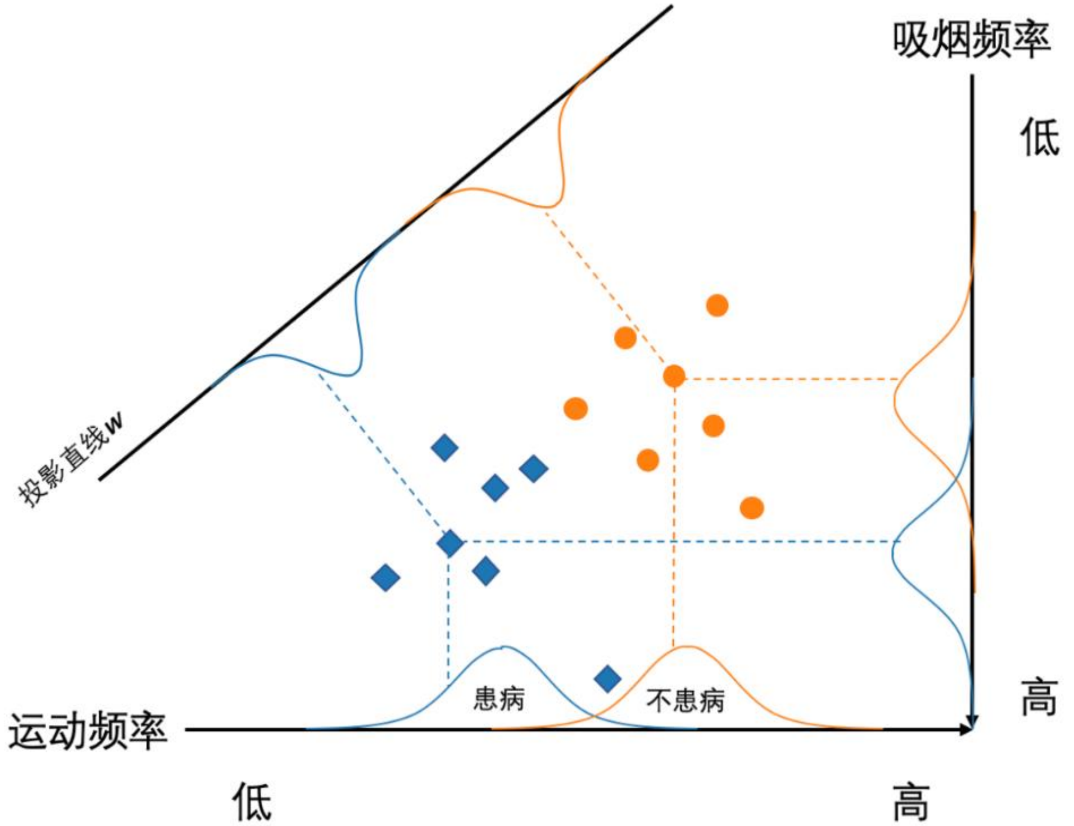
$$\boldsymbol{S}_b=(\boldsymbol{m}_2-\boldsymbol{m}_1)(\boldsymbol{m}_2-\boldsymbol{m}_1)^T$$

- **$\boldsymbol{S}_w$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:

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

线性区别分析：二分类问题

- 由于$J(w)$的分子和分母都是关于w的二项式，因此解只与w的方向有关，与w的长度无关，因此可令分母$w^T S_W w = 1$，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

线性区别分析：二分类问题

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

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

- **由于对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（了解）

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$\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_r)$，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：

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

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（了解）

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

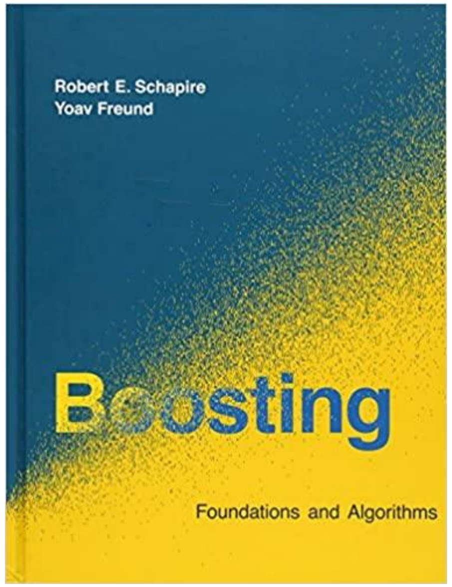
对线性判别分析的降维步骤描述如下：

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$(w_1, w_2, \ldots, w_r)$，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。

提纲

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”

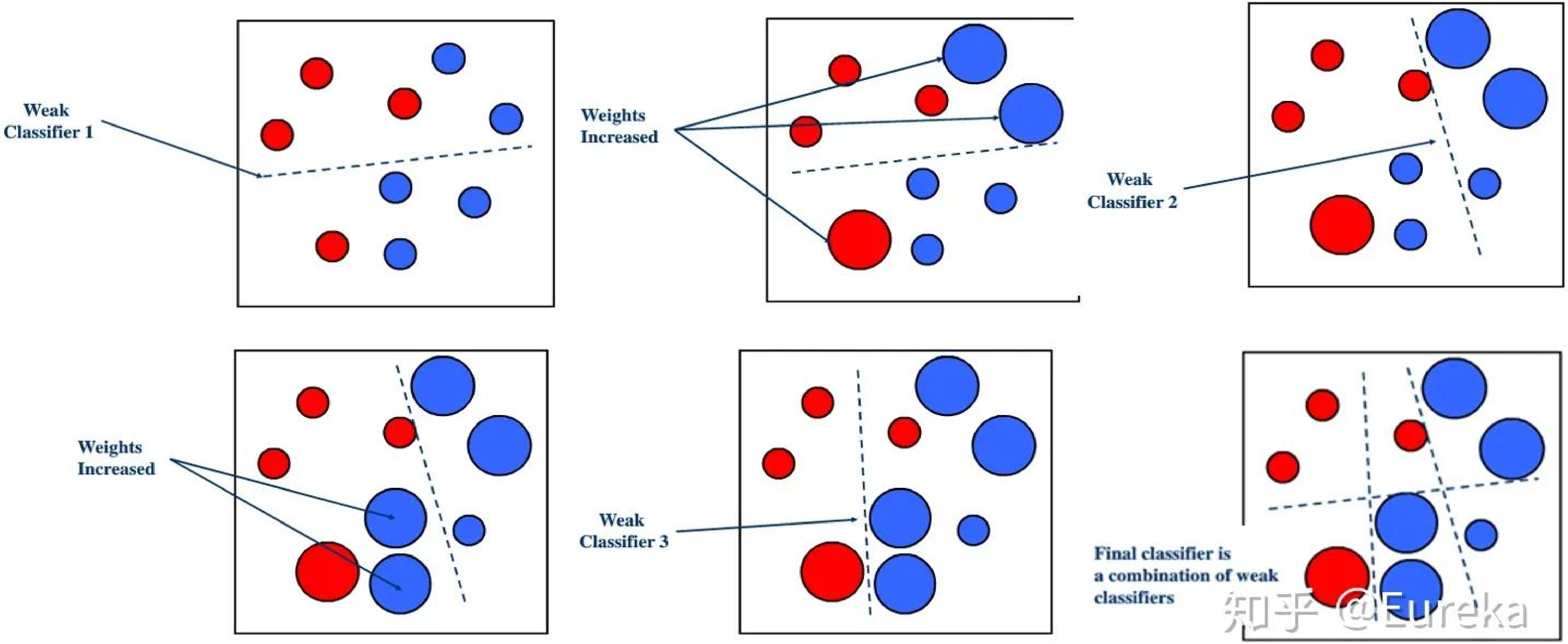
强可学习 (strongly learnable)	学习模型能够以较高精度对绝大多数样本完成识别分类任务
弱可学习 (weakly learnable)	学习模型仅能完成若干部分样本识别与分类，其精度略高于随机猜测。

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。

示例



图片来源：https://zhuanlan.zhihu.com/p/39972832

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$\Gamma = \{(x_1, y_1), ..., (x_N, y_N)\}$，$x_i(1 \leq i \leq N) \in X \subseteq R^n, y_i \in Y = \{-1,1\}$
- 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。

Ada Boosting：算法描述---数据样本权重初始化

1. 初始化每个训练样本的权重

$$D_1 = (w_{11}, \dots, w_{1i}, \dots, w_{1N}),\ \text{其中}\ w_{1i} = \frac{1}{N} (1 \leq i \leq N)$$

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

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

  - ➢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

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

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**$m = 1,2,\ldots,M$

➢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）

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

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$\boldsymbol{m}$个弱分类器$\boldsymbol{G_m(x)}$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（了解）

- $\boldsymbol{M}$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。

课上习题

- 对于如下数据，考虑使用Ada boosting方法来训练“是否出去玩”强分类器。每个弱分类器可考虑对单个属性的分类，比如对于“心情指数”这一属性，可考虑心情指数>2和心情指数<4两个方面。请问答下列问题：
- （1）Ada boosting在第一轮迭代中将会选择哪一个弱分类器？
- （2）第一轮迭代前与迭代后每个样本的权重是多少？
- （3）第二轮迭代选择的弱分类器是是哪一个？分类器权重是多少？
- （4）写出三轮迭代后的强分类器的表达式（每个弱分类器可用字母替代）

序号	出去玩	天气状况	有同伴	零花钱	特殊节日	心情指数（1 差-5 好）
1	是	好	无	多	是	5
2	是	一般	有	多	是	5
3	是	一般	有	少	否	1
4	是	一般	有	少	否	3
5	是	一般	有	少	否	5
6	是	好	无	多	是	5
7	是	好	无	多	是	5
8	否	一般	无	多	是	1
9	否	一般	有	少	否	1
10	否	一般	无	少	否	5

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

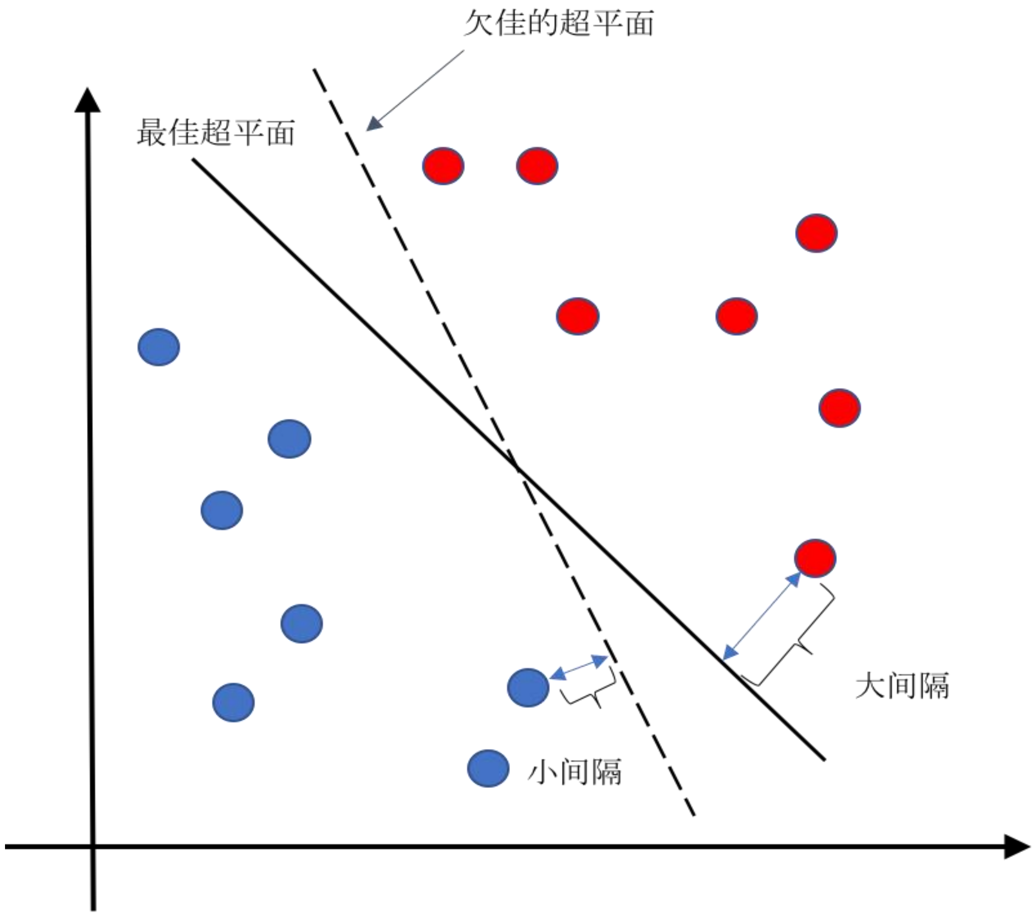
- 在回归分析中，学习一个函数将输入变量映射到连续输出空间
  - 如价格和温度等，即值域是连续空间
- 在分类模型中，学习一个函数将输入变量映射到离散输出空间
  - 如人脸和汽车等，即值域是离散空间

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

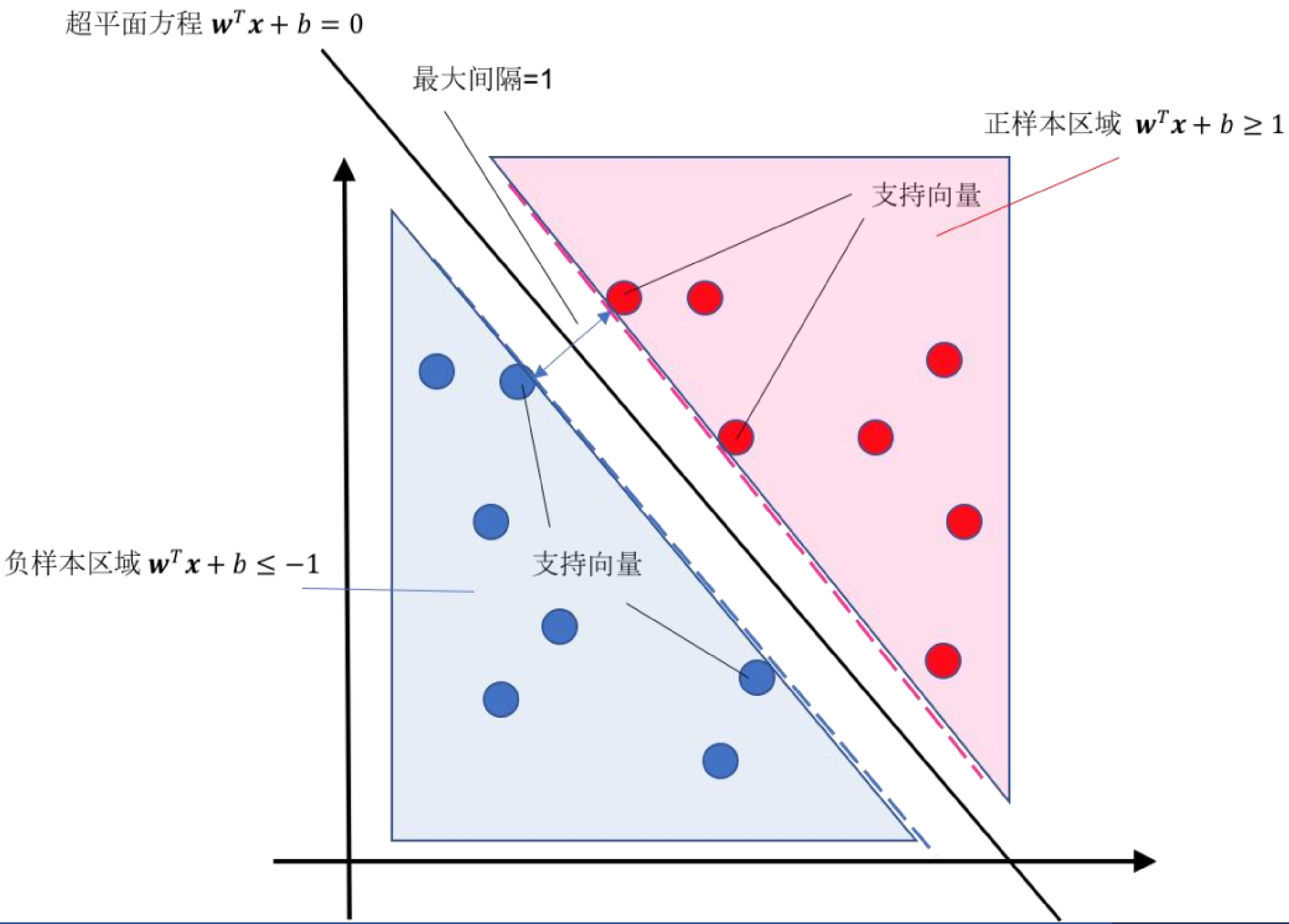
- **支持向量机（support vector machine, SVM）**
  - 通过结构风险(structural risk)最小化来解决学习问题



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。

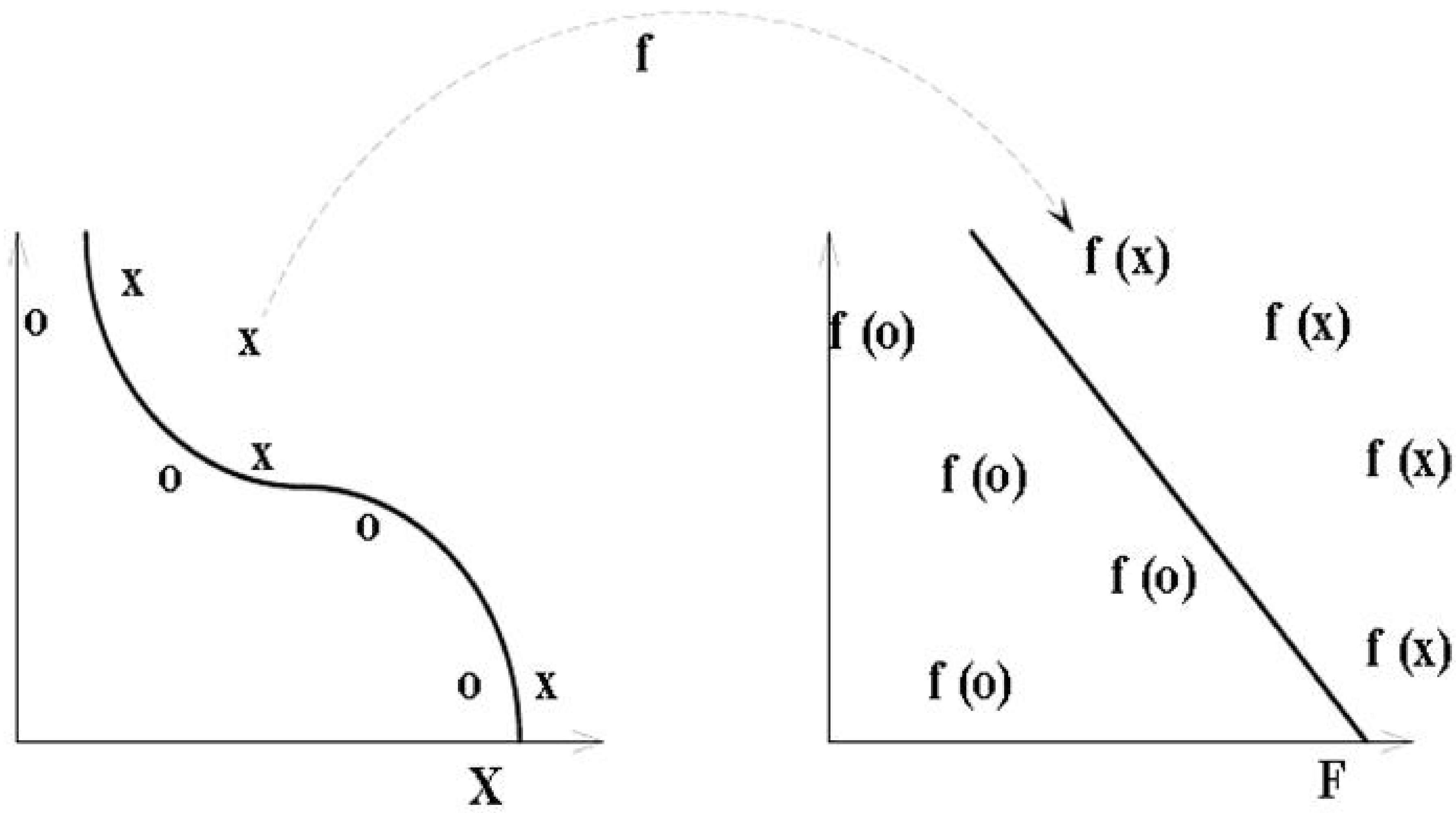
支持向量机：线性可分支持向量机

寻找一个最优的超平面，其方程为$\boldsymbol{w}^T\boldsymbol{x} + b = 0$。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：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（核函数）



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/判别式学习模型

- **常见的生成学习模型**
  - 有朴素贝叶斯
  - 隐马尔可夫模型
  - 隐狄利克雷分布(LDA)等。
- **常见的判别性模型：直接学习后验条件概率**
  - 线性回归
  - 决策树
  - 支持向量机
- **都是监督学习**

谢谢！