西南财经大学

SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

1925 - 2021

经世济民 攻坚以求

机器学习基础

感知机

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 模型
- 策略
- 算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 模型
- 策略
- 算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

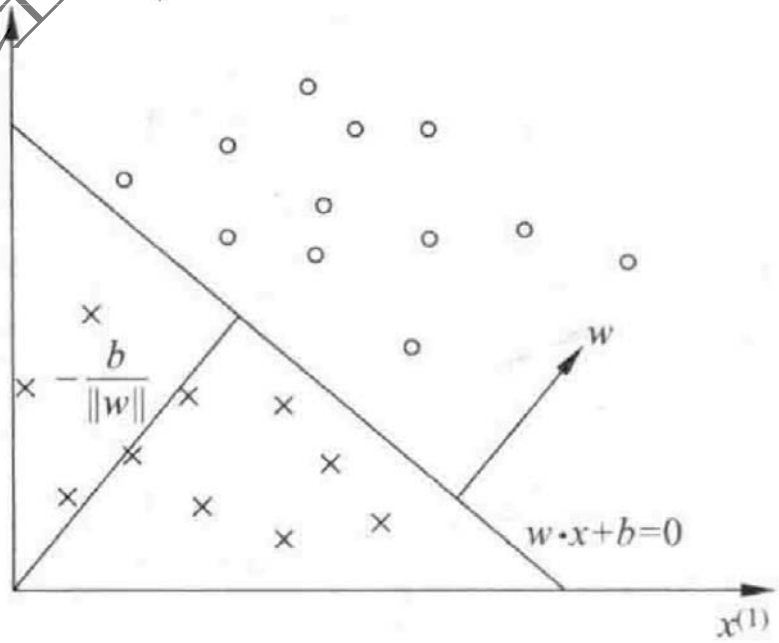
模型

- 输入空间是 $\mathcal{X} \in \mathbb{R}^n$ 表示样本的特征空间
- 输出空间是 $\mathcal{Y} = \{+1, -1\}$，表示样本的类别
- 模型 $f(x) = \text{sign}(w \cdot x + b)$
  - w 是权值向量
  - b 是偏置
  - $\text{sign}(\cdot)$ 是符号函数

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

几何解释

- 线性方程 $w \cdot x + b = 0$ 是 n 维空间中的一个超平面
  - w 是平面的法向量，b 是截距
- 超平面将特征空间分为两部分
  - 正类：$w \cdot x + b \geq 0$
  - 负类：$w \cdot x + b < 0$



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 模型
- 策略
- 算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

学习策略

- 假设数据集线性可分
- 损失函数：误分类点的数目，但不方便优化
- 另一选择：误分类点到超平面的总距离

$$-\frac{1}{\|w\|}\sum_{x_i \in M} y_i(w \cdot x_i + b)$$

  - M 为误分类点的集合

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

损失函数

- 不考虑 $\frac{1}{\|w\|}$
  - 优化困难
  - 假设$\|w\|=1$，不会改变模型的假设空间
- 损失函数 $L(w,b)=-\sum_{x_i \in M} y_i(w \cdot x_i + b)$
  - 没有误分类点，损失函数为0
  - 误分类点越少，越接近超平面，损失函数越小
  - 损失函数关于参数连续可导

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 模型
- 策略
- 算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

优化算法的原始形式

- 目标函数 $\min\limits_{w,b} L(w,b) = -\sum\limits_{x_i \in M} y_i (w \cdot x_i + b)$
- 梯度方向 $\nabla_w L(w,b) = -\sum\limits_{x_i \in M} y_i x_i$

$$\nabla_b L(w,b) = -\sum_{x_i \in M} y_i$$

- 随机梯度下降法：随机选取一个误分类点

$$w \leftarrow w + \eta y_i x_i, \quad b \leftarrow b + \eta y_i$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

为什么是随机梯度?

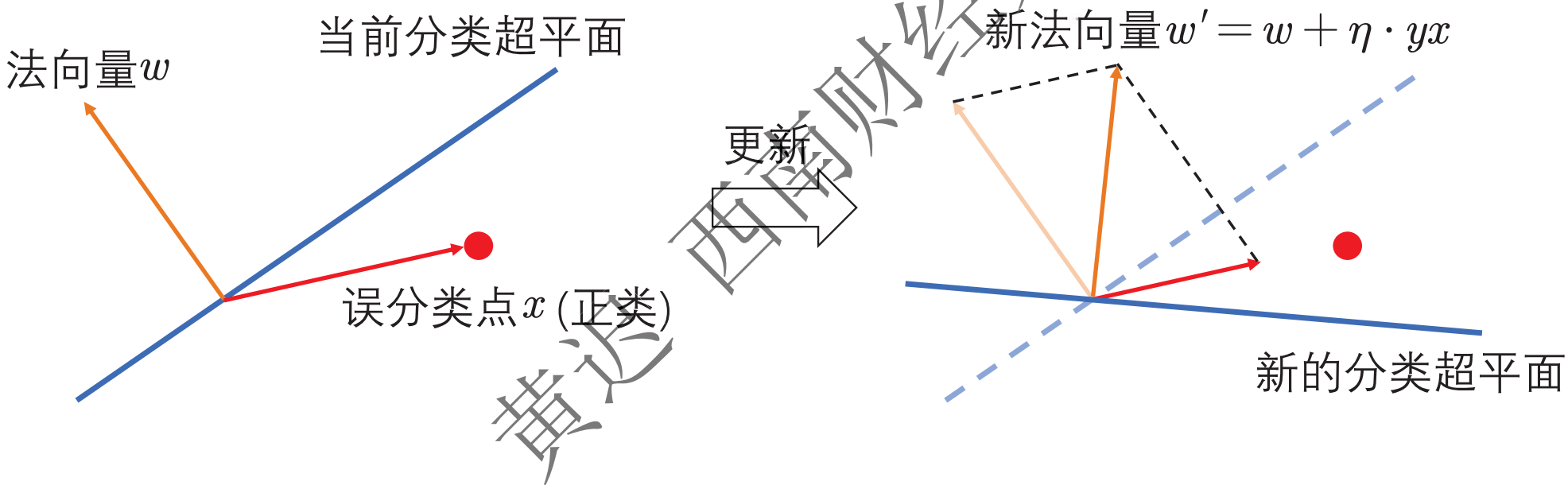
- 使用梯度方向，在求和中会干扰优化方向，影响效率
  - 例如，M 中只有两个误分类点，且分别一正一负，有

$$\nabla_b L(w,b) = -\sum_{x_i \in M} y_i = 0$$

  - 此时 b 不会更新
  - 同理 $\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i$ 可能会很小，w 更新慢
- 随机梯度确保能向一个误分类点优化

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

随机梯度几何解释



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

原始形式

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{-1,+1\}$，$i=1,2,\cdots,N$；学习率 $\eta(0<\eta\leqslant 1)$；

输出：w,b；感知机模型 $f(x)=\text{sign}(w\cdot x+b)$。

（1）选取初值 w_0,b_0；

（2）在训练集中选取数据 (x_i,y_i)；

（3）如果 $y_i(w\cdot x_i+b)\leqslant 0$，

$$w\leftarrow w+\eta y_i x_i$$

$$b\leftarrow b+\eta y_i$$

（4）转至（2），直至训练集中没有误分类点。

■

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法的收敛性

- 经过有限次迭代能得到一个将线性可分数据集完全正确划分的感知机模型
- Novikoff定理：
  - 存在 $\gamma > 0$，使得对任意样本有

$$y_i(\hat{w}_{opt} \cdot \hat{x}_i) = y_i(w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

  - 令 $R = \max\limits_{1 \leq i \leq n} \| \hat{x}_i \|$，算法的误分类错误满足不等式 $k \leq \left(\frac{R}{\gamma}\right)^2$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

定理的解释

- R是数据集中样本最大的模长，代表数据集的规模；数据集越大，算法越难收敛
- γ是样本到超平面的最近距离，代表样本中正负两类的间隔，间隔越小，算法越难收敛
- 算法的结果不唯一，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

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

优化算法的对偶形式

- 在原始形式中，参数的更新可以表示成x_i和y_i的线性组合

$$w \leftarrow w + \eta y_i x_i$$
$$b \leftarrow b + \eta y_i$$

最终结果

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (\alpha_i = n_i \cdot \eta)$$
$$b = \sum_{i=1}^{N} \alpha_i y_i$$

- 因此，在训练中只需要记录样本被选中的次数
- 对偶形式中，样本仅以内积的形式出现，可以预先计算（Gram 矩阵），节约训练成本

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

对偶形式

输入：线性可分的数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in\mathbf{R}^n$，$y_i\in\{-1,+1\}$，$i=1,2,\cdots,N$；学习率 $\eta\ (0<\eta\leqslant 1)$；

输出：α,b；感知机模型 $f(x)=\operatorname{sign}\left(\sum\limits_{j=1}^{N}\alpha_j y_j x_j\cdot x+b\right)$，其中 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)^{\mathrm{T}}$。

(1) $\alpha\leftarrow 0$，$b\leftarrow 0$；

(2) 在训练集中选取数据 (x_i,y_i)；

(3) 如果 $y_i\left(\sum\limits_{j=1}^{N}\alpha_j y_j x_j\cdot x_i+b\right)\leqslant 0$，

$$\alpha_i\leftarrow\alpha_i+\eta$$

$$b\leftarrow b+\eta y_i$$

(4) 转至 (2) 直到没有误分类数据。 ■

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 模型
- 策略
- 算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

总结

- 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，属于判别模型
- 导入基于误分类的损失函数
- 随机梯度下降法对损失函数进行极小化，分原始形式和对偶形式
- 是神经网络与支持向量机的基础

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法优点

- 简单易懂，编程实现容易
- 是支持向量机和神经网络等算法的基础，理论重要

算法缺点

- 只能处理线性可分数据集
- 无法解决回归问题
- 实际应用太少