机器学习
Machine learning

第三章 线性分类
Linear Classifier

授课人：周晓飞
zhouxiaofei@iie.ac.cn
2020-10-8

第三章 线性分类

3.1 概述

3.2 基础知识

3.3 感知机

3.4 线性鉴别分析

3.5 logistic 模型

第三章 线性分类

3.1 概述

3.2 基础知识

3.3 感知机

3.4 线性鉴别分析

3.5 logistic 模型

概述

本章讨论在向量空间中根据数据的特征向量和类别标签构造线性分类模型。

首先,介绍关于向量空间、超平面和线性决策函数的基础知识

然后，重点介绍感知机、Fisher 线性鉴别和 logistic 模型

第三章 线性分类

3.1 概述

3.2 基础知识

3.3 感知机

3.4 线性鉴别分析

3.5 logistic 模型

基础知识

向量空间

1. n维向量（Vector）

n个数组成的有序数组，称为一个n维向量，

$\mathbf{a}=(a_1, a_2, a_3, \ldots, a_n)$

其中，a_i称为第i个分量.

基础知识

向量空间

1. n维向量（Vector）

- 例子 1. 数据库表

$\mathbf{R}_1=(a_{11}, a_{12}, a_{13}, \ldots, a_{1n})$

	Item1	Item2	Item3	…	Item n
Record 1	a_{11}	a_{12}	a_{13}	…	a_{1n}
Record 2					
…					
Record k					

向量空间

- 例子 2. 人的生物特征

身高，体重，血型，口音，生日，…

a =(1.6, 110, 1, 12, 1978,…)

b =(1.78, 120, 2, 20, 1988,…)

c =(1.65, 115, 2, 31, 1980,…)

…

基础知识

向量空间

- 例子 3. 学生成绩

高数，英语，政治，专业 1，专业 2

S_1= (59，60，68，82，75)

S_2= (50，55，70，86，66)

S_3= (60，52，60，75，80)

…

基础知识

向量空间

- 例子 4. 图片

像素灰度矩阵按行（或列）展开：

Image=（30,32,34,29,……）

基础知识

向量空间

- 例子 5. 文档

> **长跑**，即长距离跑步，英文是long-distance running。路程通常在5000米以上。最初项目为4英里、6英里跑，从19世纪中叶开始，逐渐被5000米跑和10000米跑替代。田径比赛的长跑项目通常分为5000米跑、10000米跑、半程马拉松（约21100米）、马拉松（约42.195千米）等。男子项目1912年列入；女子5000米跑1996年列入，10000米跑1988年列入。据记载，现代最早的正式长跑比赛是1847年4月5日在英国伦敦举行的职业比赛，英国的杰克逊以32分35秒0的成绩夺得6英里跑冠军。

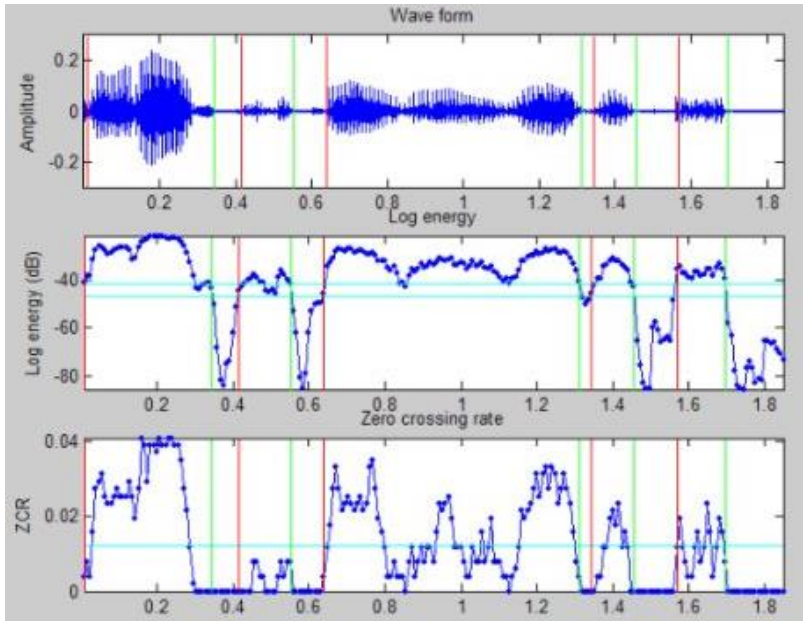
词袋模型，词频特征作为属性：

Text=（0.001, 0, 0.002, 0.005, …..）

基础知识

向量空间

- 例子 6. 音频



向量空间

2. 分类问题转化为向量的分类：

两类别： 年轻人，老年人

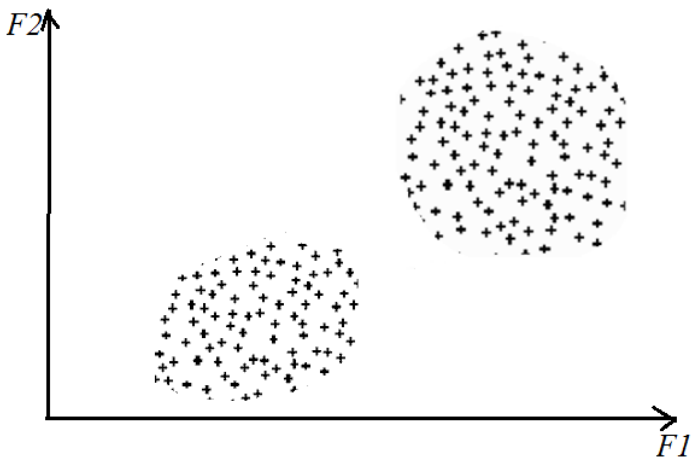
特征： (F1：黑头发的比例) 黑/黑+白；值域（0，1）

(F2：行走速度) 米/每分钟；值域（0，100）

数据：

(1，99)、(0.9，80)、(0.95，100) …

(0.2，30)、(0.5，50)、(0.4，30) …



向量空间

3. 向量空间定义：

所有分量为实数的 n 维向量构成的集合，

$$R^n = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \middle| a_i \in R, i = 1,2,\cdots,n \right\}$$

称为一个 n 维向量空间，向量空间又称线性空间。

基础知识

向量空间

线性代数中的表述：

V 是数域 P 上 n 元向量的一个非空集合，若对 V 中向量的加法和数乘仍在 V 中（运算封闭），

$$if \quad \alpha \in V, \beta \in V \Rightarrow \alpha + \beta \in V;$$

$$if \quad \alpha \in V, \lambda \in R \Rightarrow \lambda\alpha \in V.$$

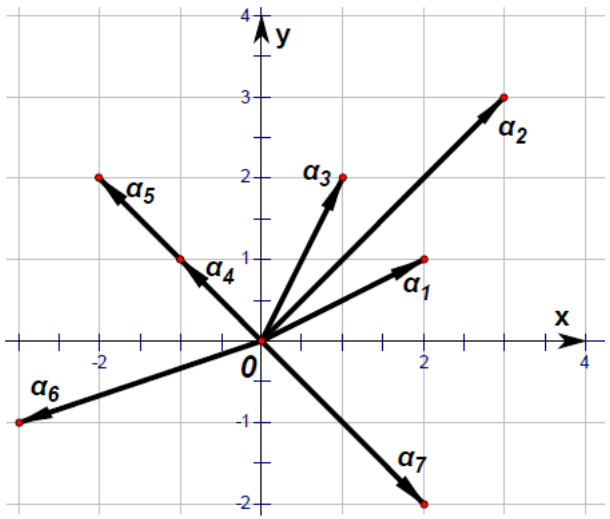
且满足运算规律，则 V 为数域 P 上的一个向量空间。（P 上 n 元向量的全体，称为 n 元向量空间。）

常说的几何意义，一般指学习问题在向量空间中的原理。

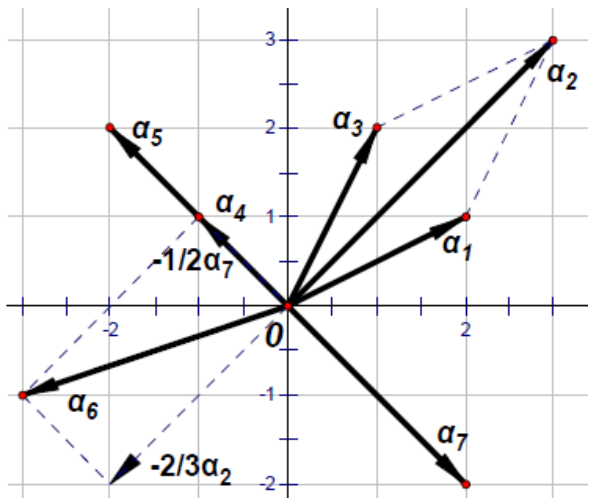
向量空间

4. 向量空间几何直观：

$\alpha_1=(2,1)$，$\alpha_2=(3,3)$，$\alpha_3=(1,2)$，$\alpha_4=(-1,1)$，$\alpha_5=(-2,2)$，$\alpha_6=(-3,-1)$，$\alpha_7=(2,-2)$。



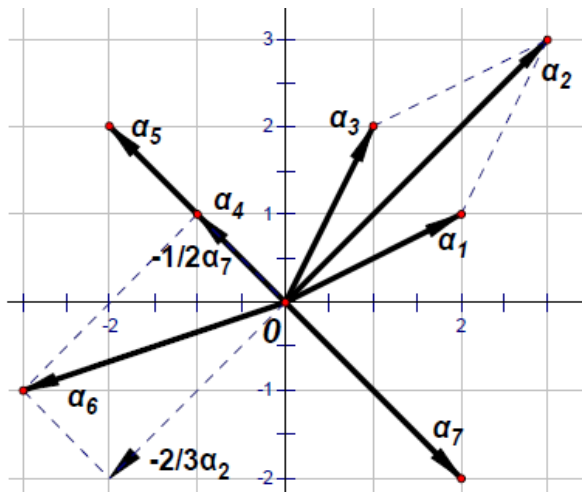
满足运算封闭：



向量空间

5. 向量的线性相关性

$$k_1\alpha_1 + k_2\alpha_2 + \cdots + k_r\alpha_r = 0$$



$$\alpha_6 = (-1/2)\alpha_7 + (-2/3)\alpha_2$$

向量空间

6. 向量的运算

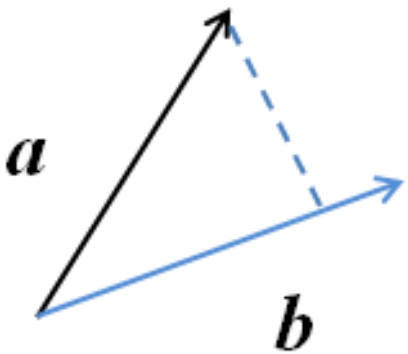
- 遵循矩阵运算
- 内积运算：

$$\boldsymbol{a}=(a_1,a_2,\cdots,a_n)^T$$

$$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^T$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle=\boldsymbol{a}^T\boldsymbol{b}=\boldsymbol{b}^T\boldsymbol{a}=|\boldsymbol{a}||\boldsymbol{b}|\cos\varphi$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle/|\boldsymbol{b}|=|\boldsymbol{a}|\cos\varphi$$



向量空间

6. 向量的运算

- 遵循矩阵运算
- 内积运算：

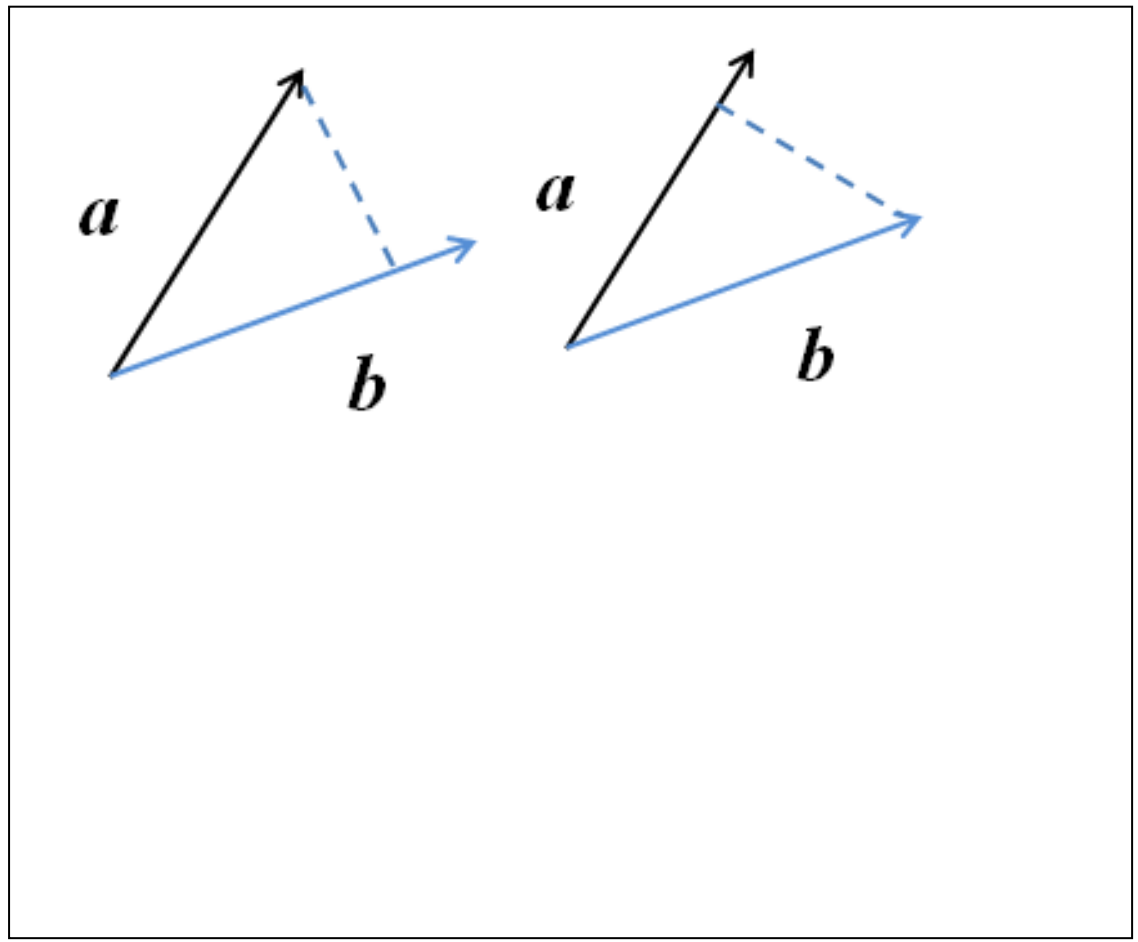
$$\boldsymbol{a}=(a_1,a_2,\cdots,a_n)^T$$

$$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^T$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle=\boldsymbol{a}^T\boldsymbol{b}=\boldsymbol{b}^T\boldsymbol{a}=|\boldsymbol{a}||\boldsymbol{b}|\cos\varphi$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle/|\boldsymbol{b}|=|\boldsymbol{a}|\cos\varphi$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle/|\boldsymbol{a}|=|\boldsymbol{b}|\cos\varphi$$



向量空间

6. 向量的运算

- 遵循矩阵运算
- 内积运算：

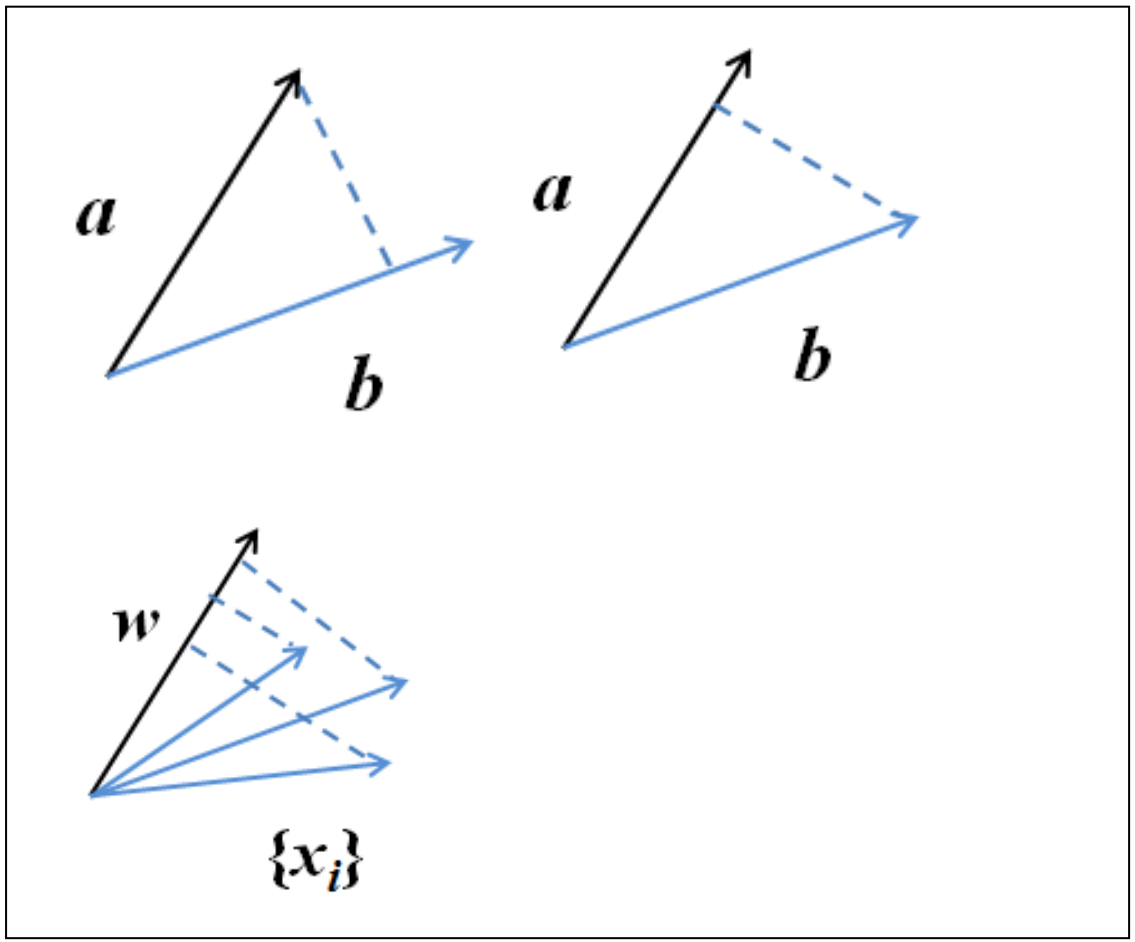
$$\boldsymbol{a}=(a_1,a_2,\cdots,a_n)^T$$

$$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^T$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle=\boldsymbol{a}^T\boldsymbol{b}=\boldsymbol{b}^T\boldsymbol{a}=|\boldsymbol{a}||\boldsymbol{b}|\cos\varphi$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle/|\boldsymbol{b}|=|\boldsymbol{a}|\cos\varphi$$

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle/|\boldsymbol{a}|=|\boldsymbol{b}|\cos\varphi$$



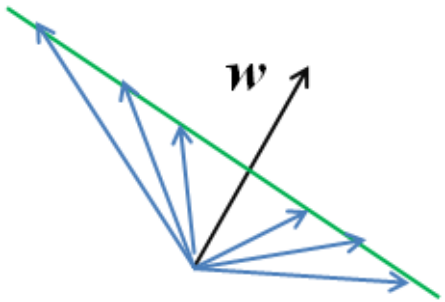
基础知识

超平面

1. 超平面（H）

超平面表达式

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



- 实例（样本）：$\boldsymbol{x}=(x_1,x_2,...,x_l)^T$
- 权向量（超平面的法向量）：$\boldsymbol{w}=(w_1,w_2,...,w_l)^T$
- 偏移量：w_0

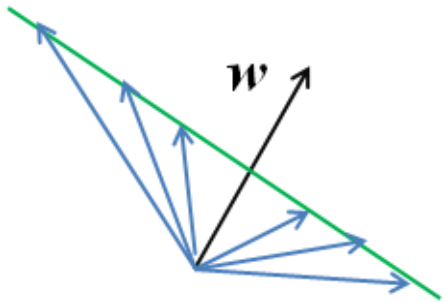
基础知识

超平面

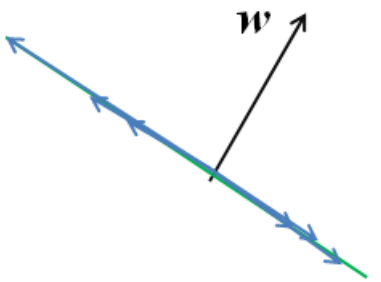
1. 超平面（H）

超平面表达式

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



- 实例（样本）：$\boldsymbol{x}=(x_1,x_2,...,x_l)^T$
- 权向量（超平面的法向量）：$\boldsymbol{w}=(w_1,w_2,...,w_l)^T$
- 偏移量：w_0

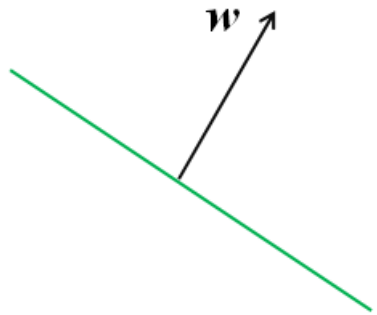


超平面

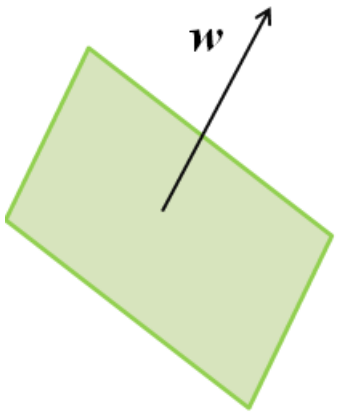
1. 超平面（H）

超平面表达式

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



- 实例（样本）：$\boldsymbol{x}=(x_1,x_2,...,x_l)^T$
- 权向量（超平面的法向量）：$\boldsymbol{w}=(w_1,w_2,...,w_l)^T$
- 偏移量：w_0



超平面

2. 线性函数的几何意义？

二维情况举例

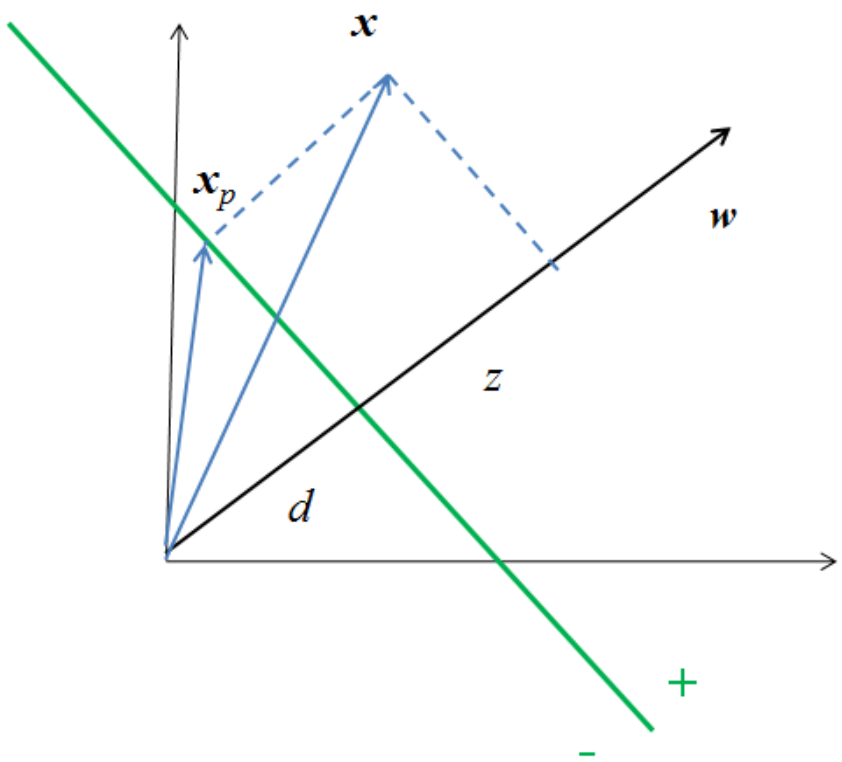
假设 x_p 是 x 在超平面 $w^T x + w_0 = 0$ 上的投影点

$$x = x_p + (x - x_p)$$

$$w^T x = w^T x_p + w^T (x - x_p)$$

$$\frac{w^T x}{\|w\|} = \frac{w^T x_p}{\|w\|} + \frac{w^T (x - x_p)}{\|w\|}$$

，



(x 向 w 投影) = (x_p 向 w 投影) + ($x - x_p$ 向 w 投影)

基础知识

超平面

解问题

$$d=\frac{\boldsymbol{w}^T\boldsymbol{x}_p}{\|\boldsymbol{w}\|};\ z=\frac{\boldsymbol{w}^T(\boldsymbol{x}-\boldsymbol{x}_p)}{\|\boldsymbol{w}\|};$$

$$\boldsymbol{w}^T\boldsymbol{x}_p+w_0=0;$$

可以得到

$$d=\frac{-w_0}{\|\boldsymbol{w}\|},\quad z=\frac{\boldsymbol{w}^T\boldsymbol{x}+w_0}{\|\boldsymbol{w}\|}=\frac{g(\boldsymbol{x})}{\|\boldsymbol{w}\|}\quad\Rightarrow\quad w_0=-d\|\boldsymbol{w}\|,\quad g(\boldsymbol{x})=z\|\boldsymbol{w}\|$$

思考：　当 $\boldsymbol{x}-\boldsymbol{x}_p$ 与 $\boldsymbol{w}$ 方向一致与否情况.

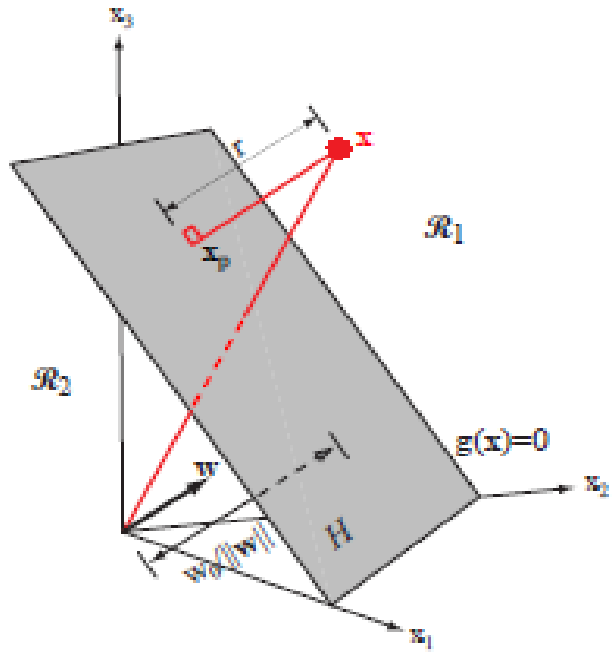
基础知识

超平面

若考虑尺度关系：

$$|z| = \frac{|g(\boldsymbol{x})|}{\|\boldsymbol{w}\|} = \frac{|g(\boldsymbol{x})|}{\sqrt{w_1^2 + w_2^2}}$$

$$|d| = \frac{|w_0|}{\|\boldsymbol{w}\|} = \frac{|w_0|}{\sqrt{w_1^2 + w_2^2}}$$



线性函数刻画了样本到超平面的距离

$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0, \quad g(x) = \|\boldsymbol{w}\| \cdot z$$

Have a break !

基础知识

相似性测度

1. 向量相似性度量

- Minkovski Metric 闵氏距离（p-范数）

$$D(\boldsymbol{x},\boldsymbol{y})=[\sum_i |x_i-y_i|^p]^{1/p}$$

- 欧氏距离（p=2）（2-范数）

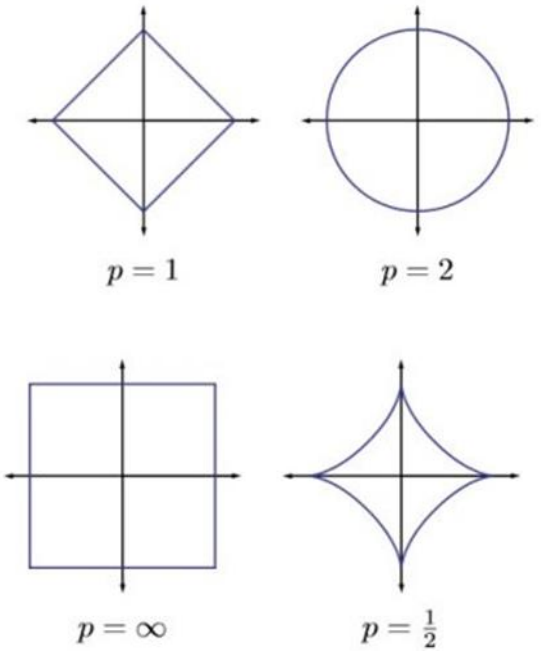
$$D(\boldsymbol{x},\boldsymbol{y})=[(\boldsymbol{x}-\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})]^{1/2}$$

- 城市块(p=1)、曼哈顿距离（1-范数）

$$D(\boldsymbol{x},\boldsymbol{y})=\sum_i |x_i-y_i|$$

- Chobychev 距离(p=inf)

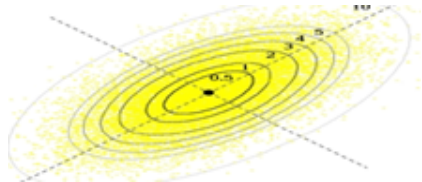
$$D(\boldsymbol{x},\boldsymbol{y})=\max_i |x_i-y_i|$$



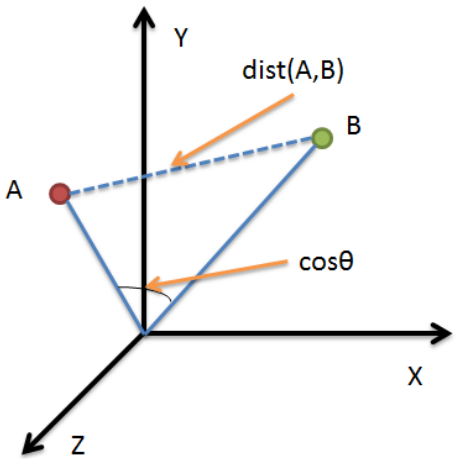
基础知识

相似性测度

- 平方距离\马氏距离 $D(x,y)=(x-y)^T Q(x-y)$



- 余弦相似性 $\cos\varphi = a^T b / |a||b|$



相似性测度

例子 1：

计算 a2, a3 之间的各种相似性

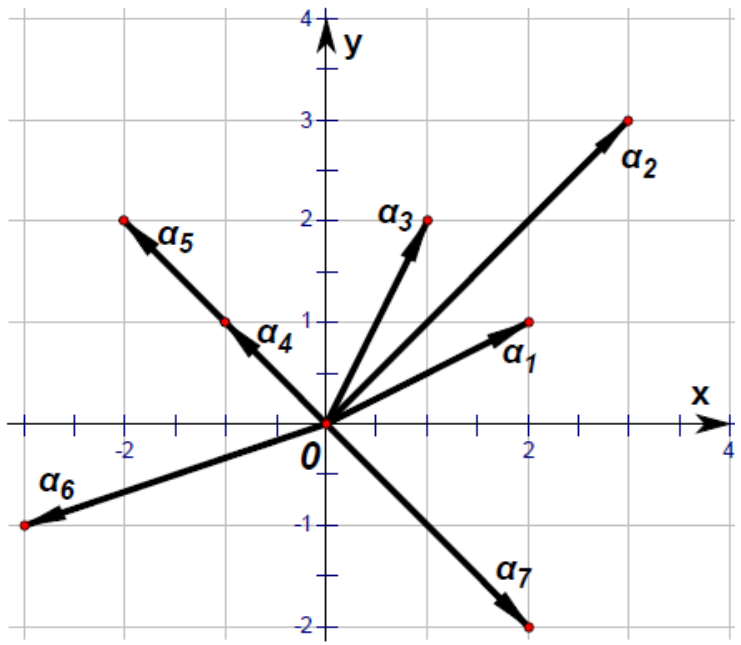
a2=（3，3）

a3=（1，2）

欧氏距离

$$d(a2,a3)=\sqrt[2]{(3-1)^2+(3-2)^2}$$

$$=\sqrt{5}=2.236$$



基础知识

相似性测度

城市距离

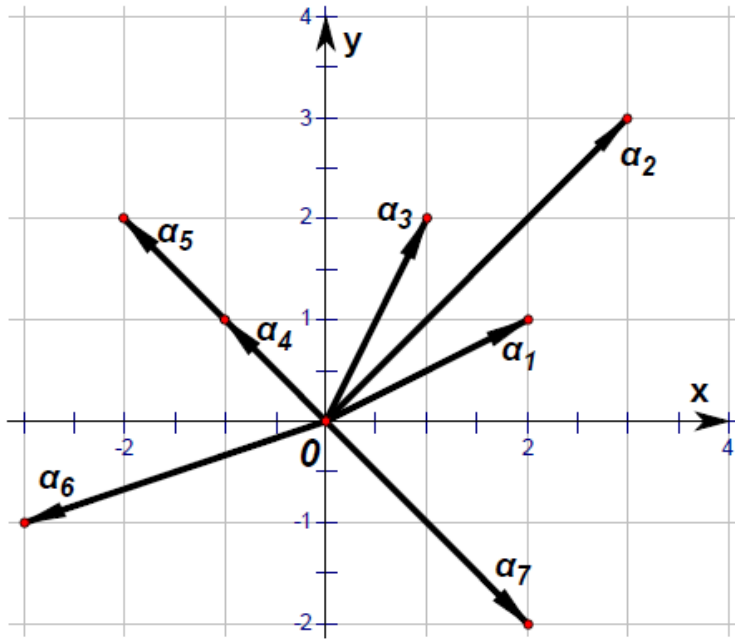
$$d(a2,a3)=|3-1|+|3-2|=3$$

Chobychev 距离

$$d(a2,a3)=\max\{|3-1|+|3-2|\}=2$$

余弦距离

$$d(a2,a3)=\frac{3\times1+3\times2}{\sqrt{3^2+3^2}\sqrt{1^2+2^2}}=9/\sqrt{90}=0.949$$



基础知识

相似性测度

例子 2： a1=[2,1]； a2=[3,3]； a3=[1,2]； a4=[-1,1]； a5=[-2,2]； a6=[-3,-1]; a7=[2,-2]

以 1 范数距离，计算与 a1 最相似性的样本

d(a1,a2) =3

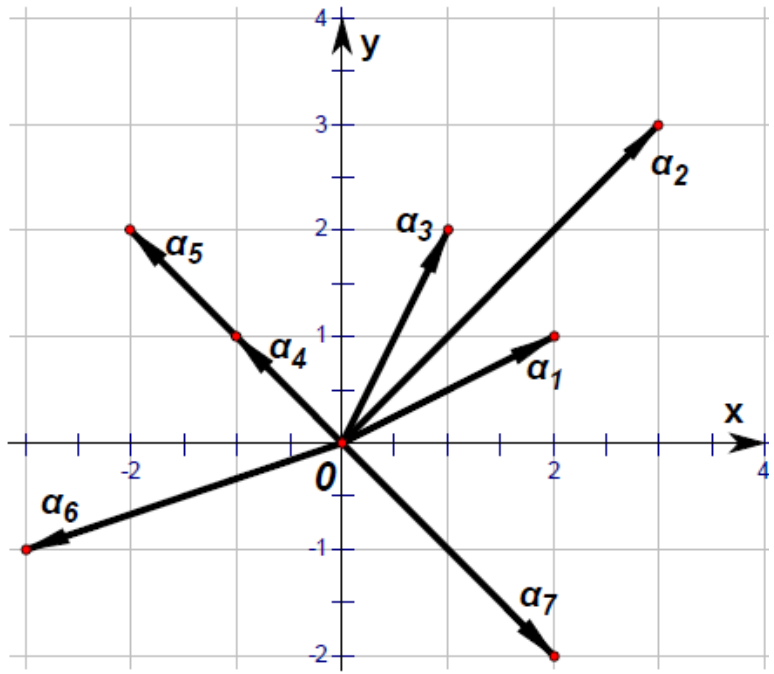
d(a1,a3) =2

d(a1,a4) =3

d(a1,a5) =5

d(a1,a6) =7

d(a1,a7) =3



基础知识

相似性测度

例子 2： a1=[2,1]； a2=[3,3]； a3=[1,2]； a4=[-1,1]； a5=[-2,2]； a6=[-3,-1]; a7=[2,-2]

以 1 范数距离，计算与 a1 最相似性的样本

d(a1,a2) =3

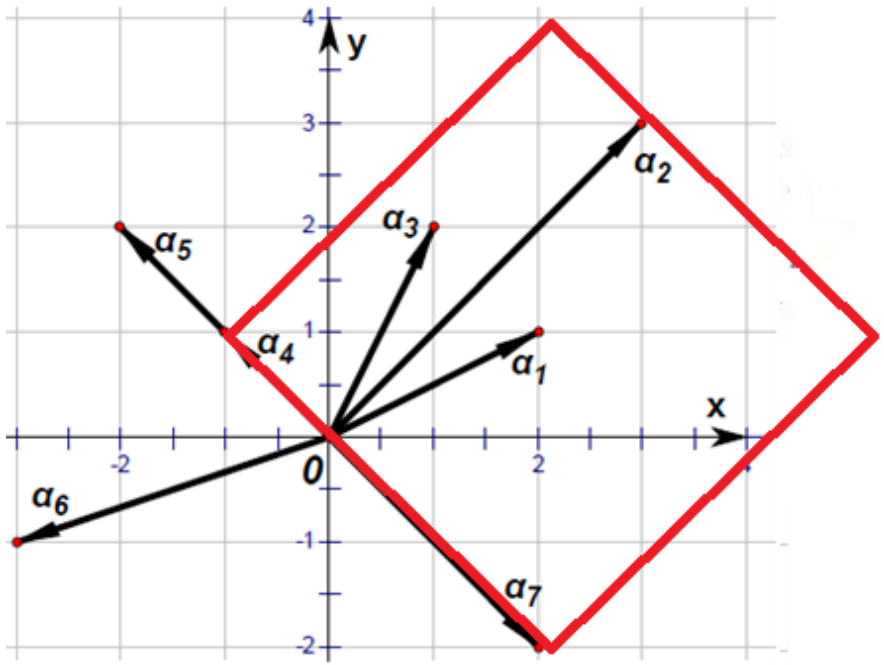
d(a1,a3) =2

d(a1,a4) =3

d(a1,a5) =5

d(a1,a6) =7

d(a1,a7) =3



常用的统计量

样本的统计量 （向量均为列向量）

- **类均值向量**

$$\boldsymbol{m}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}\boldsymbol{x}_k^{(i)}$$

- **总均值向量**

$$\boldsymbol{m} = \sum_{i=1}^{c} P_i \boldsymbol{m}_i$$

$$\boldsymbol{m} = \frac{1}{c}\sum_{i=1}^{c} \boldsymbol{m}_i$$

基础知识

常用的统计量

- 类内散度矩阵

$$\boldsymbol{S}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}(\boldsymbol{x}_k^{(i)} - \boldsymbol{m}_i)(\boldsymbol{x}_k^{(i)} - \boldsymbol{m}_i)^T$$

- 总类内离散度矩阵

$$\boldsymbol{S}_w = \sum_{i=1}^{c} P_i \boldsymbol{S}_i = \sum_{i=1}^{c} P_i \frac{1}{\mathrm{n_i}}\sum_{k=1}^{n_i}(\boldsymbol{x}_k^{(i)} - \boldsymbol{m}_i)(\boldsymbol{x}_k^{(i)} - \boldsymbol{m}_i)^T$$

- 类间散度矩阵

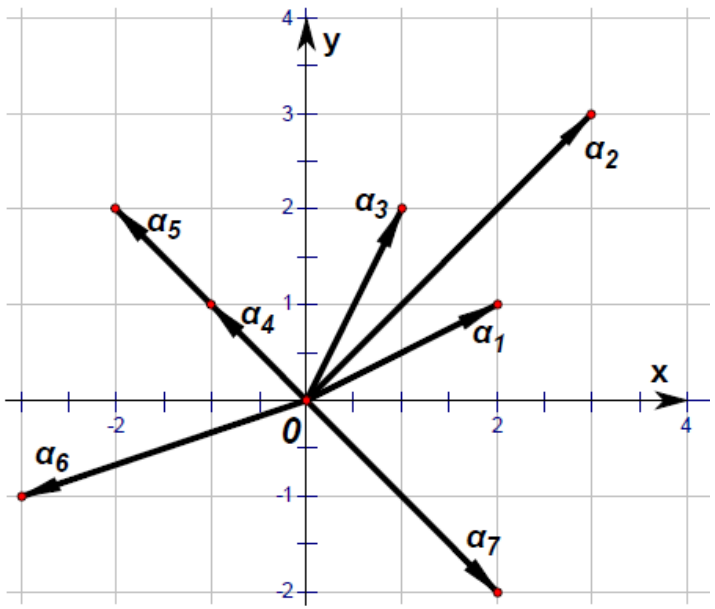
$$\boldsymbol{S}_b = \sum_{i=1}^{c}(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

基础知识

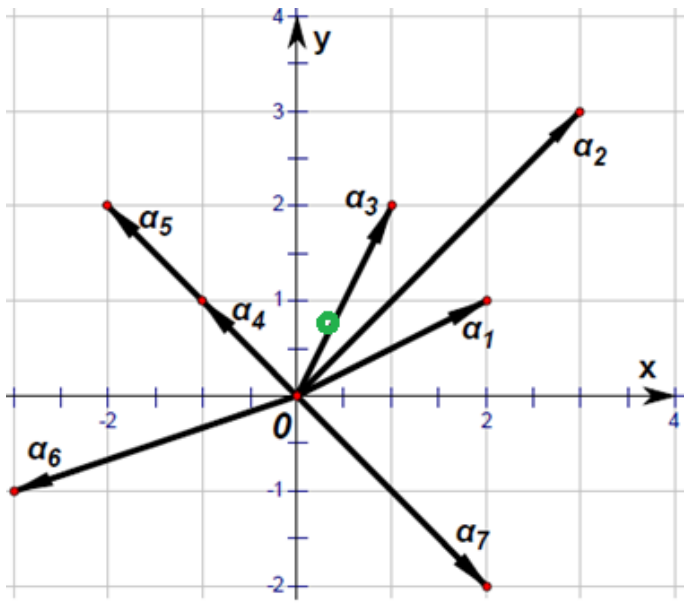
常用的统计量

例子：以 a1-a7,为一类样本

写出均值



平均=[0.2857， 0.8571]



Have a break !

基础知识

分类问题

1. 定义

根据给定的训练集 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，其中 $x_i \in C=R^n$，$y_i \in \mathrm{Y}=\{1,2,...,m\}$ $i=1,2,...,l$，要求寻找 C 上的决策函数 $g(x):\mathrm{C}\rightarrow\mathrm{Y}$

2. 评估方法

- 留出法数据集分成两类，交叉验证。
- 交叉验证法数据集分成 k 类，其中 1 类做测试，k-1 类做训练；进行 k 测实验取平均。
- 自助法 m 次随机取一个样本，共 m 个样本，放入 D'中；由 D'训练，D\D'测试。

分类问题

3. 性能评价

- 错误率与精度

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{II}(f(\boldsymbol{x}_i)\neq y_i)\quad,\qquad E(f;D)=\int_{\boldsymbol{x}\sim D}\mathrm{II}(f(\boldsymbol{x})\neq y)p(\boldsymbol{x})d\boldsymbol{x}$$

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{II}(f(\boldsymbol{x}_i)=y_i)=1-E(f;D)\quad,\qquad acc(f;D)=\int_{\boldsymbol{x}\sim D}\mathrm{II}(f(\boldsymbol{x})=y)p(\boldsymbol{x})d\boldsymbol{x}=1-E(f;D)$$

分类问题

- 查准率、查全率与 F1
- ROC 与 AUC
- 代价敏感错误率与代价曲线

4. 比较检验

- 假设检验
- 交叉验证 t 检验
- McNemar 检验
- Friedman 检验与 Nemenyi 检验

线性分类问题

1. 线性分类器描述：

- 线性判别函数:

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

- 分类界为超平面:

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

- 线性判别:

$$If \quad \boldsymbol{w}^T \boldsymbol{x} + w_0 > 0 \quad assign \quad \boldsymbol{x} \quad to \quad \omega_1$$
$$If \quad \boldsymbol{w}^T \boldsymbol{x} + w_0 < 0 \quad assign \quad \boldsymbol{x} \quad to \quad \omega_2$$

线性分类问题

2. 线性分类器的任务

通过已知的训练样本集，构造线性判别函数

训练样本 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$，$(\boldsymbol{x}_i,y_i)\in(R^d,R)$，$y_i$ 是类别标签

目标：确定 $\boldsymbol{w}$，w_0

$$\boldsymbol{w}^T\boldsymbol{x}_i+w_0\geq 0 \qquad For\ all\ i, such\ that \qquad y_i=+1$$

$$\boldsymbol{w}^T\boldsymbol{x}_i+w_0\leq 0 \qquad For\ all\ i, such\ that \qquad y_i=-1$$

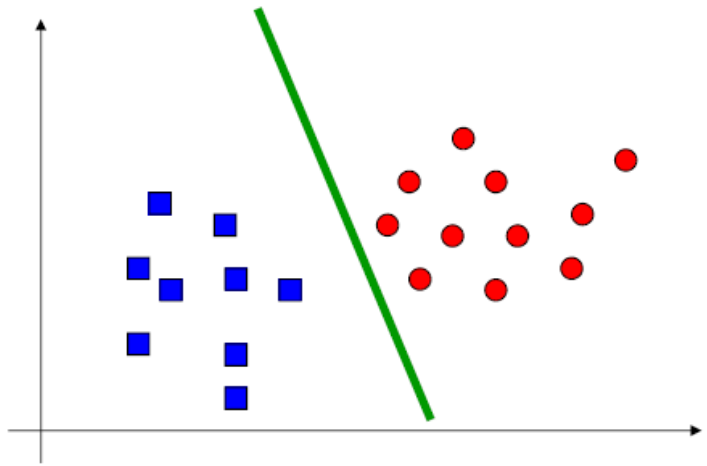
$$Together: \qquad y_i(\boldsymbol{w}^T\boldsymbol{x}_i+w_0)\geq 0$$

线性分类问题

3. 线性可分性

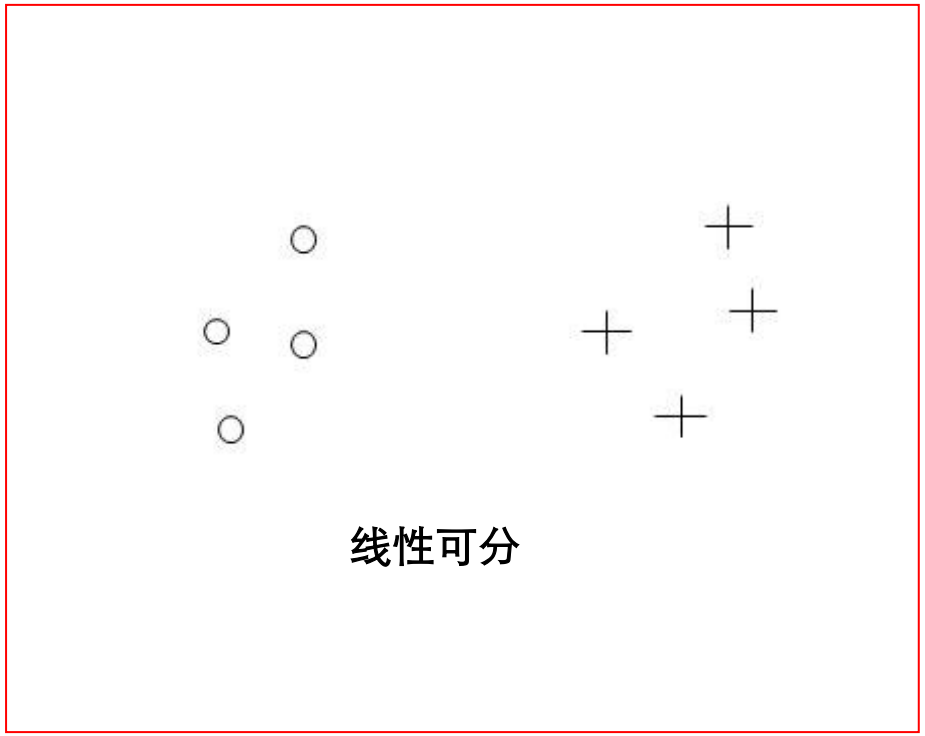
There is a **hyperplane** $\mathbf{w}^T\mathbf{x} + w_0 = 0$
that separates training instances with no error

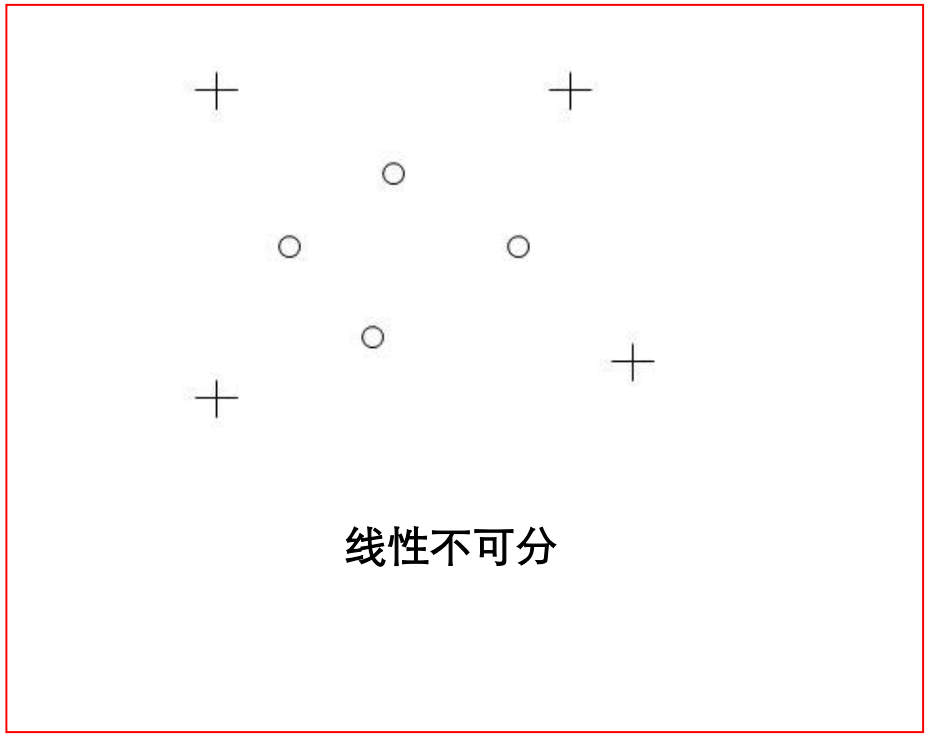
Class (+1) $\mathbf{w}^T\mathbf{x} + w_0 > 0$
Class (-1) $\mathbf{w}^T\mathbf{x} + w_0 < 0$



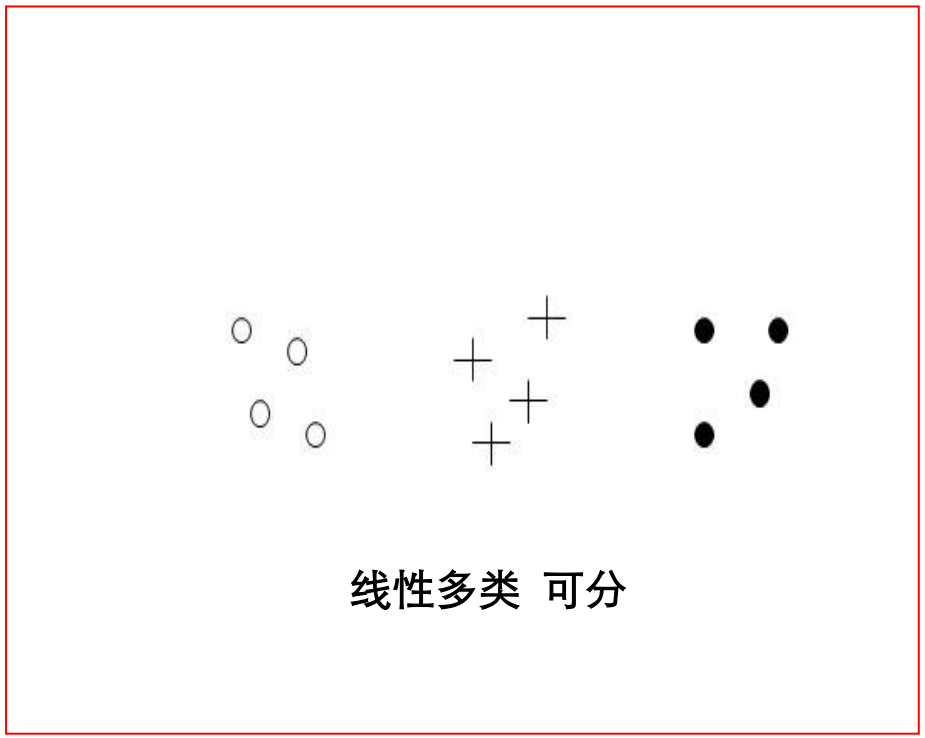
线性分类问题

举例

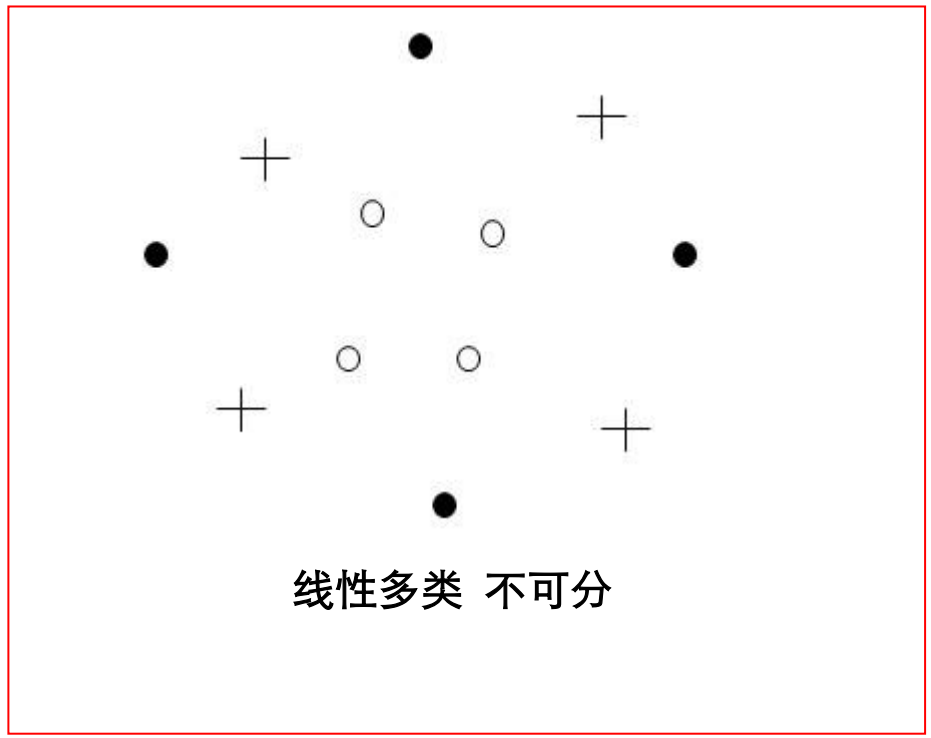




线性决策的多分类问题



线性多类 可分



线性多类 不可分

线性决策的多分类问题

- **二叉树比对**

  k 类问题，需要至少预先训练多少个二分类器?

  需要训练好 k(k-1)/2 个分类器（所有可能的分类器），
  然后采用二叉树比对测试。

线性决策的多分类问题

- **最大相似性**

Generally, $y = \{1, \cdots, C\}$, we define C discriminant functions

$$f_c(\mathbf{x}) = \mathbf{w}_c^{\mathrm{T}}\mathbf{x}, \qquad c = 1, \cdots, C,$$

where $\mathbf{w}_c$ is weight vector of class c.
Thus,

$$\hat{y} = \arg\max_{c=1}^{C} \mathbf{w}_c^{\mathrm{T}}\mathbf{x}$$

Have a break !

第三章 线性分类

3.1 概述

3.2 基础知识

3.3 感知机

3.4 线性鉴别分析

3.5 logistic 模型

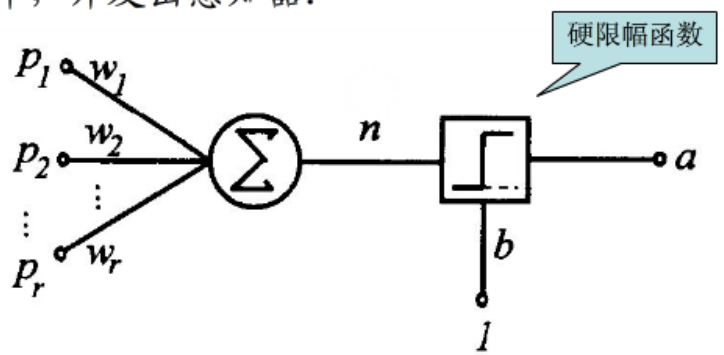
感知机

基本知识

1. 神经网络形成阶段（1943-1958），开拓性的贡献：

- McCulluch & Pitts（1943）引入神经网络的概念作为计算工具；

McCulloch和Pitts 1943年，发表第一个系统的ANN研究——阈值加权和(M-P)数学模型.
1947年，开发出感知器.



- Hebb（1949）提出自组织学习的第一个规则；
- Rosenblatt（1957）提出感知器作为有教师学习的一个模型。

感知机

基本知识

2. 线性分类

- 决策函数

$$g(\boldsymbol{x}) = \sum_{i=1}^{m} \boldsymbol{w}_i \boldsymbol{x}_i + w_0 = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

- 增广表示

$$g(x) = \sum_{i=0}^{m} w_i x_i = \dot{\boldsymbol{w}}^T \dot{\boldsymbol{x}}$$

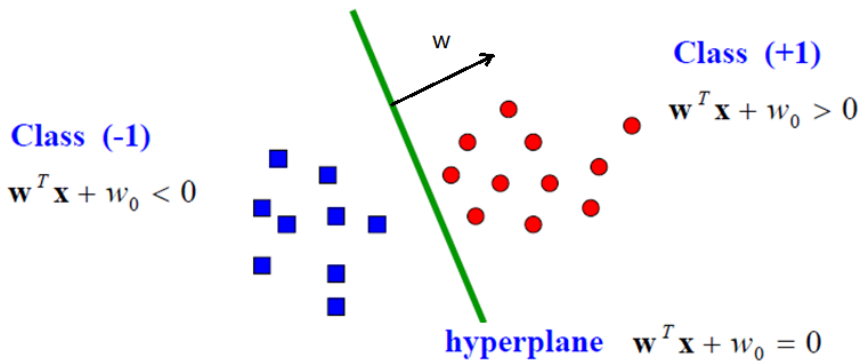
其中，$\dot{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix}$，$\dot{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix}$，$\dot{\boldsymbol{w}}, \dot{\boldsymbol{x}} \in R^{(m+1)\times 1}$

感知机

基本知识

- 决策超平面 $g(x)=\boldsymbol{w}^T\boldsymbol{x}+w_0=0$
- 分类判别

$$If \quad \boldsymbol{w}^T\boldsymbol{x}+w_0>0 \quad assign \quad \boldsymbol{x} \quad to \quad \omega_1$$
$$If \quad \boldsymbol{w}^T\boldsymbol{x}+w_0<0 \quad assign \quad \boldsymbol{x} \quad to \quad \omega_2$$



感知机

基本知识

- 决策函数几何含义

  刻画了样本到超平面的距离 $g(\boldsymbol{x}) = \|\boldsymbol{w}\| \cdot \boldsymbol{z}$

- 验证函数: $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0)$

$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \geq 0 \quad For\ all\ i, such\ that \quad y_i = +1$$

$$\boldsymbol{w}^T \boldsymbol{x}_i + w_0 \leq 0 \quad For\ all\ i, such\ that \quad y_i = -1$$

$$Together: \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0) \geq 0$$

感知机

基本知识

3. 优化方法 — 梯度下降

$$\min_{w} \ J(w)=\sum_{i} J_i(w)$$

- 梯度下降（GD）

$$w=w-\eta\frac{\partial J(w)}{\partial w}=w-\eta\nabla J(w)$$

感知机

基本知识

3. 优化方法 — 梯度下降

$$\min_{w} \ J(w)=\sum_{i} J_i(w)$$

- 梯度下降（GD）

$$w=w-\eta\frac{\partial J(w)}{\partial w}=w-\eta\nabla J(w) \ \ =w-\eta\sum_{i}\frac{\partial J_i(w)}{\partial w}= \ w-\eta\sum_{i}\nabla J_i(w)$$

感知机

基本知识

3. 优化方法 — 梯度下降

$$\min_{w} \ J(w)=\sum_{i} J_i(w)$$

- 梯度下降（GD）

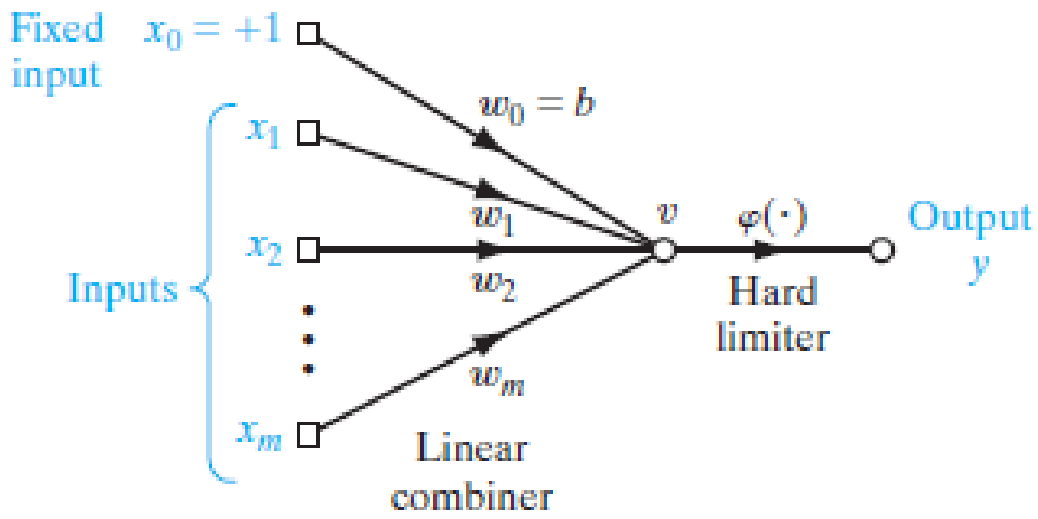
$$w=w-\eta\frac{\partial J(w)}{\partial w}=w-\eta\nabla J(w) \ \ =w-\eta\sum_{i}\frac{\partial J_i(w)}{\partial w}= \ w-\eta\sum_{i}\nabla J_i(w)$$

- 随机梯度下降（SGD）

$$w=w-\eta\frac{\partial J_i(w)}{\partial w}$$

感知机

感知机结构



信号流

- 输入

$$\boldsymbol{x}(n) = [+1, x_1(n), x_2(n), \ldots, x_m(n)]^T$$

- 神经元连接权

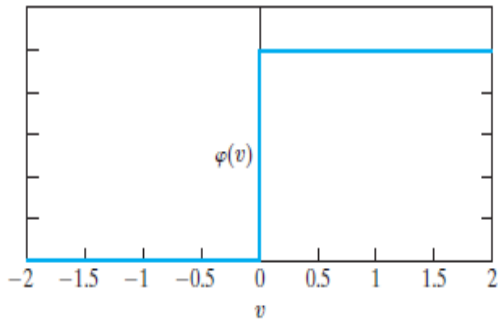
$$\boldsymbol{w}(n) = [b, w_1(n), w_2(n), \ldots, w_m(n)]^T$$

- 神经元局部感受域

$$v(n) = \sum_{i=0}^{m} w_i(n) x_i(n)$$

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

- 硬激活函数



感知机

感知机学习准则

1. 目标：最小化 **错分样本的** 误差代价。

- 代价函数（错分样本的误差函数）：

$$J(\boldsymbol{w})=\sum_{\boldsymbol{x}(n)\in E}-\boldsymbol{w}^T\boldsymbol{x}(n)d(n) \tag{1.1}$$

或者

$$J(\boldsymbol{w})=\sum_{\boldsymbol{x}(n)}-\boldsymbol{w}^T\boldsymbol{x}(n)\big(d(n)-y(n)\big) \tag{1.2}$$

其中，E 为错误分类样本集；$d(n)\in\{-1,+1\}$为$\boldsymbol{x}(n)$的已知类别标签；$y(n)\in\{-1,+1\}$为感知器的输出类别

感知机

感知机学习准则

问题：(d(n)-y(n))能否替代“错误分类样本集筛选”、(d(n)-y(n))能否替代 d(n) ？

答 1：当样本被正确分类时$(d(n)-y(n))=0$，正确分类样本被忽略，

$(d(n)-y(n))$可替代“错误分类样本集筛选”；

答 2：当样本被错误分类时，$(d(n)-y(n))\neq 0$，两种情况

$d(n)=+1,\ y(n)=-1$时，$(d(n)-y(n))=+2$，

$(d(n)-y(n))$与$d(n)$符号相同

$d(n)=-1,\ y(n)=+1$时，$(d(n)-y(n))=-2$，

$(d(n)-y(n))$与$d(n)$符号相同

$(d(n)-y(n))$能替代$d(n)$；

感知机学习准则

2. J(w)的含义：错分样本到分类超平面误差距离的总和

$$|z| = \frac{|w^T x|}{\|w\|}$$

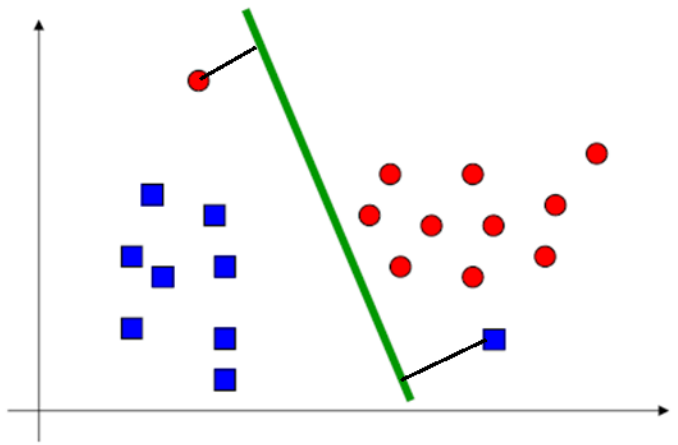
样本到超平面的距离：

正确分类样本：

$$|z| = \frac{|w^T x|}{\|w\|} = \frac{dw^T x}{\|w\|}$$

错误分类样本：

$$|z| = \frac{|w^T x|}{\|w\|} = \frac{-dw^T x}{\|w\|}$$



感知机优化

Batch Perception

$$\nabla J(\boldsymbol{w}) = \sum_{x} -(d(n) - y(n))\boldsymbol{x}(n)$$

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta(n) \sum_{x} -(d(n) - y(n))\boldsymbol{x}(n)$$

$$\nabla J(\boldsymbol{w}) = \sum_{\boldsymbol{x}(n) \in E} -\boldsymbol{x}(n)d(n)$$

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta(n) \sum_{\boldsymbol{x}(n) \in E} -\boldsymbol{x}(n)d(n)$$

感知机

感知机优化

Online Perception

$$\nabla J(\boldsymbol{w}) = -(d(n) - y(n))\boldsymbol{x}(n)$$

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta(n)[-(d(n) - y(n))]\boldsymbol{x}(n)$$

$$\nabla J(\boldsymbol{w}) = -\boldsymbol{x}(n)d(n)_{|\boldsymbol{x}(n)\in E}$$

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) - \eta(n)(-\boldsymbol{x}(n)d(n))_{|\boldsymbol{x}(n)\in E}$$

感知机

感知器算法流程

Variables and Parameters:

$\mathbf{x}(n)$ = $(m+1)$-by-1 input vector
$= [+1, x_1(n), x_2(n), \ldots, x_m(n)]^T$
$\mathbf{w}(n)$ = $(m+1)$-by-1 weight vector
$= [b, w_1(n), w_2(n), \ldots, w_m(n)]^T$
b = bias
$y(n)$ = actual response (quantized)
$d(n)$ = desired response
η = learning-rate parameter, a positive constant less than unity

1. *Initialization.* Set $\mathbf{w}(0) = \mathbf{0}$. Then perform the following computations for time-step $n = 1, 2, \ldots$.
2. *Activation.* At time-step n, activate the perceptron by applying continuous-valued input vector $\mathbf{x}(n)$ and desired response $d(n)$.
3. *Computation of Actual Response.* Compute the actual response of the perceptron as

$$y(n) = \text{sgn}[\mathbf{w}^T(n)\mathbf{x}(n)]$$

where sgn(·) is the signum function.

4. *Adaptation of Weight Vector.* Update the weight vector of the perceptron to obtain

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta[d(n) - y(n)]\mathbf{x}(n)$$

where

$$d(n) = \begin{cases} +1 & \text{if } \mathbf{x}(n) \text{ belongs to class } \mathscr{C}_1 \\ -1 & \text{if } \mathbf{x}(n) \text{ belongs to class } \mathscr{C}_2 \end{cases}$$

5. *Continuation.* Increment time step n by one and go back to step 2.

感知机

误差修正基本规则

1. 固定增量的感知机修正

- 固定增量感知器收敛定理（Rosenblatt，1962）

  若训练样本是线性可分，则感知器训练算法在有限次迭代后

可以收敛到正确的解向量 w。

误差修正基本规则

2. 误差修正自适应规则

- 增量自适应调整

设$\eta(n)$满足下式：$\eta(n)\boldsymbol{x}^T(n)\boldsymbol{x}(n) \geq \left|\boldsymbol{w}^T(n)\boldsymbol{x}(n)\right|$

$$if\ \ \mathrm{d(n)=+1},\ \eta(n)\boldsymbol{x}^T(n)\boldsymbol{x}(n) \geq \boldsymbol{w}^T(n)\boldsymbol{x}(n),\ \ 0 \geq \boldsymbol{w}^T(n)\boldsymbol{x}(n) - \eta(n)\boldsymbol{x}^T(n)\boldsymbol{x}(n)$$

$$if\ \ \mathrm{d(n)=-1},\ \eta(n)\boldsymbol{x}^T(n)\boldsymbol{x}(n) \geq -\boldsymbol{w}^T(n)\boldsymbol{x}(n),\ 0 \geq -\boldsymbol{w}^T(n)\boldsymbol{x}(n) - \eta(n)\boldsymbol{x}^T(n)\boldsymbol{x}(n)$$

感知机

误差修正基本规则

- **增量自适应调整的证明：**

错分情况修正规则

当错分样本的正确标签为 d=+1，修正 $\boldsymbol{w}(n+1)=\boldsymbol{w}(n)+\eta(n)\boldsymbol{x}(n)_{|x\in E}$

当错分样本的正确标签为 d=-1，修正 $\boldsymbol{w}(n+1)=\boldsymbol{w}(n)-\eta(n)\boldsymbol{x}(n)_{|x\in E}$

两边同乘$-\boldsymbol{x}^T(n)d(n)$，计算损失函数（错分代价）：$-\boldsymbol{x}^T(n)w(n)d(n)$

当错分样本的正确标签为 d=+1，损失函数（错分代价）：

$$-\boldsymbol{x}^T(n)\boldsymbol{w}(n+1)=-\boldsymbol{x}^T(n)\boldsymbol{w}(n)\underbrace{-\eta(n)\boldsymbol{x}^T(n)\boldsymbol{x}(n)}_{|x\in E}<=0$$

>0 <0

当错分样本的正确标签为 d=-1，损失函数（错分代价）：

$$\boldsymbol{x}^T(n)\boldsymbol{w}(n+1)=\boldsymbol{x}^T(n)\boldsymbol{w}(n)\underbrace{-\eta(n)\boldsymbol{x}^T(n)\boldsymbol{x}(n)}_{|x\in E}<=0$$

>0 <0

- 基本规则可以保证误差变小，
- 自适应规则保证误差为 0。

感知机

误差修正基本规则

- 自适应修正的几何过程：

Online Perception 为例

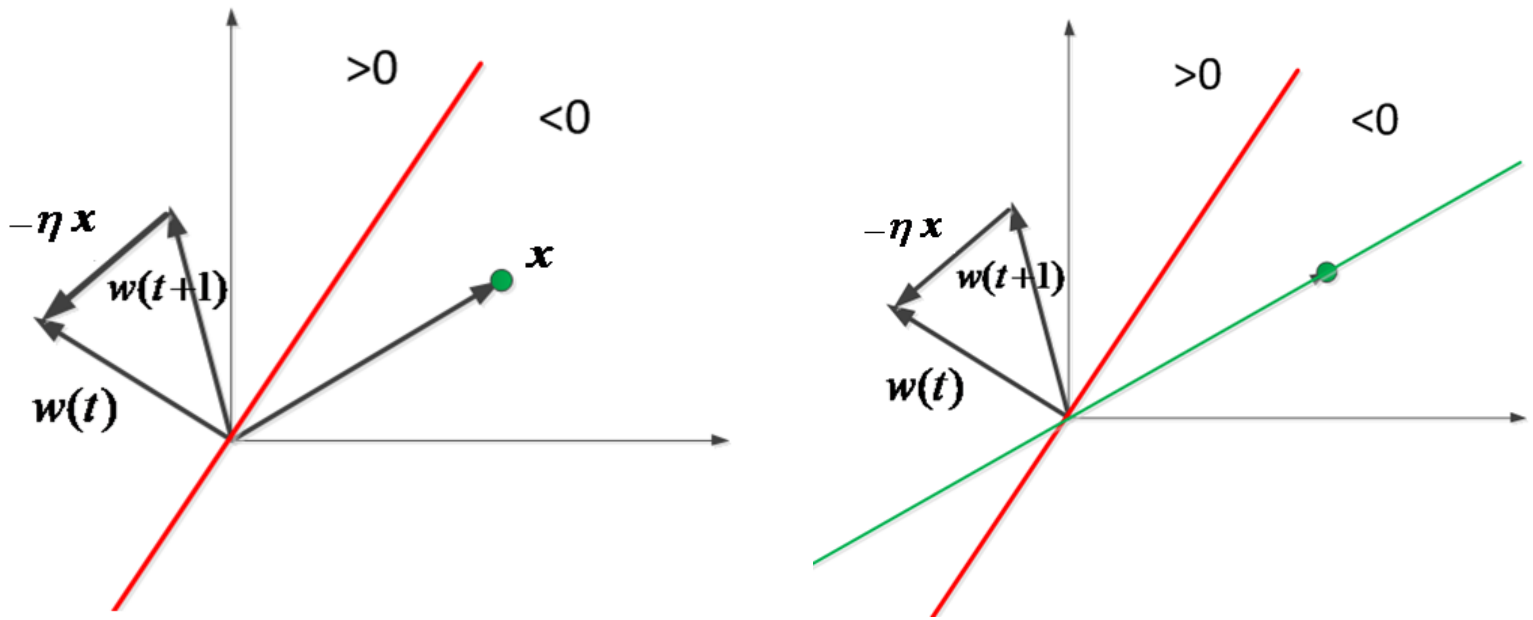
d=+1，$\boldsymbol{w}(n+1)=\boldsymbol{w}(n)+\eta(n)\boldsymbol{x}(n)_{|x\in E}$

d=-1，$\boldsymbol{w}(n+1)=\boldsymbol{w}(n)-\eta(n)\boldsymbol{x}(n)_{|x\in E}$

感知机

误差修正基本规则

当 d=+1，$\boldsymbol{w}(n+1)=\boldsymbol{w}(n)-(-\eta(n)\boldsymbol{x}(n))_{|x\in E}$

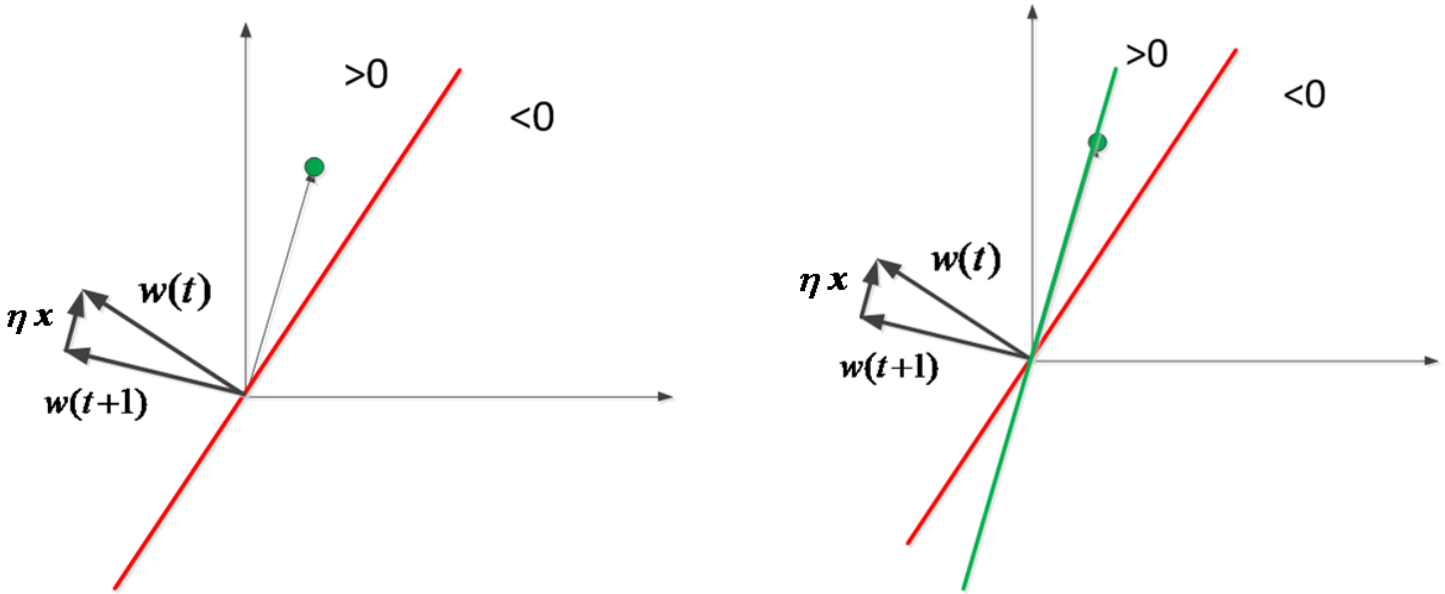


修正后的分类面（绿线）

感知机

误差修正基本规则

当 $d=-1$，$\boldsymbol{w}(n+1)=\boldsymbol{w}(n)-\eta(n)\boldsymbol{x}(n)_{|x\in E}$

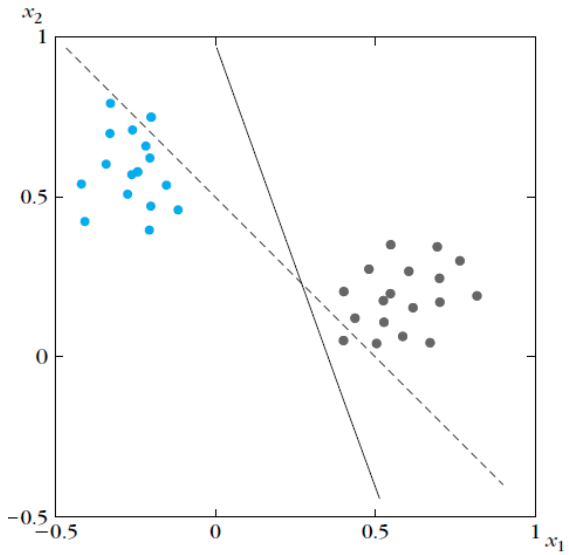


修正后的分类面（绿线）

感知机

例子--1

Initial: the dashed line $x_1 + x_2 - 0.5 = 0$



corresponding to the weight vector $[1, 1, -0.5]^T$, $\rho_t = \rho = 0.7$

感知机

例子--1

Optimization (GD): $\boldsymbol{w}(n+1)=\boldsymbol{w}(n)-\eta(n)\sum_{x\in E}-d(n)\boldsymbol{x}(n)$

all the vectors except $[0.4, 0.05]^T$ and $[-0.20, 0.75]^T$.

$$\boldsymbol{w}(t+1)=\begin{bmatrix}1\\1\\-0.5\end{bmatrix}-0.7(-1)\begin{bmatrix}0.4\\0.05\\1\end{bmatrix}-0.7(+1)\begin{bmatrix}-0.2\\0.75\\1\end{bmatrix}$$

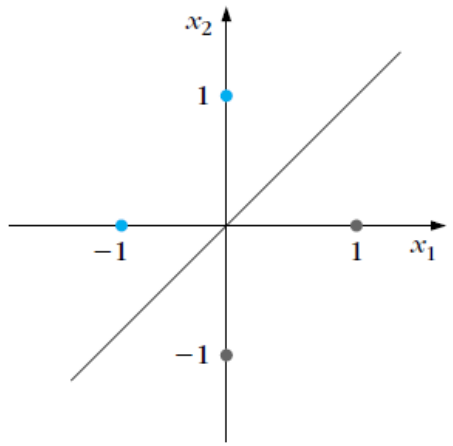
or

$$\boldsymbol{w}(t+1)=\begin{bmatrix}1.42\\0.51\\-0.5\end{bmatrix}$$

The resulting new (solid) line $1.42x_1 + 0.51x_2 - 0.5 = 0$ classifies all vectors correctly, and the algorithm is terminated.

感知机

例子--2



$(-1,0),\ (0,1)$ *belong to* C1

$(0,-1),\ (1,0)$ *belong to* C2

Initial: $\boldsymbol{w}(0)=(0,0,0)^T$

The parameter η is set equal to one.

Data:

$(-1,0,1),\ (0,1,1) \in \mathrm{C1},\ \mathrm{d}=+1,\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}>0$

$(0,-1,1),\ (1,0,1) \in \mathrm{C2},\ \mathrm{d}=-1,\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\le 0$

感知机

例子--2

Optimization (SGD):

$$\boldsymbol{w}(n+1)=\boldsymbol{w}(n)-\eta(n)\left(-d(n)\boldsymbol{x}(n)\right)_{|x\in E}=\boldsymbol{w}(n)+\eta(n)\left(d(n)\boldsymbol{x}(n)\right)_{|x\in E}$$

Step 1.

$$\boldsymbol{w}^T(0)\begin{bmatrix}-1\\0\\1\end{bmatrix}=0,\quad \boldsymbol{w}(1)=\boldsymbol{w}(0)+\begin{bmatrix}-1\\0\\1\end{bmatrix}=\begin{bmatrix}-1\\0\\1\end{bmatrix}$$

Step 2.

$$\boldsymbol{w}^T(1)\begin{bmatrix}0\\1\\1\end{bmatrix}=1>0,\quad \boldsymbol{w}(2)=\boldsymbol{w}(1)$$

Step 3.

$$\boldsymbol{w}^T(2)\begin{bmatrix}0\\-1\\1\end{bmatrix}=1>0,\quad \boldsymbol{w}(3)=\boldsymbol{w}(2)-\begin{bmatrix}0\\-1\\1\end{bmatrix}=\begin{bmatrix}-1\\1\\0\end{bmatrix}$$

感知机

例子--2

Step 4.

$$\boldsymbol{w}^T(3)\begin{bmatrix}1\\0\\1\end{bmatrix} = -1 < 0, \quad \boldsymbol{w}(4) = \boldsymbol{w}(3)$$

Step 5.

$$\boldsymbol{w}^T(4)\begin{bmatrix}-1\\0\\1\end{bmatrix} = 1 > 0, \quad \boldsymbol{w}(5) = \boldsymbol{w}(4)$$

Step 6.

$$\boldsymbol{w}^T(5)\begin{bmatrix}0\\1\\1\end{bmatrix} = 1 > 0, \quad \boldsymbol{w}(6) = \boldsymbol{w}(5)$$

Step 7.

$$\boldsymbol{w}^T(6)\begin{bmatrix}0\\-1\\1\end{bmatrix} = -1 < 0, \quad \boldsymbol{w}(7) = \boldsymbol{w}(6)$$

参考文献

1. Pattern Recognition 2nd.《模式识别》(第二版)，边肇祺，张学工等,清华大学出版社，2000.1。
2. Pattern Classification, 2nd.模式分类，第二版。
3. 周志华，机器学习，清华大学出版社，2016.