# 机器学习
# Machine Learning

**北京航空航天大学计算机学院智能识别与图像处理实验室**

**IRIP Lab, School of Computer Science and Engineering, Beihang University**

**黄 迪　刘庆杰**

**2018年秋季学期**

**Fall 2018**

部分内容来源于C. Bishop和A. NG等人的课程以及互联网资源

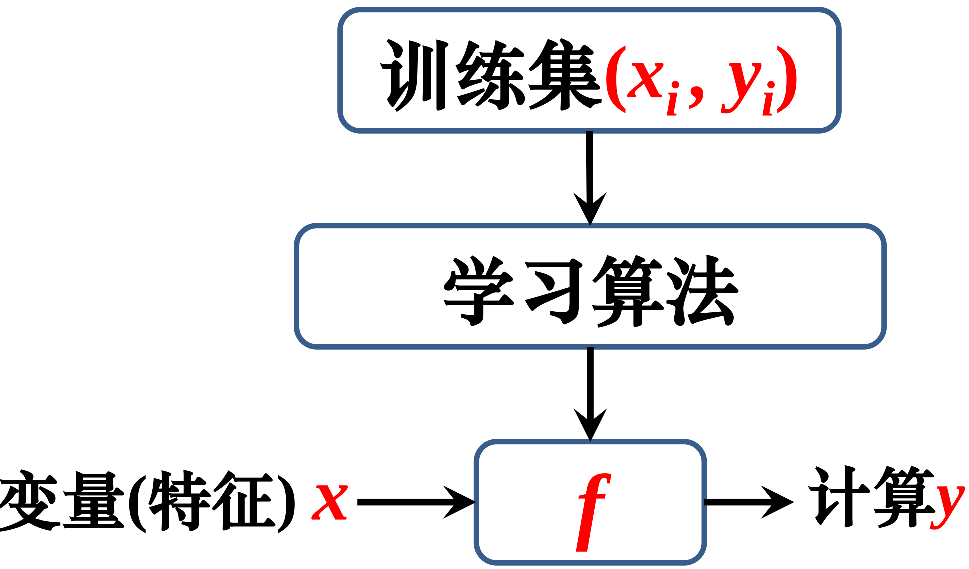
# 课前回顾

# 机器学习算法

## 机器学习主要问题

| | *Supervised Learning* | *Unsupervised Learning* |
|---|---|---|
| *Discrete* | Classification or Categorization | Clustering |
| *Continuous* | Regression | Dimensionality Reduction |

# 回归模型




$$
\begin{aligned}
f(\mathbf{x}) &= w_0 + w_1 x_1 + \ldots + w_m x_m \\
&= w_0 + \sum_{i=1}^{m} w_i x_i \\
&= \sum_{i=0}^{m} w_i x_i \qquad x_0 = 1
\end{aligned}
$$

## 如何表示 $f$？

**线性回归：**

**假设函数 $f$ 为输入$x$的线性函数**

$$
\begin{aligned}
f(\mathbf{x}) &= w_0 + w_1 x_1 + \ldots + w_m x_m \\
&= w_0 + \sum_{j=1}^{m} w_i x_i
\end{aligned}
$$

**写成向量形式:**

增加一维 $x_0=1$，表示截距项

$$f(\mathbf{x}) = \boxed{\mathbf{w}^T}\mathbf{x}$$

# 回归模型

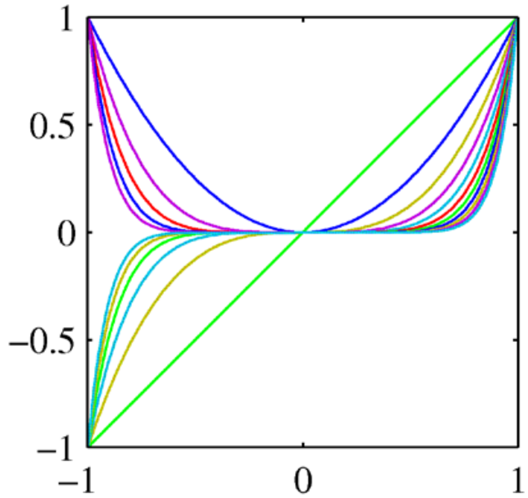
基函数

$$y(\mathbf{x},\mathbf{w}) = \sum_{j=0}^{M-1} w_j \boxed{\phi_j(\mathbf{x})} = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$$

**最简单的情况下：** $\phi_j(x) = x_j$

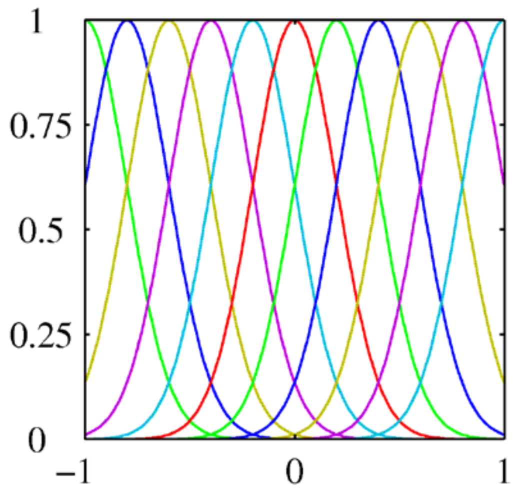
- **多项式基函数**

$$\phi_j(x) = x^j$$



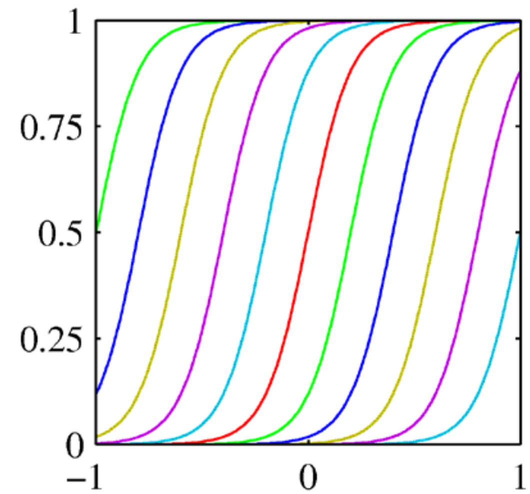
- **高斯基函数**

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$



- **Sigmoid基函数**

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) \quad \sigma(a) = \frac{1}{1+\exp(-a)}.$$



# 模型求解

- 求解问题：确定参数*w*
- 基本思想：在训练集中最小化预测值*f*与真实值*y*的差异

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w}) = \arg\min_{\mathbf{w}} \left[\frac{1}{2}\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2\right]$$

确定性方法(*w*和$t_n$都是确定量)

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

# 梯度下降 vs 标准方程组

- **梯度下降法:**
  - -需要选择 $\alpha$
  - -需要迭代多次
  - -需要数据归一化
  - -样本量非常大时也适用

- **标准方程组:**
  - -不需要选择 $\alpha$
  - -不需要迭代多次
  - -无需数据归一化
  - -样本量大时不适用
  (需要计算 $(X^TX)^{-1}$)

**样本量较小时选用标准方程组求解**
**样本量较大时选用梯度下降法求解**

# 模型求解

- **最大似然估计**：
  把待估计的参数看做是确定的量，只是其取值未知。最佳估计就是使得产生以观测到的样本的概率最大的那个值。

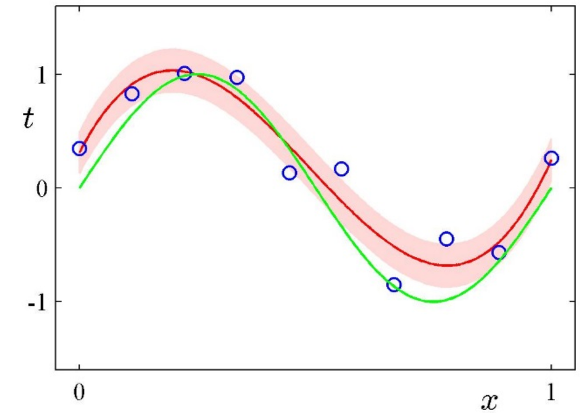
- **贝叶斯估计**：
  把待估计的参数看做是符合某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，从而利用样本信息修正了对参数的初始估计值。

# 模型求解

- 最大似然估计(*w*是确定量)

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}\left(t_n|y(x_n,\mathbf{w}),\beta^{-1}\right)$$

$$\frac{1}{\beta_{\mathrm{ML}}}=\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}_{\mathrm{ML}})-t_n\}^2$$



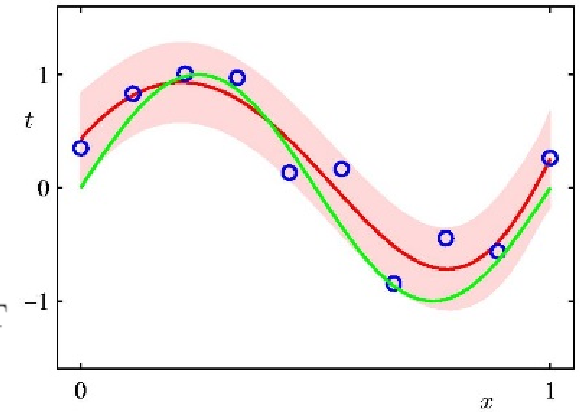

- 贝叶斯估计(*w*是随机量)

$$p(t|x,\mathbf{x},\mathbf{t})=\int p(t|x,\mathbf{w})p(\mathbf{w}|\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{w}=\mathcal{N}\left(t|m(x),s^2(x)\right)$$

$$m(x)=\beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n$$

$$s^2(x)=\beta^{-1}+\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1}=\alpha\mathbf{I}+\beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}}\quad\boldsymbol{\phi}(x_n)=\left(x_n^0,\ldots,x_n^M\right)^{\mathrm{T}}$$



# 概率密度估计

- 参数化方法:
  - 最大似然估计
  - 贝叶斯估计

总体分布形式已知
典型分布
可以写成某些参数的函数

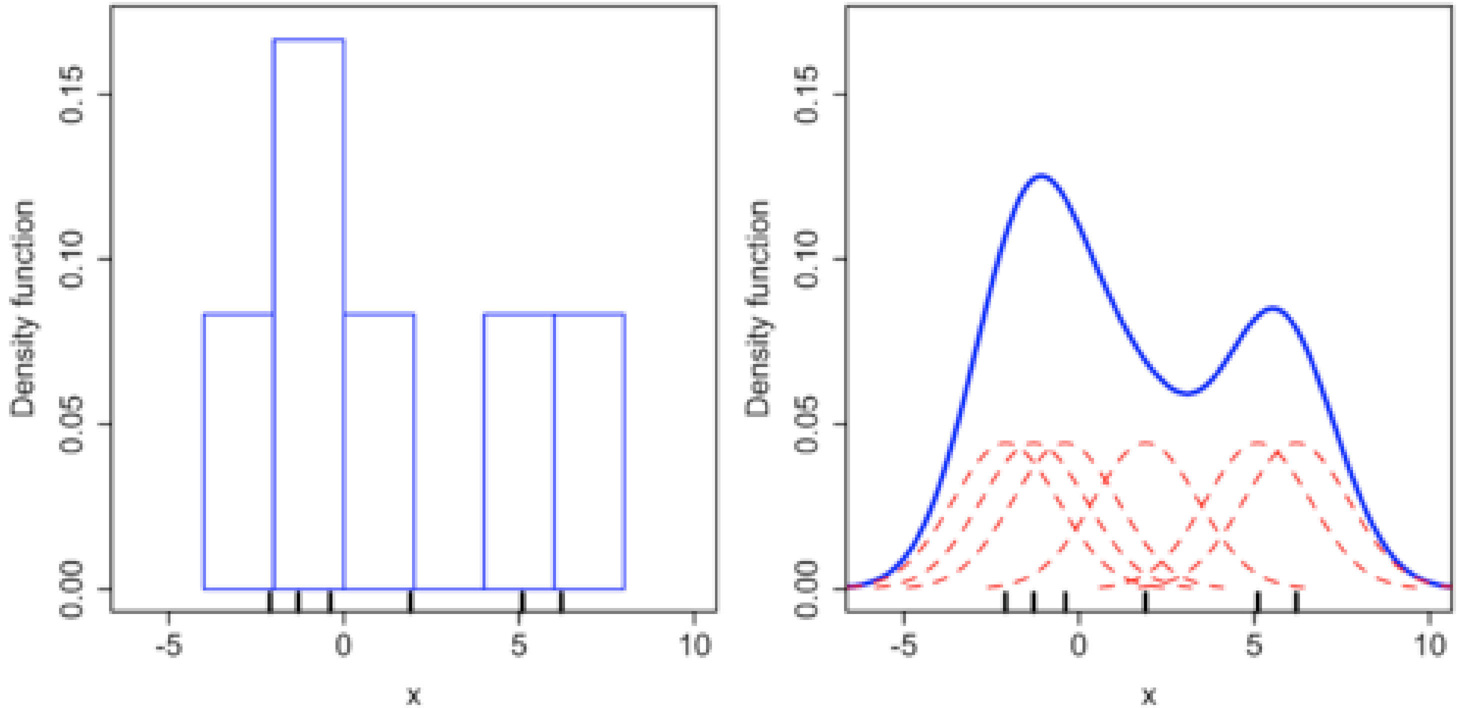
- 非参数化方法:
  - Parzen窗估计
  - $k_n$近邻估计

总体分布形式未知
非典型分布
不能写成某些参数的函数

# 非参数估计

- **直接用样本估计总体分布**

**基本思路：要估计$x_i$点的密度$p(x_i)$，可把所有样本在该点上的“贡献”相加近似作为其概率密度，进而得到 $\hat{p}(x)$。**



# 非参数估计

- **直接用样本估计总体分布**

**$p(x)$为$x$的总体概率密度函数，$N$个样本 $x=\{x_1, x_2, \ldots, x_N\}$从密度为$p(x)$的总体中独立抽取，估计 $\hat{p}(x)$ 近似 $p(x)$。**

考虑随机向量 $x$ 落入区域 $\Re$ 的概率 $P_R = \int_{\Re} p(x)dx$

$k$ 个样本落入该区域的概率符合二项分布 $P_k = C_N^k P_{\Re}^k (1-P_R)^{N-k}$

$$E[k] = NP_R \longrightarrow \hat{P}_R \doteq \frac{k}{N}$$

设$p(x)$连续，且区域体积$V$足够小，则有 $P_R = \int_{\Re} p(x)dx = p(x)V$

$$\hat{p}(x) = \frac{k}{NV}$$

**与总样本数、区域的体积及落入的样本数有关**

# 非参数估计

$$\hat{p}(x)=\frac{k}{NV} \longrightarrow \hat{p}_N(x)=\frac{k_N}{NV_N}$$

$$\lim_{N\to\infty} V_N = 0$$

$$\lim_{N\to\infty} k_N = \infty \longrightarrow \hat{p}_N(x)\text{收敛于}p(x)$$

$$\lim_{N\to\infty} \frac{k_N}{N} = 0$$

- **Parzen窗估计**

  **使区域体积序列$V_N$以$N$的某个函数的关系不断缩小**

  **同时限制$k_N$和$k_N/N$。**

- **$k_n$近邻估计**

  **使落入区域样本数$k_N$为$N$的某个函数**

  **$V_N$使区域包含$x$的$k_N$个近邻**

**有限的$N$, $V_1$选择很敏感**

↓

**动态变化$V$的取值**

# 第三章：线性分类模型

## Linear Models for Classification

# 机器学习算法

## 机器学习主要问题

| | *Supervised Learning* | *Unsupervised Learning* |
|---|---|---|
| *Discrete* | Classification or Categorization | Clustering |
| *Continuous* | Regression | Dimensionality Reduction |

# 从贝叶斯分类说起

- **样本(Sample)** $\mathbf{x} \in R^d$
- **状态(State) 第一类:** $w = w_1$ **第二类:** $w = w_2$
- **先验概率(A Priori Probability or Prior)** $P(w_1)$ $P(w_2)$
- **样本分布密度(Sample Distribution Density)** $p(\mathbf{x})$
- **类条件概率密度(Class-Conditional Probability Density)**

$$p(\mathbf{x}|w_1) \qquad p(\mathbf{x}|w_2)$$

# 从贝叶斯分类说起

- **后验概率(A Posteriori Probability or Posterior)**

$$P(w_1|\mathbf{x}), P(w_2|\mathbf{x})$$

- **错误概率(Probability of Error)**

$$P(e|\mathbf{x}) \begin{cases} P(w_2|\mathbf{x}) \text{ if } \mathbf{x} \text{ is assigned to } w_1 \\ P(w_1|\mathbf{x}) \text{ if } \mathbf{x} \text{ is assigned to } w_2 \end{cases}$$

- **平均错误率(Average Probability of Error)**

$$P(e) = \int P(e|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

- **正确率(Probability of Correctness)** $P(c)$

# 基于最小错误率的决策

$$\min \quad P(e) = \int P(e|x)p(x)dx$$

因为 $P(e|\mathbf{x}) \geq 0, \ p(x) \geq 0$

$\Rightarrow$ $\min P(e|\mathbf{x})$ for all $x$

而 $P(e|\mathbf{x}) \begin{cases} P(w_2|\mathbf{x}), & \text{if } P(w_1|\mathbf{x}) > P(w_2|\mathbf{x}) \\ P(w_1|\mathbf{x}), & \text{if } P(w_2|\mathbf{x}) > P(w_1|\mathbf{x}) \end{cases}$

$$\text{if } P(w_1|\mathbf{x}) \begin{matrix} > \\ < \end{matrix} P(w_2|\mathbf{x}) \text{ assign } \begin{matrix} x \in w_1 \\ x \in w_2 \end{matrix}$$

$$P(w|\mathbf{x}) = \max_{j=1,\ldots,c} P(w_j|\mathbf{x})$$

# 基于最小风险的决策

**条件期望损失：对于特定的x采取决策** $\alpha_i$ **的期望损失：**

$$R(\alpha_i|x) = E[\lambda(\alpha_i, w_j)|x] = \sum_{j=1}^{c} \lambda(\alpha_i, w_j)P(w_j|x), \ \ i = 1, 2, ..., k$$

**期望风险：对所有可能的x采取决策** $\alpha(x)$ **所造成的期望损失之和**

$$R(\alpha) = \int R(\alpha(x)|x)p(x)dx$$

**也称平均风险（** $R(\alpha)$**表示R依赖于决策规则**$\alpha(\cdot)$**）**

**对所有x，使** $R(\alpha(x)|x)$ **最小，则可以使** $R(\alpha)$ **最小**

**最小风险贝叶斯决策规则：**

$$\text{if } \ R(\alpha_t|x) = \min_{j=1...k} R(\alpha_j|x), \text{then } \ \alpha = \alpha_t$$

# 示例-基于最小错误率

- 假设在某个局部区域细胞中正常($w_1$)和异常($w_2$)两类的先验概率分别为：正常状态 $P(w_1)=0.9$, 异常状态 $P(w_2)=0.1$。现有一待识别细胞，其观察值为$x$，从类条件概率密度曲线上查得 $p(x|w_1)=0.2$，$p(x|w_2)=0.4$。试对该细胞$x$进行分类。

**解：利用贝叶斯公式计算$w_1$和$w_2$的后验概率**

$$P(w_1|x) = \frac{p(x|w_1)P(w_1)}{\sum_{j=1}^{2} p(x|w_j)P(w_j)} = \frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1} = 0.818$$

$$P(w_2|x) = 1 - P(w_1|x) = 0.182$$

**根据贝叶斯决策规则有**

$$P(w_1|x) = 0.818 > P(w_2|x) = 0.182$$

**把$x$归类于正常细胞。**

# 示例-基于最小风险

- **决策表**

| 状态 损失 决策 | $w_1$ | $w_2$ |
|---|---|---|
| $\alpha_1$ | 0 | 6 |
| $\alpha_2$ | 1 | 0 |

**解：** $\lambda_{11}=0,\ \lambda_{12}=6,\ \lambda_{21}=1,\ \lambda_{22}=0$ $\quad P(w_1|x)=0.818\quad P(w_2|x)=0.182$

**计算条件风险** $$R(\alpha_1|x)=\sum_{j=1}^{2}\lambda_{1j}P(w_j|x)=\lambda_{12}P(w_2|x)=1.092$$

$$R(\alpha_2|x)=\lambda_{21}P(w_1|x)=0.818$$

**由于** $R(\alpha_1|x)>R(\alpha_2|x)$

**把*x*归类于异常细胞。**

# 小结

- **已知类条件概率密度 $p(\mathbf{x}|w_i)$ 和先验概率 $P(w_i)$，计算后验概率 $P(w_i|\mathbf{x})$ 进行决策**

**若类条件概率密度参数未知?**

- **已知类条件概率密度 $p(\mathbf{x}|w_i)$ 的参数表达式，利用样本估计 $p(\mathbf{x}|w_i)$ 的未知参数，再利用贝叶斯定理将其转化成后验概率 $P(w_i|\mathbf{x})$ 进行决策**

**若类条件概率密度形式难以确定?**

- **非参数方法估计**

**需要大量样本…**

# 线性判别函数

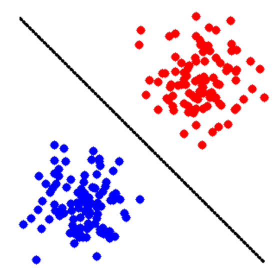
- **利用样本集直接设计分类器**

**线性回归：** $y(x) = w^T x + w_0$

**利用样本估计*w*，对于给定*x*，计算*y*。**

$$y(x) = f(w^T x + w_0) \quad f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0 \end{cases}$$

$$w^T x + w_0 \geq 0 \rightarrow C_1$$
$$w^T x + w_0 \leq 0 \rightarrow C_2$$



**将分类器设计问题转化为求准则函数极值的问题**
**准则函数：分类器设计的某些要求的函数形式**

# 线性判别函数

- **两类情况下线性判别函数** $g(x) = w^T x + w_0$

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$ **特征向量/样本向量**；$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$ **权向量**；$\boldsymbol{w_0}$**阈值权 (常数)**

令 $g(x) = g_1(x) - g_2(x)$

如果$\begin{cases} \boldsymbol{g(x) > 0, x \in w_1} \\ \boldsymbol{g(x) < 0, x \in w_2} \\ \boldsymbol{g(x) = 0,} \text{可将}\boldsymbol{x}\text{分到任意一类或拒绝} \end{cases}$

**$g(x)$=0定义了一个决策面，当$g(x)$为线性函数时，决策面就是超平面。**

# 线性判别函数

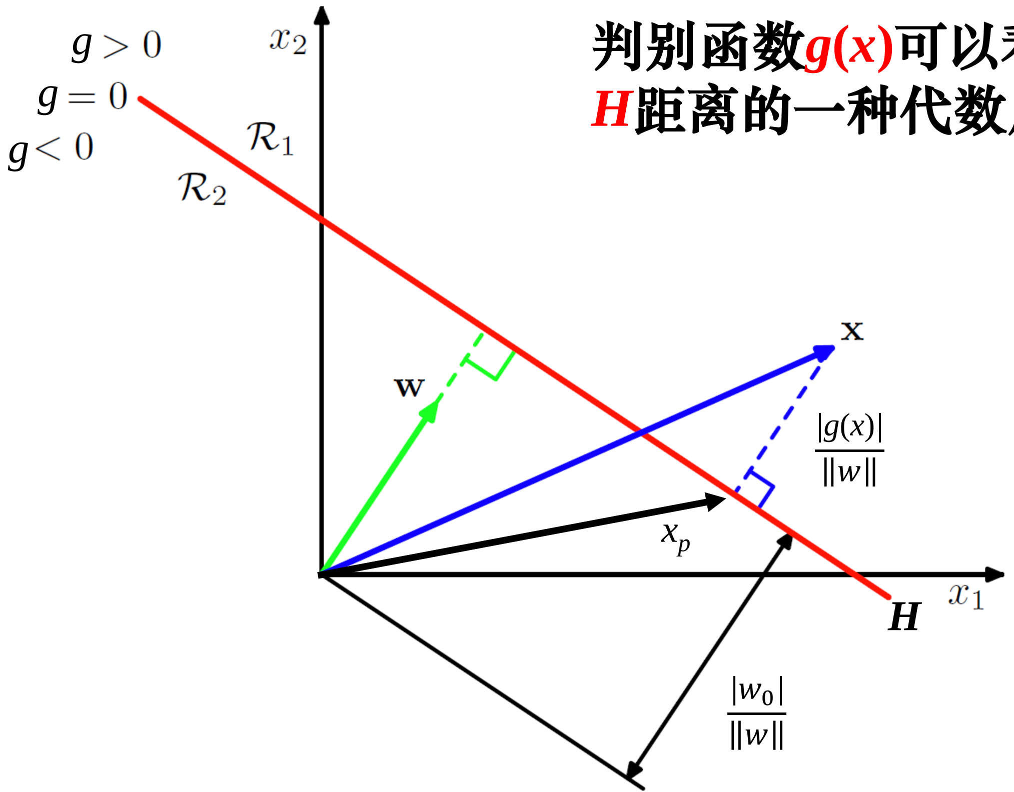
如果$x_1$和$x_2$都在决策面$H$上，则有

$$w^T x_1 + w_0 = w^T x_2 + w_0$$

即

$$w^T (x_1 - x_2) = 0$$

说明$w$和超平面$H$上任一向量正交，即$w$是$H$的法向量。




判别函数$g(x)$可以看成是特征空间中某点$x$到超平面$H$距离的一种代数度量

$$x = x_p + r\frac{w}{\|w\|}$$

$x_p$： $x$在$H$上的投影向量

$r$： $x$到$H$上的垂直距离

$\frac{w}{\|w\|}$： 是$w$方向上的单位向量

$$g(x) = w^T\left(x_p + r\frac{w}{\|w\|}\right) + w_0$$

$$= w^T x_p + r\frac{w^T w}{\|w\|} + w_0 = r\|w\|$$

$$r = \frac{g(x)}{\|w\|}$$

# 线性判别函数

若$x$为原点，则 $g(x) = w_0$

从原点到超平面$H$的距离 $r_0 = \frac{w_0}{\|w\|}$

如果 $w_0 > 0$, 则原点在$H$的正侧

如果 $w_0 < 0$, 则原点在$H$的负侧

如果 $w_0 = 0$, 则$g(x)$具有齐次形式，超平面$H$通过原点

超平面$H$的方向由权向量$w$决定；其位置由阈值权$w_0$确定。

判别函数$g(x)$正比于 $x$ 点到超平面的代数距离

当$x$在$H$的正侧时，$g(x) > 0$

当$x$在$H$的负侧时，$g(x) < 0$

# 广义线性判别函数

- **两类问题, X是一维样本空间**

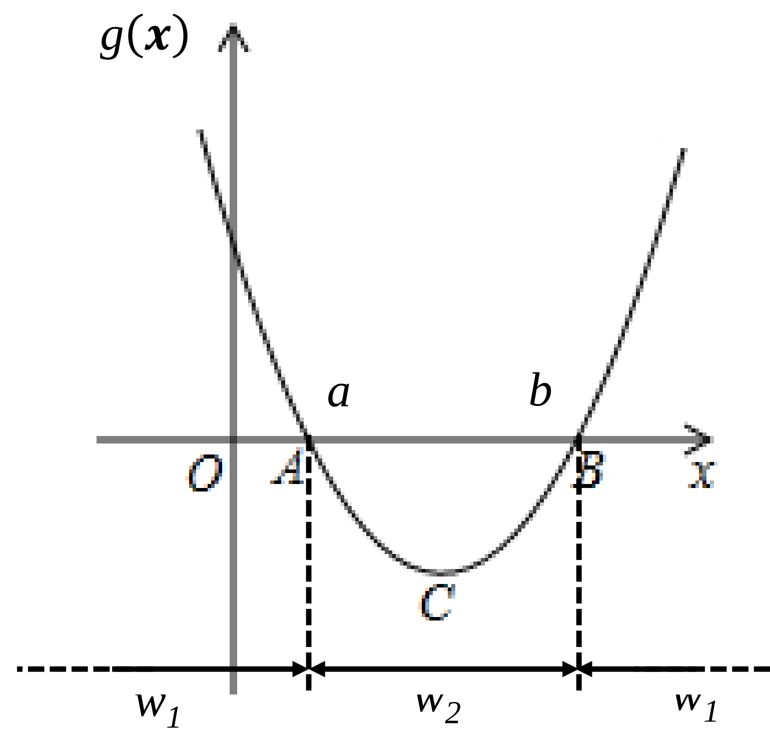
不适用于非凸和多连通区域划分

**若** $x < a$ **或** $x > b$**,** $x \in w_1$

**若** $a < x < b$**,** $x \in w_2$

**如果建立** $g(x) = (x-a)(x-b)$

**决策规则** $\begin{cases} \boldsymbol{g(x) > 0, x \in w_1} \\ \boldsymbol{g(x) < 0, x \in w_2} \end{cases}$




**一般形式** $g(x) = c_0 + c_1 x + c_2 x^2$

**可转化为** $g(x) = a^T y = \sum_{i=1}^{3} a_i y_i$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \qquad a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}$$

**称** $g(x) = a^T y$ **广义线性判别函数**

$a$ **广义权向量**

**利用线性函数的简单性解决复杂问题**

**维数大大增加→"维数灾难"**

# 线性判别函数齐次化

- **线性判别函数** $g(x) = w^T x + w_0$

**改写成** $g(x) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=1}^{d} a_i y_i = a^T y$

**称为线性判别函数的齐次简化**

$y = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ x \end{bmatrix}$ **增广样本向量；** $a = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ w \end{bmatrix}$ **增广权向量**

$\hat{d} = d + 1$; **$y$比$x$增加一维，保持样本空间欧式距离不变，变换后的样本仍全部位于$d$维子空间(原$X$空间)中。**

**方程** $a^T y = 0$ **在$Y$空间确定了一个通过原点的超平面$\hat{H}$, 它对$d$维子空间的划分与原决策面**$w^T x + w_0 = 0$**对原$X$空间的划分完全相同。$Y$空间中任意一点$y$到$\hat{H}$的距离:**

$$r = \frac{g(x)}{\|a\|} = \frac{a^T y}{\|a\|}$$

# 线性分类器设计

- **利用训练样本建立线性判别函数**

$$g(x) = \boxed{w^T} x + \boxed{w_0}$$

$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=1}^{d} a_i y_i = \boxed{a^T} y$$

**最好的结果一般出现在准则函数的极值点上，所以将分类器设计问题转化为求准则函数极值 $w^*$，$w_0^*$ 或 $a^*$ 的问题。**

**步骤1：具有类别标志的样本集$\mathcal{X} = \{x_1, x_2, \ldots x_N\}$或其增广样本集$\mathcal{Y}$。**

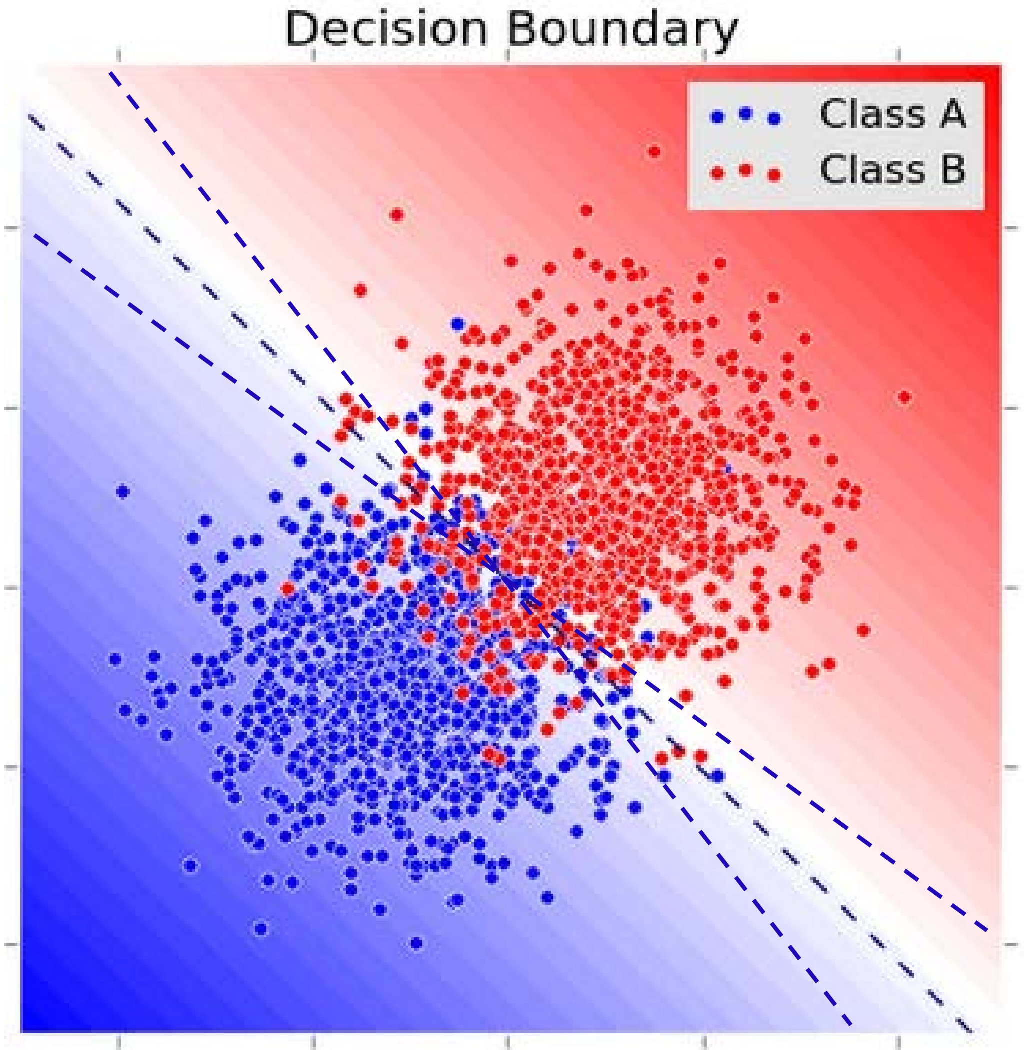
**步骤2：确定准则函数$\mathcal{J}$，满足①$\mathcal{J}$是样本集和$w$, $w_0$或$a$的函数；②$\mathcal{J}$的值反应分类器的性能，其极值对应"最好"的决策。**

**步骤3：优化求解准则函数极值 $w^*$, $w_0^*$ 或 $a^*$ 。**

**最终得到线性判别函数：$g(x) = w^{*T} x + w_0^*$ 或 $g(x) = a^{*T} y$，对于位置类别样本$x_k$，计算$g(x_k)$并通过决策规则判断其类别。**

# 准则函数

- Fisher准则
- 感知机准则
- 最小平方误差准则
- 最小错分样本数准则
- …

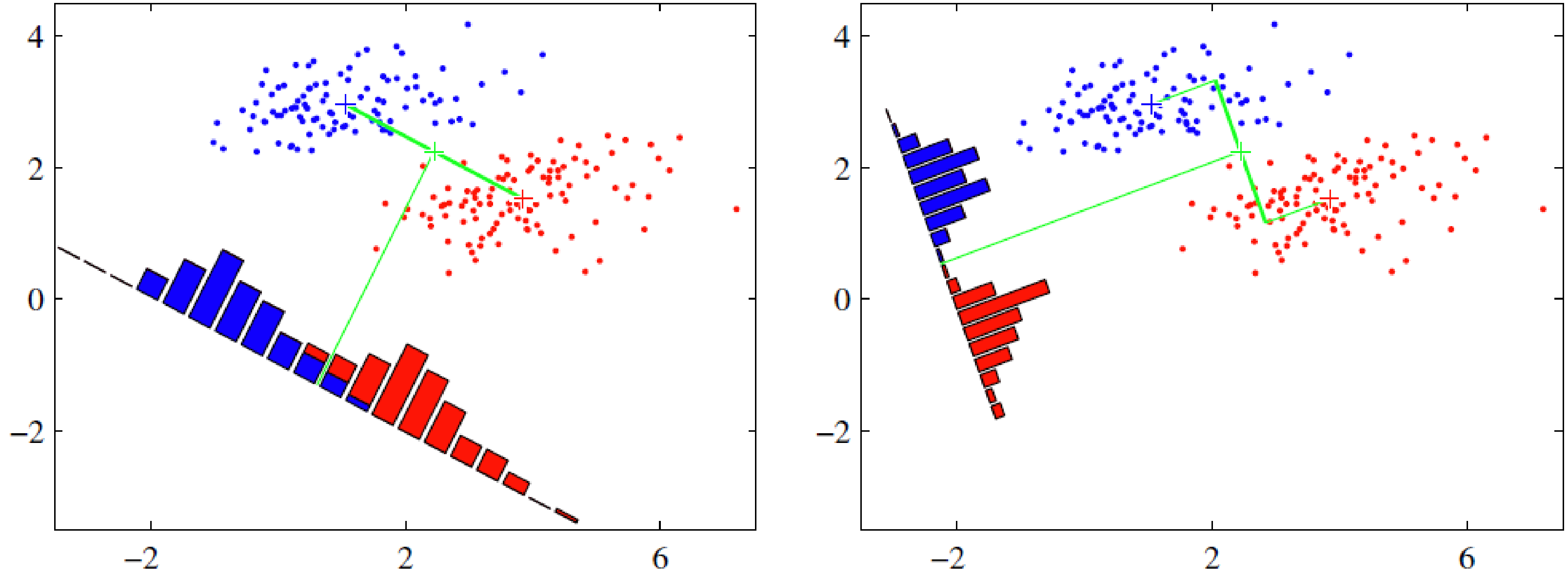


# Fisher准则

- **R. A. Fisher (1936年论文)**

**考虑把 *d* 维空间的样本投影到一条直线上形成一维空间。在一般情况下总可以找到某个方向，使样本在这个方向的直线上的投影分开得最好。**

**Fisher准则就是要解决**
**如何根据实际情况找到这条最好的、最易于分类的投影线的问题**



# Fisher准则

- **寻找最好投影方向** $w^*$

**以二分类问题为例，*d*维样本$x_1, x_2, \ldots x_N$，其中$N_1$个属于$w_1$类记为子集$X_1$，$N_2$个属于$w_2$类记为子集$X_2$。**

**在*d*维*X*空间**

**(1)各类样本的均值向量$m_i$** $m_i = \frac{1}{N_i}\sum_{x \in \chi_i} x, i = 1,2$

**(2)样本类内离散度矩阵$S_i$和总类内离散度矩阵$S_w$**

$$S_i = \sum_{x \in \chi_i}(x - m_i)(x - m_i)^T, i = 1,2$$

$$S_w = S_1 + S_2$$

$$S_w = P(w_1)S_1 + P(w_1)S_2$$

**(3)样本类间离散度矩阵$S_b$**

$$S_b = (m_1 - m_2)(m_1 - m_2)^T$$

$$S_b = P(w_1)P(w_2)(m_1 - m_2)(m_1 - m_2)^T$$

# Fisher准则

- 寻找最好投影方向 $w^*$

以二分类问题为例，*d*维样本$x_1, x_2, \dots x_N$, 其中$N_1$个属于$w_1$类记为子集$X_1$, $N_2$个属于$w_2$类记为子集$X_2$。

在一维*Y*空间 $$y_n = w^T x_n$$

(1)各类样本均值$\widetilde{m}_i$ $$\tilde{m}_i = \frac{1}{N_i} \sum_{y \in \eta_i} y, i = 1, 2$$

(2)样本类内离散度$\widetilde{S}_i^2$和总类内离散度$\widetilde{S}_w$

$$\tilde{S}_i^2 = \sum_{y \in \eta_i} (y - \tilde{m}_i)^2, i = 1, 2$$

$$\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2$$

# Fisher准则

- 希望投影后在一维Y空间中各类样本尽可能分开，即两类均值之差$(\tilde{m}_1 - \tilde{m}_2)$越大越好；同时希望各类样本内部尽量密集，即类内离散度越小越好。

Fisher准则函数： $J_F(w) = \dfrac{\boxed{(\tilde{m}_1 - \tilde{m}_2)^2}}{\boxed{\tilde{S}_1^2 + \tilde{S}_2^2}}$

# Fisher准则

- 求$J_F(w)$取得最大值的$w$ $J_F(w)=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_1^2+\tilde{S}_2^2}$

$$\tilde{m}_i=\frac{1}{N_i}\sum_{y\in\eta_i}y=\frac{1}{N_i}\sum_{x\in\chi_i}w^Tx=w^T(\frac{1}{N_i}\sum_{x\in\chi_i}x)=w^Tm_i$$

$$(\tilde{m}_1-\tilde{m}_2)^2=(w^Tm_1-w^Tm_2)^2=w^T(m_1-m_2)(m_1-m_2)^Tw=w^TS_bw$$

$$\tilde{S}_i^2=\sum_{y\in\eta_i}(y-\tilde{m}_i)^2=\sum_{x\in\chi_i}(w^Tx-w^Tm_i)^2=w^T[\sum_{x\in\chi_i}(x-m_i)(x-m_i)^T]w=w^TS_iw$$

$$\tilde{S}_1^2+\tilde{S}_2^2=w^T(S_1+S_2)w=w^TS_ww \Longrightarrow J_F(w)=\frac{w^TS_bw}{w^TS_ww}$$

**Lagrange乘子法求解:**

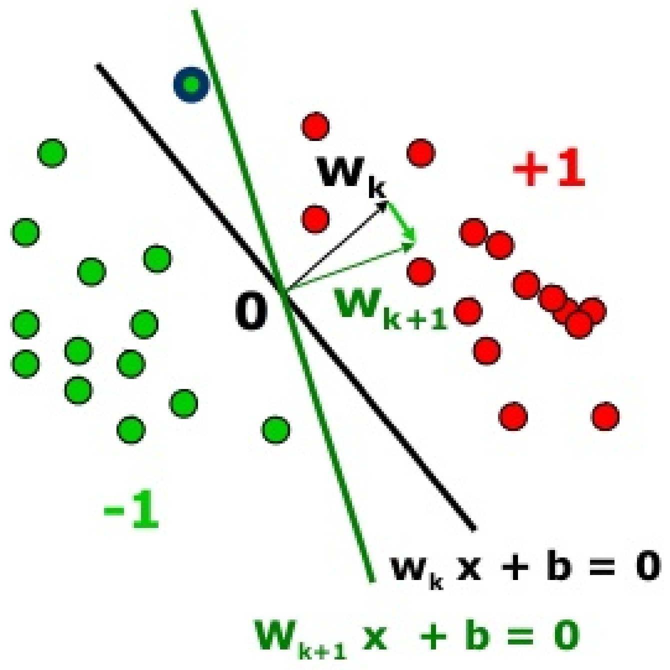
$$w^*=S_w^{-1}(m_1-m_2)$$

**分类时确定分界阈值$y_0$，与** $y=w^{*T}x$ **比较进行决策。**

# 感知机准则

- F. Rosenblatt(1950-1960年提出)

感知准则是一种自学习判别函数生成方法，由于Rosenblatt试图将其用于脑模型感知器，因此得名。该方法对随意给定的判别函数初始值，通过样本分类训练过程逐步对其修正直至最终确定。



# 几个基本概念

- **线性可分性**

一组容量为$N$的样本集$y_1, y_2, \ldots y_N$, 其中$y_n$为$\hat{d}$维增广样本向量, 分别来自$w_1$类和$w_2$类, 如果存在权向量$a$, 使得对于任何$y \in w_1$, 都有$a^T y > 0$, 而对于任何$y \in w_2$, 都有$a^T y < 0$, 则称这组样本为线性可分的, 反之亦然成立。

- **样本的规范化**

$$\begin{cases} a^T y_i > 0, y_i \in w_1 \\ \boxed{a^T y_j < 0, y_j \in w_2} \end{cases}$$

$$\Downarrow$$

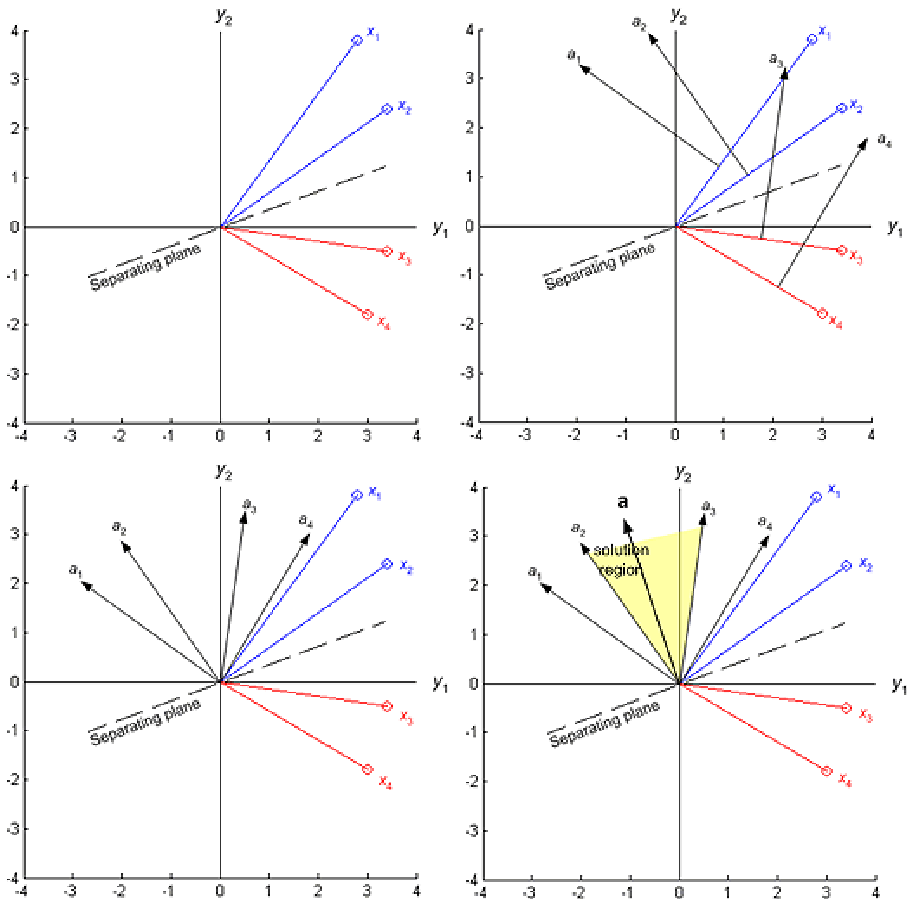
$$y_n' = \begin{cases} y_i > 0, y_i \in w_1 \\ -y_j < 0, y_j \in w_2 \end{cases}$$ 规范化增广样本向量

$$\Downarrow$$

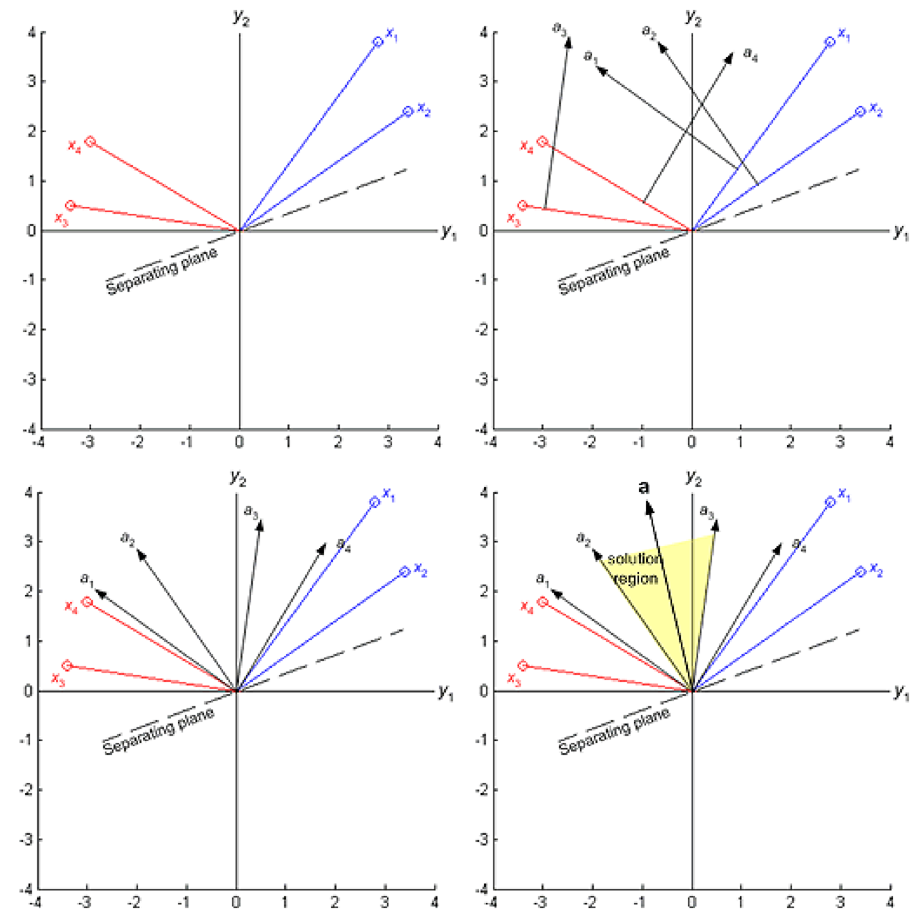
$$a^T y_n' > 0, n = 1, 2, \ldots, N$$

# 几个基本概念

- **解向量和解区**




未规范化

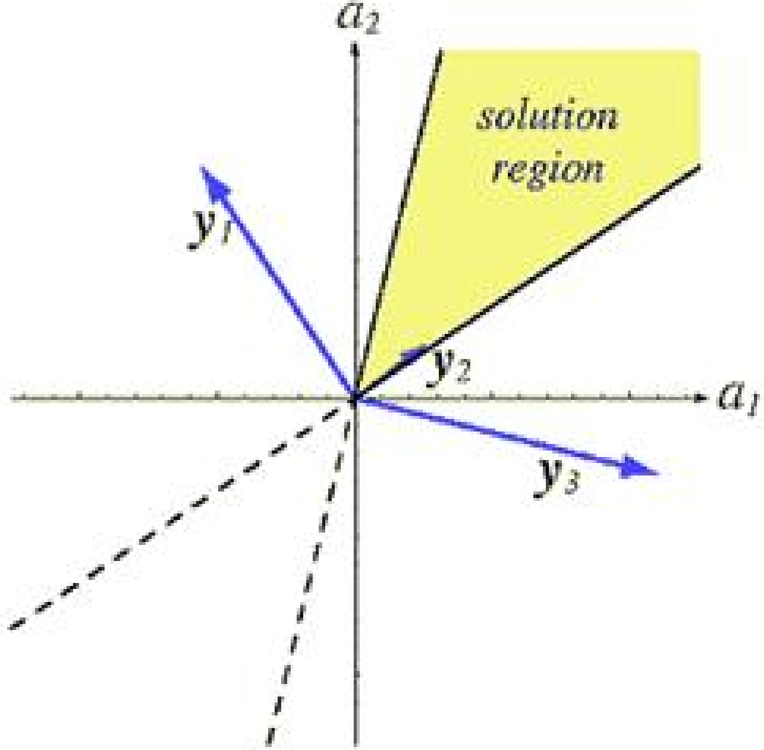


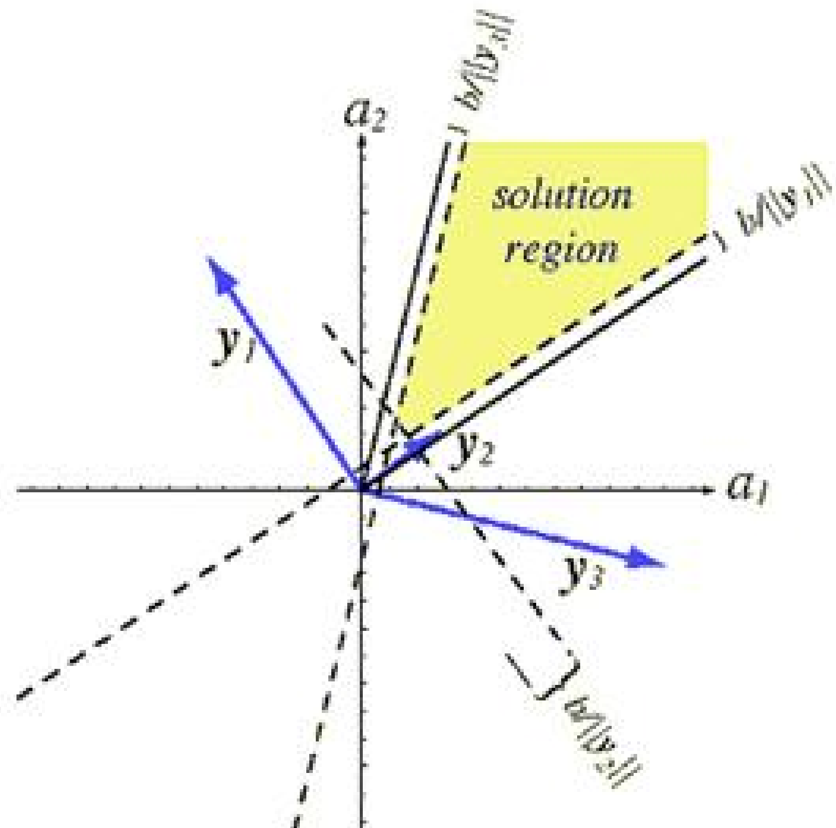

规范化

# 几个基本概念

- **对解区的限制**

**使解向量更可靠** $a^T y_n \geq b > 0$**，避免解收敛到解区边界的某点上**



# 感知机准则

- **寻找解向量*a****

对于一组样本$y_1, y_2, \ldots y_N$, 其中$y_n$是规范化增广样本向量, 使得:

$$a^T y_n > 0, n = 1, 2, ..., N$$

对于线性可分问题，构造准则函数 $J_P(a) = \sum_{y \in \eta^k} (-a^T y)$

其中 $\eta^k$ 是被权向量$a$错分的样本集合，即当$y$被错分时，就有 $a^T y_n \leq 0$

因此 $J_P(a) \geq 0$，仅当$a$为解向量或在解区边界时$J_P(a) = 0$

也就是说，当且仅当 $\eta^k$ 为空集时 $J_P^*(a) = min J_P(a) = 0$

此时无错分样本，这时的$a$就是解向量$a^*$ 。

# 感知机准则

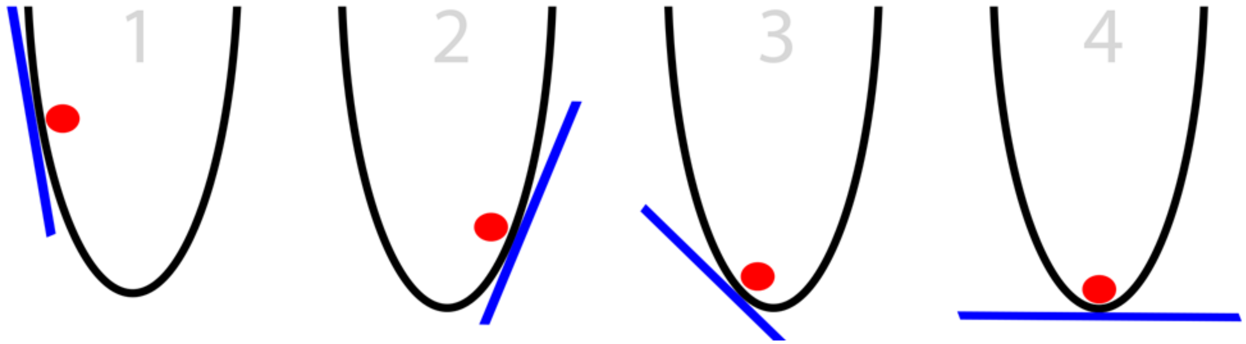
- 求使 $J_P(a)$ 达到最小值的***a****

采用梯度下降法求解 $J_P(a) = \sum_{y \in \eta^k} (-a^T y)$

$$\nabla J_P(a) = \frac{\partial J_P(a)}{\partial a} = \sum_{y \in \eta^k} (-y)$$

$$a(k+1) = a(k) - \rho_k \nabla J$$ 梯度下降法迭代公式

$$a(k+1) = a(k) + \rho_k \nabla \sum_{y \in \eta^k} y$$



# 感知机准则-示例

- Sample set for two-class case
  - Class 1

$$\mathbf{x}_1 = \begin{pmatrix} -2 \\ 2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} -2 \\ -2 \end{pmatrix} \qquad \mathbf{y}_1 = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix}$$

  - Class 2

$$\mathbf{x}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \qquad \mathbf{y}_3 = \begin{pmatrix} -1 \\ -2 \\ -1 \end{pmatrix}, \mathbf{y}_4 = \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}$$

  - Initial weight vector

$$\mathbf{a}(1) = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} \qquad \mathbf{a}(1)^t = \begin{pmatrix} 0 & 2 & 1 \end{pmatrix}$$

# 感知机准则-示例

- Example (Cont.)
  - Iterative procedure
    - (1) $\mathbf{y}^{(1)t} = \begin{pmatrix} 1 & -2 & 2 \end{pmatrix}$

$$\mathbf{a}(1)^t \mathbf{y}^{(1)} = \begin{pmatrix} 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} = -2 < 0$$

$$\mathbf{a}(2)^t = \begin{pmatrix} 0 & 2 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -2 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3 \end{pmatrix}$$

# 感知机准则-示例

- Example (Cont.)
  - Iterative procedure
    - (2) $\mathbf{y}^{(2)t} = \begin{pmatrix} 1 & -2 & -2 \end{pmatrix}$

$$\mathbf{a}(2)^t \mathbf{y}^{(2)} = \begin{pmatrix} 1 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix} = -5 < 0$$

$$\mathbf{a}(3)^t = \begin{pmatrix} 1 & 0 & 3 \end{pmatrix} + \begin{pmatrix} 1 & -2 & -2 \end{pmatrix} = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix}$$

# 感知机准则-示例

- Example (Cont.)
  - Iterative procedure
    - (3)

$$\mathbf{y}^{(3)t} = \begin{pmatrix} -1 & -2 & -1 \end{pmatrix}$$

$$\mathbf{a}(3)^t \mathbf{y}^{(3)} = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ -2 \\ -1 \end{pmatrix} = 1 > 0$$

$$\mathbf{a}(4)^t = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \ \text{(no chnage)}$$

# 感知机准则-示例

- Example (Cont.)
  - Iterative procedure
    - (4) $\mathbf{y}^{(4)t} = \begin{pmatrix} -1 & -2 & 1 \end{pmatrix}$

$$\mathbf{a}(4)^t \mathbf{y}^{(3)} = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix} = 3 > 0$$

$$\mathbf{a}(5)^t = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \quad \text{(no chnage)}$$

# 感知机准则-示例

- Example (Cont.)
  - Iterative procedure
    - (5)

$$\mathbf{y}^{(5)t} = \begin{pmatrix} 1 & -2 & 2 \end{pmatrix}$$

$$\mathbf{a}(5)^t \mathbf{y}^{(1)} = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} = 8 > 0$$

$$\mathbf{a}(6)^t = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \qquad \text{(no change)}$$

# 感知机准则-示例

- Example (Cont.)
  - Iterative procedure
    - (6)

$$\mathbf{y}^{(6)t} = \begin{pmatrix} 1 & -2 & -2 \end{pmatrix}$$

$$\mathbf{a}(6)^t \mathbf{y}^{(2)} = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix} = 4 > 0$$

$$\mathbf{a}(7)^t = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix}$$

# 感知机准则-示例

- Example (Cont.)
  - Iterative procedure
    - (7)

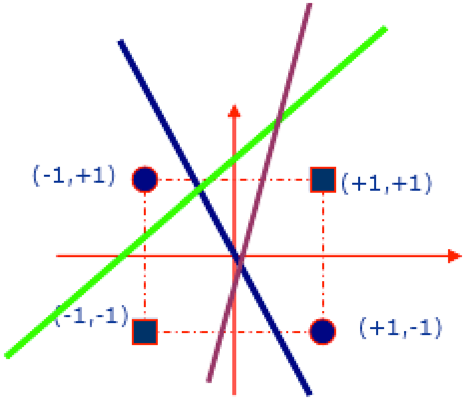
$$\mathbf{y}^{(7)t} = \begin{pmatrix} -1 & -2 & -1 \end{pmatrix}$$

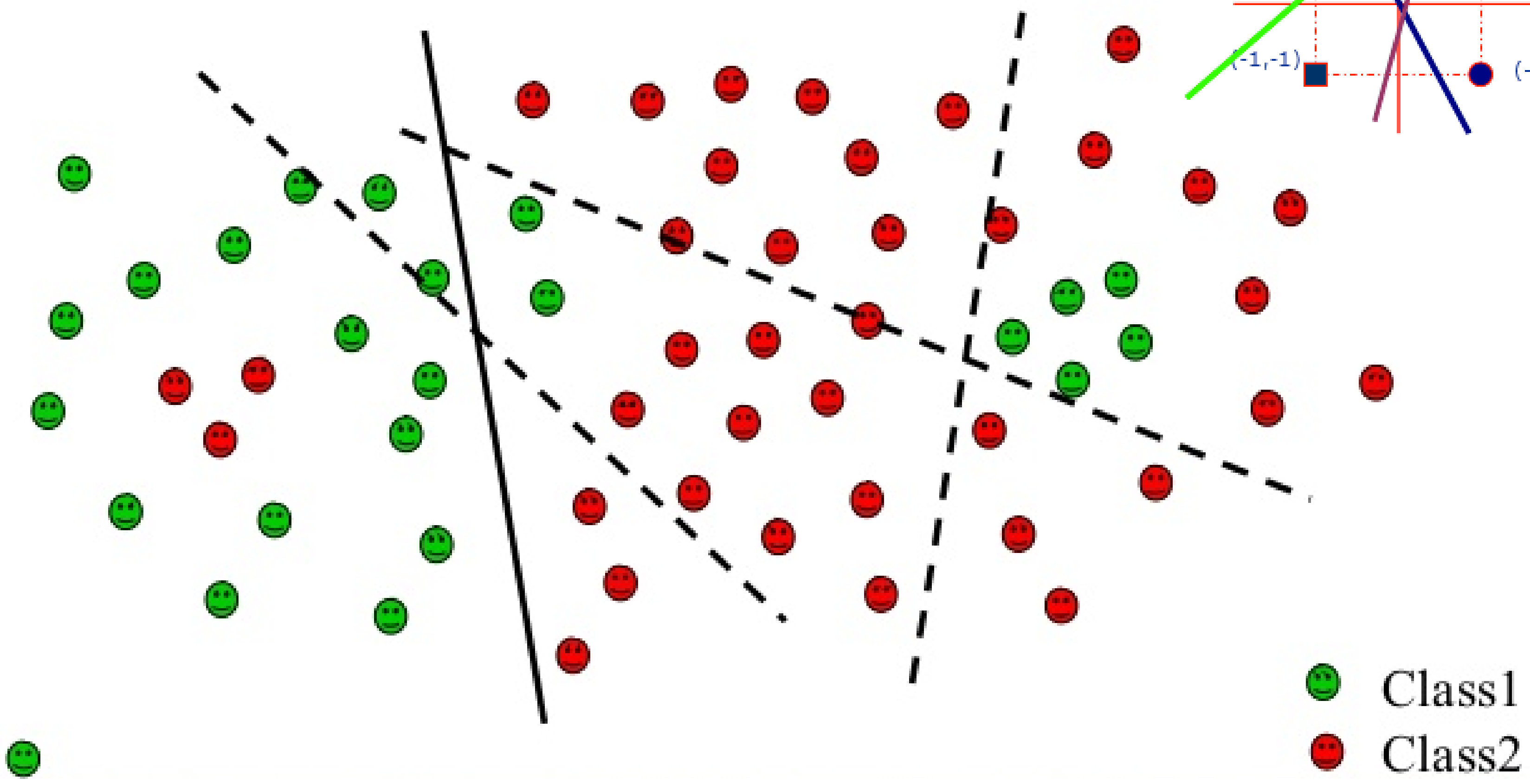
$$\mathbf{a}(7)^t \mathbf{y}^{(7)} = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ -2 \\ -1 \end{pmatrix} = 1 > 0$$

$$\mathbf{a}(8)^t = \begin{pmatrix} 2 & -2 & 1 \end{pmatrix} \ \text{(no chnage)}$$

# 感知机准则

- 结果不唯一且在线性不可分情况下不收敛

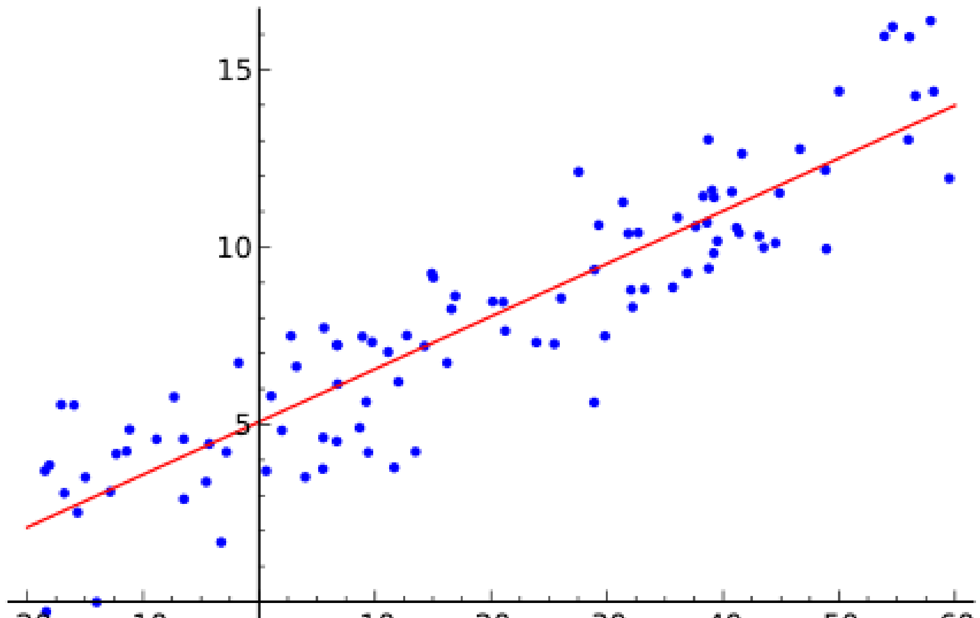


# 最小二乘准则

- A.-M. Legendre(1806提出)，C. Gauss(1809提出/1829证明)

**最小二乘法(最小平方误差法)通过最小化误差的平方和寻找数据的最佳函数匹配，即可以使求得的数据与实际数据之间误差的平方和最小。**



# 最小二乘准则

- **寻找最好投影方向** $a^*$

$$a^T y_n > 0$$

⬇

$$a^T y_n = b_n > 0$$ **$b_n$是任意给定的正常数**

**方程组形式:**

$$Ya = b$$

$$Y=\begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1\hat{d}} \\ y_{21} & y_{22} & \cdots & y_{2\hat{d}} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{N\hat{d}} \end{bmatrix}$$

**$y_n$是规范化增广向量样本**
**$Y$是$N \times \hat{d}$维矩阵，通常$N > \hat{d}$，一般为列满秩阵**

$$b = [b_1 \quad b_2 \quad \cdots \quad b_N]$$

**$b$是$N$维向量，$b_n$>0， $n$=1, 2, …, $N$**

**方程数多于未知数的矛盾方程组通常没有精确解**

**定义误差向量:** $e = Ya - b$ **及平方误差准则函数**

$$J_S(a) = \|e\|^2 = \|Ya - b\|^2 = \sum_{n-1}^{N} (a^T y_n - b_n)^2$$

# 最小二乘准则

- 求使 $J_S(a)$ 最小的$a^*$ (最小二乘近似解/伪逆解/MSE解)

采用解析法求伪逆解 $J_S(a) = \|e\|^2 = \|Ya - b\|^2 = \sum^{N} (a^T y_n - b_n)^2$

$$\nabla J_S(a) = \sum_{n-1}^{N} 2(a^T y_n - b_n) y_n = 2Y^T(Ya - b)$$

令 $\nabla J_S(a) = 0$

得 $Y^T Y a^* = Y^T b$　　矩阵 $Y^TY$是$\hat{d} \times \hat{d}$方阵一般非奇异

唯一解 $a^* = (Y^T Y)^- 1 Y^T b = Y^+ b$

其中$\hat{d} \times N$矩阵 $Y^+ = (Y^T Y)^- 1 Y^T$ 是Y的左逆矩阵

如何选$b$?

$$\boldsymbol{b} = \begin{array}{cl} \begin{bmatrix} N/N_1 \\ \cdots \\ N/N_1 \\ N/N_2 \\ \cdots \\ N/N_2 \end{bmatrix} & \begin{array}{l} N_1\text{个} \\ \\ \\ N_2\text{个} \end{array} \end{array}$$

$a^*$等价于Fisher解

$$g_0(x) = P(w_1|x) - P(w_2|x)$$

$$N \to \infty, \boldsymbol{b} = \underbrace{[\mathbf{1}, \mathbf{1}, \cdots, \mathbf{1}]^T}_{N\text{个}}$$

以最小均方误差逼近贝叶斯判别函数

# 最小二乘准则

- **求使 $J_S(a)$ 最小的$a^*$ (最小二乘近似解/伪逆解/MSE解)**

$$a^* = Y^+ b \qquad Y^+ = (Y^T Y)^- 1 Y^T$$

**问题：①要求$Y^TY$非奇异；②求$Y^+$计算量大同时可能引入较大误差。**

**采用梯度下降法求解：** $\nabla J_S(a) = 2Y^T(Ya - b)$

$$\begin{cases} a(1), Random \\ a(k+1) = a(k) - \rho_k Y^T(Ya - b) \end{cases}$$

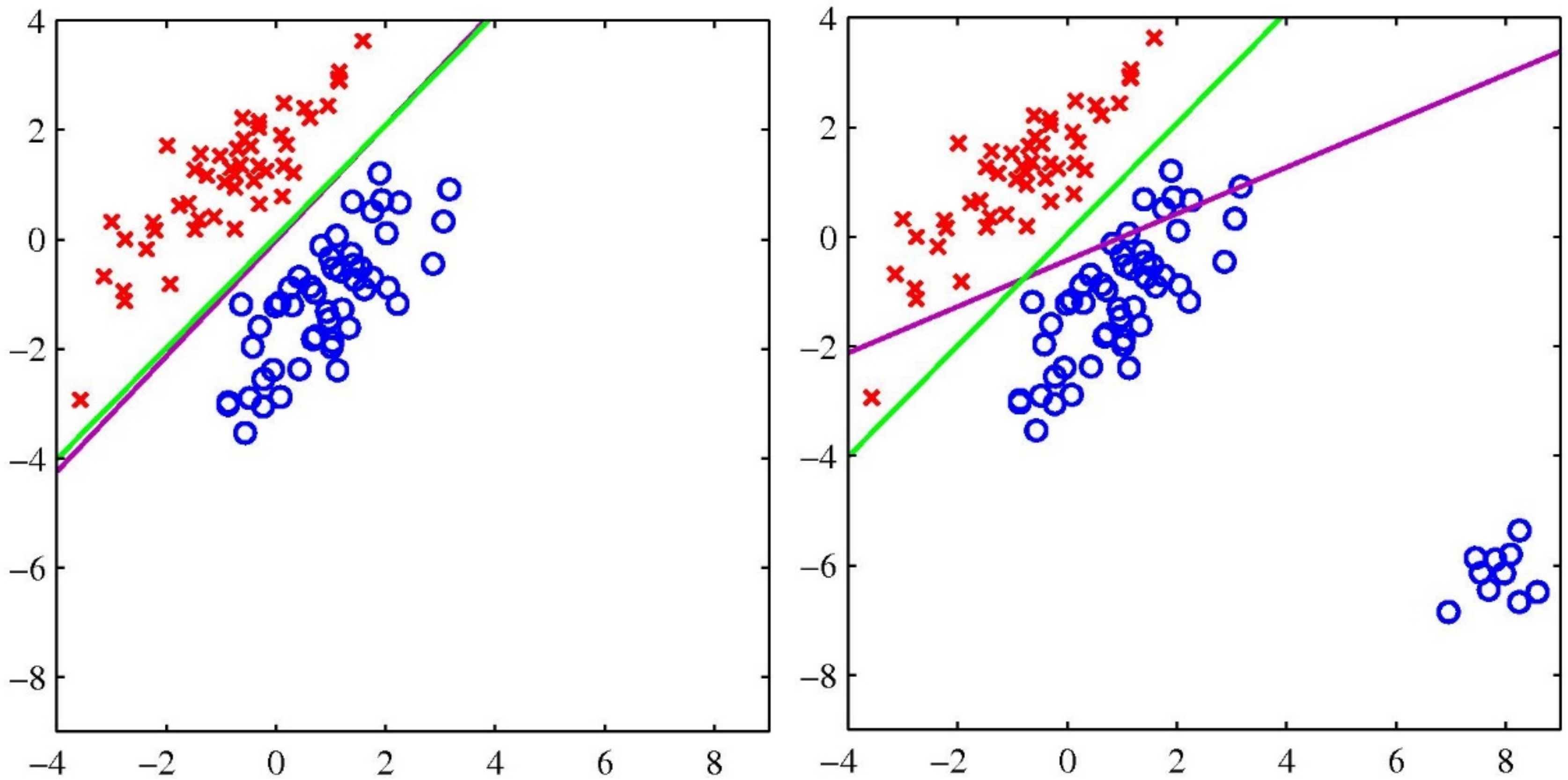
**可以证明，选择** $\rho_k = \dfrac{\rho_1}{k}$ **，**$\rho_1$**是任意常数**

**该算法权向量收敛于使** $\nabla J_S(a) = 2Y^T(Ya - b) = 0$ **的权向量$a^*$**

**不要求$Y^TY$奇异与否，只计算$\hat{d} \times \hat{d}$方阵$Y^TY$，比$\hat{d} \times N$阵$Y^+$计算量小**

# 最小二乘准则

- 对于异常值(Outlier)非常敏感



# 多分类问题

- **1 *vs.* (N-1) or 1 *vs.* 1**