机器学习

李成龙

安徽大学人工智能学院

“多模态认知计算”安徽省重点实验室

内容安排

- 什么是机器学习
- 机器如何学习
- 如何让机器学习的更好
- 为什么机器能学习

内容安排

- **机器如何学习**
  - 有监督学习
    - 感知机
    - 支持向量机
    - 朴素贝叶斯分类
    - 决策树
    - 集成学习（Bagging算法与随机森林、Boosting算法）
    - 线性回归
    - 逻辑回归
    - Softmax回归
    - 神经网络与深度学习
  - 无监督学习
    - 聚类
    - 主成分分析

本节目录

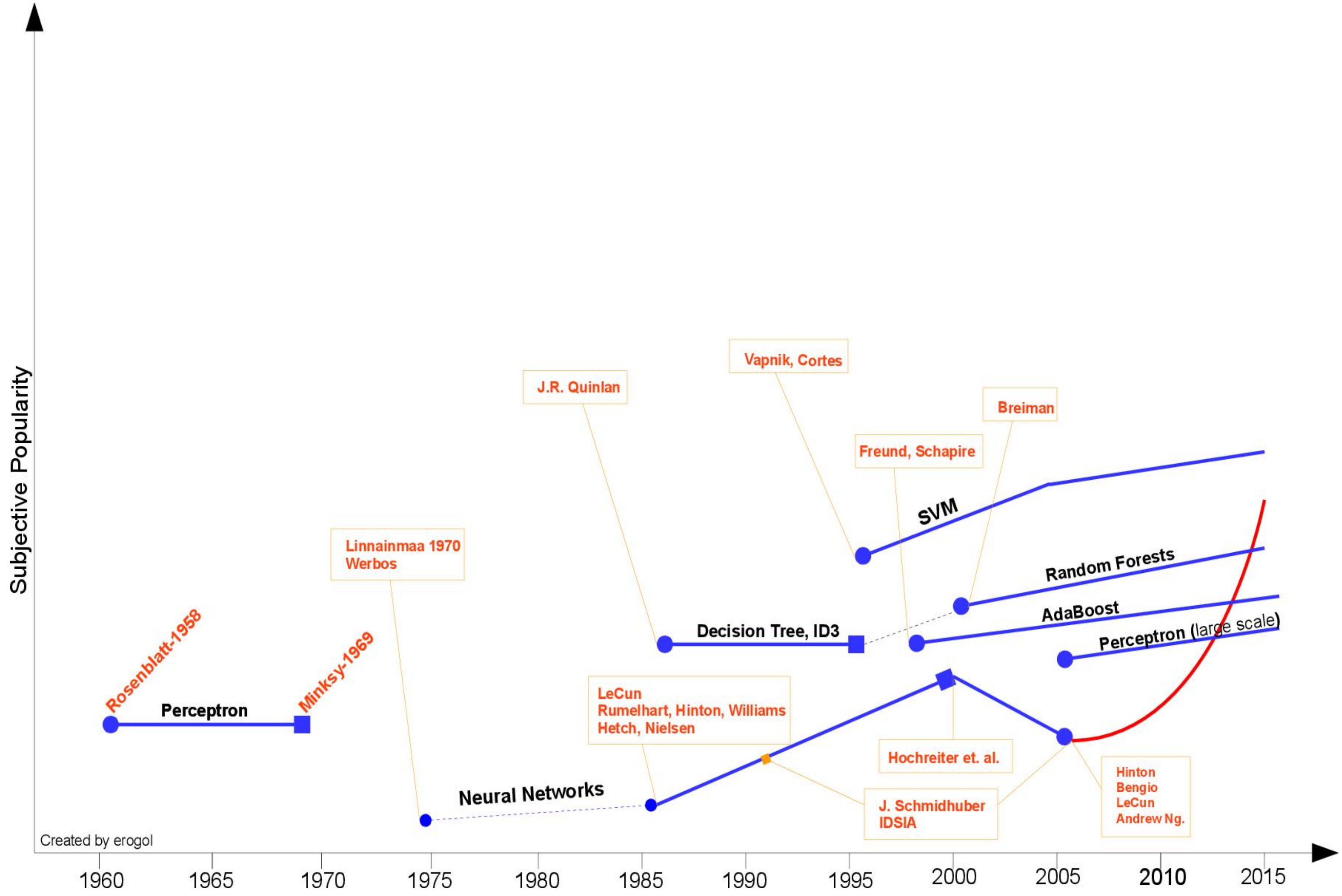
- 背景
- 感知机模型
- 学习策略
- 学习算法
- 存在问题

本节目录

- **背景**
- 感知机模型
- 学习策略
- 学习算法
- 存在问题

背景

- 主要发展历程



Subjective Popularity
Rosenblatt-1958
Perceptron
Minksy-1969
Linnainmaa 1970
Werbos
Neural Networks
J.R. Quinlan
Decision Tree, ID3
LeCun
Rumelhart, Hinton, Williams
Hetch, Nielsen
Vapnik, Cortes
SVM
Freund, Schapire
Breiman
Random Forests
AdaBoost
Perceptron (large scale)
Hochreiter et. al.
J. Schmidhuber
IDSIA
Hinton
Bengio
LeCun
Andrew Ng.
Created by erogol
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015

背景

- **特点**
  - 输入为实例的特征向量，输出为实例的类别，取+1和-1
  - 感知机对应于输入空间中将实例划分为正负两类的分离超平面
  - 导入基于误分类的损失函数
  - 利用梯度下降法对损失函数进行极小化
  - 感知机学习算法具有简单而易于实现的优点
  - 1957年由Rosenblatt提出，是神经网络与支持向量机的基础

本节目录

- 背景
- 感知机模型
- 学习策略
- 学习算法
- 存在问题

感知机模型

- **线性二分类问题**

http://d03.res.meilishuo.net/pic/l/b4/ee/13e7e3e906e6e1962da77bfdab9c_800_800.jpeg

http://www.lvmama.com/guide/2010/1213-100455.html

感知机模型

- **线性二分类问题**

http://www.lvmama.com/guide/2010/1213-100455.html

感知机模型

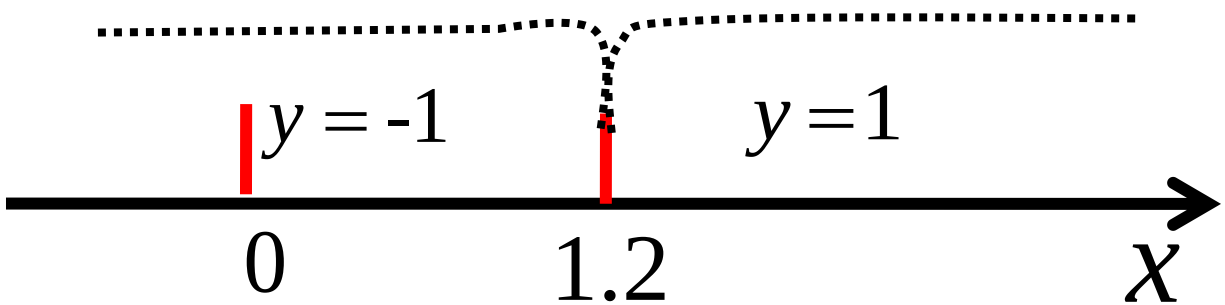
- **问题建模**
  - 例1：儿童免票乘车

x：身高，y：$\{$-1：免票　1：购票$\}$

标记（label）

特征（feature）

$$y=\begin{cases}1 & x \geq 1.2m \\ -1 & x < 1.2m\end{cases}$$

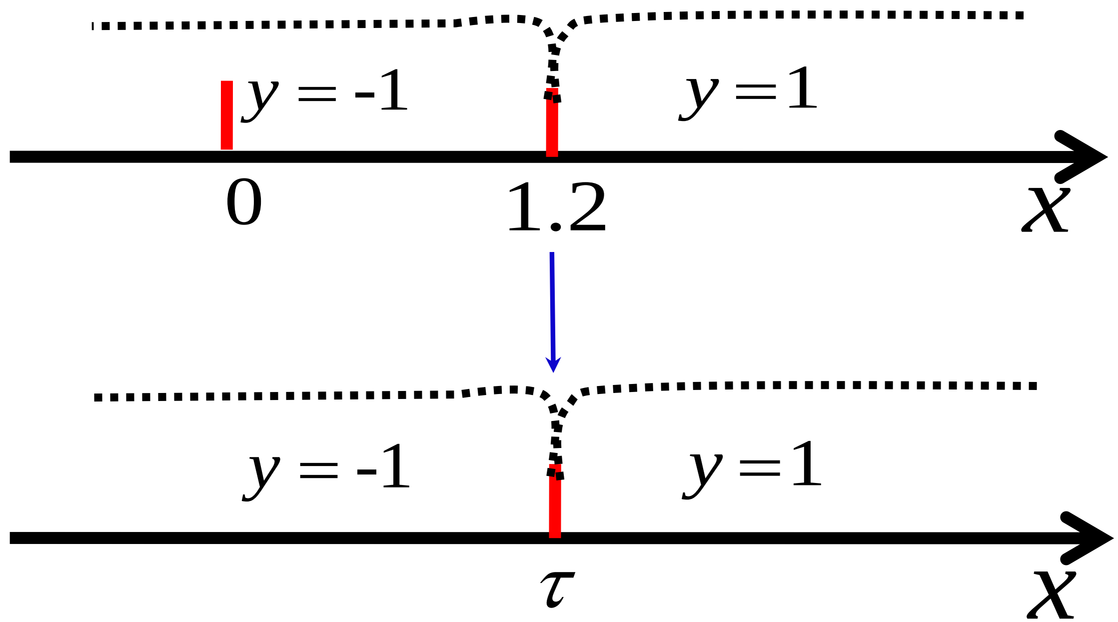


数轴上的分类

感知机模型

- **问题建模**
  - 例1：儿童免票乘车

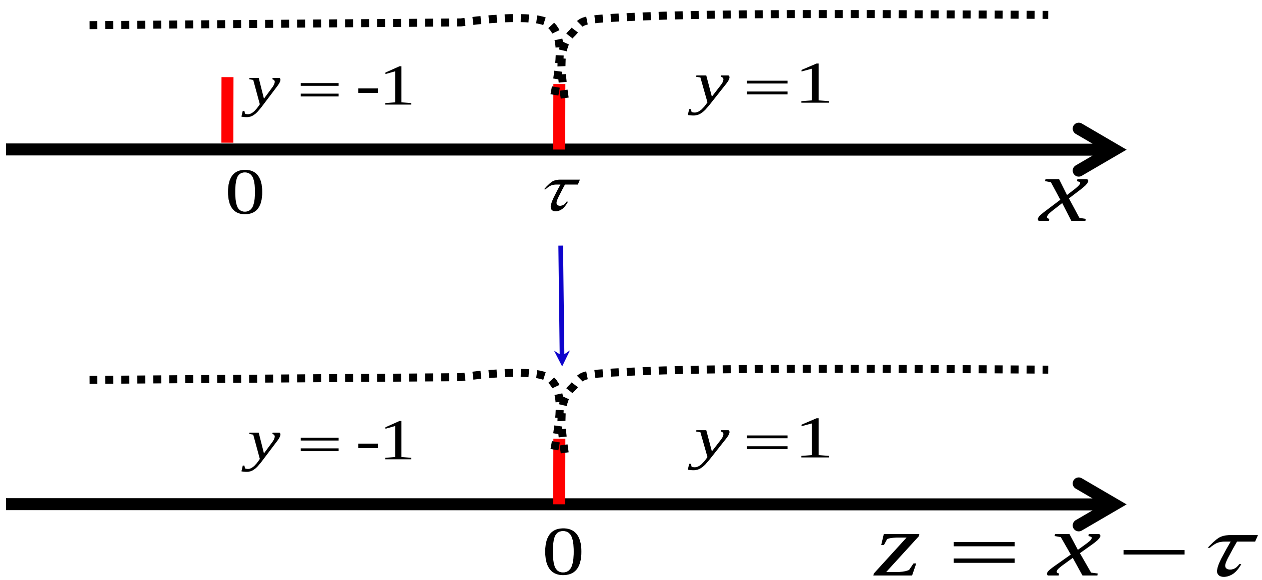
$$y = f(x;\tau) = \begin{cases} 1 & x \geq \tau \\ -1 & x < \tau \end{cases} \qquad \tau = 1.2$$



感知机模型

- **问题建模**
  - 例1:儿童免票乘车

$$y = h(x;\tau) = \begin{cases} 1 & z = x - \tau \geq 0 \\ -1 & z = x - \tau < 0 \end{cases}$$

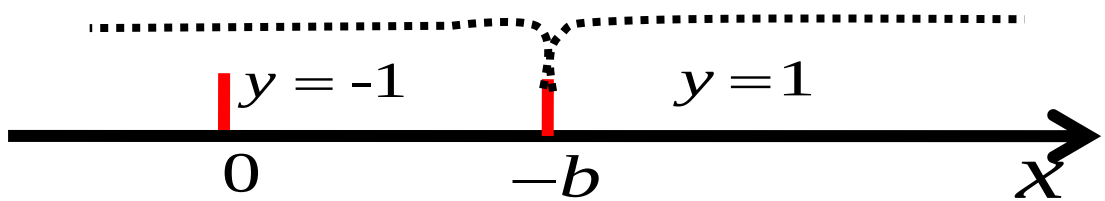


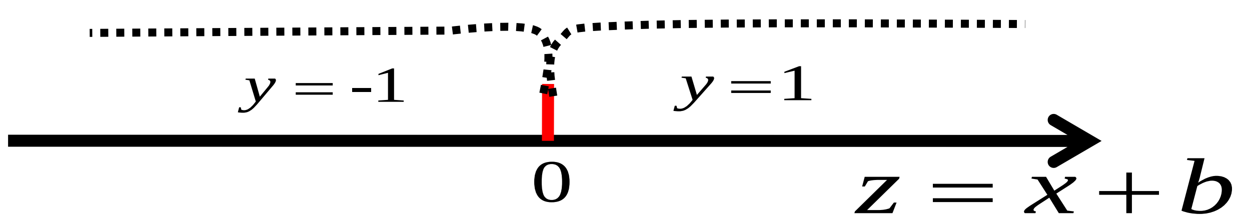
感知机模型

- **问题建模**
  - 例1:儿童免票乘车

$$y = f(x;b) = \begin{cases} 1 & z = x + b \geq 0 \\ -1 & z = x + b < 0 \end{cases}$$

偏置(Bias)





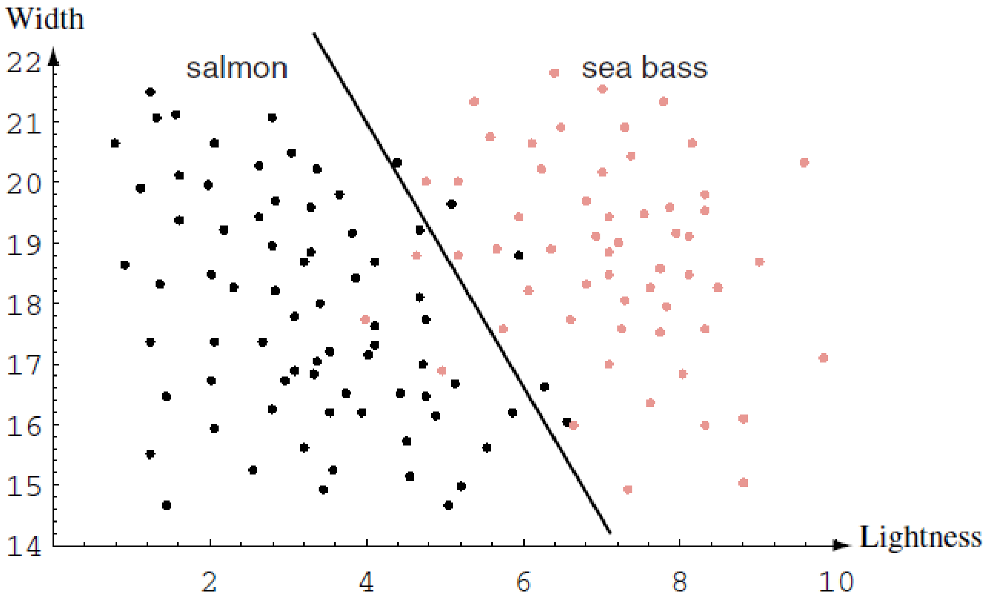
感知机模型

- **问题建模**
  - 数轴上的分类
    - 单个特征x（一维特征、实数）
    - 一个参数b（偏置项）

$$f(x;b)=\begin{cases}1 & x+b\geq 0\\ \text{-}1 & x+b<0\end{cases}$$

感知机模型

- **问题建模**
  - 例2:鱼的分类



Duda et.at: Pattern Classification,2nd eds.

感知机模型

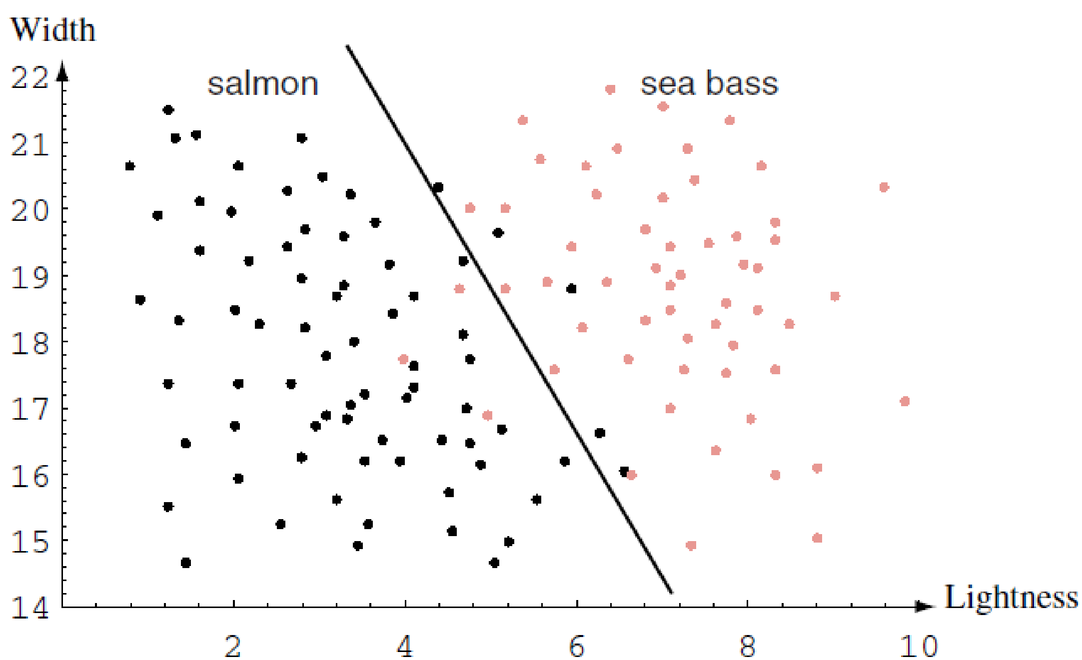
- **问题建模**
  - 例2: 鱼的分类

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

x_1 : lightness

x_2 : width

特征向量
(feature vector)



$y : \{ \text{-1} : salmon \quad 1 : seabass \}$

标记 (label)

$$y = f(\mathbf{x}) = ?$$

感知机模型

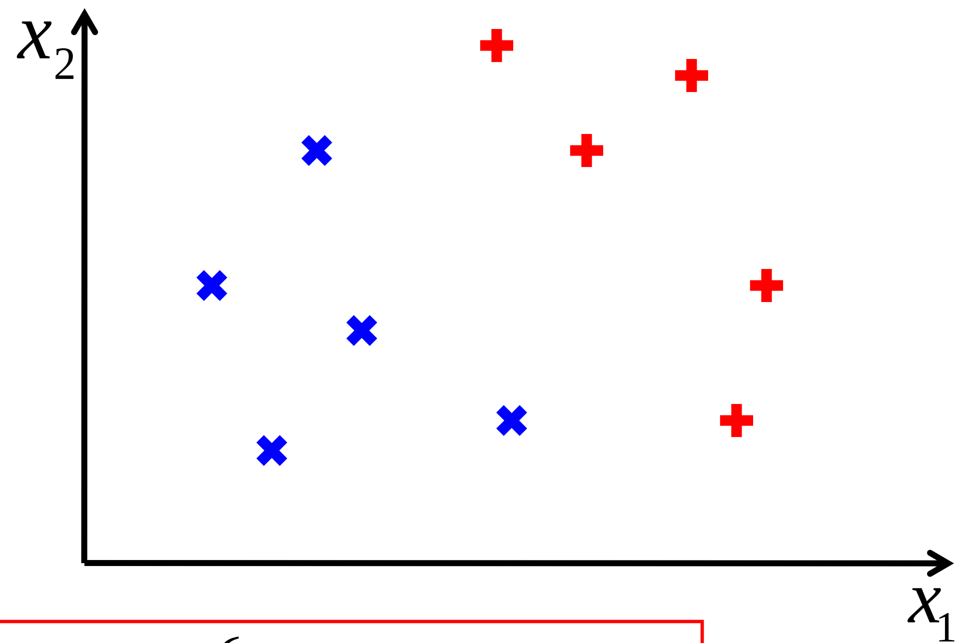
- **问题建模**
  - 平面上的分类：**转换成数轴上的分类**

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow$$

$$x' = w_1 x_1 + w_2 x_2$$

$$x' = \mathbf{w}^T \mathbf{x}$$

$$\mathbf{w} = \begin{pmatrix} w_1 & w_2 \end{pmatrix}^T$$



$$y = f(\mathbf{x}; b) = \begin{cases} 1 & x' + b \geq 0 \\ -1 & x' + b < 0 \end{cases}$$

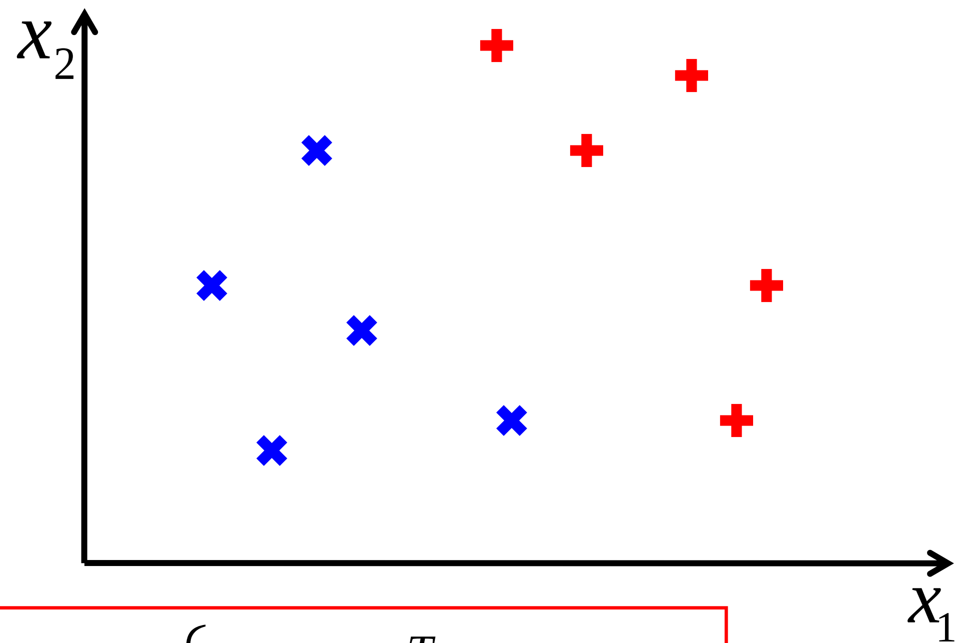
- **问题建模**
  - 平面上的分类：**转换成数轴上的分类**

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow$$

$$x' = w_1 x_1 + w_2 x_2$$

$$x' = \mathbf{w}^T \mathbf{x}$$

$$\mathbf{w} = \begin{pmatrix} w_1 & w_2 \end{pmatrix}^T$$



$$\boxed{y = f(\mathbf{x}; \mathbf{w}, b) = \begin{cases} 1 & \mathbf{w}^T \mathbf{x} + b \geq 0 \\ \text{-}1 & \mathbf{w}^T \mathbf{x} + b < 0 \end{cases}}$$

感知机模型

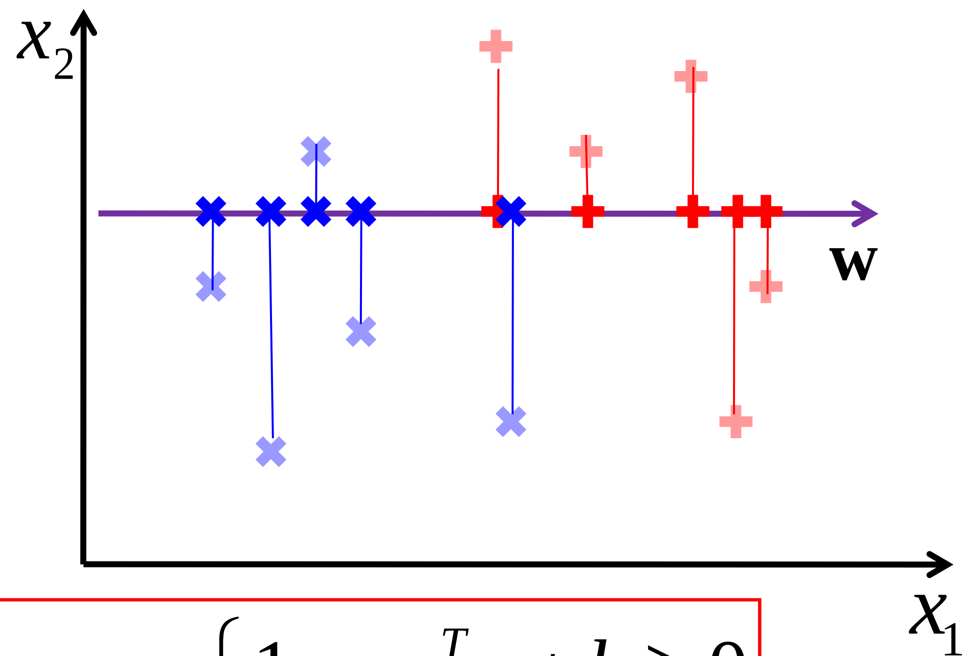
- **问题建模**
  - 平面上的分类：**转换成数轴上的分类**

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to$$

$$x' = w_1 x_1 + w_2 x_2$$

$$x' = \mathbf{w}^T \mathbf{x}$$

$$\mathbf{w} = \begin{pmatrix} w_1 & w_2 \end{pmatrix}^T$$



$$\boxed{y = f\left(\mathbf{x}; \mathbf{w}, b\right) = \begin{cases} 1 & \mathbf{w}^T \mathbf{x} + b \geq 0 \\ \text{-}1 & \mathbf{w}^T \mathbf{x} + b < 0 \end{cases}}$$

感知机模型

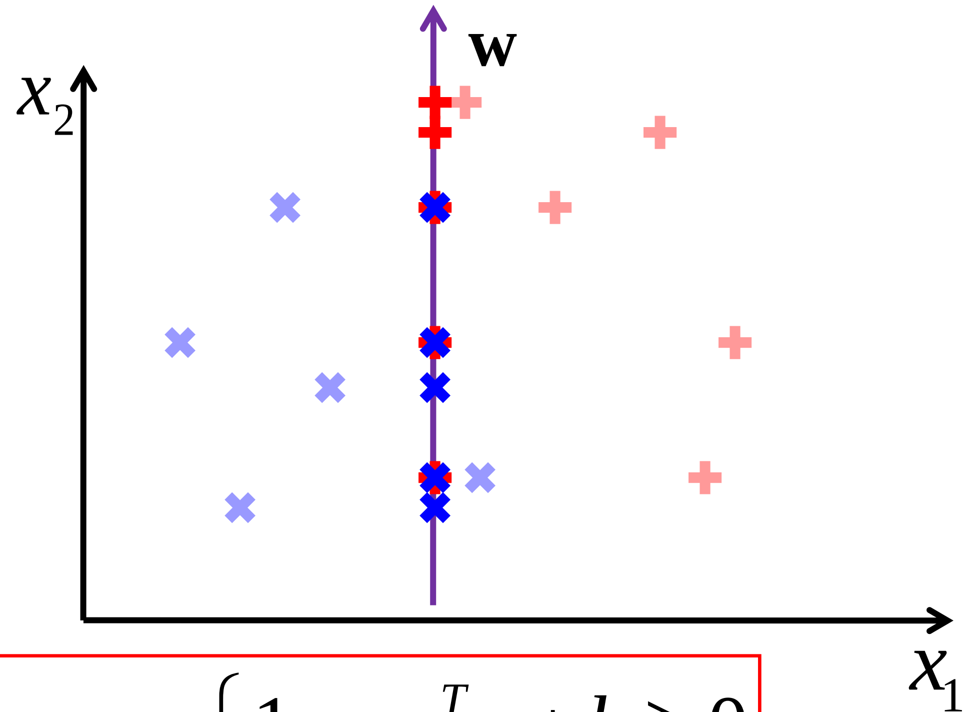
- **问题建模**
  - 平面上的分类：**转换成数轴上的分类**

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow$$

$$x' = w_1 x_1 + w_2 x_2$$

$$x' = \mathbf{w}^T \mathbf{x}$$

$$\mathbf{w} = \begin{pmatrix} w_1 & w_2 \end{pmatrix}^T$$



$$y = f(\mathbf{x}; \mathbf{w}, b) = \begin{cases} 1 & \mathbf{w}^T \mathbf{x} + b \geq 0 \\ -1 & \mathbf{w}^T \mathbf{x} + b < 0 \end{cases}$$

感知机模型

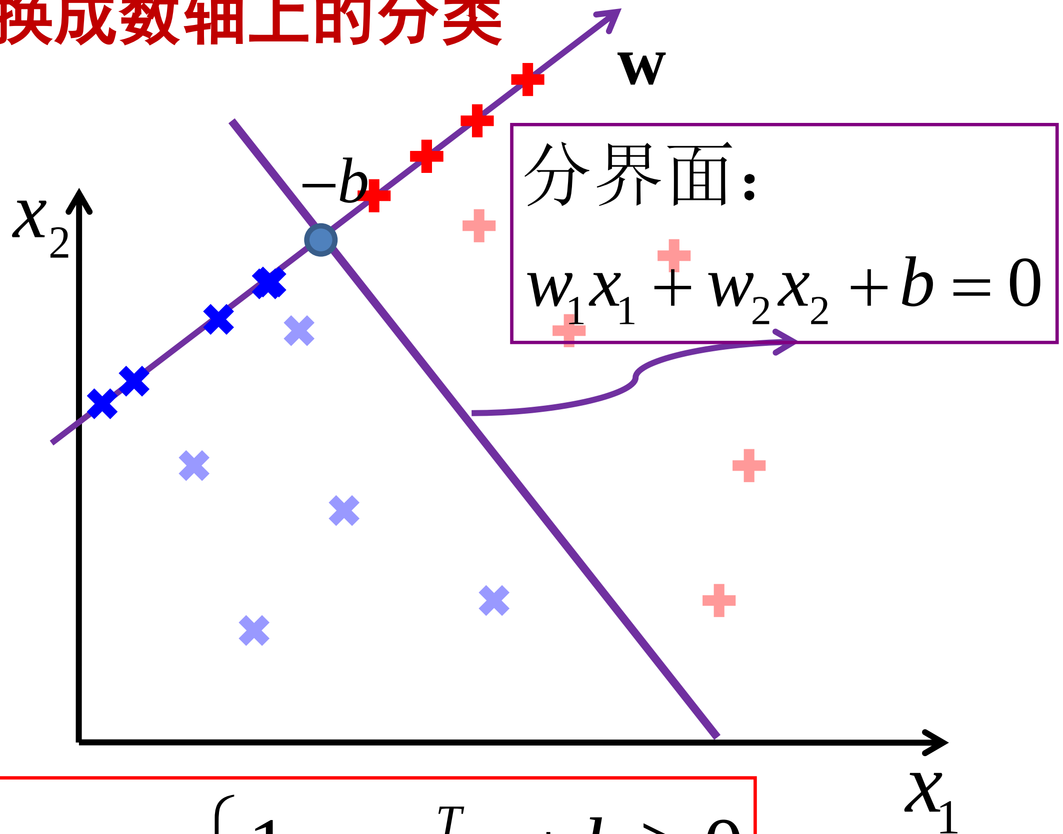
- **问题建模**
  - 平面上的分类：**转换成数轴上的分类**

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow$$

$$x' = w_1 x_1 + w_2 x_2$$

$$x' = \mathbf{w}^T \mathbf{x}$$

$$\mathbf{w} = \left(w_1 \quad w_2 \right)^T$$



$$y = f\left(\mathbf{x}; \mathbf{w}, b\right) = \begin{cases} 1 & \mathbf{w}^T \mathbf{x} + b \geq 0 \\ \text{-}1 & \mathbf{w}^T \mathbf{x} + b < 0 \end{cases}$$

感知机模型

- **问题建模**
  - 平面上的分类
    - 特征向量 $\mathbf{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^T$
    - 参数：w_1, w_2, b

变换到一维空间（数轴）：

$$\mathbf{x} \rightarrow x' = w_1 x_1 + w_2 x_2$$

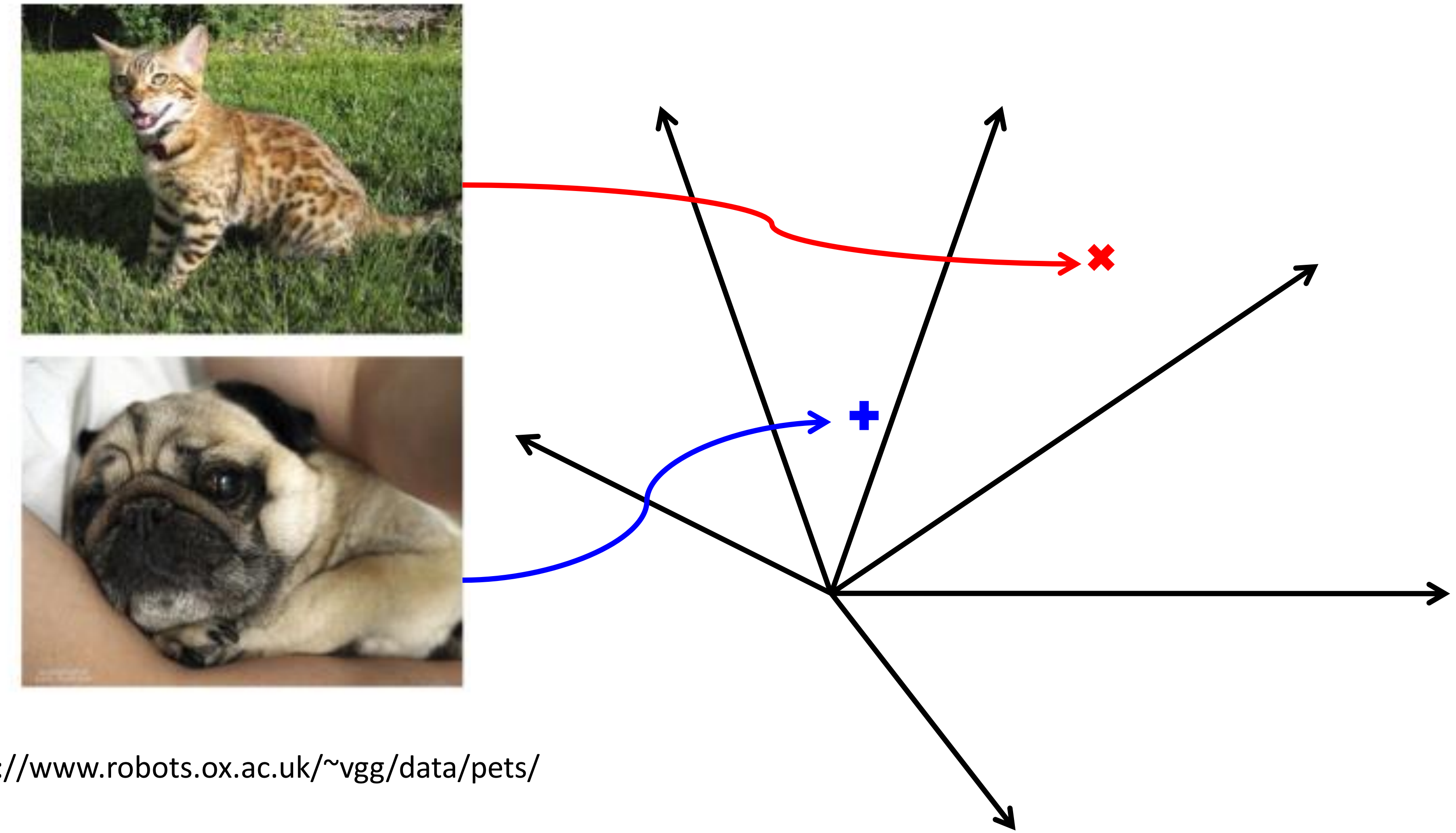
$$y = f(\mathbf{x}; \mathbf{w}, b) = \begin{cases} 1 & \mathbf{w}^T \mathbf{x} + b \geq 0 \\ -1 & \mathbf{w}^T \mathbf{x} + b < 0 \end{cases}$$

分界面：$\mathbf{w}^T \mathbf{x} + b = 0$

$\mathbf{w}$: 分界面的法向

感知机模型

- **问题建模**
  - 例3：高维空间上的分类



http://www.robots.ox.ac.uk/~vgg/data/pets/

感知机模型

安徽大学 ANHUI UNIVERSITY

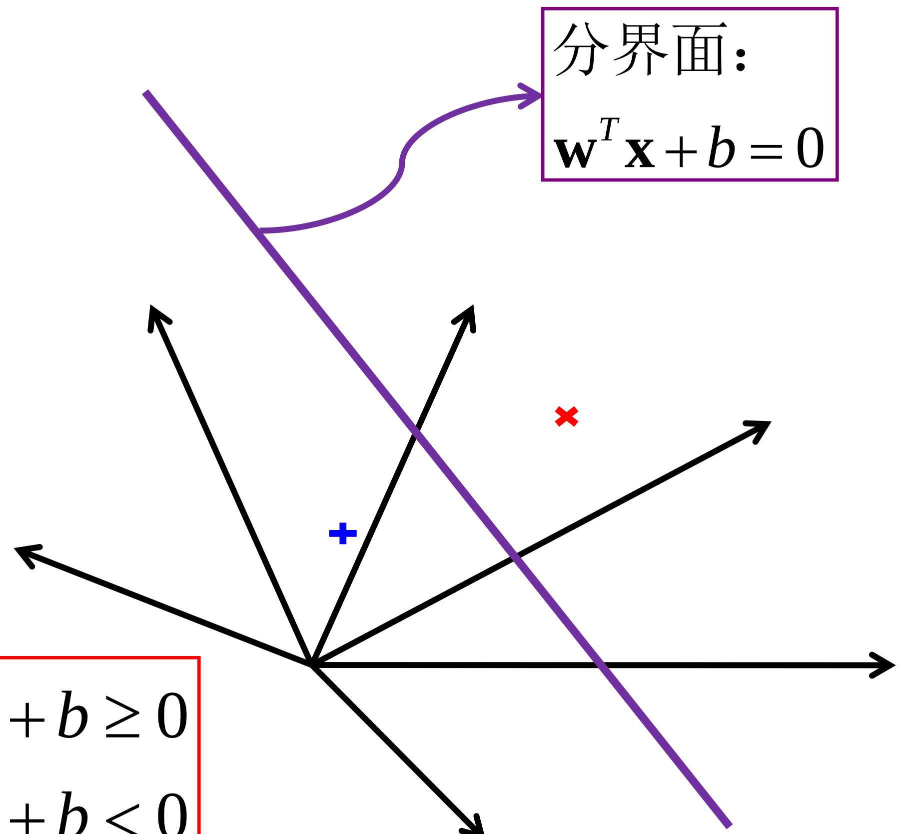
- **问题建模**
  - 高维空间上的分类：**转换成数轴上的分类**

$$\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}^T \rightarrow$$

$$x' = w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

$$x' = \mathbf{w}^T\mathbf{x}$$

$$\mathbf{w} = \begin{pmatrix} w_1 & w_2 & \cdots & w_n \end{pmatrix}^T$$



$$\boxed{y = f(\mathbf{x};\mathbf{w},b) = \begin{cases} 1 & \mathbf{w}^T\mathbf{x} + b \geq 0 \\ -1 & \mathbf{w}^T\mathbf{x} + b < 0 \end{cases}}$$

感知机模型

- **问题建模**
  - 高维空间上的分类
    - 特征向量 $\mathbf{x}=\begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}$
    - 参数：$\mathbf{w}=\begin{pmatrix} w_1 & w_2 & \cdots & w_n \end{pmatrix}^T, b$

$$\boxed{y=f(\mathbf{x};\mathbf{w},b)=\begin{cases} 1 & \mathbf{w}^T\mathbf{x}+b\ge 0 \\ -1 & \mathbf{w}^T\mathbf{x}+b<0 \end{cases}}$$

感知机模型

- **问题建模**
  - 感知机定义

假设输入空间（特征空间）是 $\mathcal{X} \in R^n$，输出空间是$\mathcal{Y} \in \{-1,1\}$，其中输入$\mathbf{x} \in \mathcal{X}$ 表示实例的特征向量，对应于输入空间（特征空间）的点，输出 $y \in \mathcal{X}$表示实例的类别，由输入空间到输出空间的函数：

$$f\left(\mathbf{x};\mathbf{w},b\right) = \operatorname{sgn}(\mathbf{w}^T\mathbf{x}+b)$$

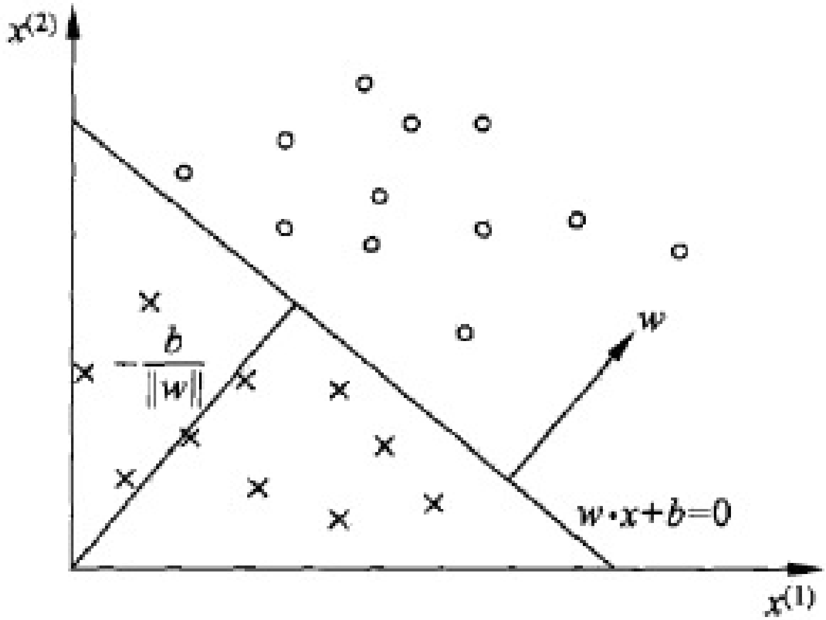
称为感知机。

符号函数：

$$\operatorname{sgn}\left(\mathbf{x}\right) = \begin{cases} 1 & \mathbf{x} \geq 0 \\ -1 & \mathbf{x} < 0 \end{cases}$$

感知机模型

- **问题建模**
  - 感知机几何解释
    - 线性方程：$\mathbf{w}^T\mathbf{x}+b=0$
    - 对于分类超平面，**w**为法向量，b为截距，用于分离正、负类

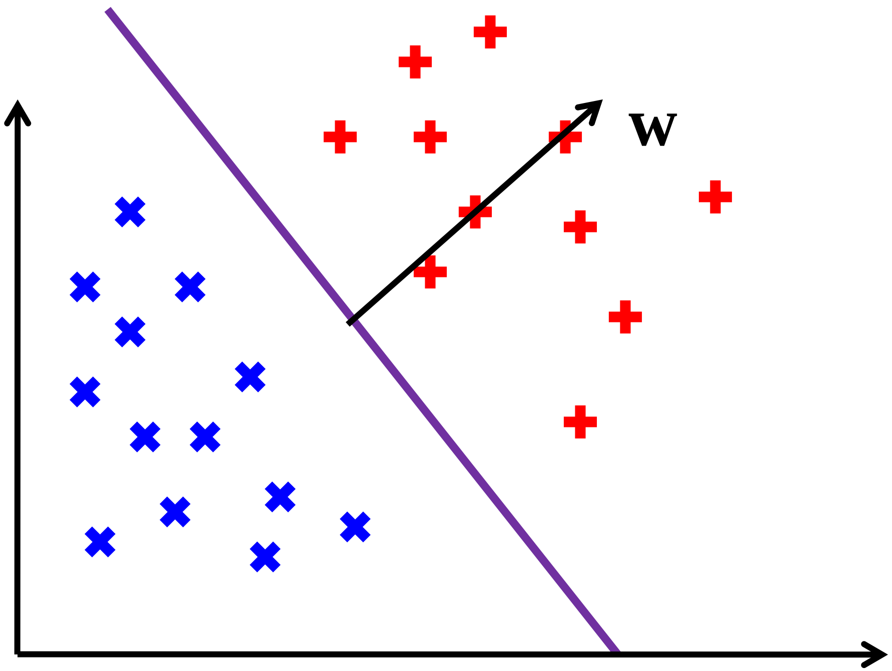


本节目录

- 背景
- 感知机模型
- 学习策略
- 学习算法
- 存在问题

学习策略

- **两个问题**
  - 准则：什么样的$\mathbf{w}$, b是好的
    - 经验风险最小化准则(ERM, Empirical Risk Minimization)
  - 算法：如何找到这个好的$\mathbf{w}$, b
    - 学习/优化算法(e.g. 梯度下降法)



学习策略

- **损失函数**
  - 自然选择：误分类点的数目，但损失函数不是$\mathbf{w}, b$ 连续可导，不宜优化
  - 另一选择：误分类点到超平面的距离
    - 超平面方程：$\mathbf{w}^T\mathbf{x}+b=0$
    - 样本数据点$\mathbf{x}$到该超平面的距离表示为：$|\mathbf{w}^{\mathrm{T}}\mathbf{x}+b|$
    - 判断$\mathbf{w}^T\mathbf{x}+b$ 与标记y符号是否一致：若一致，则分类正确；否则，分类错误
    - 可用 $y(\mathbf{w}^T\mathbf{x}+b)$ 替代$|\mathbf{w}^{\mathrm{T}}\mathbf{x}+b|$ 表示样本数据点到超平面的距离
    - 假设M是误分类样本点集合，则经验风险可以定义为误分类点到超平面的总距离：

$$L(\mathbf{w},b)=-\sum_{\mathbf{x}_i\in M} y_i(\mathbf{w}^T\mathbf{x}_i+b)$$

思考： $\mathbf{w}, b$如果进行同比例的缩放，超平面方程不变，但上述距离发生改变，如何解决？

学习策略

- **优化目标**
  - 最小化经验风险，即求解最优化问题：

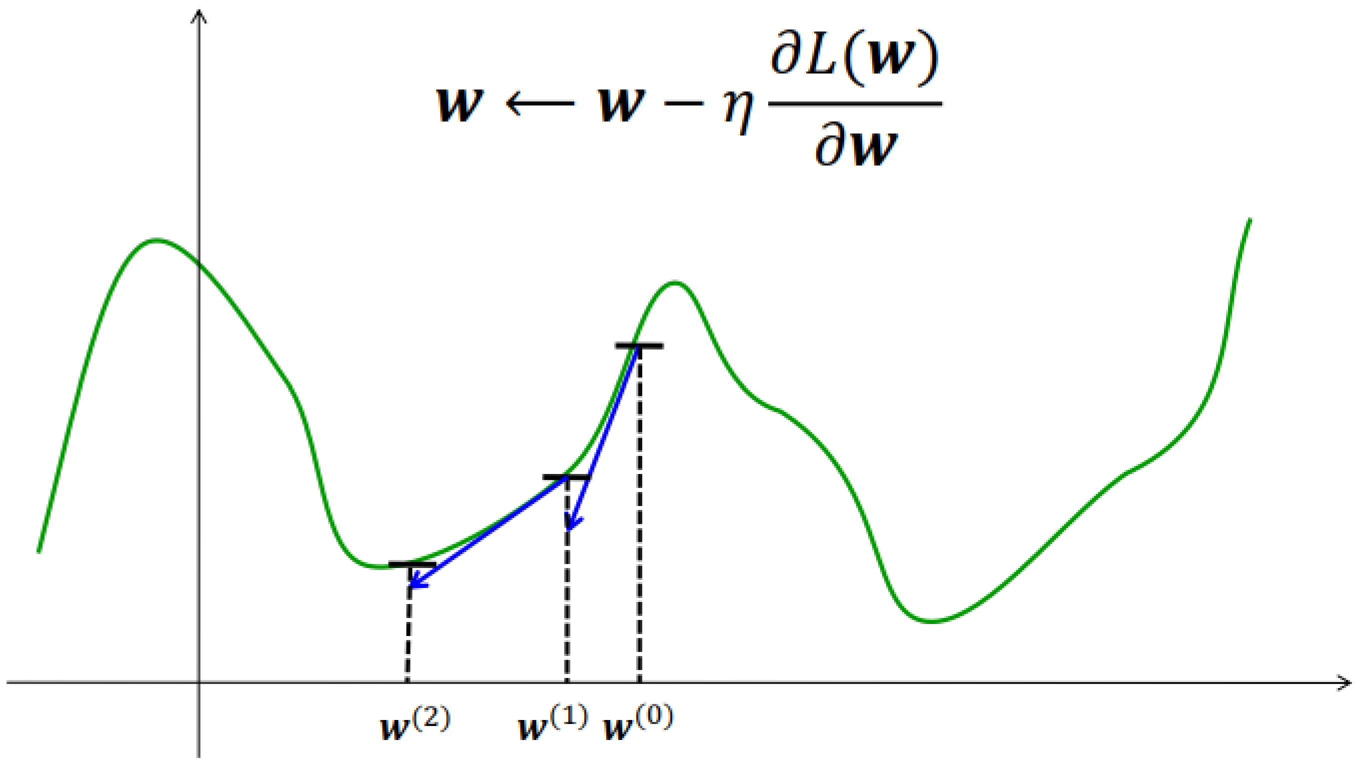
$$\min_{\mathbf{w},b} L(\mathbf{w},b) = \min_{\mathbf{w},b} -\sum_{\mathbf{x}_i \in M} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

本节目录

- 背景
- 感知机模型
- 学习策略
- 学习算法
- 存在问题

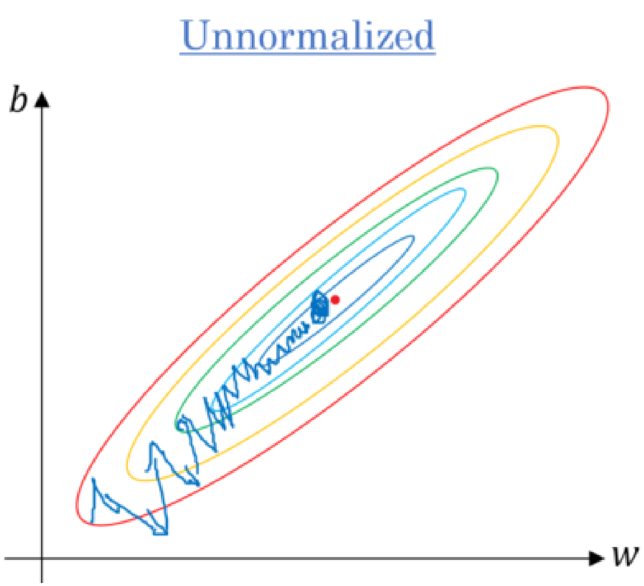
学习算法

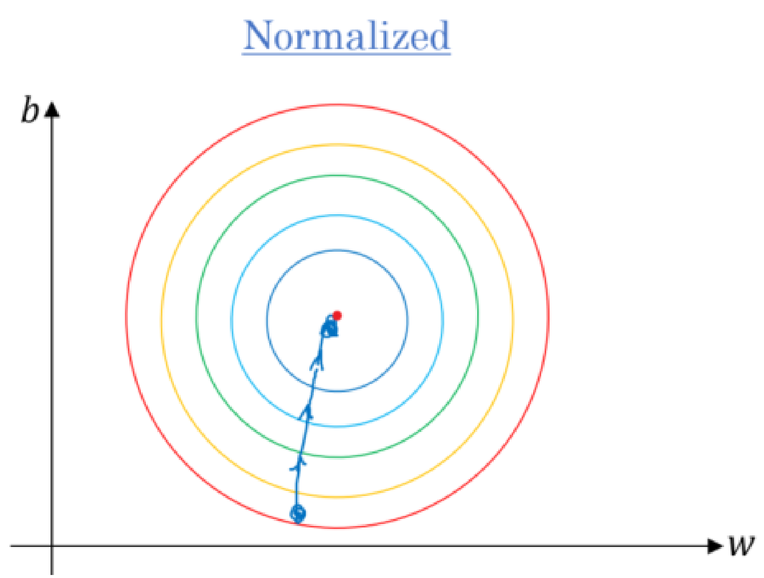
- **梯度下降法**
  - 算法原理



学习算法

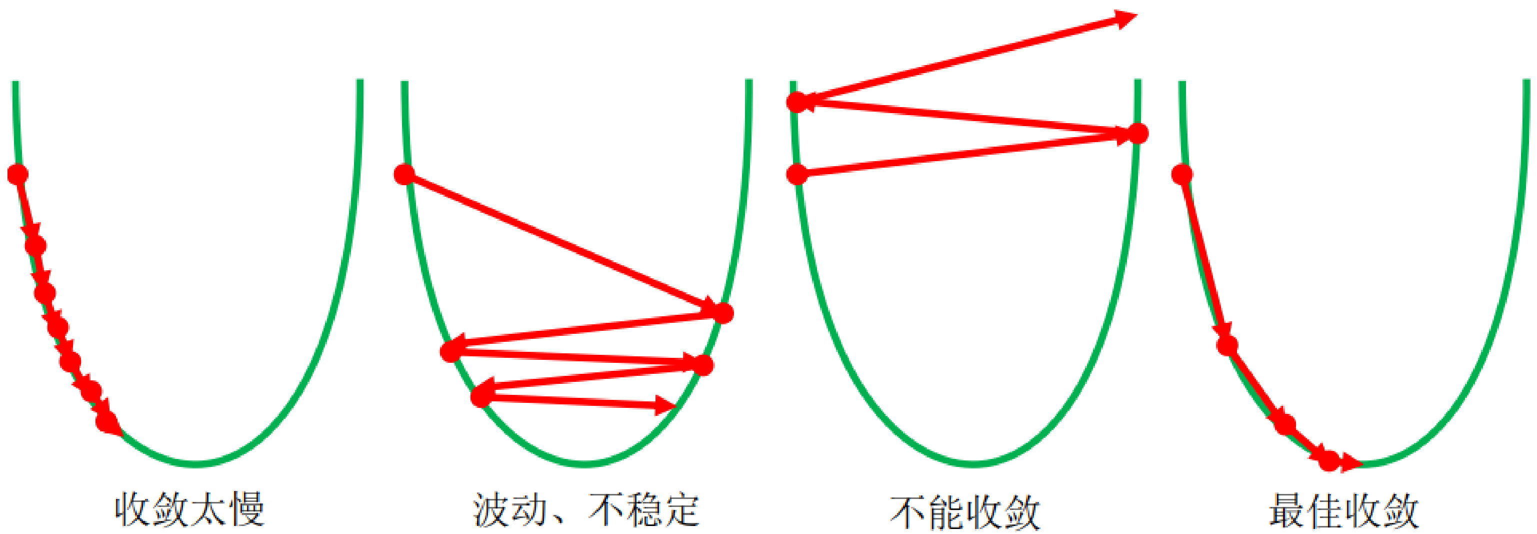
- **梯度下降法**
  - 几个影响因素
    - 特征缩放





学习算法

- **梯度下降法**
  - 几个影响因素
    - 特征缩放
    - 学习率



学习算法

- **感知机学习算法**
  - 优化目标

$$\min_{\mathbf{w},b} L(\mathbf{w},b) = \min_{\mathbf{w},b} -\sum_{\mathbf{x}_i \in M} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

  - 随机梯度下降法
    - 首先任意选择一个超平面，即初始化**w**, b，然后不断极小化目标函数,损失函数L的梯度：

$$\nabla_{\mathbf{w}} L(\mathbf{w},b) = -\sum_{\mathbf{x}_i \in M} y_i \mathbf{x}_i \qquad \nabla_b L(\mathbf{w},b) = -\sum_{\mathbf{x}_i \in M} y_i$$

    - 逐个选取误分类点更新：

$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i \qquad b \leftarrow b + \eta y_i$$

补充：批梯度下降法、随机梯度下降法和小批梯度下降法的联系与区别

学习算法

- **感知机学习算法**
  - 算法流程

输入：训练样本集$D = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$，最大迭代次数$T$

初始化：$\mathbf{w}^{(0)} = \mathbf{0}, b^{(0)} = 0$

$for(t = 1 : T)\{$

| 随机排列样本

| $flag = True$

| $for(i = 1 : n)\{$

| | $\hat{y} = h(\mathbf{x}^{(i)}; \mathbf{w}, b)$

| | $if(\hat{y} \neq y^{(i)})\{$

| | | $\mathbf{w} = \mathbf{w} + y^{(i)}\mathbf{x}^{(i)}$

| | | $b = b + y^{(i)}$

| | | $flag = False$

| | $\}$

| $\}$

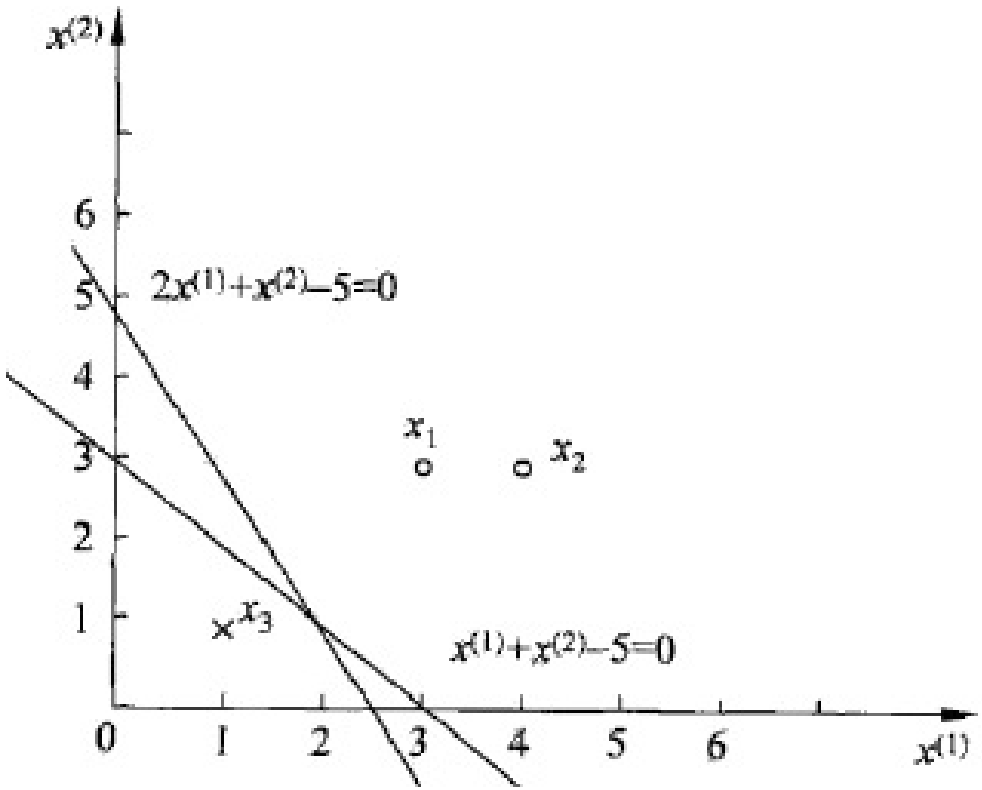
| $if(flag) \quad break$

$\}$

输出：$\mathbf{w}, b$

学习算法

- **感知机学习算法**
  - 练习：给定正例$\boldsymbol{x}_1=(3,3)$，$\boldsymbol{x}_2=(4,3)$，负例$\boldsymbol{x}_3=(1,1)$，请计算初值$\mathbf{w}=0$、$b=0$、学习率为1情况下的分类超平面



学习算法

- **感知机学习算法**
  - 练习：给定正例$\boldsymbol{x}_1$=(3,3)，$\boldsymbol{x}_2$=(4,3)，负例$\boldsymbol{x}_3$=(1,1)，请计算初值**w**=0、b=0、学习率为1情况下的分类超平面
    - 具体迭代过程如下表所示
    - 分类超平面：$x^{(1)}+x^{(2)}-3=0$
    - 感知机模型：$f(\boldsymbol{x})=x^{(1)}+x^{(2)}-3$

迭代次数	误分类点	w	b	$w \cdot x+b$
0		0	0	0
1	x_1	$(3,3)^{\mathrm{T}}$	1	$3x^{(1)}+3x^{(2)}+1$
2	x_3	$(2,2)^{\mathrm{T}}$	0	$2x^{(1)}+2x^{(2)}$
3	x_3	$(1,1)^{\mathrm{T}}$	−1	$x^{(1)}+x^{(2)}-1$
4	x_3	$(0,0)^{\mathrm{T}}$	−2	−2
5	x_1	$(3,3)^{\mathrm{T}}$	−1	$3x^{(1)}+3x^{(2)}-1$
6	x_3	$(2,2)^{\mathrm{T}}$	−2	$2x^{(1)}+2x^{(2)}-2$
7	x_3	$(1,1)^{\mathrm{T}}$	−3	$x^{(1)}+x^{(2)}-3$
8	0	$(1,1)^{\mathrm{T}}$	−3	$x^{(1)}+x^{(2)}-3$

学习算法

- **感知机学习算法**
  - 算法收敛性

可以证明经过有限次迭代可以得到一个将训练数据集完全正确划分的分类超平面及感知机模型（证明略）

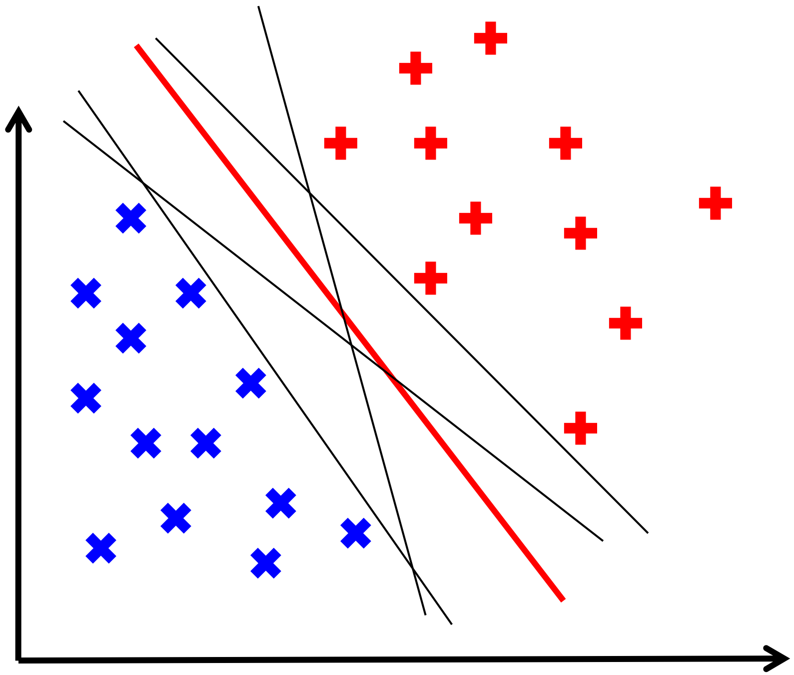
- 误分类的次数是有上界的，当训练数据集线性可分时，感知机学习算法原始形式迭代是收敛的

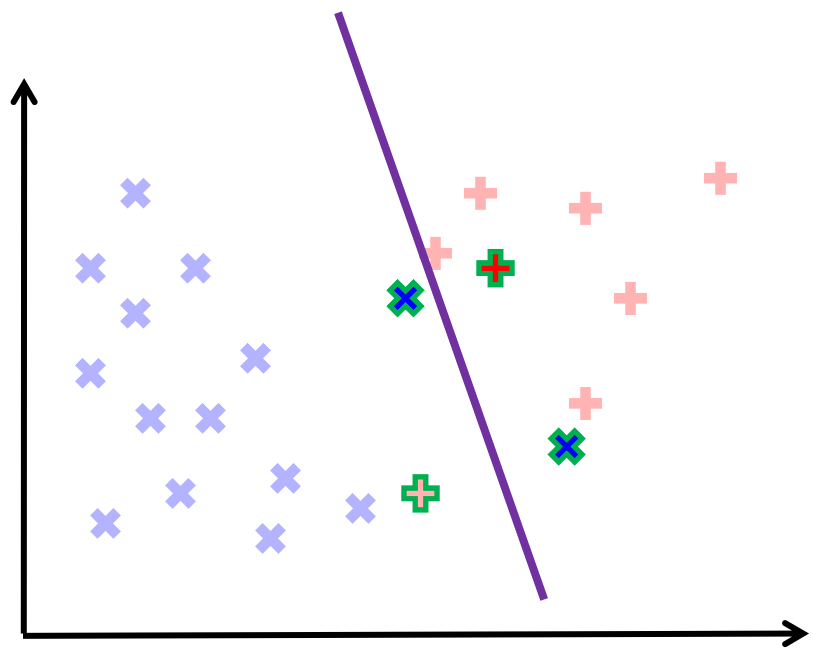
本节目录

- 背景
- 感知机模型
- 学习策略
- 学习算法
- 存在问题

存在问题

- 感知机算法存在许多解，既依赖于初值，也依赖迭代过程中误分类点的选择顺序
- 线性不可分数据集，迭代震荡





总结

- **背景**
- **感知机模型**
  - 数轴上的分类、平面上的分类、高维空间上的分类
  - 模型定义
- **学习策略**
  - 经验风险最小化
  - 误分类点到分类超平面的距离
- **学习算法**
  - 随机梯度下降法
- **存在问题**
  - 不唯一
  - 无法解决非线性问题

练习题

编程实现感知机学习算法，并设计简单的线性可分数据集和线性不可分数据集，观察优化过程。

思考题

针对感知机存在的问题，思考如下两个问题：

- 如何得到唯一的分类超平面？
- 如何解决非线性问题？