数据挖掘与机器学习

潘斌

panbin@nankai.edu.cn

范孙楼227

1

上节回顾

- 模型的评估与选择
- 损失函数
- 准确度的局限性
- PR曲线和ROC曲线
- 交叉验证

本节提要

- 偏差-方差分解
- 线性分类器
  - 垂直平分分类器
  - Fisher投影准则
  - 感知准则

实验1：手写LBP特征

- 第9周 实验课
- 给定10张图片，用LBP特征把它们的特征描述出来，并画成特征直方图
- 助教负责简单讲解python的使用、如何读图

- McNemar检验
  - 考察二学习器

两学习器分类差别列联表

算法 B	算法 A 正确	算法 A 错误
正确	e_{00}	e_{01}
错误	e_{10}	e_{11}

  - 若二学习器性能相同，则$e_{01}=e_{10}$

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \sim \chi^2(1)$$

修正系数

（《机器学习》有误）

- McNemar检验

		宣讲后		合计
		有必要	无必要	
宣讲前	有必要	28	6	34
	无必要	49	17	66
合计		77	23	100

- 计算统计量；查表找临界值

- Friedman检验和Nemenyi检验
  - 用于多算法的比较（H_0：所有算法性能相同）
  - 基于算法排序
    - 使用交叉验证法得到每个算法在每个数据集上的测试结果
    - 在每个数据集上根据测试性能好坏排序，并赋序值1，2，……。若算法性能相同，则平分序值
    - 计算平均序值

算法比较序值表

数据集	算法 A	算法 B	算法 C
D_1	1	2	3
D_2	1	2.5	2.5
D_3	1	2	3
D_4	1	2	3
平均序值	1	2.125	2.875

- 若算法性能相同，则平均序值应相同。

- k:算法个数
- N: 数据集个数
- r_i: 第i个算法的平均序数

$$\tau_{\chi^2} = \frac{k-1}{k} \cdot \frac{12N}{k^2-1} \sum_{i=1}^{k} \left(r_i - \frac{k+1}{2} \right)^2$$

$$= \frac{12N}{k(k+1)} \left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4} \right)$$

在 k 和 N 都较大时，服从自由度为 $k-1$ 的 χ^2 分布.

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1) - \tau_{\chi^2}} \sim F(k-1,(k-1)(N-1))$$

- 若H0被拒绝，则算法性能显著不同。需要进一步区分各算法。

Nemenyi 检验计算出平均序值差别的临界值域

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}},$$

若两个算法的平均序值之差超出了临界值域 CD，则以相应的置信度拒绝“两个算法性能相同”这一假设.

q_α 的值可以查看下表获得：

α	算法个数 k								
	2	3	4	5	6	7	8	9	10
0.05	1.960	2.344	2.569	2.728	2.850	2.949	3.031	3.102	3.164
0.1	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

- 例（续）
  - K = 3， N = 4， α = 0.05

算法比较序值表

数据集	算法 A	算法 B	算法 C
D_1	1	2	3
D_2	1	2.5	2.5
D_3	1	2	3
D_4	1	2	3
平均序值	1	2.125	2.875

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} \sim F(k-1,(k-1)(N-1))$$

$$\boxed{\tau_F = 24.429}$$

F 检验的常用临界值

$\alpha = 0.05$		
数据集个数 N	算法个数 k = 2	3
4	10.128	5.143
5	7.709	4.459
8	5.591	3.739
10	5.117	3.555
15	4.600	3.340
20	4.381	3.245

拒绝原假设！

Nemenyi 检验中常用的 q_α 值

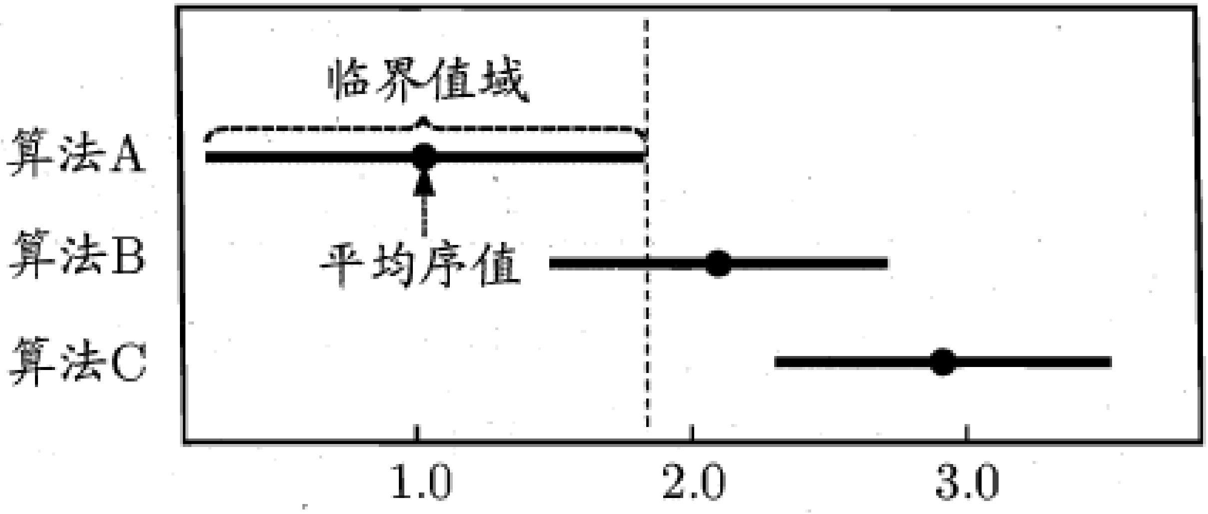
α	算法个数 k = 2	3	4
0.05	1.960	2.344	2.569
0.1	1.645	2.052	2.291

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

$$\boxed{CD = 1.657}$$

A、C性能显著不同，AB\BC否

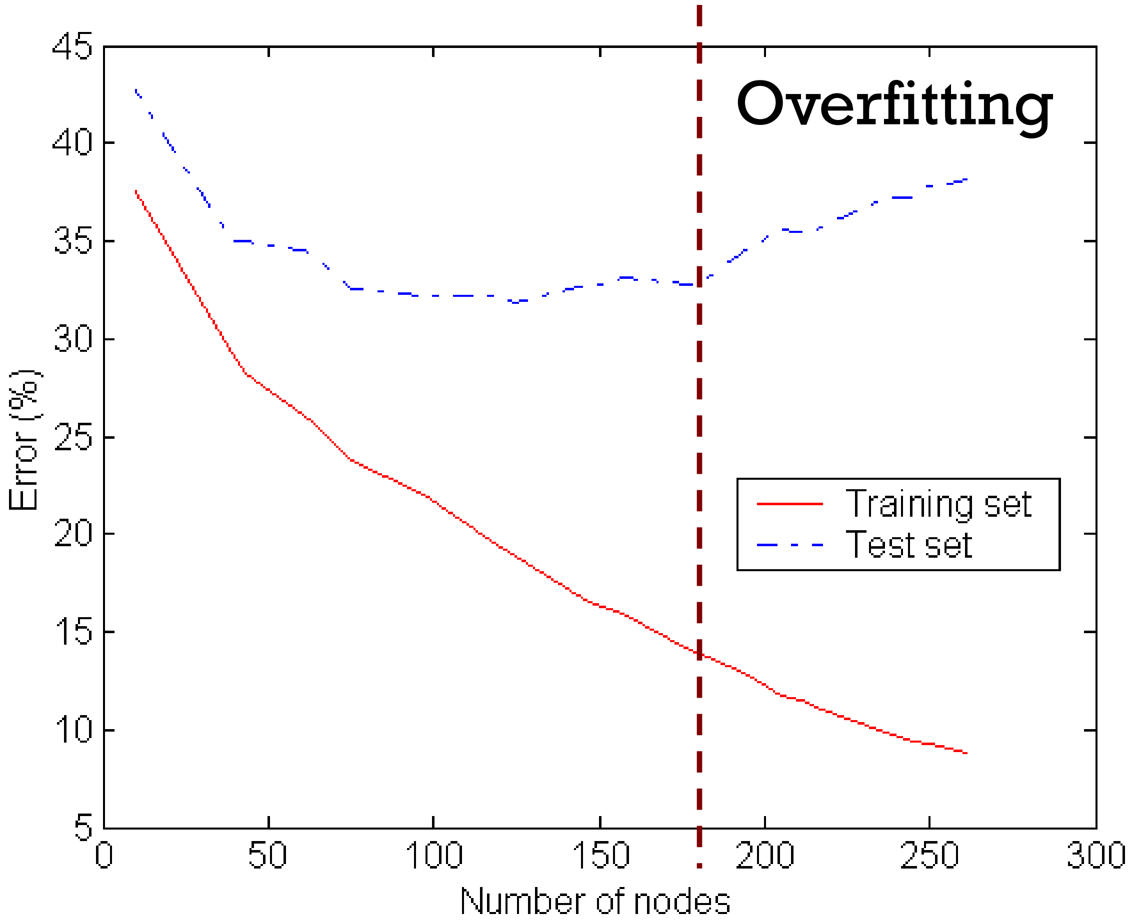
- Friedman检验图
  - 横轴：平均序值；纵轴：各算法
  - 点：每个算法的平均序值；横线段：临界值域
  - 若两个算法的横线段有交叠，则算法无显著差别；否则说明有显著差别。



Friedman 检验图

偏差-方差分解

- 过拟合与欠拟合



Underfitting: when model is too simple, both training and test errors are large

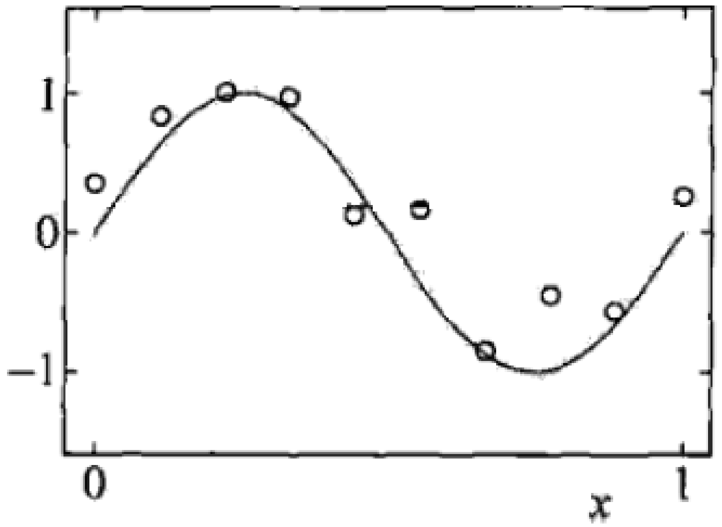
模型M1：锯齿∩绿色→树叶

模型M2：绿色→树叶

例 假设给定一个训练数据集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

在 M 次多项式函数中选择一个对已知数据以及未知数据都有很好预测能力的函数.



设 M 次多项式为

$$f_M(x,w)=w_0+w_1x+w_2x^2+\cdots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

求以下经验风险最小化：

$$L(w)=\frac{1}{2}\sum_{i=1}^{N}(f(x_i,w)-y_i)^2=\frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=0}^{M}w_jx_i^j-y_i\right)^2$$

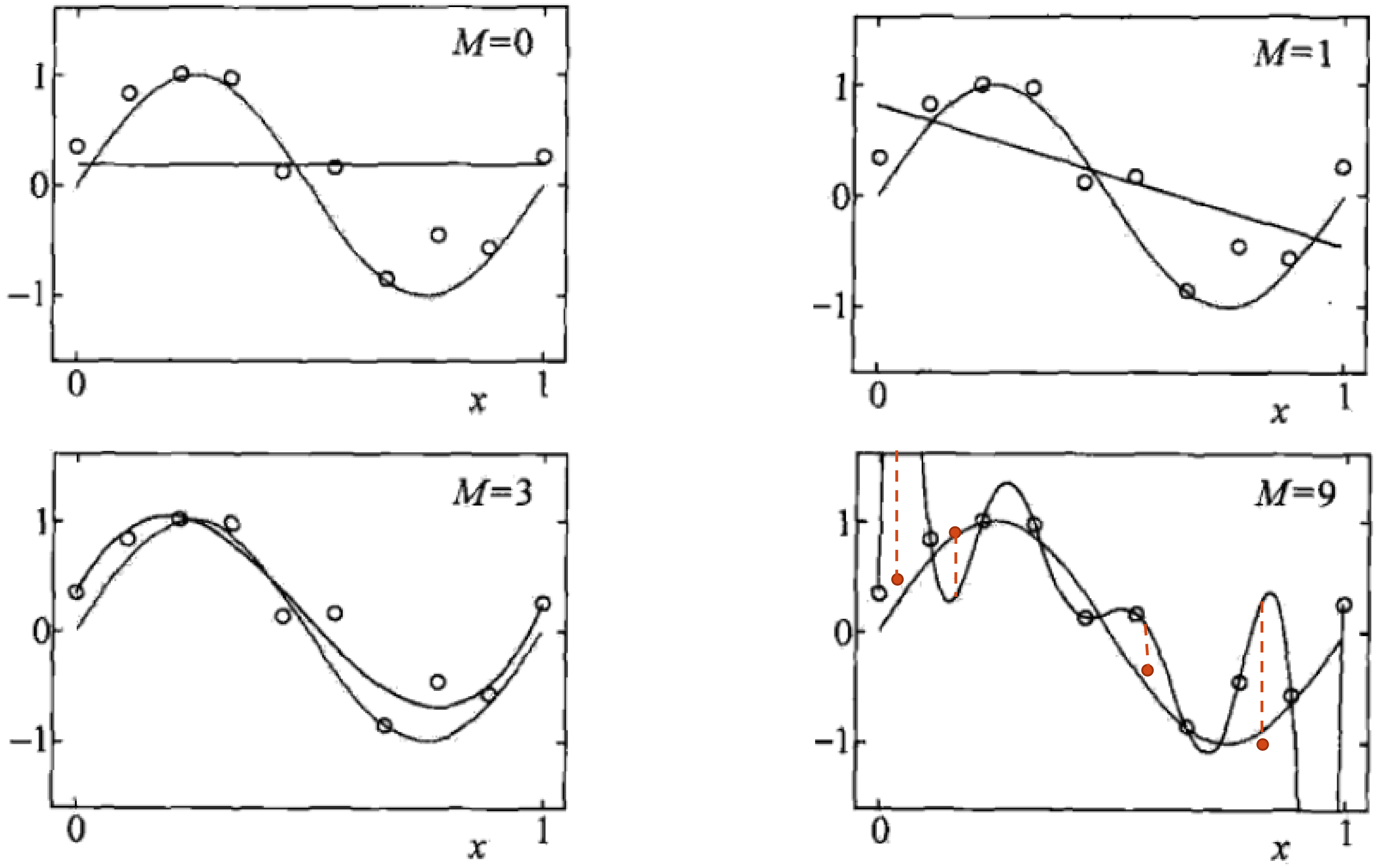
平方损失

对 w_j 求偏导数并令其为 0，可得

$$w_j = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^{\,j+1}}, \quad j = 0,1,2,\cdots,M$$

于是求得拟合多项式系数 $w_0^*, w_1^*, \cdots, w_M^*$.

$$f_M(x,w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



噪声（标注错误）期望为0

噪声与模型无关

- # 偏差-方差分解

符号	涵义
$\mathbf{x}$	测试样本
D	数据集（多个）
y_D	$\mathbf{x}$ 在数据集中的标记
y	$\mathbf{x}$ 的真实标记
f	训练集 D 学得的模型
$f(\mathbf{x}; D)$	由训练集 D 学得的模型 f 对 $\mathbf{x}$ 的预测输出
$\bar{f}(\mathbf{x})$	模型 f 对 $\mathbf{x}$ 的 **期望预测** 输出

假定$\mathbb{E}_D[y_D - y] = 0.$ $\quad \bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)]$

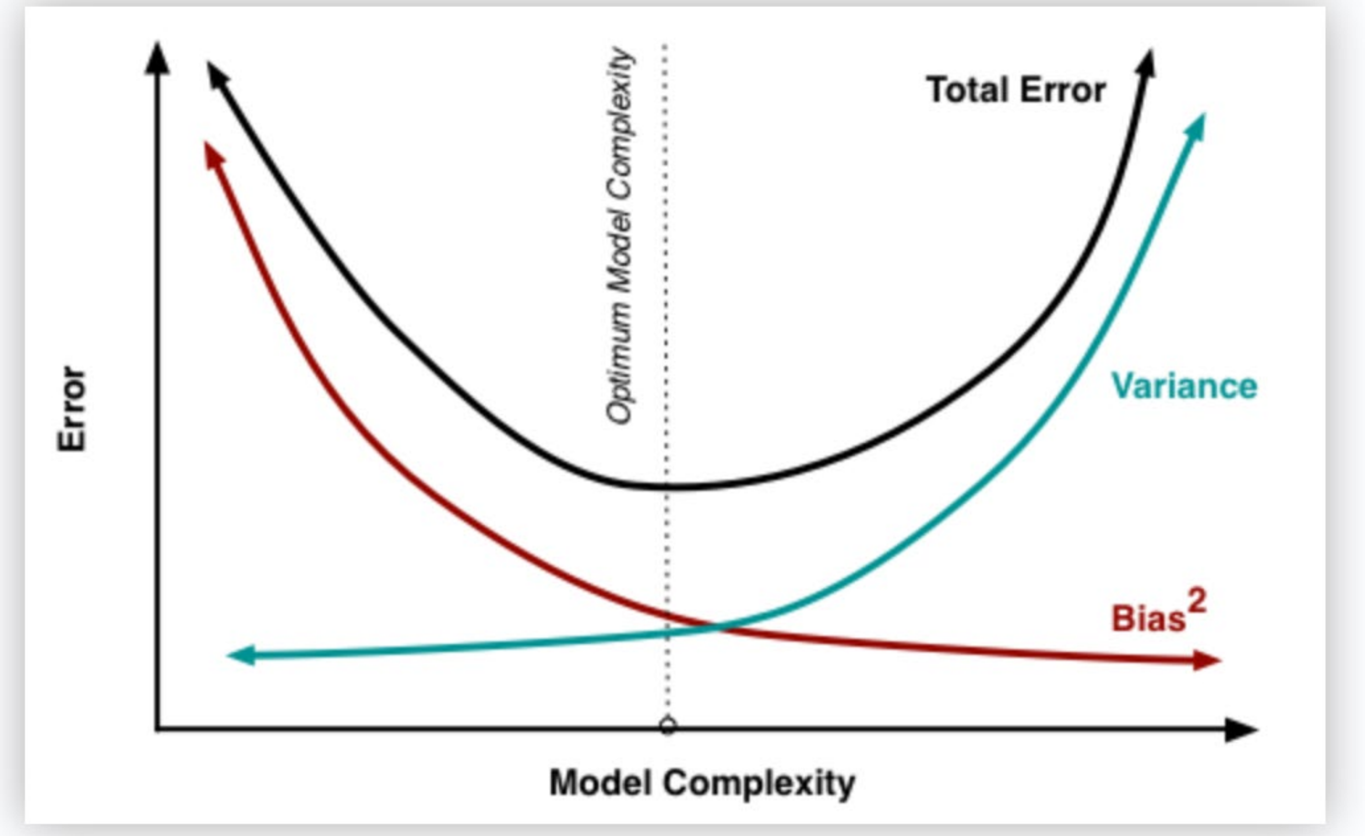
$$
\begin{aligned}
E(f; D) &= \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&\quad + \mathbb{E}_D\left[2\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y + y - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)^2\right] + \mathbb{E}_D\left[\left(y - y_D\right)^2\right] \\
&\quad + 2\mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)\left(y - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x}; D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y\right)^2\right] + \mathbb{E}_D\left[\left(y - y_D\right)^2\right] \\
&= Variance + Bias + Noise
\end{aligned}
$$

$$Err(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \left(\bar{f}(\boldsymbol{x}) - y\right)^2 + \mathbb{E}_D\left[(y_D - y)^2\right]$$

- variance
- bias^2
- noise

欠拟合

- 拟合能力不足
- 偏差占主导
- 训练数据的扰动不足以使学习器发生变化
- 增加模型参数



过拟合

- 拟合能力过强
- 方差占主导
- 训练数据的轻微扰动会导致模型变化
- 减少参数，正则化
- 集成学习

线性分类器

- 1　线性分类器基础
- 2　垂直平分分类器
- 3　Fisher投影准则
- 4　感知准则
- 5　最小错分样本数准则
- 6　最小平方误差准则

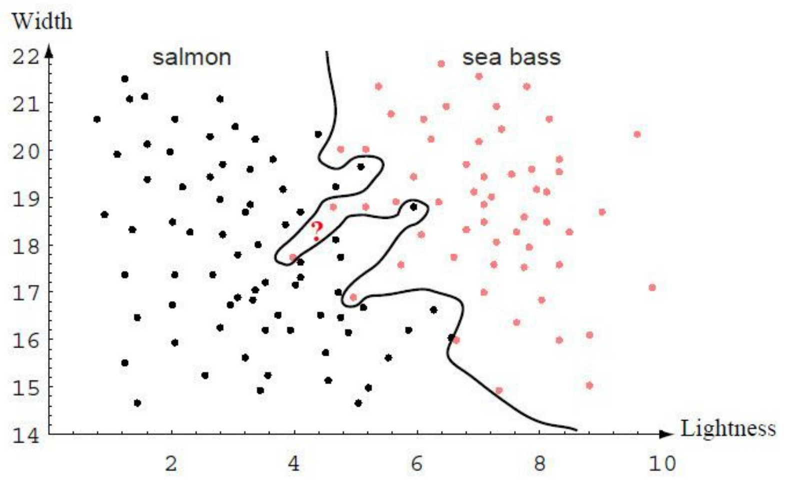
1　线性分类器基础

- 1.1　数学基础知识
- 1.2　线性分类器概念
- 1.3　线性判别函数
- 1.4　增广变换
- 1.5　相关概念归纳
- 1.6　线性分类器设计概述

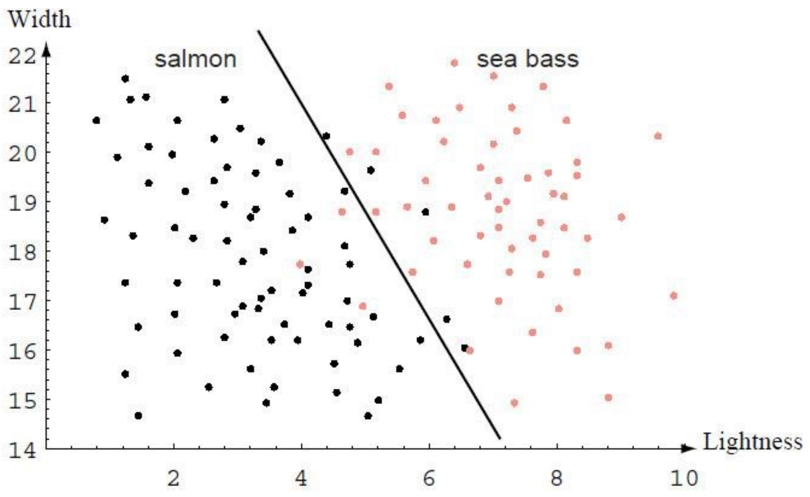
1.1　数学基础知识

- **相关的数学基础包括**
  - 矩阵
  - 向量
  - 矩阵和向量的转置
  - 向量运算
  - 矩阵运算

1.2 线性分类器概念



Width
salmon
sea bass
?
Lightness
22
21
20
19
18
17
16
15
14
2
4
6
8
10



Width
salmon
sea bass
Lightness
22
21
20
19
18
17
16
15
14
2
4
6
8
10

1.2 线性分类器概念

- [线性分类器] 对于两类的分类问题，采用线性判别函数划分特征空间（即采用直线或平面等将两类样本在特征空间中的区域划分开），这样的分类器是线性分类器。

- 线性分类器特点：特征空间一分为二，适合于解决两类的分类问题

1.3 线性判别函数

- **两类二维问题**
  - C——类别数，D——维数，N——样本数
  - C = 2，D =2

  - 直线方程
    - 代数形式$w_1x_1 + w_2x_2 + w_0 = 0$
    - 向量形式$w^Tx + w_0 = 0$

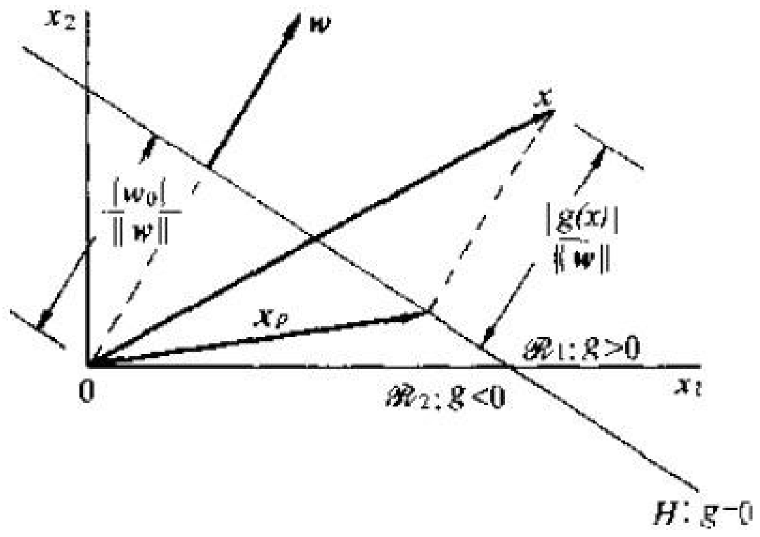
  - 定义线性判别函数
    - $g(x) = w^Tx + w_0$

1.3 线性判别函数

- **两类多维问题**
  - C = 2, D任意
  - 定义线性判别函数
    - $g(x) = w^T x + w_0$
    - w ——权向量
    - w_0——阈值权

1.3 线性判别函数

- **线性判别函数的几何性质**
  - 法向量方向
  - 原点距离



1.4 增广变换

- **线性判别函数的增广变换**
  - 定义增广变换

$$g(\boldsymbol{x}) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=0}^{d} a_i y_i = \boldsymbol{a}^T \boldsymbol{y}$$

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix}, \boldsymbol{a} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} w_0 \\ \boldsymbol{w} \end{bmatrix},$$

1.4 增广变换

- **线性判别函数的增广变换**
  - 则线性判别函数为
    - $g(x) = a^T y$

    - y ——增广样本向量
    - a ——增广（或广义）权向量

1.4 增广变换

- **线性判别函数的增广变换**
  - 增广变换的特点
    - 维数增加了一维：$D_G = D + 1$
    - 样本向量实际还是位于原D维子空间中
    - 样本间欧式距离保持不变
    - $a^T y = 0$ 是过原点的超平面H_G

1.5　相关概念归纳

- **概念回顾**
  - 线性判别函数，记为g(x)
  - 线性决策面，记为H
  - 线性决策面方程，令g(x) = 0

1.5 相关概念归纳

- **线性决策面法向量方向**
  - 线性决策面将特征空间分为两个区域。
    - 其中法向量方向区域称为正侧区域（简称正侧）
    - 法向量反方向的区域称为负侧区域（简称负侧）。
  - 设计时，通常使
    - 正侧对应ω_1类（甲类或A类）
    - 负侧对应ω_2类（乙类或B类）。

1.5 相关概念归纳

- **决策规则**
  - 已知判别函数
    - $g(x) = w^T x + w_0$， 或$g(x) = a^T y$
  - 则决策规则为
    - 对于未知样本x，若$g(x) > 0$，则x决策为ω_1类
    - 若$g(x) < 0$，则x决策为ω_2类

1.6　线性分类器设计概述

- **线性分类器的理论设计**
  - 设计线性分类器
  - ↓
  - 设计决策规则
  - ↓
  - 设计线性判别函数$g(x) = w^T x + w_0$，或$g(x) = a^T y$
  - ↓
  - 求解权向量w和阈值权w_0，或增广权向量a

1.6　线性分类器设计概述

- **线性分类器设计常规步骤**
  - 给定类别已知的样本——训练样本集
  - ↓
  - 选择一个准则函数J，其值反映分类器性能（分类结果优劣）
  - ↓
  - 采用求最优解的数学方法求准则函数J的极值解，从而求得权向量w和阈值权w_0，或增广权向量a

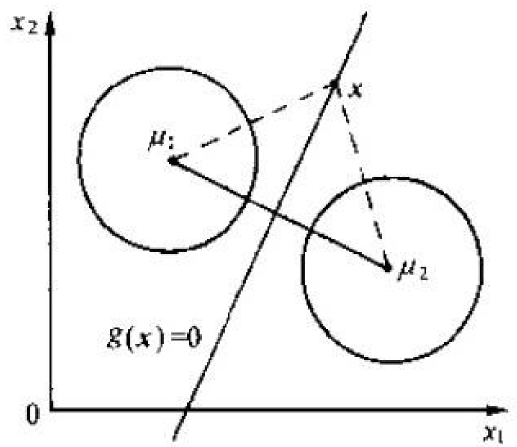
2　垂直平分分类器

- 2.1　问题与思路
- 2.2　垂直平分形式
- 2.3　最小距离形式
- 2.4　实例
- 2.5　特点

2.1　问题与思路

- 垂直平分分类器又称为最小距离分类器。

- 设计思路
  – 基于两类样本均值点作垂直平分线



2.1 问题与思路

- 已知
  - 给定类别已知的训练样本集Z有N个样本，
    - 其中ω_1类样本有N_1个，样本集用Z_1表示；
    - ω_2类样本有N_2个，样本集用Z_2表示；
  - 显然
    - $N_1 + N_2 = N$
    - $Z_1 + Z_2 = Z$
- **试求垂直平分分类器**

2.2　垂直平分形式

- **判别函数与决策面方程**
  - 对于两类二维问题
  - C = 2，D =2
  - 垂直平分线性判别函数
    - $g(x) = w^T x + w_0$
  - 垂直平分直线方程
    - $g(x) = 0$ 即 $w^T x + w_0 = 0$

2.2 垂直平分形式

- **求解权向量与阈值权**
  - 先求均值向量
    - m_1 和 m_2
  - 利用垂直几何关系，设权向量
    - $w = (m_1 - m_2)$
  - 则直线方程为
    - $(m_1 - m_2)^T x + w_0 = 0$

（注意正侧在m_1这边）

2.2 垂直平分形式

- **求解权向量与阈值权**
  - 再利用平分几何关系，中点x_0在直线上
    - $x_0 = (m_1 + m_2) / 2$
  - 代入方程求得
    - $w_0 = -(m_1 - m_2)^T (m_1 + m_2) / 2$

2.2 垂直平分形式

- **最终结果**
  - 线性判别函数
    - $g(x) = (m_1 - m_2)^T x - (m_1 - m_2)^T (m_1 + m_2) / 2$
    - $= (m_1 - m_2)^T (x - (m_1 + m_2) / 2)$
  - **决策面方程**
    - $(m_1 - m_2)^T (x - (m_1 + m_2) / 2) = 0$

2.2 垂直平分形式

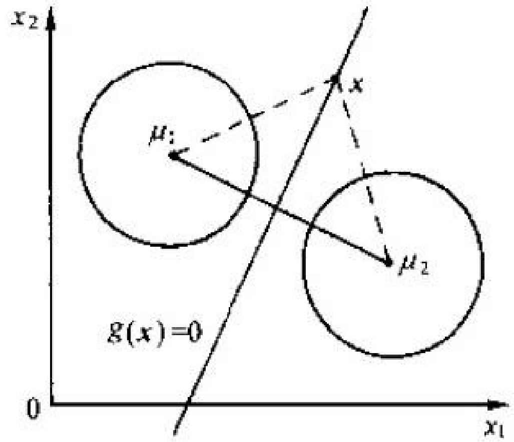
- **决策规则**
  - 已知垂直平分判别函数
    - $g(x) = (m_1 - m_2)^T (x - (m_1 + m_2)/2)$
  - 垂直平分决策规则为
    - 对于未知样本x，若$g(x) > 0$，则x决策为ω_1类
    - 若$g(x) < 0$，则x决策为ω_2类

2.2 垂直平分形式

- **判别函数与决策面方程**
  - 很容易推广到两类多维问题
  - C = 2, D任意
  - 垂直平分线性判别函数
    - $g(x) = w^T x + w_0$
  - 垂直平分决策面方程
    - $g(x) = 0$ 即 $w^T x + w_0 = 0$

2.3 垂直平分分类器的最小距离形式

- **最小距离等价形式的由来**
  - 定义欧式距离（非线性）为判别函数
    - $G_1(x) = d_1(x) = \|x - m_1\|$
    - $G_1(x) = d_2(x) = \|x - m_2\|$



2.3 最小距离形式

- **决策规则**
  - 等价的最小距离决策规则为
    - 对于未知样本x，若$d_1(x) < d_2(x)$，则x决策为ω_1类
    - 若$d_1(x) > d_2(x)$，则x决策为ω_2类

2.4 实例

- **已知**
  - 甲类：$[0\ \ 3]^T$、$[2\ \ 4]^T$、$[1\ \ 3]^T$、$[2\ \ 3]^T$、$[0\ \ 2]^T$
  - 乙类：$[4\ \ 1]^T$、$[3\ \ 2]^T$、$[2\ \ 1]^T$、$[3\ \ 0]^T$、$[3\ \ 1]^T$
- **试问**
  - 待分类样本为$x = [5\ \ 0]^T$，问x应决策为哪一类？

2.5 特点

- **最小距离分类器的主要特点**
  - 解决两类分类问题的线性分类器
  - 原则上对样本集无特殊要求
  - 未采用准则函数求极值解（非最佳决策）
  - 算法最简单，分类器设计最容易

3　Fisher投影准则

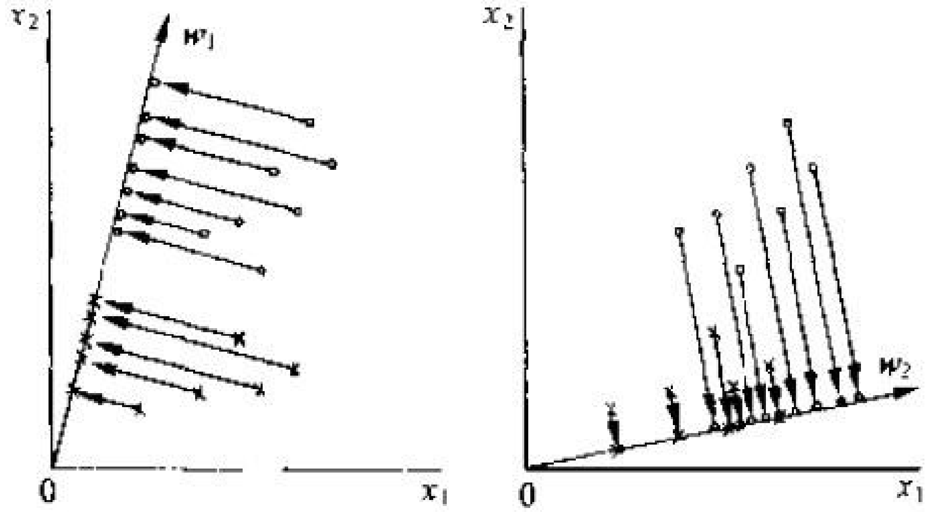
- 3.1　问题与思路
- 3.2　Fisher准则函数
- 3.3　准则函数化简
- 3.4　求极值解
- 3.5　特点
- 3.6　后续研究

3.1 问题和思路

- **原因**
  - 高维问题——特征个数太多
  - （经典理论）分类器设计困难
  - 分类困难

3.1 问题和思路

- **设计思路**
  - 通过投影对高维分类问题降维
  - Fisher将高维特征空间的样本投影到一维直线上



3.1　问题和思路

- **问题**
  - 已知C = 2，D维分类问题的样本集
  - 设投影向量为p
  - 则一维投影方程为$y = p^T x$
  - 求最佳投影向量p（的方向）

3.2 Fisher准则函数

- **Fisher定义的准则函数**
  - 定义各类均值m_1和m_2
  - 定义各类离散度S_1和S_2
  - 定义总离散度$S_W = S_1 + S_2$
  - 定义类间离散度S_B

1. 在 d 维 X 空间

(1) 各类样本均值向量 $\boldsymbol{m}_i$

$$\boldsymbol{m}_i = \frac{1}{N_i}\sum_{x \in \mathscr{X}_i} \boldsymbol{x}$$

(2) 样本类内离散度矩阵 S_i 和总类内离散度矩阵 S_w

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

$$S_w = S_1 + S_2$$

(3) 样本类间离散度矩阵 S_b①

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

2. 在一维 Y 空间

(1) 各类样本均值 $\tilde{m}_i$

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y \in \mathscr{Y}_i} y$$

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

$$\tilde{S}_i^2 = \sum_{y \in \mathscr{Y}_i} (y - \tilde{m}_i)^2$$

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

(3) 样本的类间离散度：

$$(\tilde{m}_1 - \tilde{m}_2)^2$$

3.2 Fisher准则函数

- **Fisher定义的准则函数**
  - 定义Fisher投影准则

$$J_F(p) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^{\,2} + \tilde{S}_2^{\,2}}$$

- **Fisher投影准则的物理含义**
  - 投影后异类样本尽量远离
  - 投影后同类样本尽量靠近

3.3 准则函数化简

- **化简Fisher准则函数**
  - 分子的化简

$$\tilde{m}_i = \frac{1}{N_i}\sum_{y\in \mathscr{Y}_i} y = \frac{1}{N_i}\sum_{x\in \mathscr{X}_i} w^T x$$

$$= w^T\left(\frac{1}{N_i}\sum_{x\in \mathscr{X}_i} x\right) = w^T m_i$$

分子便成为

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

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

3.3　准则函数化简

- **化简Fisher准则函数**
  - 分母的化简

$$\tilde{S}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2 = \sum_{x \in \mathcal{X}_i} (w^T x - w^T m_i)^2$$

$$= w^T \Big[\sum_{x \in \mathcal{X}_i} (x - m_i)(x - m_i)^T \Big] w = w^T S_i w$$

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

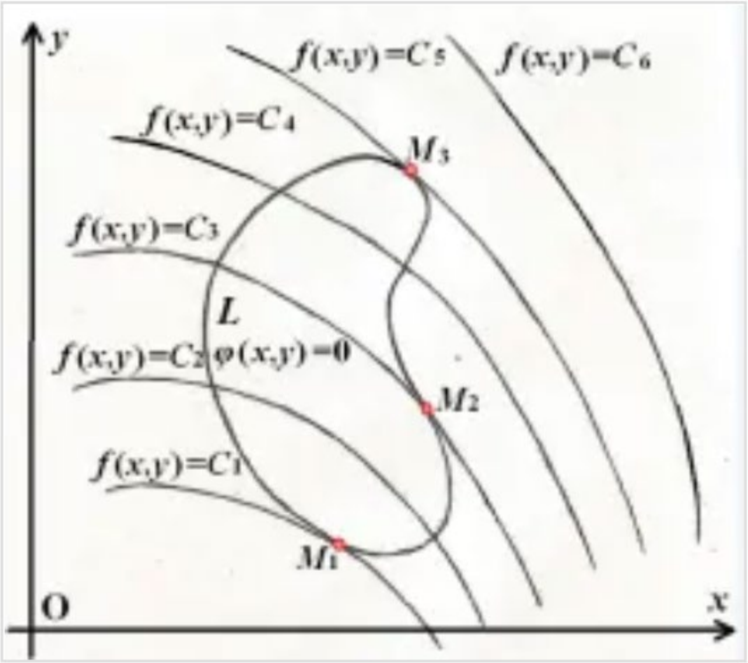
3.3 准则函数化简

- **Fisher准则函数**
  - 化简的结果

$$J_F(p) = \frac{p^T S_b p}{p^T S_W p}$$

3.4 求极值解

曲线 L 为约束条件 $\varphi(x,y)=0$， $f(x,y)=C$ 为目标函数的等值线



- **求Fisher函数的极值解**
  - 采用Lagrange乘子法求极值
    - 等式约束条件：令分母为常数
    - 目标函数：分子

$$w^T S_w w = c \neq 0$$

定义 Lagrange 函数为

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

式中 λ 为 Lagrange 乘子。将式(4-28)对 w 求偏导数，得

$$\frac{\partial L(w,\lambda)}{\partial w} = S_b w - \lambda S_w w$$

令偏导数为零，得

$$S_b w^* - \lambda S_w w^* = 0$$

即

$$S_b w^* = \lambda S_w w^*$$

拉格朗日函数

$$F(x,y,\lambda) = f(x,y) + \lambda\varphi(x,y)$$

3.4 求极值解

- **求Fisher函数的极值解**

其中 w^* 就是 $J_F(w)$ 的极值解。因为 S_w 非奇异，式(4-29)两边左乘 S_w^{-1}，可得

$$S_w^{-1}S_b w^* = \lambda w^* \tag{4-30}$$

解式(4-30)为求一般矩阵 $S_w^{-1}S_b$ 的本征值问题。但在我们这个特殊情况下，利用式(4-19)S_b 的定义，式(4-30)左边的 $S_b w^*$ 可以写成

$$S_b w^* = (m_1 - m_2)(m_1 - m_2)^T w^* = (m_1 - m_2)R$$

式中

$$R = (m_1 - m_2)^T w^*$$

为一标量，所以 $S_b w^*$ 总是在向量 $(m_1 - m_2)$ 的方向上。由于我们的目的是寻找最好的投影方向，w^* 的比例因子对此并无影响，因此，从式(4-30)可得

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

从而可得

$$w^* = \frac{R}{\lambda}S_w^{-1}(m_1 - m_2) \tag{4-31}$$

忽略比例因子 R/λ，得

$$w^* = S_w^{-1}(m_1 - m_2) \tag{4-32}$$

3.4 求极值解

- **求Fisher函数的极值解**
  - 极值解（极大值）

$$p^* = S_W^{-1}(m_1 - m_2)$$

3.5 特点

- **Fisher投影的特点**
  - 解决两类问题的线性投影
  - 原则上对样本集无特殊要求（S_w矩阵可逆）
  - 采用Fisher投影准则函数求极值解（最佳决策）
  - 分类器设计较容易

3.6 后续研究

- 1936 年，Fisher发表经典论文，提出投影准则。Wilks和Duda分别提出判别向量集概念，由判别向量集构成子空间，对原始样本在子空间中的投影向量进行分类判别。
- 1970年，Sammon提出基于Fisher准则的最佳判别平面。Foley和Sammon提出采用正交条件下的最佳判别向量集进行特征提取的方法。
- 1988年，Duchene等给出多类情况最佳判别向量集的计算公式。
- ……
- Linear Discriminent Analysis (LDA)

实验2：垂直平分分类器

- 给定训练数据，学习得到一个垂直平分分类器
- 对测试样本进行分类

- Python编程实现

4　感知准则

- 4.1　样本集线性可分
- 4.2　解向量和解区
- 4.3　感知准则函数
- 4.4　求极值解
- 4.5　特点
- 4.6　后续研究

4.1 样本集线性可分

- **样本集的线性可分性**
  - [线性可分] 若训练样本集可以被某个线性分类器完全正确分类，则该样本集是线性可分的。
  - 样本集是线性可分的——至少存在一个权向量，能将该样本集中的每个样本都正确分类；
  - 否则就是线性不可分的（异或问题）。

4.1 样本集线性可分

- **问题**
  - 已知C = 2，D维分类问题的样本集（其它略）
  - 设该样本集是线性可分的
  - 提出感知准则（因此称为感知器）
  - 求能够对样本集正确分类的解（某个线性分类器）
  - 感知器用来解决线性可分样本集分类问题

4.1　样本集线性可分

- **线性可分性样本集的规范化**
  - 感知准则采用增广向量形式

    判别函数$g(x) = a^T y$

    对于未知样本x，若$g(x) > 0$，则x决策为ω_1类

    若$g(x) < 0$，则x决策为ω_2类
  - 规范化
    - 对ω_2类样本的增广向量全部乘以-1
  - 规范化之后的分类结果
    - $a^T y_i > 0$——正确分类
    - $a^T y_i < 0$——错误分类

4.2　解向量和解区

- **概念**
  - 解向量——能将线性可分样本集中的每个样本都正确分类的权向量。
  - 解区——解向量往往不是一个，而是由无穷多个解向量组成的（角度）区域，称为解区。

4.3 感知准则函数

- **Rosenblatt定义感知准则函数**
  - 对于规范化的增广样本集
    - $a^T y_i < 0$——错误分类
  - 定义感知准则函数，作为优化准则函数

$$\min J_p(a) = \sum_{y \in Z_E} (-a^T y)$$

  - 求解向量（或解区）

4.3 感知准则函数

- 图示法求解区
- 解区可以直接画图求出（二维条件时）

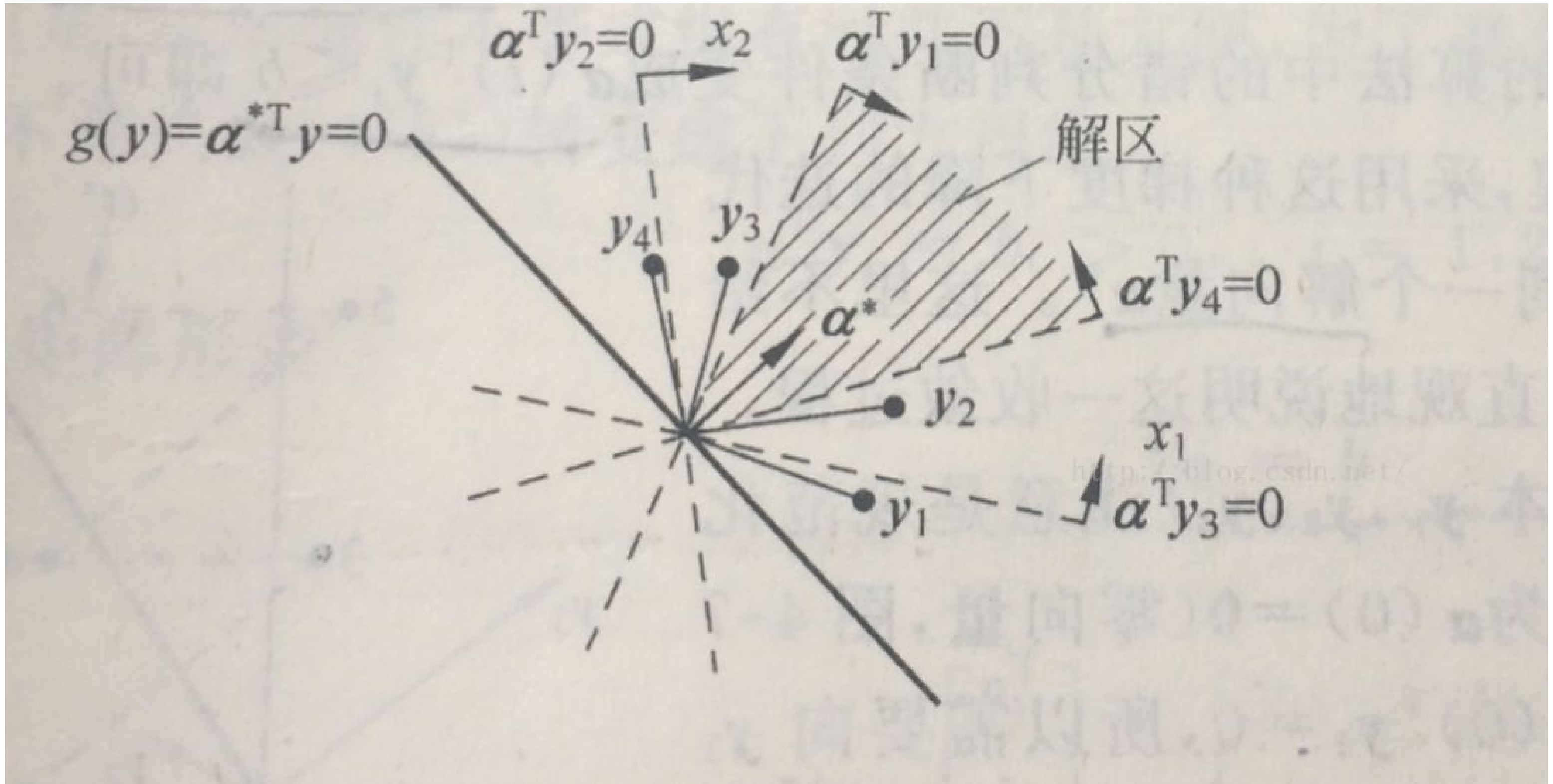


图 4-5　解向量和解区

4.4 求极值解

- **求解感知器**
  - 采用梯度下降法求优化准则函数极值（极小值）
    - 先求梯度方向
    - 计算参数改变量
    - 得到迭代公式

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

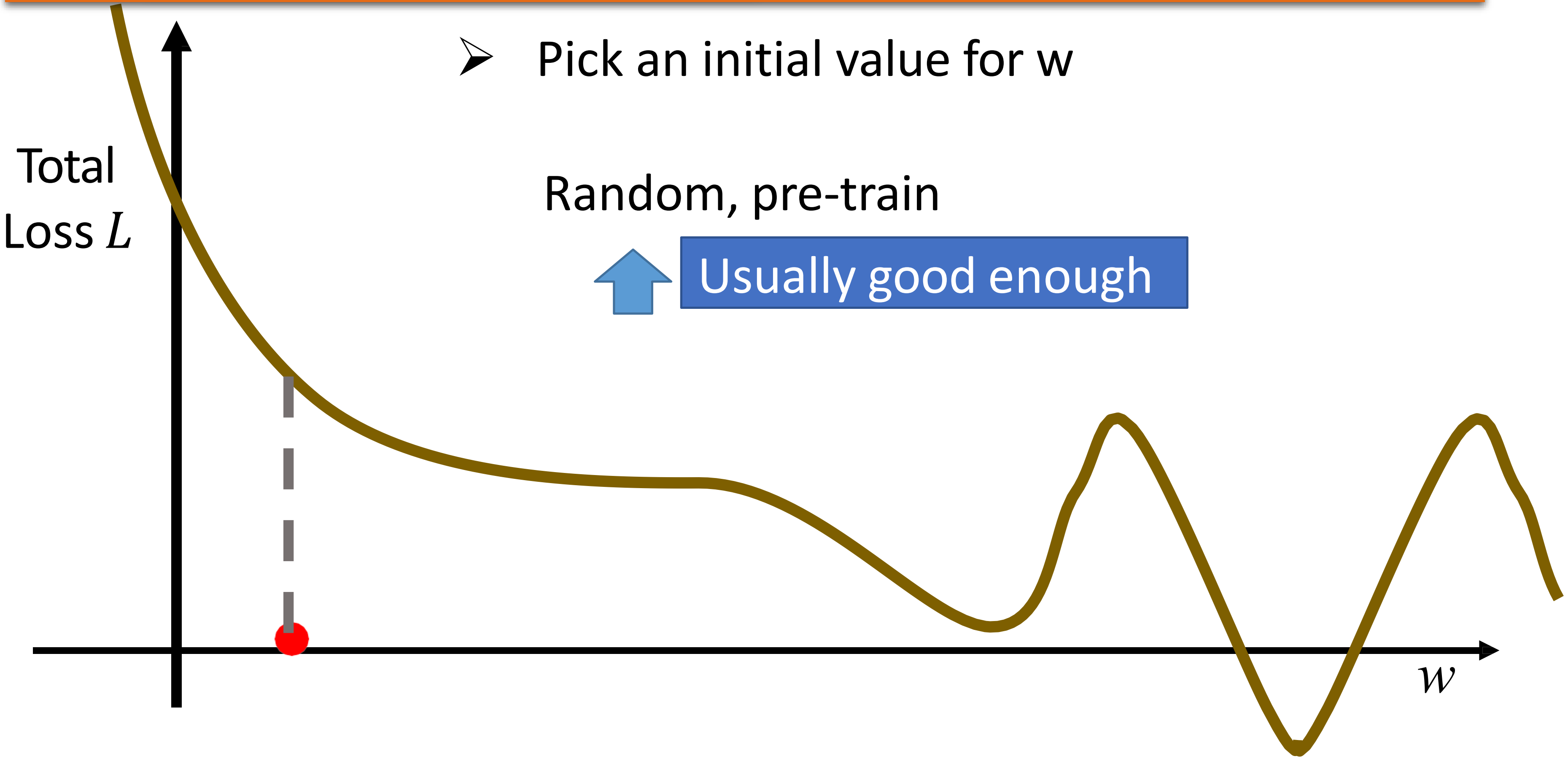
$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \rho_k \nabla J$$

$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) + \rho_k \sum_{y \in \mathscr{Y}^k} \boldsymbol{y}$$

优化算法：梯度下降

Parameters $\theta = \{w_1, w_2, \cdots, b_1, b_2, \cdots\}$

Find *parameters θ^** that minimize total loss L



- Pick an initial value for w

Random, pre-train

Usually good enough

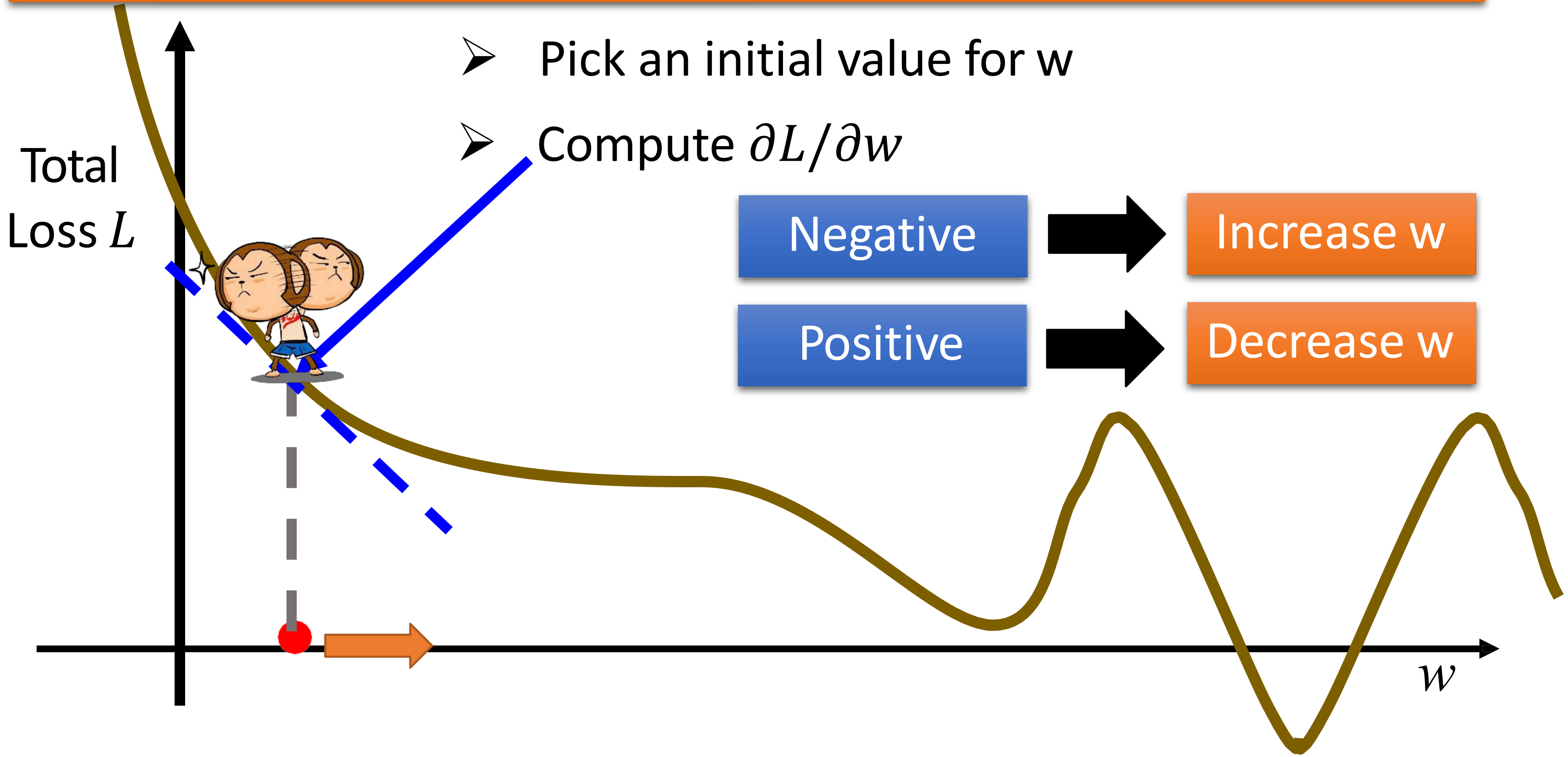
梯度下降

parameters $\theta = \{w_1, w_2, \cdots, b_1, b_2, \cdots\}$

Find *parameters* $\boldsymbol{\theta^*}$ that minimize total loss L

- Pick an initial value for w
- Compute $\partial L / \partial w$

Negative	→	Increase w
Positive	→	Decrease w



梯度下降

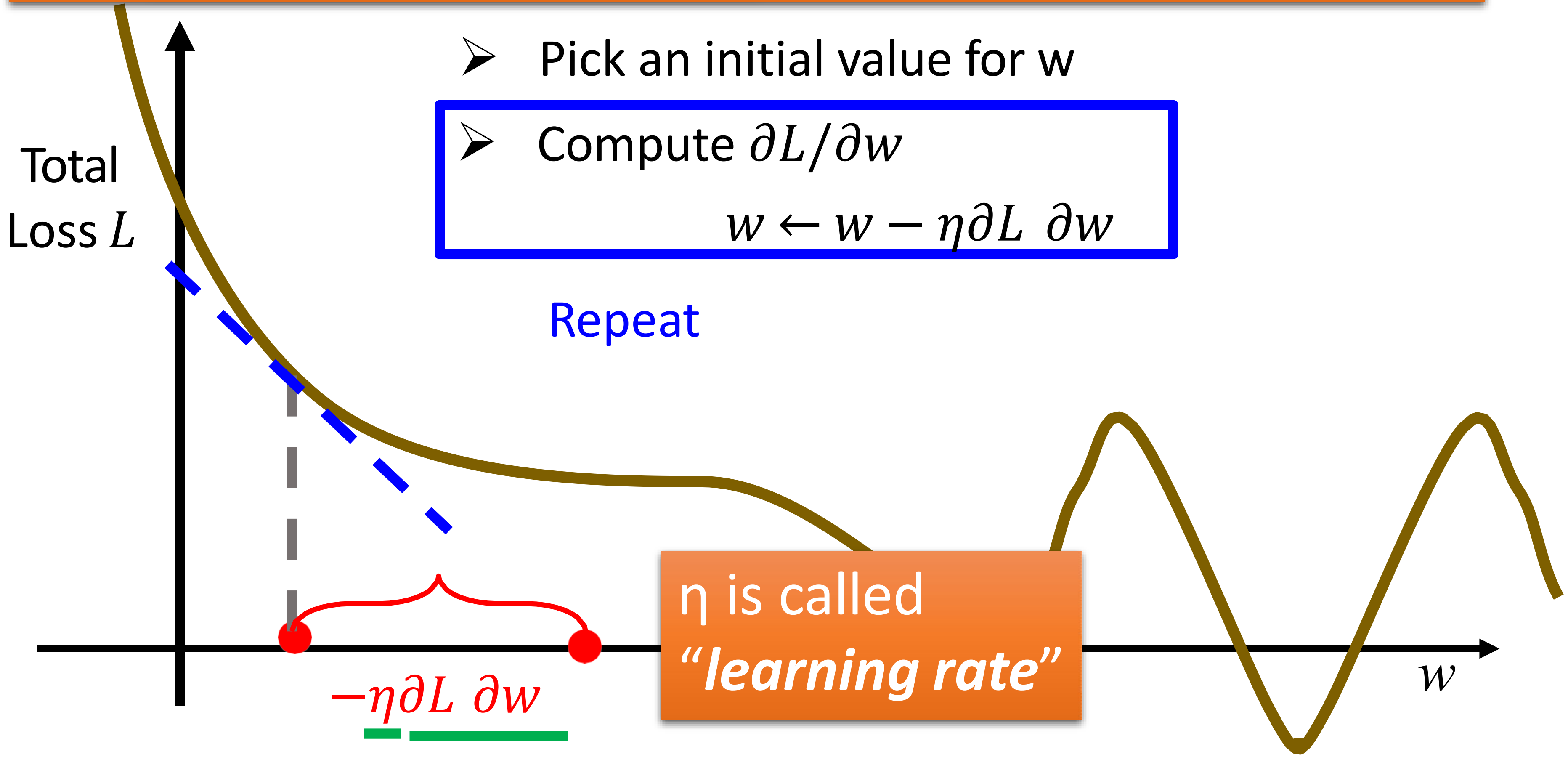
parameters $\theta = \{w_1, w_2, \cdots, b_1, b_2, \cdots\}$

Find ***parameters*** $\boldsymbol{\theta}^*$ that minimize total loss L

- Pick an initial value for w
- Compute $\partial L / \partial w$

  $w \leftarrow w - \eta \partial L \ \partial w$

Repeat

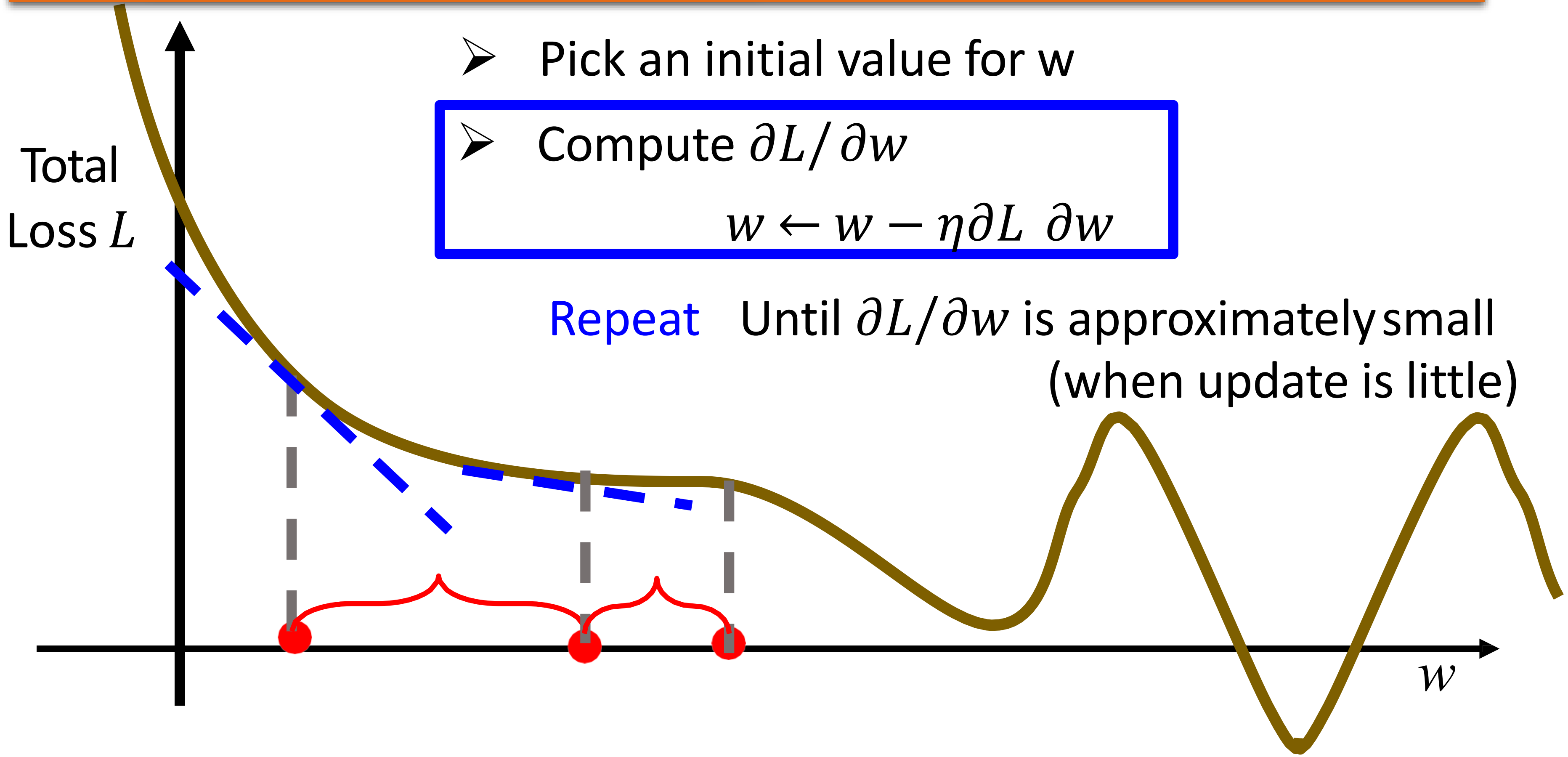


η is called
"learning rate"

梯度下降

parameters $\theta = \{w_1, w_2, \cdots, b_1, b_2, \cdots\}$

Find *parameters θ^** that minimize total loss L



- Pick an initial value for w
- Compute $\partial L / \partial w$

  $w \leftarrow w - \eta \partial L \;\; \partial w$

Repeat Until $\partial L / \partial w$ is approximately small (when update is little)

4.4 求极值解

- **求解感知器**
  - 梯度下降法求极值的问题
    - 收敛性
    - 步长的选择

4.5　特点

- **感知准则（分类器）的特点**
  - 解决两类问题的线性分类器
  - 样本集必须是线性可分的
  - 采用感知准则函数求极值解（最优决策）
  - 分类器设计过程复杂

THE END !

78