河北师范大学软件学院
Software College of Hebei Normal University

分类模型

第一部分 线性分类模型

张朝晖

2018-2019学年 20180911-0912

1.分类模块

产生式分类模型 **A.**贝叶斯分类模型

判别式分类模型 **线性分类模型**
- **B. Fisher**判别分类
- **C.** 感知器分类模型
- **D.** 大间隔分类模型(线性***SVM***)

非线性分类模型
- **E.** 核***SVM***(非线性***SVM***)
- **F.** 核**Fisher**判别分类
- **G.** 神经网络

其它分类模型
- ***H*.KNN**分类模型
- ***I***.决策树分类模型
- ***J*.*Logistic***回归
- ***K*.*Softmax***回归

2.聚类模块
- ***L*.K-**均值聚类
- ***M***.高斯混合聚类
- ***N*.DBSCAN**聚类
- ***O***.层次聚类

3.回归模块
- ***P*.KNN**回归
- ***Q***.回归树
- ***R***.最小二乘线性回归
- ***S***.岭回归
- ***T***.LASSO回归

4.集成学习
- ***U*.*Bagging***
- ***V***.随机森林
- ***W*.*Boosting***

5.特征工程
- ***X***.主成分分析***(PCA)***
- ...

6.评价模块 混淆矩阵(及其相关指标)、ROC曲线、交叉验证

主要内容

A.引言

B. Fisher线性判别分析

C.感知器

D. 线性支持向量机(线性SVM)

A.引言

1. 有关约定

关键词：

- 线性运算、非线性运算
- 函数、方程
- 线性/非线性
- 特征空间及其划分
- 决策域、决策边界(决策面、分类边界)
- 产生式分类模型、判别式(鉴别式)分类模型
- 线性/非线性

  分类器、分类模型、分类边界、判别函数
- 两类别分类、多类别分类

2.基于模型的分类器设计 -- 代表模型：贝叶斯分类模型

基本思路：

(1)基于训练集，估计描述每个类别的概率密度(参数)

(2)观测样本的类别决策

训练样本 → 估计$p(x|\omega_j)$ → 贝叶斯判决

如何估计了类条件概率密度函数

(1)**参数估计法**

需明确类条件概率密度函数形式；

参数估计需要大量样本

(2)**非参数估计法**

不足：$p(x|\omega_j)$的估计需要大量样本；

随特征空间维数增加，占用大量存储空间

==>**贝叶斯决策尽管最优，但实现困难**

3. 基于样本（数据）的直接分类器设计

例：鉴别式分类模型，由训练样本集，直接确定决策域划分

这里我们首先关注利用样本数据设计鉴别式分类模型！！！

基本思路

明确
- 训练样本集
- 分类器（判别函数）类型
  - 线性判别函数
  - 非线性判别函数
- 分类器设计的目标准则
- 算法

⟶ **利用样本，估计判别函数的未知参数使所选目标准则最优**

分类器设计 ⇐ **判别函数参数估计** ⇒ 准则函数极值解求取

4. 鉴别式分类模型 { 线性分类模型 / 非线性分类模型

先从简单情况开始：
面向两类别分类的“线性分类模型”

线性分类模型的分类边界是关于$\boldsymbol{x}$的线性方程.

线性判别函数

- 对于多类：$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + \omega_{i0}$　　$i = 1,\ldots,c$
  - c个判别函数
- 对于两类：$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + \omega_0$　　1个判别函数

两类问题 $c = 2$

- d维特征向量(样本向量)　$\boldsymbol{x} = [x_1 \quad x_2 \quad \cdots \quad x_d]^T$
- 判别函数　$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + \omega_0$　　$[\boldsymbol{g}(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})]$
- 决策边界　$\boldsymbol{w}^T \boldsymbol{x} + \omega_0 = 0$
  - 法向量(权向量)$\boldsymbol{w} = [w_1 \quad w_2 \quad \cdots \quad w_d]^T$
  - 阈值权ω_0
- 决策域：第1类决策域、第2类决策域

两类别分类模型--决策边界$\boldsymbol{H}$(分类超平面)的确定：

$g(\boldsymbol{x})=0 \quad (g(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}_1+w_0)$

对于$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \boldsymbol{H}$并且$\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ 有 $g(\boldsymbol{x}_1)=0,\ g(\boldsymbol{x}_2)=0$

即 $\boldsymbol{w}^T\boldsymbol{x}_1+w_0=\boldsymbol{w}^T\boldsymbol{x}_2+w_0$

$\longrightarrow \quad \boldsymbol{w}^T(\boldsymbol{x}_1-\boldsymbol{x}_2)=0$

权向量w与超平面H的任意向量正交(w垂直于H)

权向量w是超平面H的法向量

对于$\boldsymbol{x} \in \Re_1$，有$g(\boldsymbol{x})>0$,所以 **决策面H的法向量指向$\Re_1$**

$\longrightarrow \begin{cases} \text{决策面}\boldsymbol{H}\text{的正侧：}\ \Re_1 \\ \text{决策面}\boldsymbol{H}\text{的负侧：}\ \Re_2 \end{cases}$

两类别分类模型

决策规则

$$\text{对于观测}\boldsymbol{x}\quad\begin{cases}\text{若}g(\boldsymbol{x})>0, & \text{则决策}\boldsymbol{x}\text{为}\omega_1\text{类}\\ \text{若}g(\boldsymbol{x})<0, & \text{则决策}\boldsymbol{x}\text{为}\omega_2\text{类}\\ \text{若}g(\boldsymbol{x})=0, & \text{则拒绝, 或结合实际问题分到某1类}\end{cases}$$

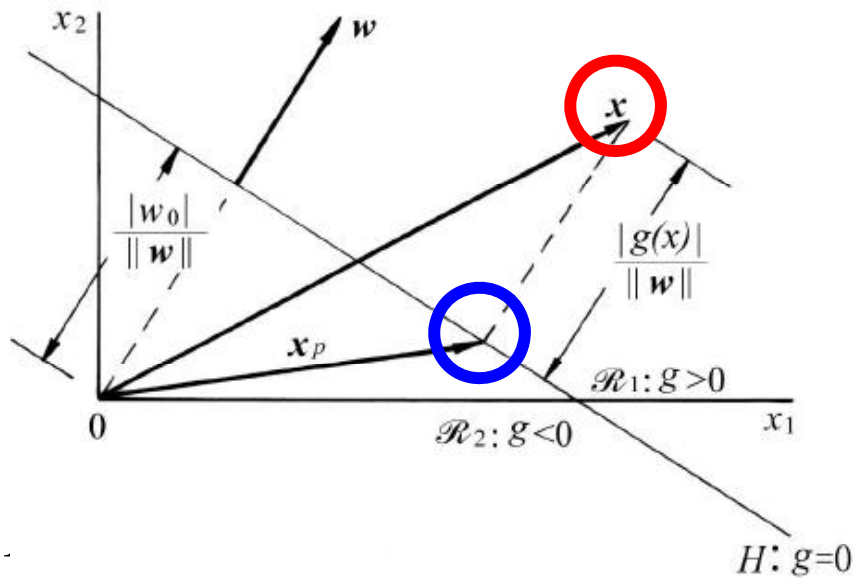
决策面(决策边界) H　　$g(\boldsymbol{x})=0$　(超平面)

$$\textbf{决策域}\begin{cases}\Re_1=\{\boldsymbol{x}:\boldsymbol{x}\in\Re\text{且}g(\boldsymbol{x})>0\}\\ \Re_2=\{\boldsymbol{x}:\boldsymbol{x}\in\Re\text{且}g(\boldsymbol{x})<0\}\end{cases}$$

可将判别函数$g(x)$视为观测x到决策面H距离的代数度量。

判别函数 $g(x)=w^T x+\omega_0$

三角形三边关系 $x=x_p+r\dfrac{w}{\|w\|}$



其中：

x_p x在H的投影向量，$x_p \in$

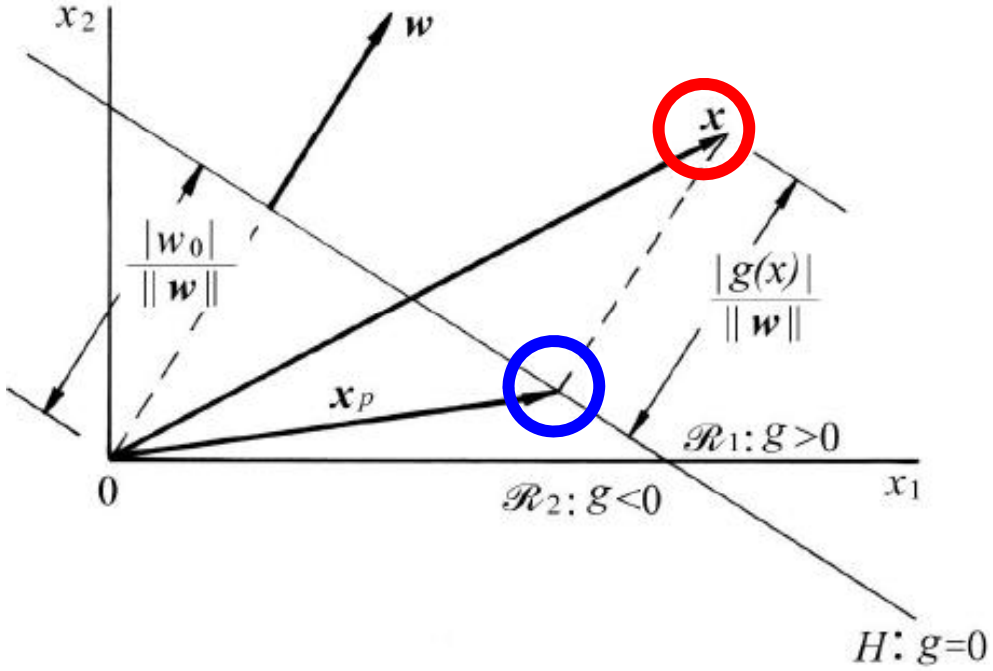
$g(x_p)=0=w^T x_p+\omega_0$

r x点到H面的代数垂直距离

若$r>0$,则x在H正侧。

$\dfrac{w}{\|w\|}$ w方向的单位向量

$$\begin{cases} g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + \omega_0 \\ g(\boldsymbol{x}_p) = 0 = \boldsymbol{w}^T \boldsymbol{x}_p + \omega_0 \\ \\ \boldsymbol{x} = \boldsymbol{x}_p + r\dfrac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \end{cases}$$



$$\Rightarrow g(\boldsymbol{x}) = \boldsymbol{w}^T \overbrace{\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)}^{\boldsymbol{x}} + \omega_0$$

$$= \boldsymbol{w}^T \boldsymbol{x}_p + r\boldsymbol{w}^T \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \omega_0 = g(\boldsymbol{x}_p) + r\|\boldsymbol{w}\| = r\|\boldsymbol{w}\|$$

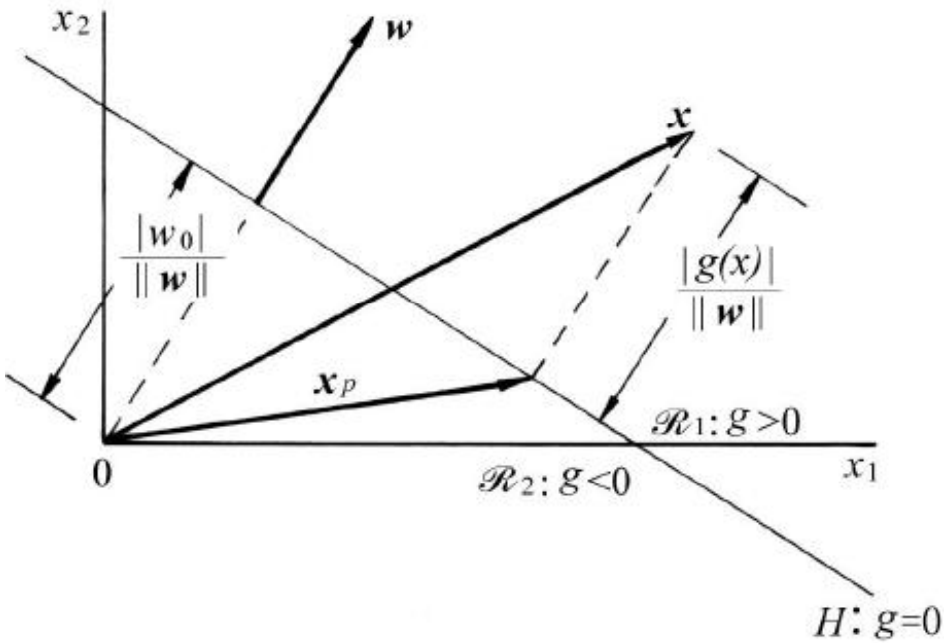
即： $g(\boldsymbol{x}) = r\|\boldsymbol{w}\|$ 或 $r = \dfrac{g(\boldsymbol{x})}{\|\boldsymbol{w}\|}$ $\|\boldsymbol{x} - \boldsymbol{x}_p\|_2 = \dfrac{|g(\boldsymbol{x})|}{\|\boldsymbol{w}\|}$

即：判别函数$g(x)$正比于x到超平面的代数距离r

问题：如何确定任意观测样本到分类边界的距离？

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + \omega_0 = r\|\boldsymbol{w}\|$$

$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$



对于原点 $\boldsymbol{x} = \boldsymbol{0}$，有：

$$g(\boldsymbol{x} = \boldsymbol{0}) = \omega_0, \quad r = \frac{\omega_0}{\|\boldsymbol{w}\|}$$

$$\begin{cases} \omega_0 > 0, & \text{原点}\boldsymbol{x} = \boldsymbol{0}\text{在}\boldsymbol{H}\text{正侧}; \\ \omega_0 < 0, & \text{原点}\boldsymbol{x} = \boldsymbol{0}\text{在}\boldsymbol{H}\text{负侧} \\ \omega_0 = 0, & \boldsymbol{H}\text{经过原点，判别函数}g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}。 \end{cases}$$

问题：如何确定原点到分类边界的距离？

超平面$\left(\text{分类面，决策面}\right)\boldsymbol{H}$的法向量为$\boldsymbol{w}$;

超平面$\boldsymbol{H}$的偏移$\left(\text{超平面位置}\right)$由$\omega_0$决定;

判别函数$\boldsymbol{g}\left(\boldsymbol{x}\right)$正比于观测$\boldsymbol{x}$到超平面$\boldsymbol{H}$的代数距离;

决策域$\Re_1,\Re_2$分别位于超平面$\boldsymbol{H}$的正负侧。

设计线性分类器的关键：

基于何种准则，估计判别函数的参数$\boldsymbol{w},\omega_0$;

主要内容

A.引言

B. Fisher线性判别分析

C.感知器

D. 线性支持向量机(线性SVM)

B. Fisher判别分类模型

掌握：

- Fisher判别分类模型设计的基本思想
- 准则函数
- Fisher判别分类模型的实现步骤
- Fisher法进行监督式特征提取的前提条件

问题：

你能否将Fisher判别分类模型的“Fisher 判别比”用于分类任务中的特征选择？

作业：

以鸢尾花数据集为例，采用Fisher判别分类模型(先监督式降维，再设计分类模型(如朴素贝叶斯))，进行分类。

主要内容

第一部分 类别数目C=2时 FLDA分类模型

关键：确定最佳投影直线

第二部分 类别数目C>2时FLDA分类模型

许多机器学习问题涉及样本的“降维”

〞降维〞的目的不止一个，如：

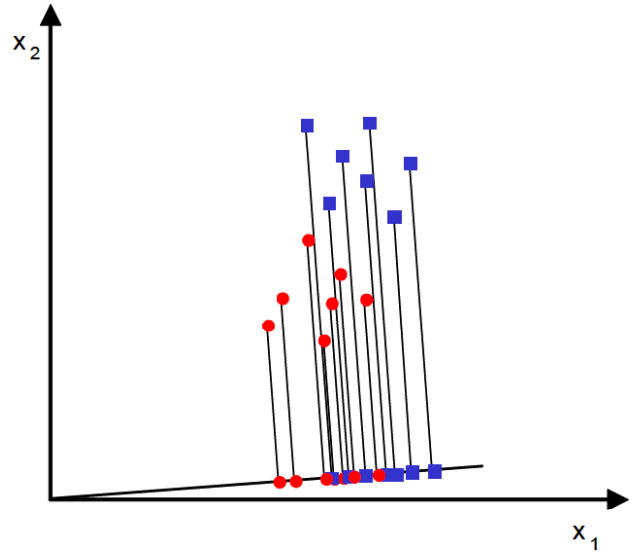
(1) ***PCA*--数据描述**(信息压缩)

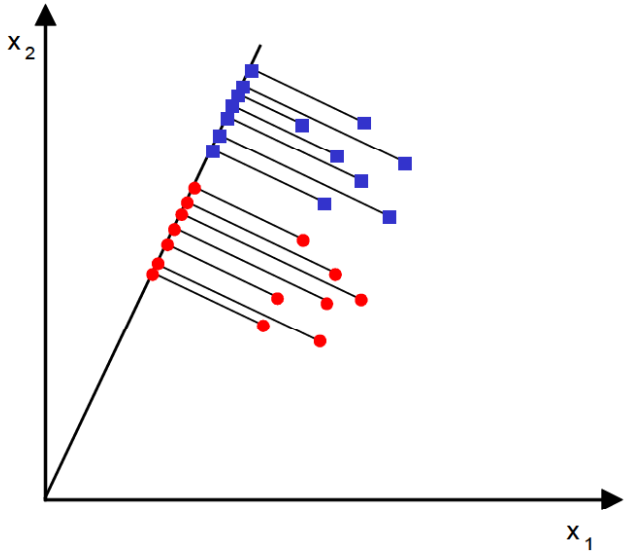
目的：寻找有效表示特征的主轴方向

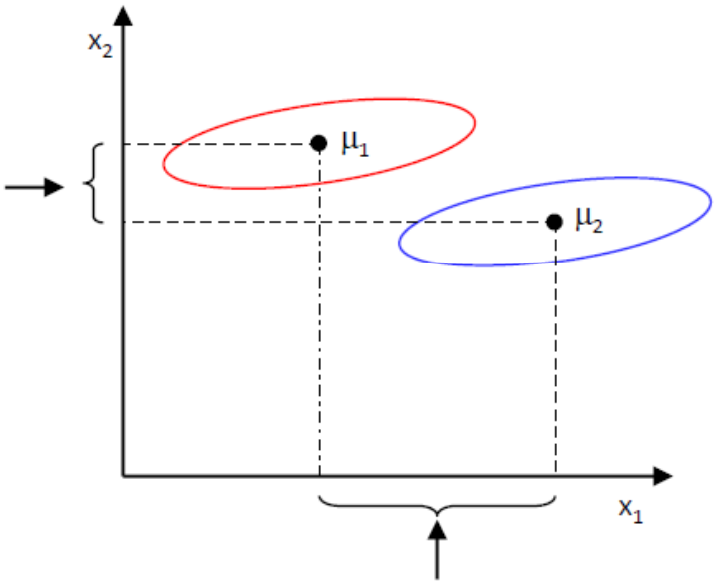
(2) ***Fisher*线性判别--分类器设计**

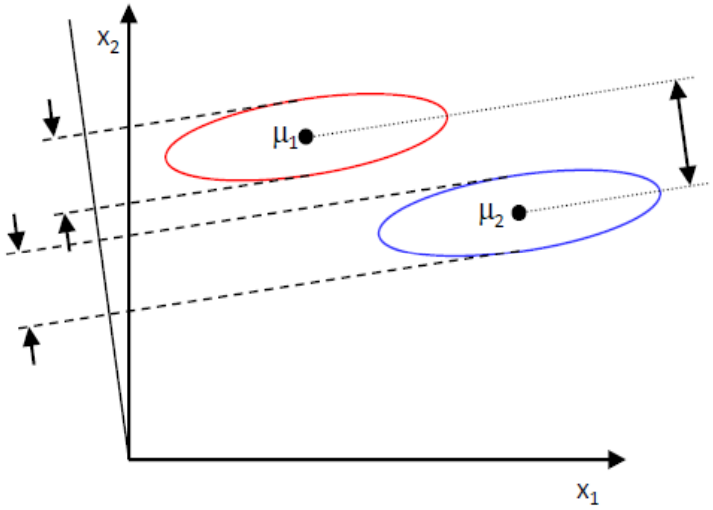
目的：寻找能有效分类的方向

什么样的投影方向更有助于分类?









1 基本思想

将d维特征空间的样本向某低维子空间做投影(两类分类对应1维空间)；在该投影空间，能最大限度区分各类数据点,以有效分类.

$$\begin{cases} \textbf{类间样本} & \text{尽可能远离} \\ \textbf{类内样本} & \text{尽可能聚集} \end{cases}$$

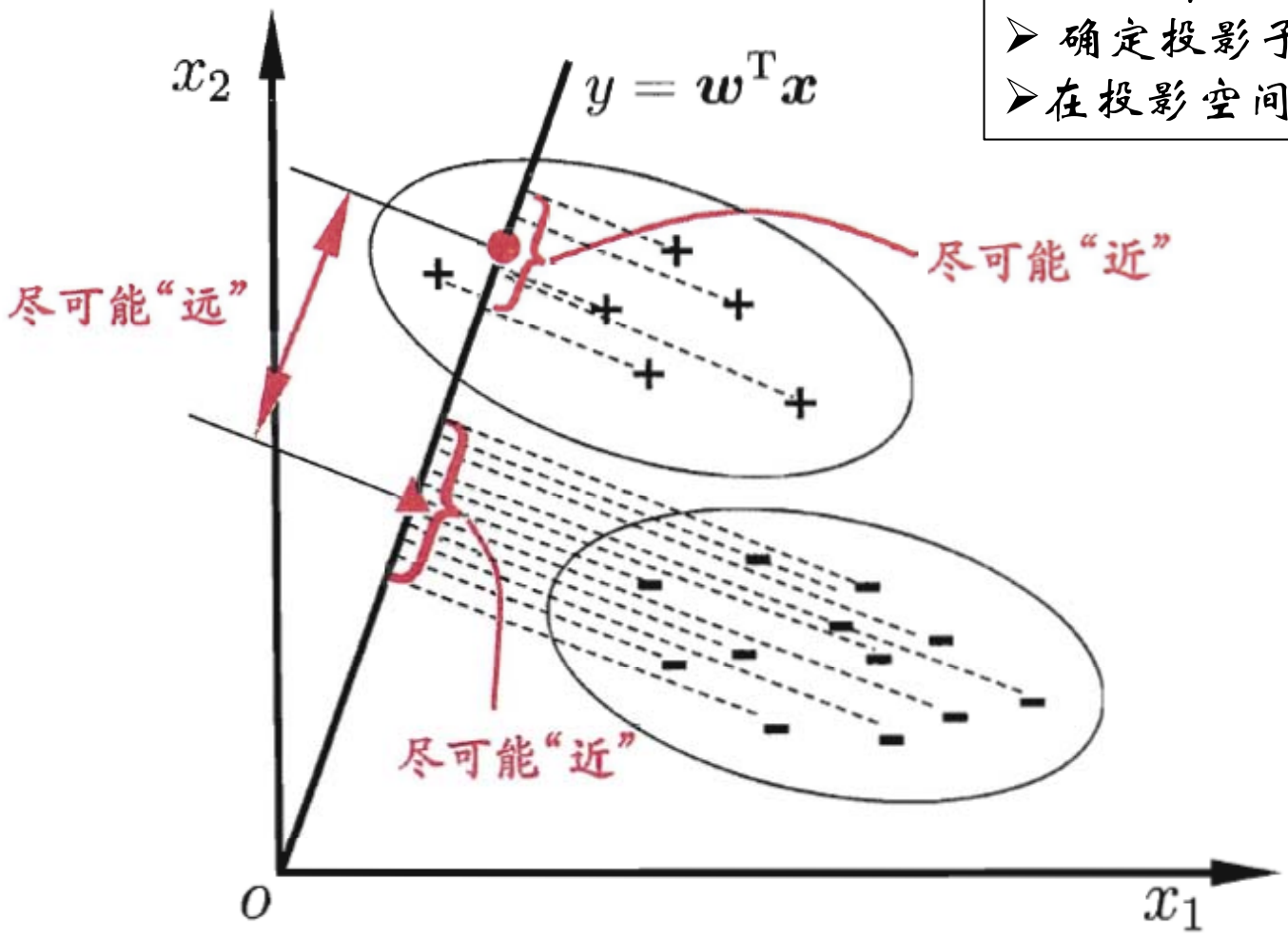
分类器的设计 $\Rightarrow$ **寻找投影直线的最佳方向** $(C=2)$

类间样本 尽可能远离

类内样本 尽可能聚集

二维特征、两类别Fisher判别分析 示意
引自周志华《机器学习》page60

任务分解：
- 确定投影子空间
- 在投影空间，设计分类模型



2 问题描述

已知：d维训练样本$\mathcal{X}=\{x_1,\ldots,x_N\}$

其中 $\begin{cases}\omega_1\text{类：}\ \mathcal{X}_1=\{x_1^{\,1},\ldots,x_{N_1}^{\,1}\}\\ \omega_2\text{类：}\ \mathcal{X}_2=\{x_1^{\,2},\ldots,x_{N_2}^{\,2}\}\end{cases}$

$\mathcal{X}=\mathcal{X}_1\cup\mathcal{X}_2 \qquad N=N_1+N_2$

$\xrightarrow{\text{对}x\text{投影：}y=w^Tx}\mathcal{Y}=\{y_1,\ldots,y_N\}$ $\begin{cases}\omega_1\text{类}\ \mathcal{Y}_1,\ N_1\\ \omega_2\text{类}\ \mathcal{Y}_2,\ N_2\end{cases}$

其中 $y_i=w^Tx_i, i=1,2,...,N$

求解：最佳投影方向w^*，使两类分类效果最好。

3.准备工作(定义几个必要的参量)

[1]投影前——d维特征空间

各类均值向量

$$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{x\in\mathcal{X}_i}\boldsymbol{x} \qquad i=1,2$$

总体均值向量(样本集的中心)

$$\boldsymbol{m} = \frac{\boldsymbol{m}_1 N_1 + \boldsymbol{m}_2 N_2}{N_1 + N_2}$$

类内散度矩阵$(within-class\ \ scatter\ \ matrix)$

$$\boldsymbol{S}_i = \frac{1}{N_i}\sum_{x\in\mathcal{X}_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^T \qquad i=1,2$$

[1]投影前——d维特征空间

总类内散度矩阵$(pooled\ \ within-class\ \ scatter\ \ matrix)$

$$S_w=\frac{1}{N_1+N_2}\sum_{i=1}^{2}N_iS_i$$

S_w对称半正定；若$N>d$，S_w通常非奇异

类间散度矩阵$(between-class\ \ scatter)$

$$\begin{cases}\text{不考虑先验概率} & S_b=(m_1-m_2)(m_1-m_2)^T\\ \text{若考虑先验概率} & S_b=P(\omega_1)P(\omega_2)(m_1-m_2)(m_1-m_2)^T\end{cases}$$

$$S_b=\frac{1}{N_1+N_2}\sum_{i=1}^{2}N_i(m_i-m)(m_i-m)^T=\sum_{i=1}^{2}P(\omega_i)(m_i-m)(m_i-m)^T$$

或者 $$S_b=\frac{1}{2}\sum_{i=1}^{2}(m_i-m)(m_i-m)^T$$

[2]投影后 –1维空间

样本均值

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in\mathcal{Y}_i} y = \frac{1}{N_i}\sum_{x\in\mathcal{X}_i} w^T x = w^T\left(\frac{1}{N_i}\sum_{x\in\mathcal{X}_i} x\right) = w^T m_i$$

$$i = 1,2$$

总体样本均值

$$\tilde{m} = \frac{\tilde{m}_1 N_1 + \tilde{m}_2 N_2}{N_1 + N_2} = w^T m$$

类内离散度

$$\tilde{S}_i^2 = \frac{1}{N_i}\sum_{y_j\in\mathcal{Y}_i}\left(y_j - \tilde{m}_i\right)^2$$

$$= \frac{1}{N_i}\sum_{x_j\in\mathcal{X}_i} w^T\left(x_j - m_i\right)\left(x_j - m_i\right)^T w = w^T S_i w \qquad i = 1,2$$

[2]投影后 –1维空间

总类内离散度

$$\widetilde{S}_w = \widetilde{S}_1^2 + \widetilde{S}_2^2 = w^T S_1 w + w^T S_2 w = w^T \left(S_1 + S_2 \right) w = w^T S_w w$$

$$\widetilde{S}_w = \frac{1}{N_1 + N_2} \sum_{i=1}^{2} N_i \widetilde{S}_i^2 = \frac{1}{N_1 + N_2} \sum_{i=1}^{2} N_i w^T S_i w = w^T S_w w$$

类间离散度

$$\widetilde{S}_b^2 = \left(\widetilde{m}_1 - \widetilde{m}_2 \right)^2 = \left(w^T m_1 - w^T m_2 \right)^2$$

$$= \left(w^T m_1 - w^T m_2 \right) \left(m_1^T w - m_2^T w \right)$$

$$= w^T \left(m_1 - m_2 \right) \left(m_1 - m_2 \right)^T w = w^T S_b w$$

$$\widetilde{S}_b^2 = \frac{1}{N_1 + N_2} \sum_{i=1}^{2} N_i \left(\widetilde{m}_i - \widetilde{m} \right) \left(\widetilde{m}_i - \widetilde{m} \right)^T$$

$$= \frac{1}{N_1 + N_2} \sum_{i=1}^{2} N_i w^T \left(m_i - m \right) \left(m_i - m \right)^T w = w^T S_b w$$

4.构造*Fisher*准则函数

高维$(d$维$)$样本投影到1维空间

$$\begin{cases}\textbf{类间样本尽量分开} \Rightarrow \textbf{投影后类间离散度} \uparrow \\ \textbf{类内样本尽量聚集} \Rightarrow \textbf{投影后总类内离散度} \downarrow\end{cases}$$

$$\Rightarrow \boxed{\begin{array}{l}\textit{Fisher}\text{准则函数} \quad J_F(w)=\dfrac{\left|\tilde{m}_1-\tilde{m}_2\right|^2}{\tilde{S}_w}=\dfrac{\left(\tilde{m}_1-\tilde{m}_2\right)^2}{\tilde{S}_w} \\ \textit{Fisher}\text{准则} \quad \max_w J_F(w) \\ w^*=\arg\max_w J_F(w) \quad \textbf{如何求解}?\end{array}}$$

$$\textit{Fisher}\text{线性分类器设计}\begin{cases}(1)\text{最佳投影空间的估计} \\ (2)\text{基于训练样本在最佳投影空间} \\ \quad\quad \text{投影设计最优边界。}\end{cases}$$

5.确定最优投影方向--求解 $w^* = \arg\max_{w} J_F(w)$

[1]确定 $J_F(w)$ 关于 w 的显式函数形式

$$J_F(w) = \frac{\left(\tilde{m}_1 - \tilde{m}_2\right)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$

$$\left(\tilde{m}_1 - \tilde{m}_2\right)^2 = \left(w^T m_1 - w^T m_2\right)^2$$

$$= w^T \left(m_1 - m_2\right)\left(m_1 - m_2\right)^T w = w^T S_B w$$

$$\tilde{S}_w = \tilde{S}_1^2 + \tilde{S}_2^2 = w^T S_1 w + w^T S_2 w = w^T \left(S_1 + S_2\right) w = w^T S_w w$$

所以 $\boxed{J_F(w) = \dfrac{w^T S_B w}{w^T S_w w}}$ ----**广义瑞利(*Rayleigh*)商**.

[2]估计 $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \boldsymbol{J}_F(\boldsymbol{w}) = \arg\max_{\boldsymbol{w}} \dfrac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}$

$\Rightarrow$

$$\begin{aligned} &\max_{\boldsymbol{w}} \quad \boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w} \\ &\boldsymbol{s.t.} \quad \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} = \boldsymbol{c} \neq 0 \end{aligned}$$

采用***Lagrange***条件极值法，引入辅助函数：

$$L(\boldsymbol{w}, \lambda) = \boldsymbol{J}_F(\boldsymbol{w}) - \lambda\left(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - \boldsymbol{c}\right)$$

$$= \boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w} - \lambda\left(\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - \boldsymbol{c}\right)$$

λ为标量***Lagrange***乘子

$$L(w,\lambda)=w^T S_B w-\lambda\left(w^T S_w w-c\right)$$

$$\frac{\partial L}{\partial w}=2\left(S_B w-\lambda S_w w\right)$$

$$\frac{\partial\left(x^T A x\right)}{\partial x}=\left(A+A^T\right)x$$

A是对称矩阵 $=2Ax$

令 $\dfrac{\partial L}{\partial w}=0$

得 $S_B w^*=\lambda S_w w^*$ $\xrightarrow{N>d,\ S_w\text{非奇异}}$ $\boxed{S_w^{-1}S_B w^*=\lambda w^*}$

相应地：$J(w^*)=\dfrac{w^{*T}S_B w^*}{w^{*T}S_w w^*}=\dfrac{\lambda w^{*T}S_w w^*}{w^{*T}S_w w^*}=\lambda$

思路1 要使$J(w^*)=\lambda$最大，

应使λ取$S_w^{-1}S_B$矩阵的最大本征值，

w^*为$S_w^{-1}S_B$矩阵相应本征列向量

[2]**确定** $w^* = \arg\max_w J_F(w)$

思路2：直接找到 w^*

类间散度矩阵 $S_B = (m_1 - m_2)(m_1 - m_2)^T$

$$S_B w^* = (m_1 - m_2)\underbrace{(m_1 - m_2)^T w^*}_{\text{标量}R} = (m_1 - m_2)R$$

所以 $S_B w^*$ 方向与向量 $(m_1 - m_2)$ 方向一致。

$$\boxed{\lambda w^* = S_w^{-1} S_B w^* = S_w^{-1}(m_1 - m_2)R}$$

$$\Rightarrow \quad \boxed{w^* = \frac{R}{\lambda} S_w^{-1}(m_1 - m_2)}$$

忽略标量 $\frac{R}{\lambda}$，得**最优投影方向** $\boxed{w^* = S_w^{-1}(m_1 - m_2)}$

[3] $w^* = \arg\max_w J_F(w)$ 计算流程

STEP1. 获取来自两类 ω_1 / ω_2 的训练样本集 $\mathcal{X}$

$$\begin{cases} \omega_1\text{类：}\ \mathcal{X}_1,\ N_1\text{个样本} \\ \omega_2\text{类：}\ \mathcal{X}_2,\ N_2\text{个样本} \end{cases}$$

STEP2. 计算各类样本均值 $\quad m_i I = \dfrac{1}{N_i}\sum_{x\in\mathcal{X}_i} x \quad i = 1,2$

STEP3. 类内散度矩阵

$$S_i = \frac{1}{N_i}\sum_{x\in\mathcal{X}_i}(x - m_i)(x - m_i)^T \qquad i = 1,2$$

STEP4. 总类内散布矩阵 S_w 及逆 S_w^{-1}.

$$S_w = \frac{1}{N_1 + N_2}\sum_{i=1}^{2} N_i S_i$$

为避免 S_w 的奇异，
引入一个小正数 ε
计算 $w^* = (S_w + \varepsilon I)^{-1}(m_1 - m_2)$

STEP5. 计算 w^* $\qquad w^* = S_w^{-1}(m_1 - m_2)$

利用上述过程，可直接确定投影直线 w^*.

6.确定最优分类超平面

$$\left(\boldsymbol{w}^*\right)^T \boldsymbol{x} + \omega_0 = 0$$

明确 $\begin{cases} \text{法向量} & \boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \boldsymbol{J}_F\left(\boldsymbol{w}\right) \\ \omega_0 = ? \end{cases}$

一维最佳投影空间的分类边界

基于最小错误率贝叶斯决策

第1，两类样本正态分布，且分布形状一致

--线性分类器

第2，若维数$\boldsymbol{d}$及样本数目$\boldsymbol{N}$足够大时，$y = \boldsymbol{w}^{*T}\ \boldsymbol{x}$近似正态分布，可在投影空间利用"**两步贝叶斯决策**"法。

A.确定阈值点(分类边界) y_0

先验概率相等 $y_0 = \dfrac{\widetilde{m}_1 + \widetilde{m}_2}{2} = \dfrac{(m_1 + m_2) S_w^{-1} (m_1 - m_2)}{2}$

先验概率不等

$$y_0 = \frac{\widetilde{m}_1 + \widetilde{m}_2}{2} + \log \frac{P(\omega_2)}{P(\omega_1)} = \frac{(m_1 + m_2) S_w^{-1} (m_1 - m_2)}{2} + \log \frac{P(\omega_2)}{P(\omega_1)}$$

投影后样本数据中心

$$y_0 = \frac{\widetilde{m}_1 N_1 + \widetilde{m}_2 N_2}{N_1 + N_2} = \frac{(m_1 N_1 + m_1 N_2) S_w^{-1} (m_1 - m_2)}{N_1 + N_2}$$

B.决策

对于任意观测$\boldsymbol{x}$

$$\begin{cases} \text{投影空间：} \quad y = \boldsymbol{w}^{*T}\,\boldsymbol{x} \\ \text{判决：} \quad \text{若} y > y_0, \text{则} \boldsymbol{x} \text{为} \omega_1 \text{类；} \\ \qquad\quad \text{若} y < y_0, \text{则} \boldsymbol{x} \text{为} \omega_2 \text{类。} \end{cases}$$

分类边界满足：$y - y_0 = 0$，即 $\boldsymbol{w}^{*T}\,\boldsymbol{x} - y_0 = 0$

原始特征空间

判别函数：$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{w}^{*T}\,\boldsymbol{x} + \omega_0 = \boldsymbol{w}^{*T}\,\boldsymbol{x} - y_0$

分类边界：$\boldsymbol{w}^{*T}\,\boldsymbol{x} - y_0 = 0$

决策规则：$\begin{cases} \text{若} \boldsymbol{g}(\boldsymbol{x}) > 0, \text{则} \boldsymbol{x} \text{为} \omega_1 \text{类} \\ \text{若} \boldsymbol{g}(\boldsymbol{x}) < 0, \text{则} \boldsymbol{x} \text{为} \omega_2 \text{类} \end{cases}$

主要内容

第一部分 类别数目C=2时 FLDA分类模型

关键:确定最佳投影直线

第二部分 类别数目C>2时FLDA分类模型

关键:确定最佳投影矩阵

注意： 线性投影子空间的维数最高为C-1

1 问题描述

已知：d维训练样本$\mathcal{X}=\{x_1,\ldots,x_N\}=\mathcal{X}_1\cup\mathcal{X}_2\cup\cdots\cup\mathcal{X}_C$

其中 ω_i类， $\mathcal{X}_i$， N_i个

$$N=\sum_{i=1}^{C}N_i$$

$$\xrightarrow{\text{对}x\text{投影：}y=W^Tx}\mathcal{Y}=\{y_1,\ldots,y_N\}\quad\begin{cases}\omega_1\text{类}\ \mathcal{Y}_1,\ N_1\\ \vdots\\ \omega_C\text{类}\ \mathcal{Y}_C,\ N_C\end{cases}$$

其中 $y_j=W^Tx_j, i=1,2,...,N$

求解：最佳投影子空间W^*，使C类分类效果最好。

2.准备工作

[1]投影前——d维特征空间

各类均值向量 $$m_i = \frac{1}{N_i}\sum_{x\in\mathcal{X}_i} x \qquad i = 1,\cdots,C$$

总体均值向量 $$m = \sum_{i=1}^{C}\frac{N_i}{N}m_i$$

类内散度矩阵 $(within-class\ \ scatter\ \ matrix)$

$$S_i = \frac{1}{N_i}\sum_{x\in\mathcal{X}_i}(x-m_i)(x-m_i)^T \qquad i = 1,\cdots,C$$

总类内散度矩阵 $(pooled\ \ within-class\ \ scatter\ \ matrix)$

$$S_w = \sum_{i=1}^{C}\frac{N_i}{N}S_i$$

S_w对称半正定；若$N > d$, S_w通常非奇异

类间散度矩阵 $(between-class\ \ scatter)$ $$S_b = \sum_{i=1}^{C}\frac{N_i}{N}(m_i-m)(m_i-m)^T$$

[2]投影后 – 低维空间 $\left(\text{维数} \le C-1\right)$

总类内散度矩阵 $\widetilde{S}_w = W^T S_w W$

$$\widetilde{S}_w = W^T S_w W = W^T \left(\sum_{i=1}^{C} \frac{N_i}{N} S_i \right) W$$

类间散度矩阵 $\widetilde{S}_b = W^T S_b W$

$$\widetilde{S}_b = W^T S_b W = W^T \left[\sum_{i=1}^{C} \frac{N_i}{N} \left(m_i - m \right) \left(m_i - m \right)^T \right] W$$

3.*Fisher*准则函数

*Fisher*准则函数 $J_F(W)=\dfrac{\det(\widetilde{S}_b)}{\det(\widetilde{S}_w)}=\dfrac{\det(W^T S_b W)}{\det(W^T S_w W)}$

或者 $J_F(W)=\dfrac{tr(\widetilde{S}_b)}{tr(\widetilde{S}_w)}=\dfrac{tr(W^T S_b W)}{tr(W^T S_w W)}$

类间远离

类内紧致

或者 $J_F(W)=\dfrac{\prod\limits_{diag}\widetilde{S}_b}{\prod\limits_{diag}\widetilde{S}_w}=\dfrac{\prod\limits_{diag}W^T S_b W}{\prod\limits_{diag}W^T S_w W}=\dfrac{\prod\limits_{j}w_j^T S_b w_j}{\prod\limits_{j}w_j^T S_w w_j}=\prod\limits_{j}\dfrac{w_j^T S_b w_j}{w_j^T S_w w_j}$

$\prod\limits_{diag}A$ —矩阵A的主对角线元素乘积

S_b的秩最大值为$C-1$

*Fisher*准则 $\max\limits_{W\in R^{d\times(C-1)}} J_F(W)$

最优投影子空间 $W^*=\arg\max\limits_{W\in R^{d\times(C-1)}} J_F(W)$

并且 $W^*=[w_1\cdots w_{C-1}]_{d\times(C-1)}$

4.确定最优投影子空间--求解 $W^* = \arg\max_W J_F(W)$

例： $J_F(W) = \dfrac{tr(\widetilde{S}_b)}{tr(\widetilde{S}_w)} = \dfrac{tr(W^T S_b W)}{tr(W^T S_w W)}$

$$\max_W J_F(W) \quad \Rightarrow \quad \begin{cases} \max\limits_W & tr(W^T S_b W) \\ s.t. & tr(W^T S_w W) = 1 \neq 0 \end{cases}$$

$$\Rightarrow \begin{cases} L(W,\lambda) = tr(W^T S_b W) - \lambda\left(tr(W^T S_w W) - 1\right) \\ \max\limits_{W,\lambda} \; L(w,\lambda) \end{cases}$$

$$L(W,\lambda) = tr(W^T S_b W) - \lambda\left(tr(W^T S_w W) - 1\right)$$

$$\frac{\partial L(W,\lambda)}{\partial W} = 2\left(S_b W - \lambda S_w W\right) = 0$$

$$\frac{\partial \mathrm{tr}(X^{\mathrm{T}} A X)}{\partial X} = \left[X^{\mathrm{T}}(A^{\mathrm{T}} + A)\right]^{\mathrm{T}} = (A + A^{\mathrm{T}})X$$

$$S_b W = \lambda S_w W = S_w W \Lambda \qquad S_b w_i = \lambda S_w w_i \qquad i = 1, \ldots, C-1$$

$$S_b W = \lambda S_w W$$

$$S_b w_i = \lambda S_w w_i \qquad i = 1, \ldots, C-1$$

$$S_w^{-1} S_b w_i = \lambda w_i$$

为避免S_w的奇异，
引入一个小正数ε
计算$(S_w + \varepsilon I)^{-1} S_b$

若λ为矩阵$S_w^{-1} S_b$的某个本征值，则w_i就是相应的本征列向量.

若有$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{C-1} > 0$

相应$w_1^*, w_2^*, \ldots, w_{C-1}^*$构成最佳投影矩阵

$$W^* = \left[w_1^* \quad w_2^* \quad \cdots \quad w_{C-1}^* \right]$$

这就是$C-1$维投影子空间.

任意观测样本$x \in \Re$，特征提取结果：$y = W^{*T} x$

主要内容

A.引言

B. Fisher线性判别分析

C.感知器

D. 线性支持向量机(线性SVM)

B. 感知器分类模型

要求：

- 理解感知器准则函数的构造思想
- 掌握感知器准则函数的形式、意义、
- 感知器分类模型的学习
- 基于感知器的类别决策
- 感知器分类模型学习的前提条件

关键词：

- 解向量、解空间
- 训练样本的增广、规范化
- 权向量的增广
- 线性可分的训练样本
- 梯度下降法(固定步长/变步长；单样本/批量样本)

1. 基本思想

针对两类别线性可分问题，引入解向量；

基于错分的训练样本，构造感知器准则函数；

采用梯度下降法，估计解向量，得线性分类模型。

2.几个基本概念

(1)增广的特征向量 / 权向量

线性判别函数 $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + \omega_0 = \boldsymbol{w}^T\boldsymbol{x} + 1\omega_0$

定义

增广的特征向量 $\boldsymbol{y} = [1 \quad x_1 \quad \cdots \quad x_d]^T = \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix}$

增广的权向量 $\boldsymbol{a} = [\omega_0 \quad w_1 \quad \cdots \quad w_d]^T = \begin{bmatrix} \omega_0 \\ \boldsymbol{w} \end{bmatrix}$

⟶

线性判别函数 $g(\boldsymbol{y}) = \boldsymbol{a}^T\boldsymbol{y}$

决策规则 若$g(\boldsymbol{y}) > 0$,则决策$\boldsymbol{y} \in \omega_1$

若$g(\boldsymbol{y}) < 0$,则决策$\boldsymbol{y} \in \omega_2$

(2)训练样本集的线性可分性

考虑 $\begin{cases} \text{两类：}\ \omega_1,\ \omega_2 \\ \text{增广样本集}\mathcal{Y}=\{y_1,\cdots,y_N\} \end{cases}$

若至少存在一个线性分类器$g(y)=a^T y$，能使所有训练样本正确分类，即

$$\exists a,\begin{cases} \text{若}y_i\in\omega_1\text{类，则有}a^T y_i>0 \\ \text{若}y_i\in\omega_2\text{类，则有}a^T y_i<0 \end{cases} \qquad i=1,2,...,N$$

则上述训练样本是**线性可分**的；否则，**线性不可分**。

若样本集$\mathcal{Y}$**线性可分**，则必存在向量a，对各训练样本正确分类。

(3)规范化增广样本向量

对于**线性可分的增广样本集**$\mathcal{Y}=\{y_1,\ldots,y_N\}$，

定义**规范化增广样本向量**y'，满足

$$y_i'=\begin{cases} y_i & \text{若} y_i\in\omega_1\text{类} \\ -y_i & \text{若} y_i\in\omega_2\text{类} \end{cases} \qquad i=1,\ldots,N$$

则必存在权向量a，使得

$$a^T y_i'>0, \quad i=1,\ldots,N$$

目的在于：模型学习过程中，有助于区分正确分类的训练样本与错分的训练样本

可记**规范化增广样本向量**y'为y

(4)解向量和解区

对于线性可分的训练样本集，设其**规范化增广样本集**为$\mathcal{Y}=\{y_1,\cdots,y_N\}$，若存在权向量$a$满足

$$a^T y_i>0,\quad i=1,\ldots,N$$

则称权向量a为1个**解向量**，记为a^*.

权值空间中，所有解向量组成的区域称为**解区**。

(4)解向量和解区--续

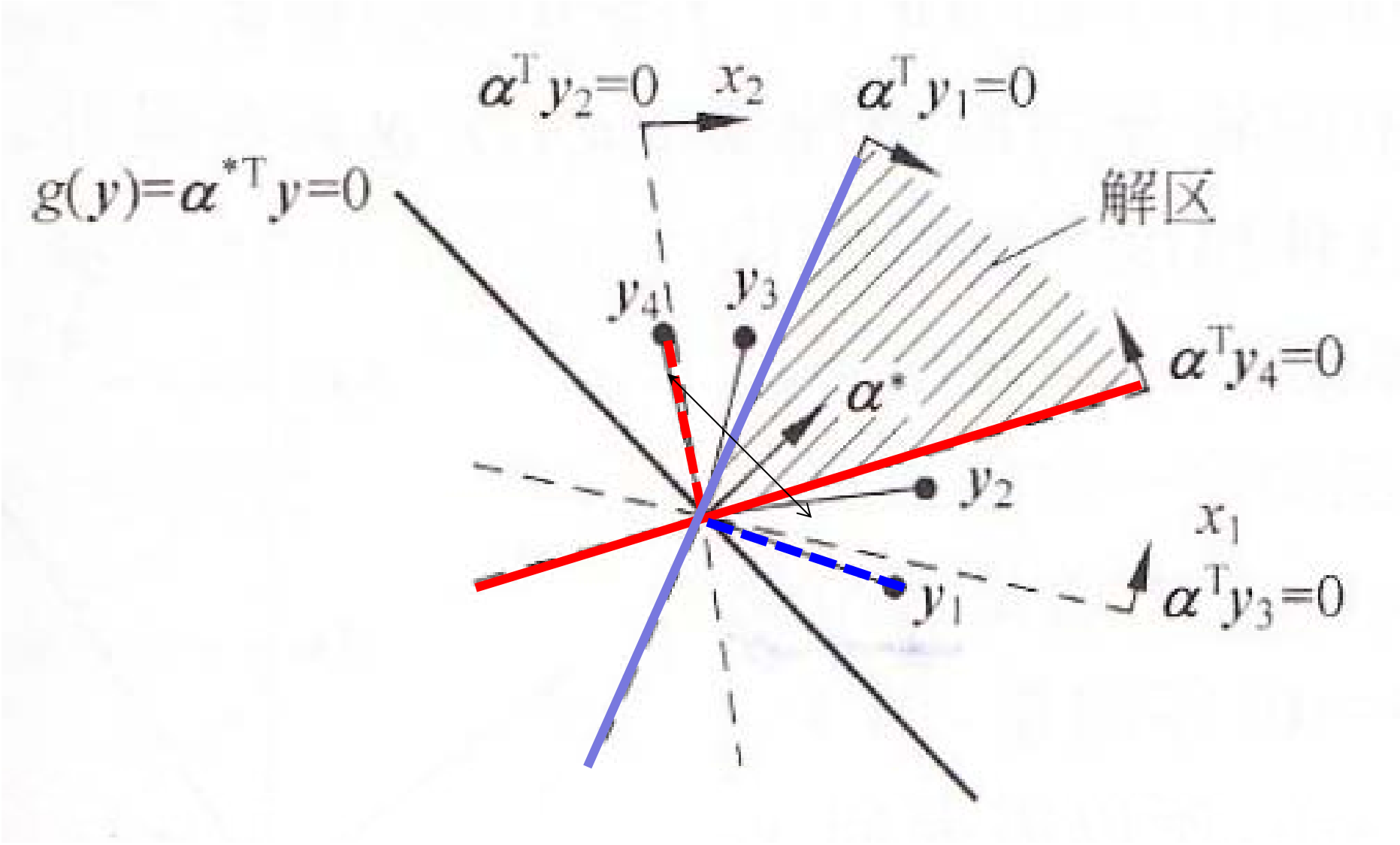
对于每个样本y_i

$$\begin{cases} a^T y_i = 0 \text{是在}\textbf{权值空间}\text{过原点的超平面}\widehat{H}_i, \\ \text{超平面}\widehat{H}_i\text{的法向量是}y_i; \\ \textbf{解向量}a^* \text{在超平面正侧}\left(\textbf{权值空间}\text{正侧}\ \ a^T y_i > 0\right) \end{cases}$$

N个样本产生N个超平面

解区是**权值空间**N个超平面**正侧区域的交集**。

规范化增广样本集、解向量和解区



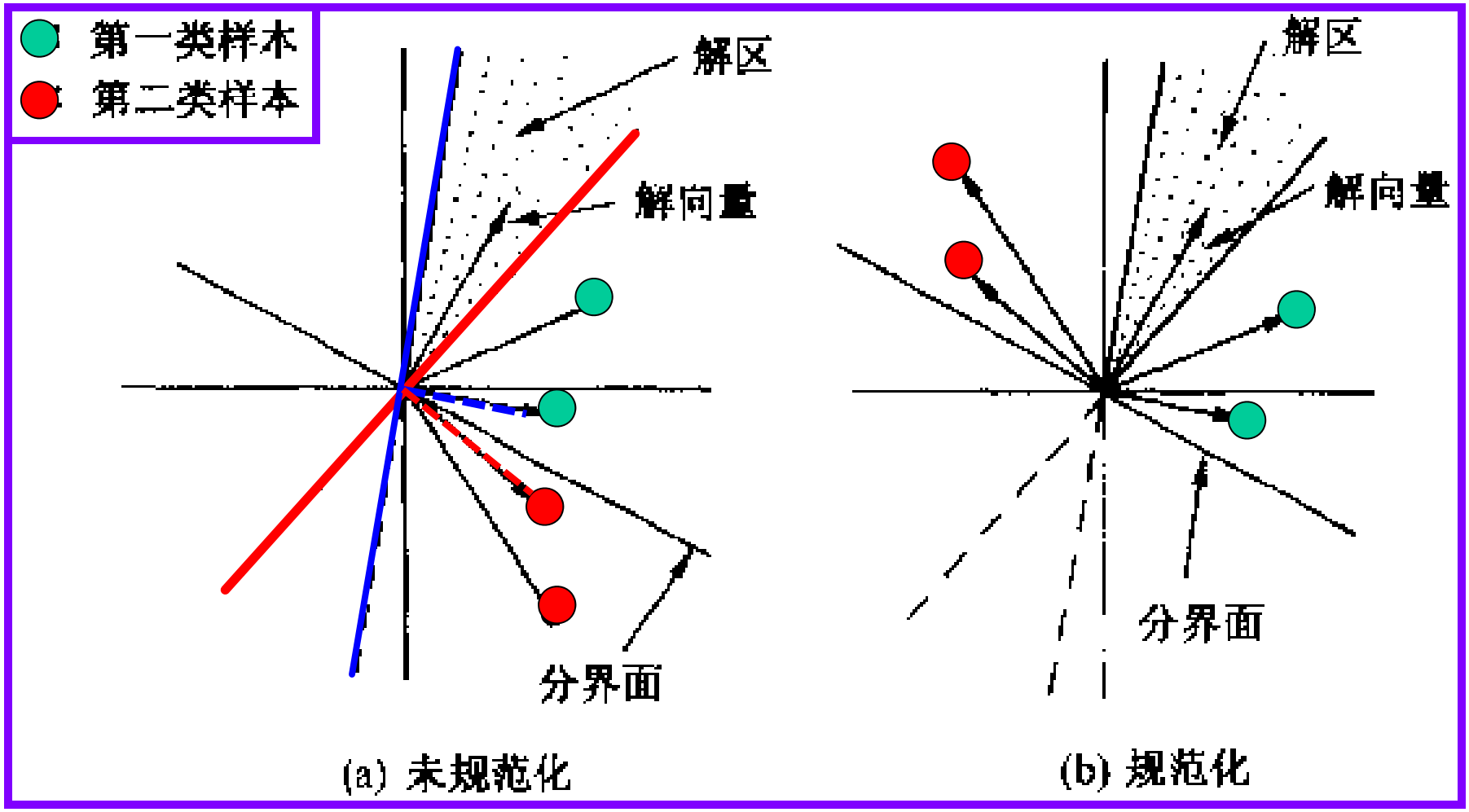


图　二维空间、两类情况下，增广训练样本规范化前后，训练样本及其特征空间(与权空间维数一致)的解区域及解向量示意图。

(5) 对解区的限制 / 余量

若解向量存在，通常不是惟一的。

解区所有的解向量都满足 $a^T y_i > 0.$

通常，解向量越靠近解区中间

→ 分类面离各类样本越远

→ 它到各样本的最近距离越远

→ 对新样本错分的可能性越小

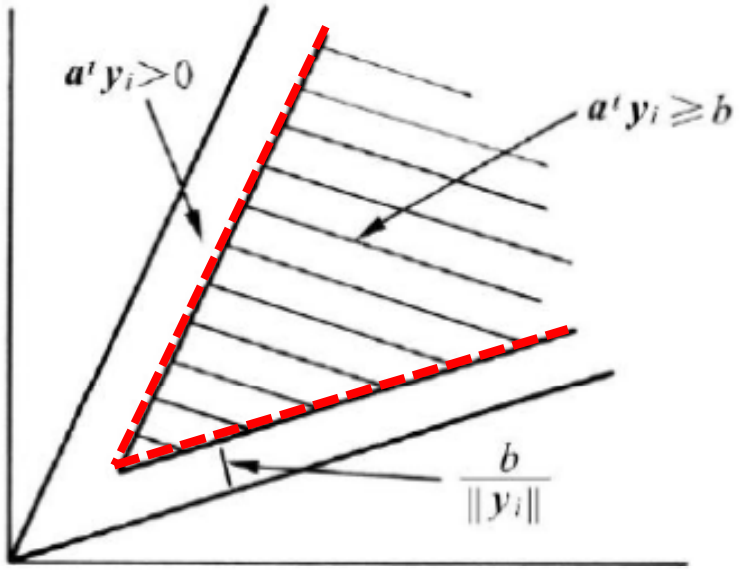
⇒ 为使解向量更可靠，需要**限制解区**.

收缩解区，只考虑**靠近解区中间**的解向量。

引入"**余量**$(margin)$"$b>0$,在解区中寻找解向量$\boldsymbol{a}^*$,

满足$\boldsymbol{a}^T\boldsymbol{y}_i > b > 0, \quad i = 1,\ldots,N$

好处
- 解向量更可靠
- 推广性更强
- 防止算法收敛至解区边界



引入余量的解区

3.感知器准则函数的构建 / 解向量求解

(1)构建感知器准则函数$J_p(a)$

给定**规范化增广样本集** $\mathcal{Y}=\{y_1,\cdots,y_N\}$

若权向量a使得训练样本y_k被错分，则有 $a^T y_k<0$

对于**所有错分样本**，定义**感知器准则函数**

$$J_p(a)=\sum_{a^T y_k\le 0}\left(-a^T y_k\right)$$

函数值越大，训练集的错分程度越高

当且仅当$J_p(a^*)=\min J_p(a)=0$，错分样本数目为0，

a^*为解向量.

线性分类模型的学习目的，就是如何利用规范增广的训练集，估计这个解向量

(2)确定解向量 $a^* = \arg\min_a J_p(a) = \sum_{a^T y_k \le 0} \left(-a^T y_k\right)$

采用梯度下降法 $(gradient\ decent\ method)$

梯度 $\nabla J_p(a) = \dfrac{\partial J_p(a)}{\partial a} = \sum_{a^T y_k \le 0} \left(-y_k\right)$

迭代形式 $a(t+1) = a(t) - \rho_t \cdot \nabla J_p\left(a(t)\right)$

$$= a(t) + \rho_t \cdot \sum_{a^T(t)\cdot y_k \le 0} y_k$$

$a(t+1)$处准则函数取值：

$$J_p\left(a(t+1)\right) = \sum_{a^T(t+1)\cdot y_k \le 0} \left[-a^T(t+1)\cdot y_k\right]$$

由$a(0)$开始得权向量序列：$a(1), a(2), \cdots, a(t), a(t+1), \ldots$

方法1：固定增量单样本法

$\left(Fixed-Increment\ \ Single\ \ Sample\ \ Method\right)$

将样本集$\{y_1,\ldots,y_N\}$视为一个不断重复出现的样本序列, 逐样本考察。

学习率$\left(\text{决定步长}\right)$固定为常数　　$\rho_t=\rho\left(=1\right)>0$

$$a\left(t+1\right)=a\left(t\right)-\rho\cdot\nabla J_p\left(a\left(t\right)\right)=a\left(t\right)+y_k$$

STEP1. 任意选择初始权向量$a\left(0\right),\ t\leftarrow 0;\ \ \rho\left(=1\right)>0$

STEP2. 顺次考察各样本.

若$\left[a\left(t\right)\right]^T y_j\le 0,$则修正$a\left(t\right)$为 $a\left(t+1\right)=a\left(t\right)+y_j$

STEP3. 重复***STEP2***, 直到$J_p\left(a\right)=0,$得解向量a^*

$$\left[\boldsymbol{a}\left(t+1\right)\right]^T \mathbf{y}_k = \left[\boldsymbol{a}\left(t\right)+\mathbf{y}_k\right]^T \mathbf{y}_k$$

$$= \left[\boldsymbol{a}\left(t\right)\right]^T \mathbf{y}_k + \left\|\mathbf{y}_k\right\|^2 \geq \left[\boldsymbol{a}\left(t\right)\right]^T \mathbf{y}_k$$

只要解区存在，逐渐修正权向量，不断接近超平面正侧

对于线性可分样本集,经有限步迭代，总可找到解向量a^*。

方法2-带余量的可变增量单样本修正算法

考虑余量$b>0$ $\begin{cases}\text{严格错分条件} & \left[a(t)\right]^T y_k \le b \\ \text{学习率} & \rho_t>0 \ \text{可变}\end{cases}$（如何自适应调整$\rho_t$?）

对于$a(t)$, 若$\left[a(t)\right]^T y_k \le b$, 则修正$a(t)$

得 $a(t+1)=a(t)+\rho_t y_k$

要使 $\left[a(t+1)\right]^T y_k > b > 0$

则 $\left[a(t+1)\right]^T y_k = \left[a(t)+\rho_t y_k\right]^T y_k$

$$=\left[a(t)\right]^T y_k + \rho_t \cdot \|y_k\|^2 > b > 0$$

$\Rightarrow$ **可变步长** $\rho_t > \dfrac{b-\left[a(t)\right]^T y_k}{\|y_k\|^2}$

思考：若余量为0，如何自适应调整ρ_t?

4.基于感知器的样本分类实现步骤

判别函数
$$\begin{cases} g(x; w, \omega_0) = w^T x + \omega_0 = w^T x + 1\omega_0 \\ g(y; a) = a^T y = \begin{bmatrix} \omega_0 & w^T \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} \end{cases}$$

若训练集不是线性可分，将不能找到解向量

STEP1.原始特征空间训练样本集的**增广、规范化**.

STEP2.分类模型的学习。

利用**增广、规范化**训练集估计**解向量**$a^* = \begin{bmatrix} \omega_0^* \\ w^* \end{bmatrix}$

得w^*，ω_0^*，判别函数为：$g(x) = w^{*T} x + \omega_0^*$

STEP3.分类模型的使用。

对于原始特征空间的任意观测样本x，决策规则

$$\begin{cases} \text{若} g(x) = w^{*T} x + \omega_0^* > 0, \text{则判断} x \text{为第1类} \\ \text{若} g(x) = w^{*T} x + \omega_0^* < 0, \text{则判断} x \text{为第2类} \end{cases}$$