02 感知机

从线性预测到误分类驱动训练

本讲的学习目标

- 解释感知机的二分类任务、线性分数、决策边界与预测规则
- 从有符号分数和感知机损失说明误分类区域的参数更新
- 按原始形式执行判断、更新、重复访问与停止检查
- 用完整二维例子连接训练输出与新样本预测
- 说明收敛的条件、结果边界及训练结果与新数据评价的区别

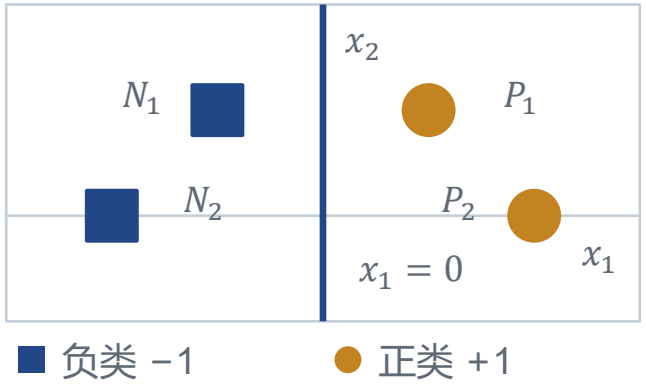
本讲另介绍对偶表示、样本内积与 Gram 矩阵的基本思想，不要求理解、推导或独立使用

1 线性模型、训练与预测

线性分数的符号给出类别判断；训练形成参数，预测使用固定参数。参数根据标签怎样调整，是后续训练部分要解决的问题。

从带标签样本到分类规则

训练样本有标签，新输入的标签却未知。记住已有点还不够，需要一条按特征判断新输入的规则。



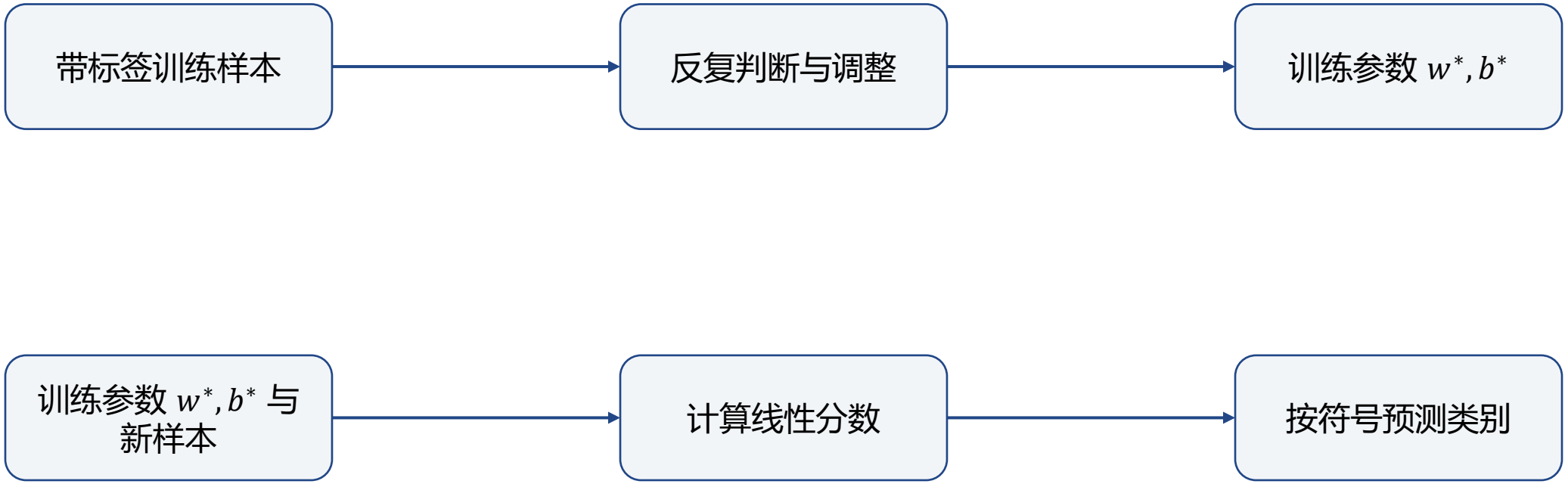
- 将分割线两侧分别判为正、负类，把已知点的标签关系扩展为区域判断规则
- 示意规则：$f(x) = x_1$，参数尚未由训练得到
- **感知机的思路**：选用可调整的线性分数，把分割边界两侧的位置关系转为分数正负；根据已知标签调整参数，再用这条规则预测新输入
- 训练集 $D = \{(x_i, y_i)\}_{i=1}^{n}$：$n$ 为样本数，x_i 为特征向量，$y_i \in \{-1, +1\}$ 为已知标签

线性分数与类别预测

- 二维中，直线两侧可以用一个线性表达式的正负表示；改变系数和常数项，就能调整分割方向与位置。将这个思路用于 d 个特征，得到可计算、可调整的**线性分数**
  - 输入 $x \in \mathbb{R}^d$，d 为特征维数；分数是特征的加权和再加偏置
  - $f(x) = w^T x + b = \sum_{j=1}^{d} w_j x_j + b$
  - $w \in \mathbb{R}^d$ 是**权重**，$b \in \mathbb{R}$ 是**偏置**，j 是特征下标
  - w_j 决定第 j 个特征对分数的贡献，b 给分数增加统一偏移；这些参数共同决定判断规则
- **类别预测**由分数符号给出：$f(x) > 0$ 预测 $+1$，$f(x) < 0$ 预测 -1
  - 记作 $\hat{y} = sign(f(x))$；分数为零时，预测类别需要另定约定
- 分数描述模型的线性判断，与类别概率具有不同含义

训练形成参数，预测使用参数

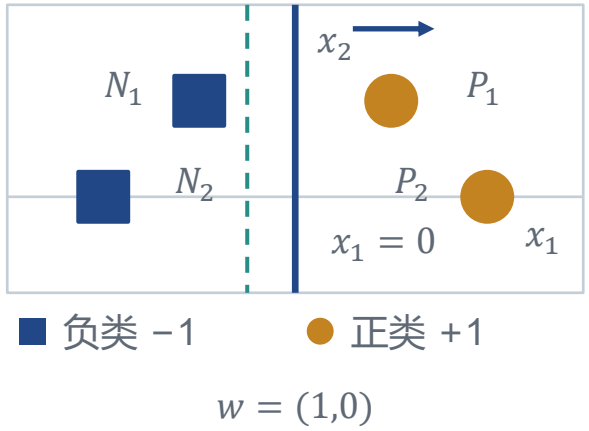
- **训练调整参数，预测固定参数**；新样本真实标签不参与预测
- 图中 w^*, b^* 的星号表示训练结果



线性分类器的边界与可分性

- 回看导入中的两侧规则：**线性分类器**按 $f(x) = w^T x + b$ 的符号划分类别，正分数区域预测为 $+1$，负分数区域预测为 -1
- 当 $w \neq 0$ 时，两个区域由**决策边界** $w^T x + b = 0$ 分隔
  - 边界是分数为零的点集，二维中为直线，高维中为超平面
  - w 给出垂直于边界的法向方向
  - 固定 w 时，b 改变边界位置，使边界可以不经过原点
    - 边界满足 $w^T x = -b$：$b = 0$ 时经过原点，$b \neq 0$ 时不经过原点
- **线性可分**：存在一组参数，使全部正类训练点严格位于正分数区域，全部负类训练点严格位于负分数区域
  - 代数条件为全部训练样本满足 $y_i(w^T x_i + b) > 0$
- 当 $w = 0, b = 0$ 时，所有分数为零，尚未形成有效的分类边界

可分性说明分离参数存在，训练还需根据数据找到这样的参数



固定 w；$b = 0.5$，$x_1 = -0.5$

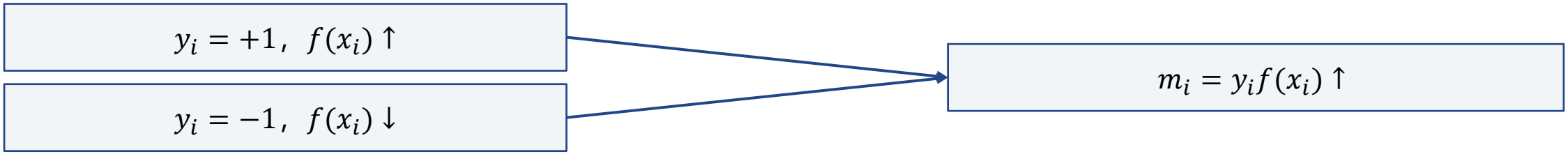
实线：$f(x) = x_1$；均为示意，非训练结果

2 感知机损失与参数更新

从分类对错进入连续训练准则，解释更新方向与当前样本作用。

有符号分数与感知机损失

修正待更新样本时，正类增大分数、负类减小分数，怎样统一这两个方向?



- **有符号分数** $m_i = y_i(w^T x_i + b)$，正确侧满足 $m_i > 0$；两类修正都转为提高当前样本的 m_i
- 错误侧 $m_i < 0$ 时，提高 m_i 就减小惩罚 $-m_i$；正确侧 $m_i > 0$ 时不惩罚
  - **感知机损失** $\ell_i(w, b) = max(0, -m_i)$
- **训练判定**：$m_i > 0$ 时保持参数；$m_i \leq 0$ 时更新
  - $m_i = 0$ 时损失为零且有折点，仍按训练规则更新
- **0–1 损失**：预测正确取 0，错误取 1，累计为错分数；感知机损失利用连续分数，两种损失的最小化不等价

梯度与参数更新

错误侧要提高 m_i，等价于减小损失 $-m_i$。下面用负梯度实现这一修正方向。

- 在 $m_i < 0$ 区域，固定当前样本 x_i, y_i，把损失看作待调整参数 w, b 的函数：
  $\ell_i(w, b) = -y_i(w^T x_i + b)$
- **梯度**由各参数的偏导数组成；每个偏导表示其余参数固定时，损失随该参数变化的局部变化率。梯度指向局部增加最快的方向，负梯度指向减小方向
  - 损失表达式中 w_j 的系数为 $-y_i x_{ij}$，因此 $\partial \ell_i / \partial w_j = -y_i x_{ij}$，合起来是 $\nabla_w \ell_i = -y_i x_i$
  - 偏置的导数为 $\partial \ell_i / \partial b = -y_i$
- 参数沿**负梯度更新**；**学习率** $\eta > 0$ 控制更新步幅，本讲固定其值
  - $w \leftarrow w - \eta \nabla_w \ell_i = w + \eta y_i x_i$
  - $b \leftarrow b - \eta \partial \ell_i / \partial b = b + \eta y_i$
- 普通梯度推导适用于 $m_i < 0$；$m_i = 0$ 时按训练约定执行同一更新

一次更新对当前样本的作用

更新是否实现了“提高当前样本的有符号分数”？回代检查，再判断它能否同时改善整个训练集。

- 令 $w' = w + \eta y_i x_i$、$b' = b + \eta y_i$，代回同一训练样本
  - 代入得到 $m_i' = y_i[(w + \eta y_i x_i)^T x_i + b + \eta y_i]$；由 $y_i^2 = 1$，化为 $m_i' = m_i + \eta(\|x_i\|^2 + 1) > m_i$
    - $\|x_i\|^2 = \sum_j x_{ij}^2$ 为特征平方和；常数 1 来自偏置更新
- 对当前样本：有符号分数增大，但不一定一次就变为正值；若更新前 $m_i < 0$，感知机损失减小
  - 正类 $y_i = +1$ 时 $m_i = f(x_i)$，更新使当前分数增大；负类 $y_i = -1$ 时 $m_i = -f(x_i)$，更新使当前分数减小
- **局部改善不保证全体改善**：样本共用同一组参数，其他样本的有符号分数可能增大或减小；当前点改善不保证全体改善，因此还需重复访问并检查全部样本

3 原始训练与新样本预测

把单点更新组织成完整训练，再用所得的固定参数预测新样本。

感知机原始形式的训练循环

一次更新只保证当前样本的有符号分数增大，不保证全体改善。训练需要重复访问全部样本，并记录整轮是否发生更新，再决定能否停止。

- 给定训练集 D、初值 w, b、固定学习率 $\eta > 0$ 与访问顺序

重复：

本轮发生更新 ← 否

按固定顺序访问每个训练样本 (x_i, y_i)：

$m_i \leftarrow y_i(w^T x_i + b)$

若 $m_i \leq 0$：

$w \leftarrow w + \eta y_i x_i$

$b \leftarrow b + \eta y_i$

本轮发生更新 ← 是

若本轮发生更新为否：输出 w^*, b^* 并停止

- **一轮**是全部训练样本各访问一次
- **停止**要求完整一轮没有更新
  - 整轮无更新时参数保持不变，所有训练样本均满足 $m_i > 0$
  - **零损失不代表停止**：初值 $w = 0, b = 0$ 时全部 $m_i = 0$。各点损失为零仍须更新

示例：感知机的参数更新

训练数据与约定

- 课程构造数据，按表中顺序访问

样本	特征 x	标签 y
P_1	$(1,1)$	$+1$
N_1	$(-1,1)$	-1
P_2	$(2,0)$	$+1$
N_2	$(-2,0)$	-1

- 初值 $w=(0,0), b=0$，学习率 $\eta=1$
- 判断量 $m=y(w^T x+b)$；$m\le 0$ 时更新
  - $w\leftarrow w+yx$，$b\leftarrow b+y$
- 完整一轮无更新时停止

参数更新

- 第一轮的两次更新

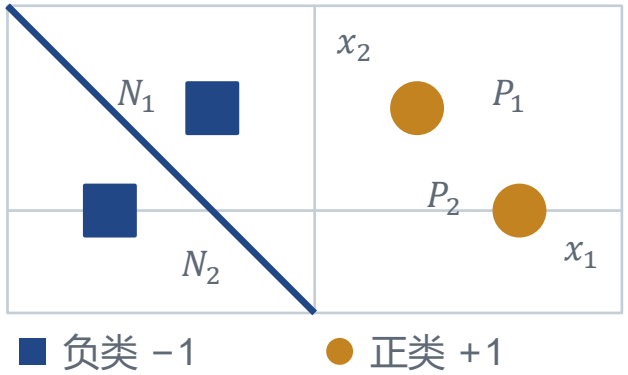
样本	更新前 f, m	更新后 w, b
P_1	$0,0$	$w^{(1)}=(0,0)+(1,1)=(1,1)$ $b=1$
N_1	$1,-1$	$w^{(2)}=(1,1)-(-1,1)=(2,0)$ $b=0$

- 两次更新后得到 $w=(2,0), b=0$；能否停止，还需检查本轮余下样本和下一完整轮

示例：更新影响与整轮检查

第二次更新修正了 N_1，同一组参数下的其他样本会怎样变化？

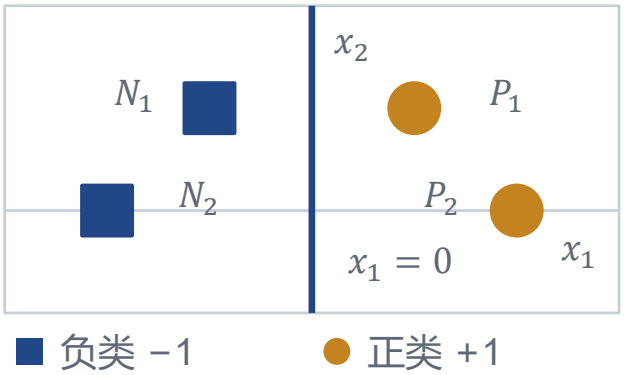
$w = (1,1), b = 1$



$x_1 + x_2 + 1 = 0$

$m_{N_1} = -1,\ m_{P_1} = 3$

$w = (2,0), b = 0$



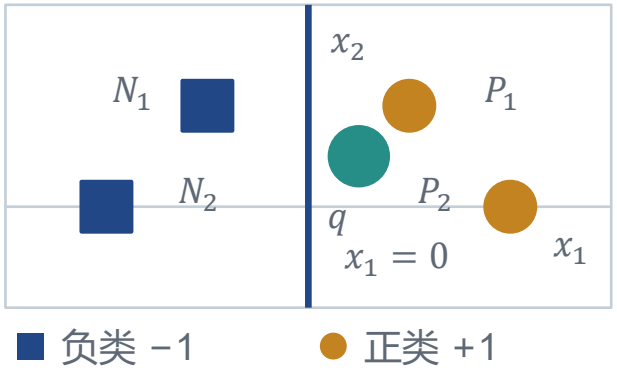
$m_{N_1} = 2,\ m_{P_1} = 2$

- **局部改善不保证全体改善**：P_1 的有符号分数下降，仍在正确侧
- 第一轮余下 P_2, N_2：$f = 4, -4$，$m = 4,4$，无更新；本轮已更新，不能停止
- 下一整轮：$f = 2, -2, 4, -4$，$m = 2,2,4,4$，无更新，停止
- 输出 $w^* = (2,0), b^* = 0$，模型为 $f^*(x) = 2x_1$

感知机推理：从固定参数到类别预测

使用训练输出预测新样本

- 上一页输出 $w^* = (2,0), b^* = 0$，模型为 $f^*(x) = w^{*T}x + b^* = 2x_1$；新样本为 $q = (0.5,0.5)$
- **推理**固定这组训练参数，沿用“先算线性分数，再按符号预测类别”的规则
- 代入得到 $f^*(q) = 2\times0.5 + 0\times0.5 + 0 = 1$，分数为正，因此 $\hat{y}(q) = +1$



预测 +1
真实标签未知

预测与正确性评价

- 判断这次预测是否正确，需要与 q 的真实标签比较
- 评价模型的新数据表现，需要独立评价数据及相应使用条件；单次预测结果不足以完成这项评价

4 收敛条件、结果边界与对偶表示

限定训练结果的条件与评价意义，再介绍同一模型的另一种表示。

感知机的收敛条件与结论

感知机的收敛保证

- 前例在两次更新后得到分离参数；单个成功实例只说明这组数据上的执行结果。算法能否保证有限停止，还取决于数据条件与更新规则
- 本讲的**收敛**指经过有限次更新达到严格分离训练集的状态，随后整轮无更新并停止
- **收敛条件**
  - 训练集有限且线性可分；采用固定正学习率，循环访问全部样本，并按 $m_i \le 0$ 执行标准更新
- **收敛结论**
  - 上述条件下，经过有限次更新得到使全部训练样本 $m_i > 0$ 的参数
  - 这些参数使全部样本保持在正确一侧，此后完整访问一轮无更新，满足已建立的停止规则
- **偏置的增广表示**
  - 把偏置 b 看作常数特征 1 的权重，将分数与两项更新写成统一的向量形式
  - 令增广特征为 $\tilde{x} = (x, 1)$，增广参数为 $\tilde{w} = (w, b)$，则 $\tilde{w}^T \tilde{x} = w^T x + b$
  - 权重与偏置的更新合为 $\tilde{w} \leftarrow \tilde{w} + \eta y_i \tilde{x}_i$，即同时加入 $\eta y_i x_i$ 与 ηy_i
  - 分数和更新规则都与原模型一致，因此含偏置模型也纳入上述保证

感知机训练的适用边界

收敛保证的范围

- 上述条件下，算法保证获得严格分离训练集的参数；数据条件、解的选择与新数据评价分别限定这一结果
- **数据不可分时的停止限制**
  - **线性不可分**指不存在参数使全部训练点满足 $y_i(w^T x_i + b) > 0$
  - 整轮无更新要求同一组参数使全部训练点的有符号分数为正；不可分数据无法满足这一状态，因此本讲标准循环会持续更新
- **可分训练集的分离解**
  - 同一训练集可能有多条分离边界，初始化与访问顺序可能影响所得参数和边界
  - 收敛结论保证得到分离解；比较边界的优劣还需指定准则
- **新数据上的评价**
  - 严格分离只说明全部训练点满足分类要求；新数据表现需要独立评价数据及相应使用条件支持

感知机的对偶表示与内积

同一模型的另一种表示

- 原始形式直接维护权重 w 与偏置 b；对偶形式把权重表示为训练样本的累积贡献。本讲采用零初值与固定学习率：
- $w = \sum_{i=1}^{n} \alpha_i y_i x_i$，其中 $\alpha_i \geq 0$ 记录样本 i 的累积更新量，该系数在样本每次触发更新时增加 η；n 是训练样本数。偏置 b 仍按原始规则更新并保留在模型中
- 相应分数为 $f(x) = \sum_{i=1}^{n} \alpha_i y_i (x_i^T x) + b$，$x_i^T x$ 是训练样本与当前输入的内积；得到分数后仍按符号预测类别

内积与 Gram 矩阵

- **Gram 矩阵**保存训练点的两两内积：$G_{ij} = x_i^T x_j$
- 训练特征固定时，G 不随参数更新变化，可以与当前系数和偏置配合计算训练点分数。新输入与训练点的内积不在训练 Gram 中，需要另外计算
- 这种表示改变参数与计算的组织方式，分类任务和预测规则保持不变

总结

从训练数据到新样本预测

- 感知机从有限带标签样本学习线性二分类模型，训练形成参数，推理使用固定参数预测类别

模型与推理

- 线性分数为 $f(x) = w^T x + b$，在 $w \neq 0$ 时，零分数点集构成决策边界
- 推理固定训练参数，对新样本计算分数，再按符号预测类别

更新与停止

- 有符号分数 $m_i = y_i f(x_i)$ 统一两类的训练判断，$m_i \leq 0$ 时更新
- 在误分类区域沿感知机损失的负梯度更新；零分数按算法约定更新。遍历全部训练样本，整轮无更新后输出参数

收敛与评价

- 有限且线性可分的训练集、固定正学习率与标准循环更新支持有限停止
- 严格不可分时无法达到整轮无更新的停止状态；可分时所得分离解可能非唯一
- 严格分开训练点只说明训练集上的分类结果；新数据表现需要独立评价数据及相应使用条件支持

对偶表示

- 本讲还介绍了按样本贡献表示同一组参数、用内积及 Gram 矩阵组织计算的另一种视角；核心预测规则保持不变