01 从线性模型到神经网络

机器学习 · 深度学习入门

学习目标

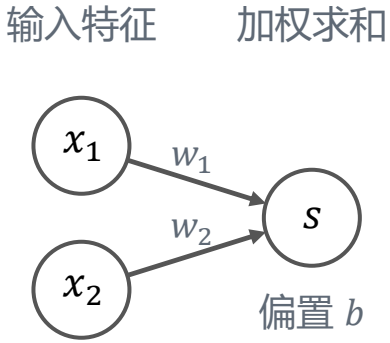
- 解释隐藏层、参数与非线性激活的不同作用
- 根据给定参数，跟踪输入、隐藏表示、输出与损失
- 区分网络结构、预测规则与训练损失
- 解释二分类、多分类输出与标签的对应
- 初步区分网络表达能力、实际训练与泛化表现

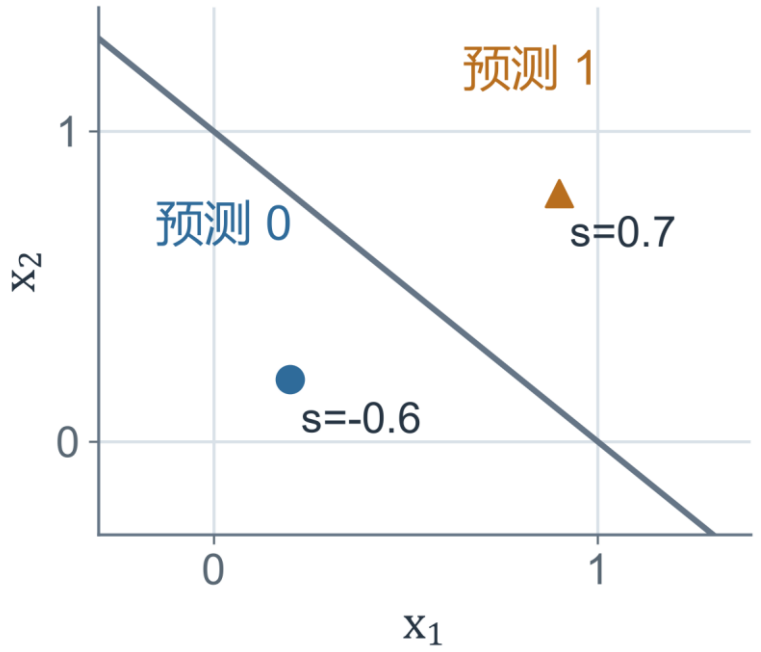
1 从线性模型到非线性表示

线性限制、隐藏特征与非线性响应

线性得分与决策边界

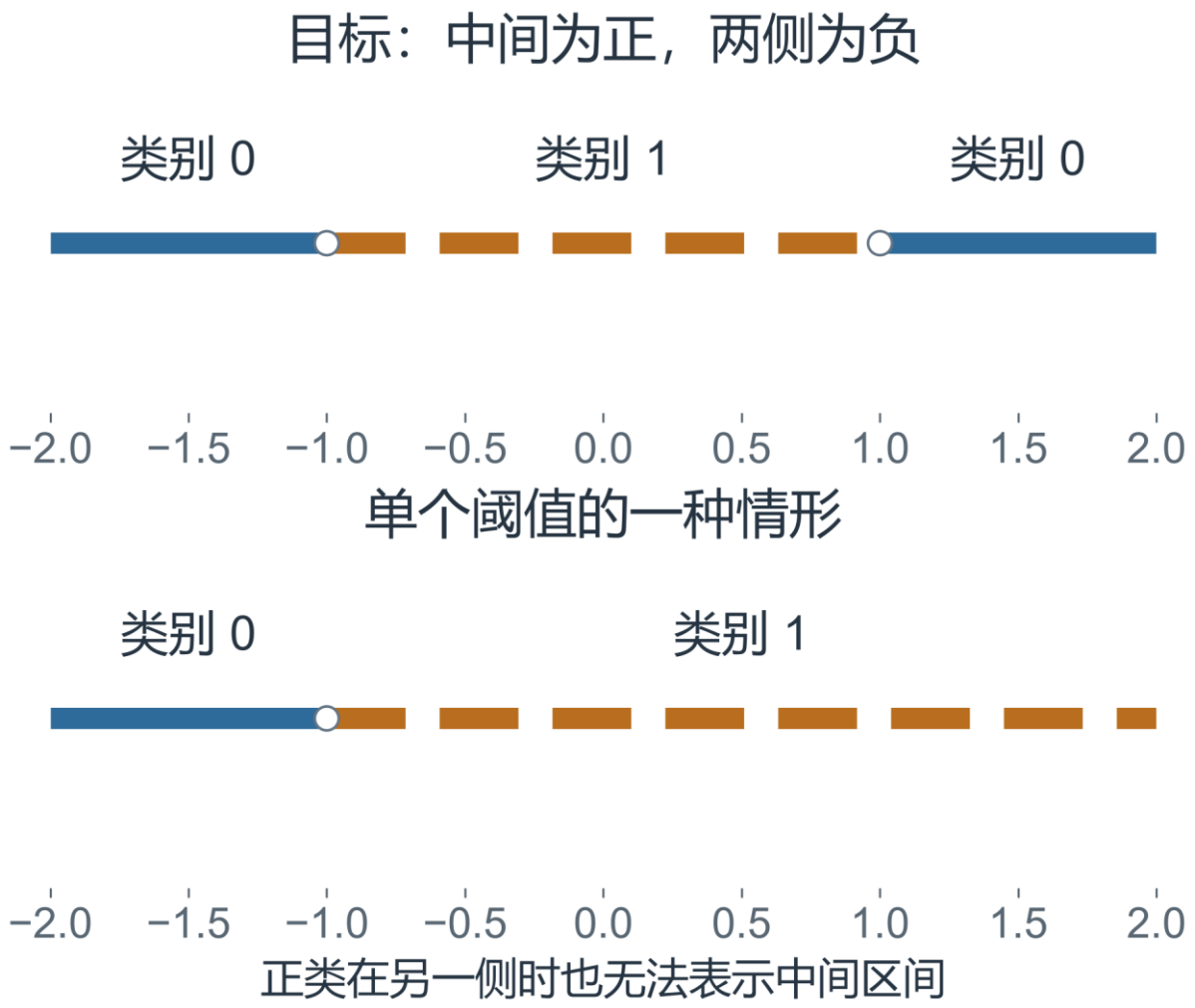
- 线性模型加权输入并加偏置，形成得分
  - $s = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b$，$\boldsymbol{w}$ 为权重向量，b 为标量偏置
  - 含偏置的得分计算称为**仿射变换**
- 得分之后作阈值判定
  - 本页按 $s \geq 0$ 预测类别 1，否则预测 0
- 边界满足 $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b = 0$
  - 权重非零时，二维边界是直线；高维为超平面
  - 右图采用 $s = x_1 + x_2 - 1$；直线两侧分别预测 0 与 1





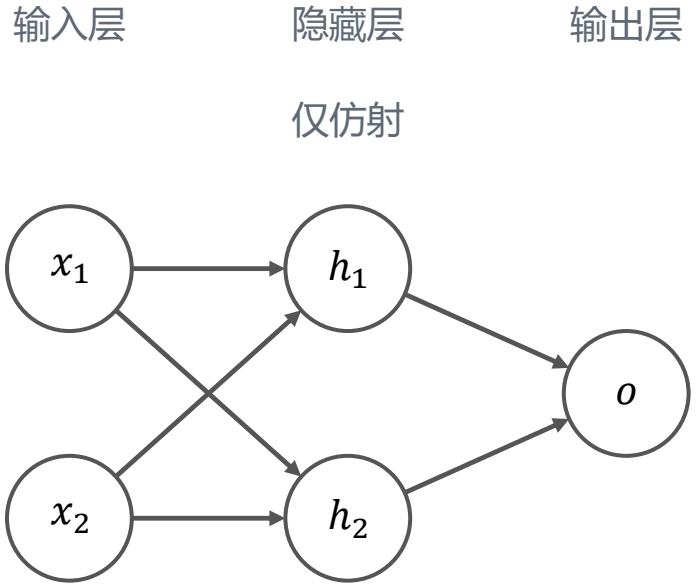
单个阈值的分类限制

- 一维分类任务：$-1 < x < 1$ 为正类，两侧及端点为负类
- 仿射得分 $s = wx + b$ 配合固定阈值
  - $w \neq 0$ 时，正类只能位于一个阈值的一侧
  - $w = 0$ 时，只能给所有输入相同预测
- 中间正、两侧负需要两次类别切换
  - 单个仿射得分的阈值分类无法满足这一关系



仿射变换复合的表达限制

- 在输入与单个输出之间加入**隐藏层**，其单元接收输入并将计算结果交给输出层
  - 单隐藏层网络有两个带参数层，输入层不计入
- 先看各层只有加权求和与偏置的情况
  - $\boldsymbol{h} = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1$，$o = \boldsymbol{W}_2\boldsymbol{h} + b_2$
  - 单输出得分 o 与偏置 b_2 为标量
- 代入隐藏表示，合并计算
  - $o = \boldsymbol{W}_2(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + b_2$
  - $o = (\boldsymbol{W}_2\boldsymbol{W}_1)\boldsymbol{x} + (\boldsymbol{W}_2\boldsymbol{b}_1 + b_2)$
- 合并后仍是输入的仿射得分，配合单个阈值仍不能完成中间正、两侧负的分类
- 非线性激活使隐藏表示有可能超出这一限制

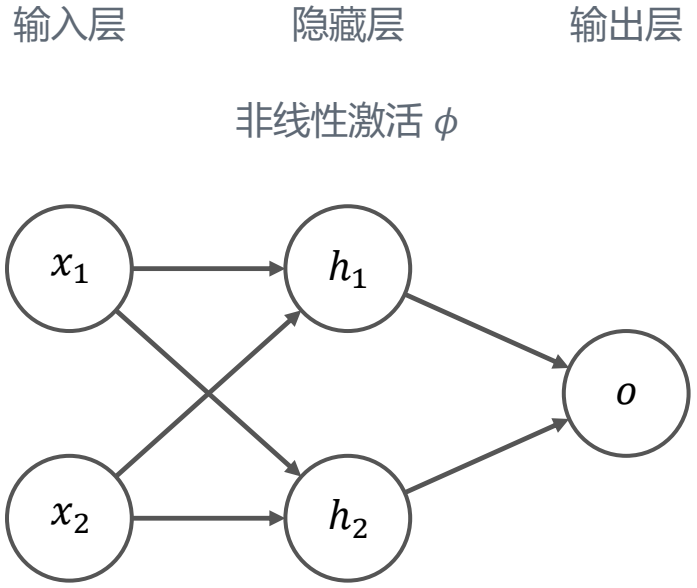


输入→隐藏：$\boldsymbol{W}_1, \boldsymbol{b}_1$

隐藏→输出：$\boldsymbol{W}_2, b_2$

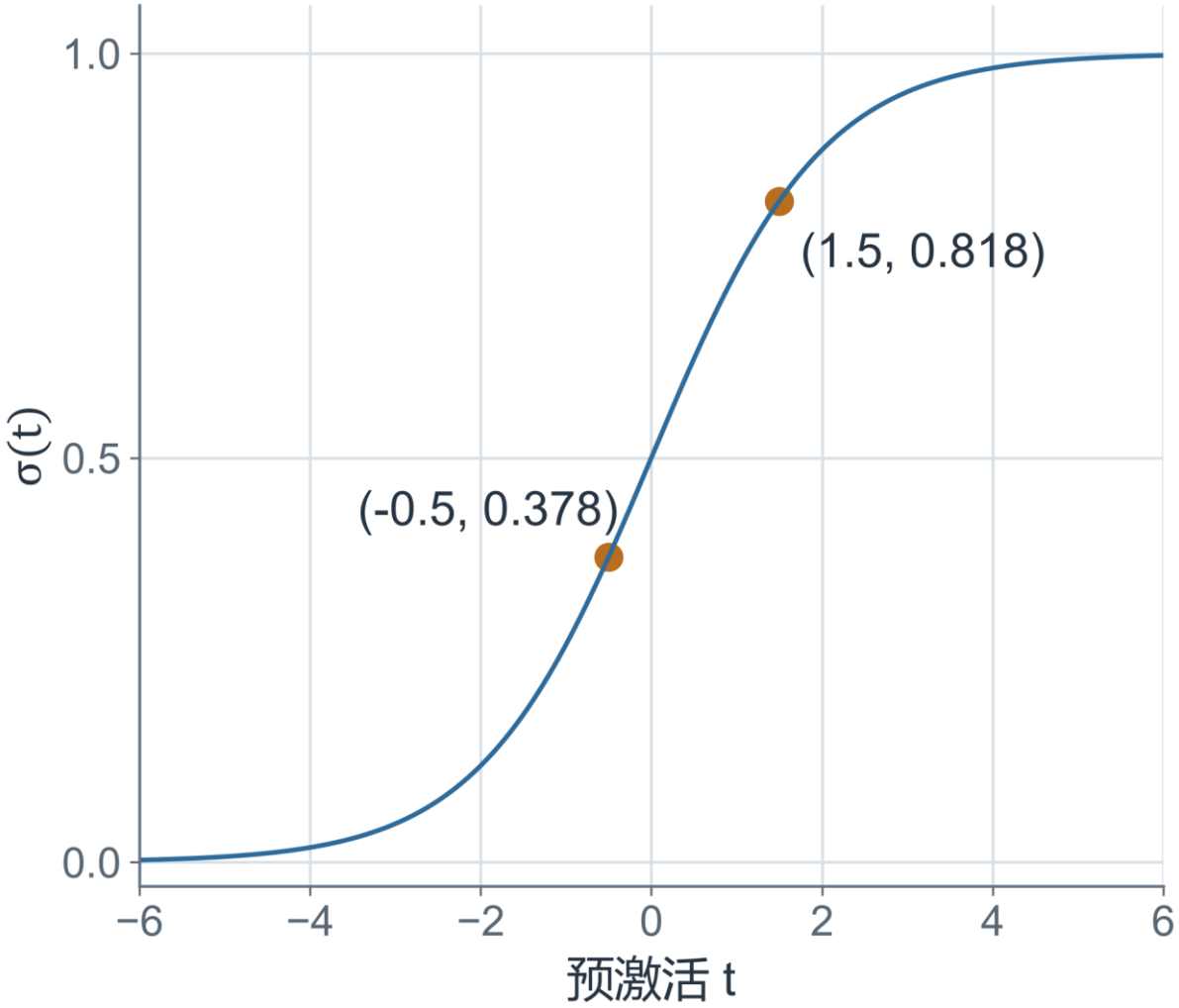
隐藏表示的形成与使用

- **隐藏表示**是同一样本的一组新特征
  - 原始特征 $\boldsymbol{x}$ 经网络计算得到 $\boldsymbol{h}$，各分量由隐藏单元产生
  - 这些特征由输入计算得到，不是额外采集的数据；通常没有直接给定的分量标签
- 输出层组合这些新特征，形成得分
  - 多个隐藏单元产生不同的非线性响应，输出层将它们加权组合
  - 组合后的得分可以在中间为正、两侧非正，回应前面的区间分类任务
  - 隐藏层参数决定特征生成规则，训练可以调整这些规则
- 从加权和到隐藏表示
  - **预激活量** $\boldsymbol{a} = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1$，是加入偏置后的加权和
  - **激活函数** ϕ 逐元素作用：$h_j = \phi(a_j)$，即 $\boldsymbol{h} = \phi(\boldsymbol{a})$
  - $o = \boldsymbol{W}_2\boldsymbol{h} + b_2$；本页单输出，$o, b_2$ 为标量



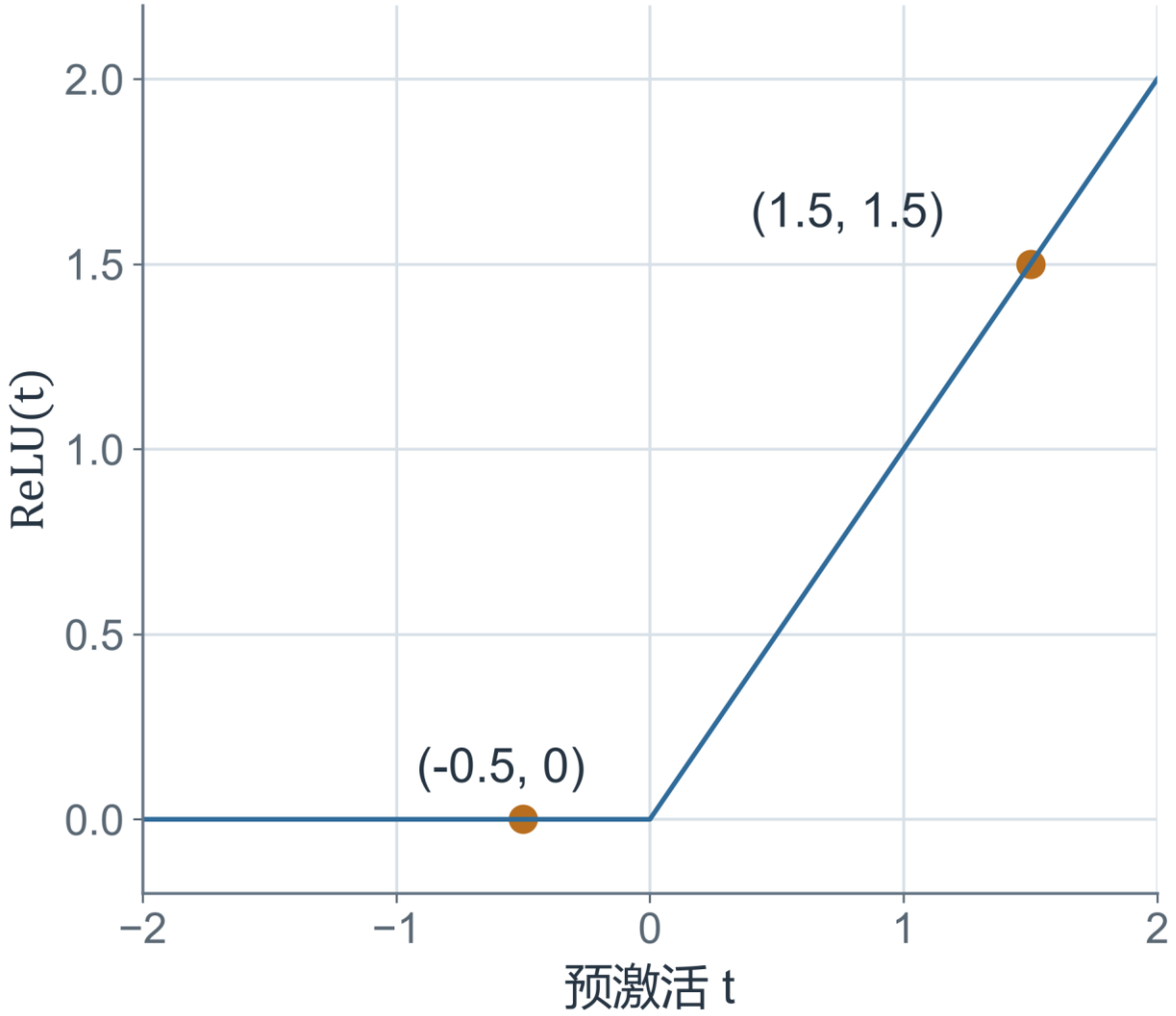
Sigmoid 的函数形式与逐元素作用

- **sigmoid 函数**为 $\sigma(t) = 1/(1 + \mathrm{e}^{-t})$
  - 实数输入，输出在 $(0, 1)$ 内；$\sigma(0) = 0.5$
- 单调递增的 S 形曲线，整体非线性
  - 两端分别趋近 0 与 1
- 隐藏层对各分量分别计算 sigmoid
  - $\boldsymbol{a} = (-0.5, 1.5)^{\mathsf{T}}$ 时，$\boldsymbol{h} \approx (0.3775, 0.8176)^{\mathsf{T}}$
- 隐藏输出不自动表示类别概率

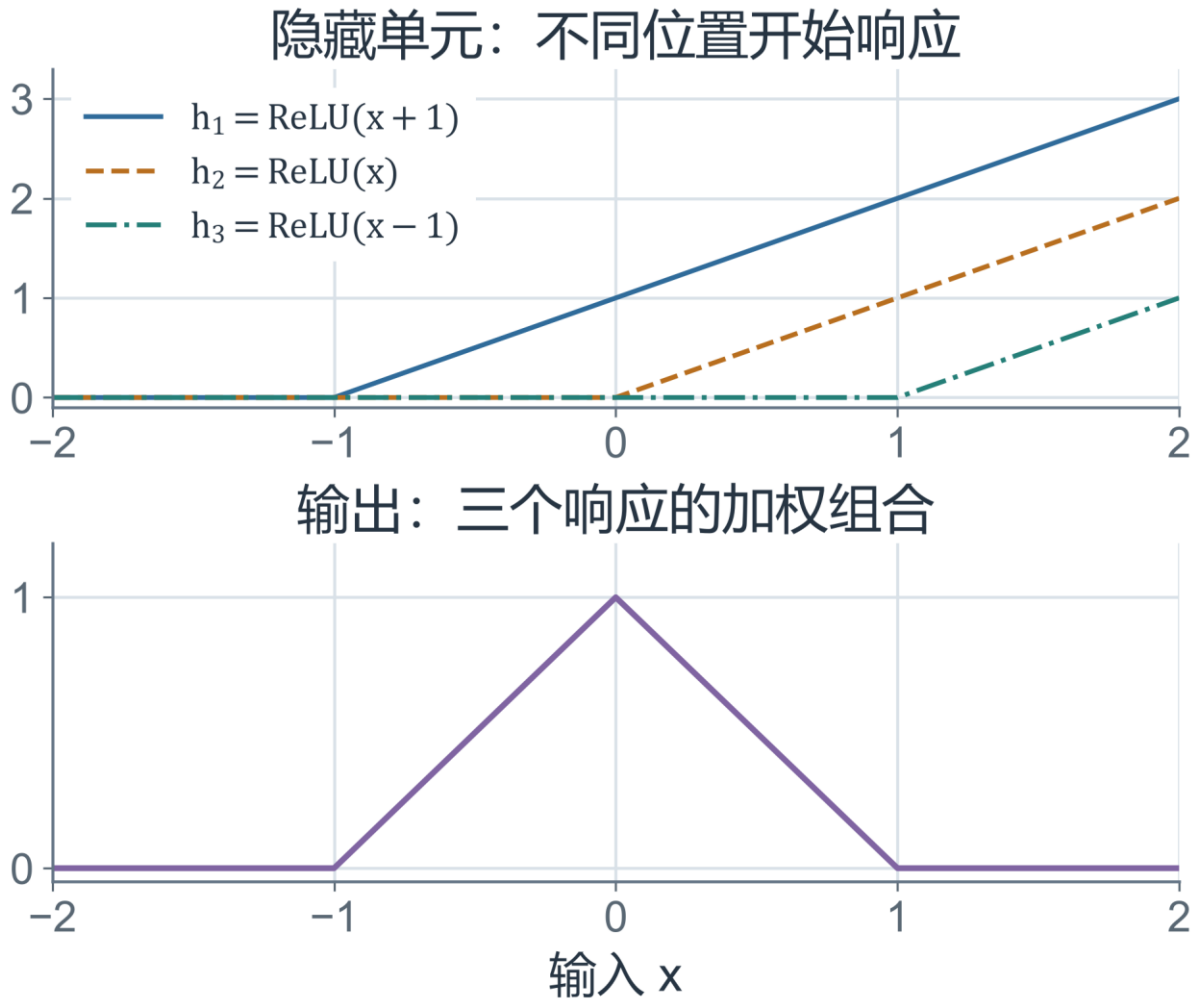


ReLU 的逐元素作用

- **修正线性单元（ReLU）**
  - 定义：$\phi(t) = \max(0, t)$
  - 负输入变为 0，非负输入保持原值
- 逐元素作用于预激活向量
  - $\boldsymbol{a} = (-0.5, 1.5)^{\top}$，$\boldsymbol{h} = (0, 1.5)^{\top}$
- 函数分段线性，整体为非线性
- 不同隐藏单元可响应不同输入区域



ReLU 隐藏响应的组合



- 本例由1个输入、3个隐藏单元和1个输出组成
  - 三个单元在不同位置开始响应；$\boldsymbol{h} = (h_1, h_2, h_3)^\top$ 是同一输入 x 的新特征表示
- 输出层组合这些响应，形成中间高、两侧为零的形状
  - 按 $f(x) > 0$ 判为正类，恰好得到 $-1 < x < 1$；端点归负类
- 图中构造的数学描述
  - $h_1 = \max(0, x+1)$，$h_2 = \max(0, x)$，$h_3 = \max(0, x-1)$
  - $f(x) = h_1 - 2h_2 + h_3$，输出权重为 $(1, -2, 1)$，偏置为 0

补充查阅：常见激活函数

函数	定义	输出范围	当前用途
ReLU	$\max(0, t)$	非负实数	本讲隐藏计算
sigmoid	$1/(1+\mathrm{e}^{-t})$	$(0, 1)$	二分类输出或隐藏激活
tanh	$(\mathrm{e}^{t}-\mathrm{e}^{-t})/(\mathrm{e}^{t}+\mathrm{e}^{-t})$	$(-1,1)$	隐藏激活比较

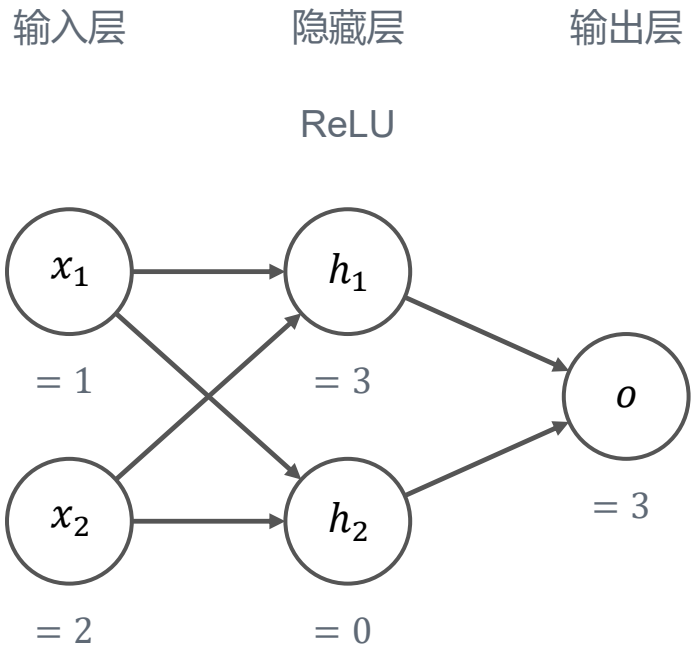
- 函数形式与所在位置共同决定职责

2 单隐藏层网络的前向计算

独立实例的前向计算与矩阵维度。

从输入到得分的前向计算

- **前向计算**沿输入、隐藏层、输出层依次求值；本页算到得分
  - 独立计算例：输入 $\boldsymbol{x} = (1,2)^\top$；参数人工给定，尚未训练
  - $\boldsymbol{W}_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$，$\boldsymbol{b}_1 = (0,0)^\top$
  - $\boldsymbol{W}_2 = [1 \quad -1]$，$b_2 = 0$
- 每个隐藏单元计算加权和，再作 ReLU
  - $a_1 = 1 + 2 = 3$，$a_2 = 1 - 2 = -1$
  - $\boldsymbol{h} = \mathrm{ReLU}(\boldsymbol{a}) = (3,0)^\top$
- 输出组合隐藏分量
  - $o = 1\times 3 - 1\times 0 = 3$
- 标签尚未参与计算；得分在后面转换为预测与损失

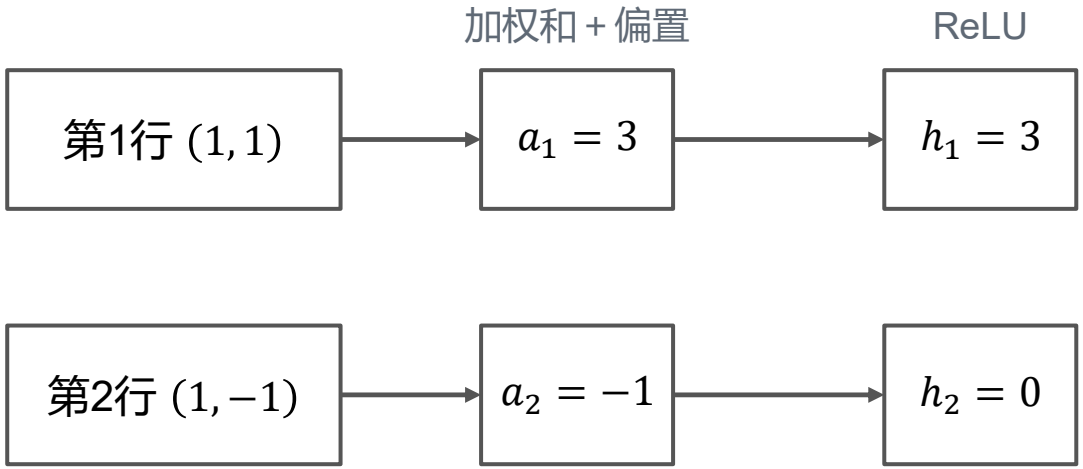


前向计算的矩阵表示与维度

- 单样本列向量：$\boldsymbol{x} \in \mathbb{R}^d$，隐藏宽度 m 为隐藏单元数
  - 本例为**全连接**：每个接收单元对上一层全部分量加权求和
- 隐藏层：$\boldsymbol{a} = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1$，$\boldsymbol{h} = \phi(\boldsymbol{a})$
  - $\boldsymbol{W}_1 \in \mathbb{R}^{m\times d}$，$\boldsymbol{b}_1, \boldsymbol{a}, \boldsymbol{h} \in \mathbb{R}^m$
- 输出宽度 k 为得分个数：$\boldsymbol{o} = \boldsymbol{W}_2\boldsymbol{h} + \boldsymbol{b}_2$
  - $\boldsymbol{W}_2 \in \mathbb{R}^{k\times m}$，$\boldsymbol{b}_2, \boldsymbol{o} \in \mathbb{R}^k$
- 前例 $d = 2, m = 2, k = 1$；单输出写作 o, b_2
  - 按带参数层计数为两层，输入不计层

前例 $\boldsymbol{W}_1$：行对应隐藏单元

两列依次对应 $x_1 = 1, x_2 = 2$



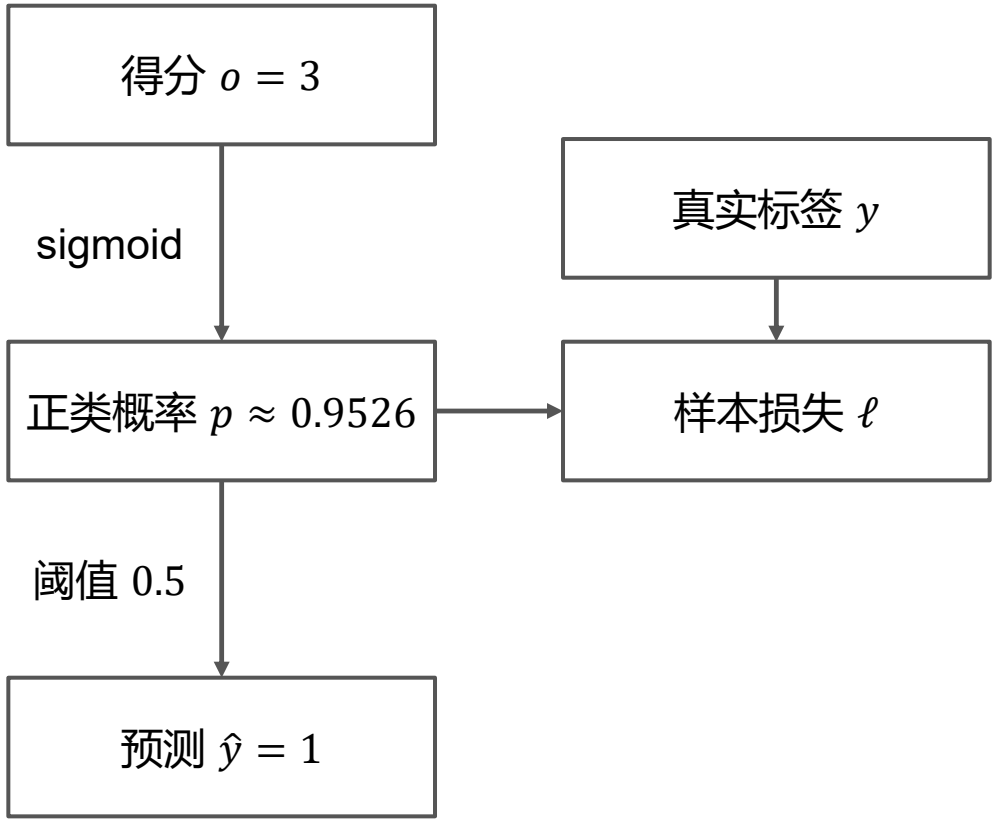
$(m\times d)(d\times 1) \rightarrow m\times 1$

3 预测与学习目标

从得分到损失，再到参数学习

二分类的得分、概率与预测

- 同一样本先得到得分，再形成概率与预测
  - sigmoid：$p = \mathrm{P}_\theta(y = 1 \mid \boldsymbol{x}) = 1/(1 + \mathrm{e}^{-o})$
  - 正类概率为 p，负类概率为 $1 - p$
- 按 $p \geq 0.5$ 预测 1，否则预测 0
  - 等价于按 $o \geq 0$ 判为 1
- 预测不使用真实标签；评价预测才需要标签
  - 概率 p 与真实标签 y 共同进入损失计算

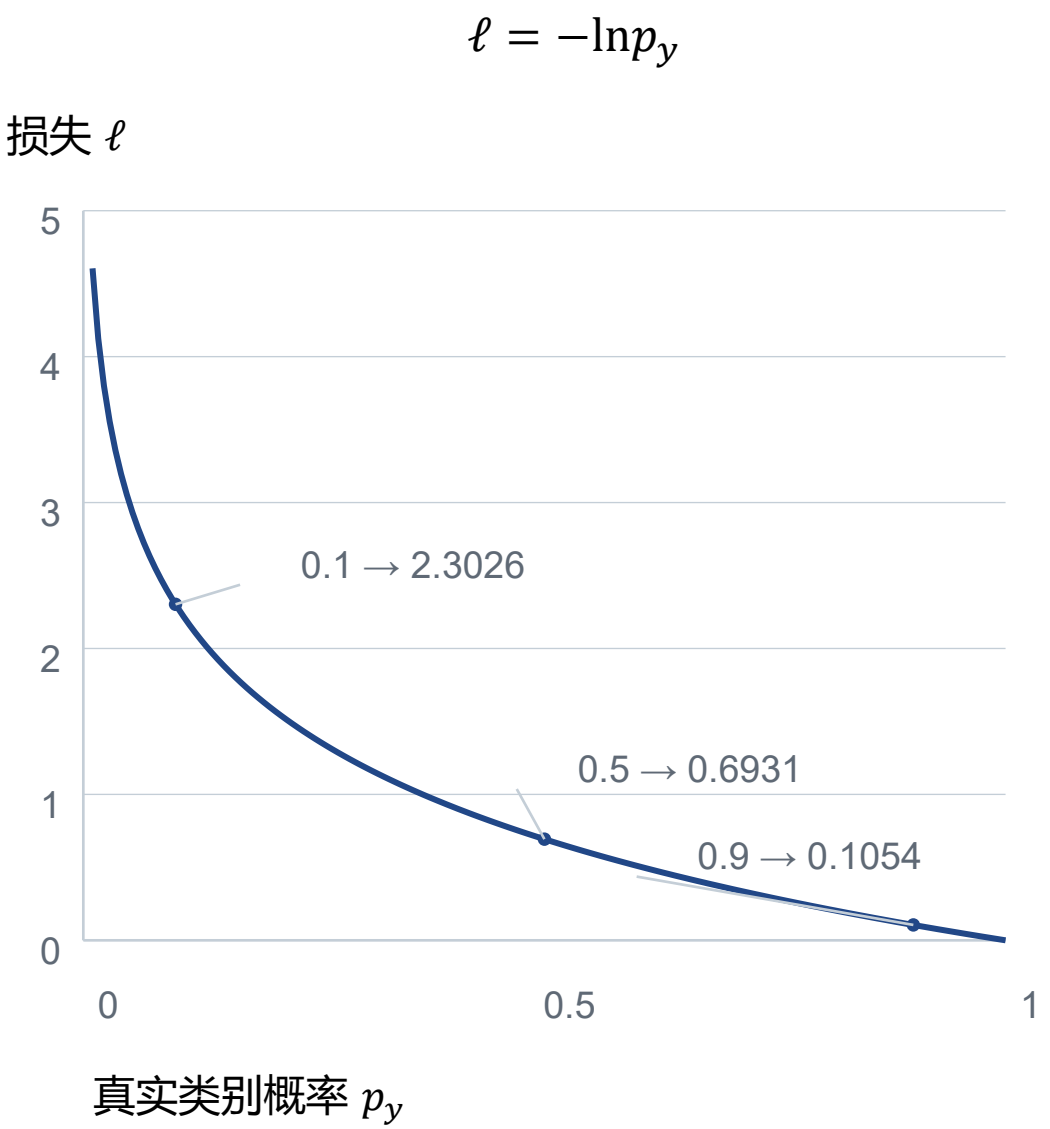


预测路径：不需要真实标签

评价路径：需要真实标签

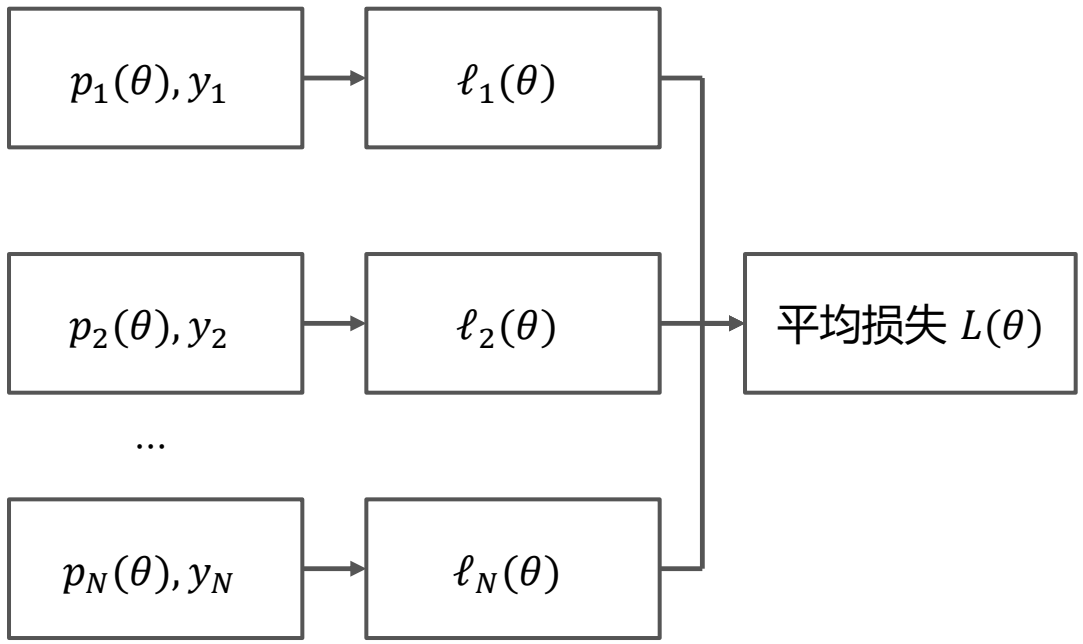
二分类交叉熵

- 预测标签相同，损失仍可能不同
  - 核心前向例给定 $y=1$：$p\approx 0.9526$，预测 1，损失约为 0.0486
  - 若正类概率为 0.9，仍预测 1，损失约为 0.1054
- **交叉熵损失**评价真实类别获得的概率 p_y
  - $y=1$ 时 $p_y=p$；$y=0$ 时 $p_y=1-p$
  - 图中 $\ell=-\ln p_y$：真实类别概率越低，损失越大
- 二分类标签 $y\in\{0,1\}$，交叉熵定义为
  - $\ell=-y\ln p-(1-y)\ln(1-p)$，ln 为自然对数
  - $y=1$ 时取 $-\ln p$；$y=0$ 时取 $-\ln(1-p)$



平均损失与参数学习目标

- 固定训练集，用同一组参数 θ 评价各样本
  - $\theta = (\boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{W}_2, \boldsymbol{b}_2)$
- 各样本的模型输出与真实标签共同决定损失
  - 对 N 个损失取平均：$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\theta)$
- 学习根据数据寻找使平均损失较小的参数
  - 隐藏参数改变表示，输出参数改变表示的使用方式
  - 前向求值与损失评价当前参数；更新方法在下一讲展开



所有样本共用当前参数 θ

二分类示意：下标为样本编号

求和后除以样本数 N

单标签多分类：softmax 与预测

- K 个类别互斥，每个输入有一个真实类别
- 输出层为每类给出一个实数得分 o_j，**softmax** 将整组得分转换为类别概率
  - $p_j = \mathrm{e}^{o_j} / \sum_{r=1}^{K} \mathrm{e}^{o_r}$，各概率为正且总和为 1
- 预测采用概率最大的类别：$\hat{y} = \mathrm{argmax}_j p_j$
  - 概率与得分的大小顺序相同，最大得分类别也相同
- 标签可用类别索引表示，也可用 one-hot 向量表示
  - **one-hot 向量**在真实类别位置为 1，其余位置为 0

多分类交叉熵与计算示例

- 独立的三类输出例；输出层宽度为 3
  - $\boldsymbol{o} = (2,1,0)^{\mathsf{T}}$，$\boldsymbol{p} \approx (0.6652,0.2447,0.0900)^{\mathsf{T}}$
- 预测取最大概率，固定为类别 1
  - 最大得分类别也相同；右图只改变真实标签
- 损失取真实类别对应的概率
  - $\ell = -\ln p_y$；真实类别为 1 时约为 0.4076，为 3 时约为 2.4076
  - one-hot 标签为 $\boldsymbol{t}$ 时，$\ell = -\sum_j t_j \ln p_j$
- 归一化数值核对
  - 指数值为 $(7.3891, 2.7183, 1)$，分母约为 11.1074

同一分布：分别比较两种标签

橙框为真实类；预测均为类别1

若 $y = 1$，$\ell \approx 0.4076$

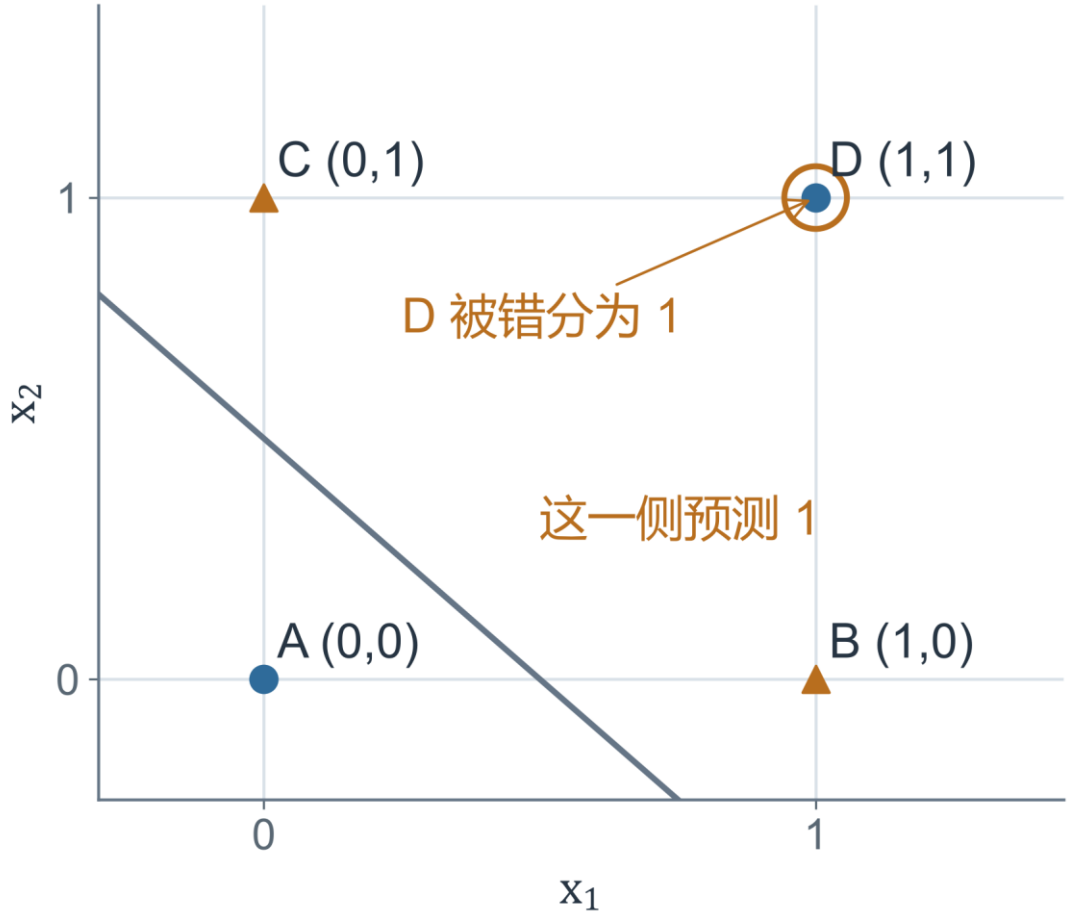


若 $y = 3$，$\ell \approx 2.4076$



迁移验证：二维 XOR 的表示问题

- 选讲验证；可整段跳至“表达能力、训练过程与泛化表现”
- 前面已用隐藏响应解决一维区间分类；现在检验它能否解决二维任务
- **异或（XOR）**：两个 0/1 输入不同则标签为 1，相同则为 0
- 若单一直线能分开四点，其得分需满足
  - 正类：$w_1 + b \geq 0$、$w_2 + b \geq 0$
  - 负类：$b < 0$、$w_1 + w_2 + b < 0$
  - 前两式相加与后两式相加，对同一量给出相反要求

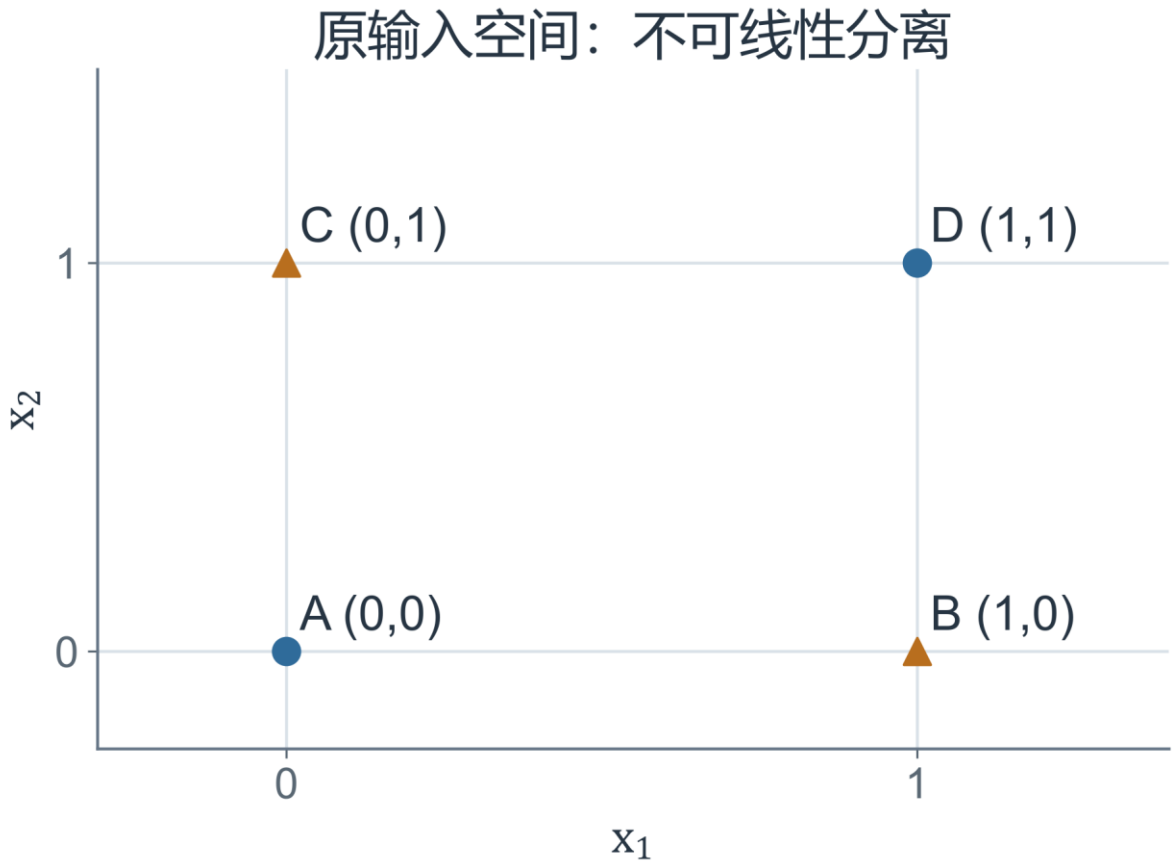


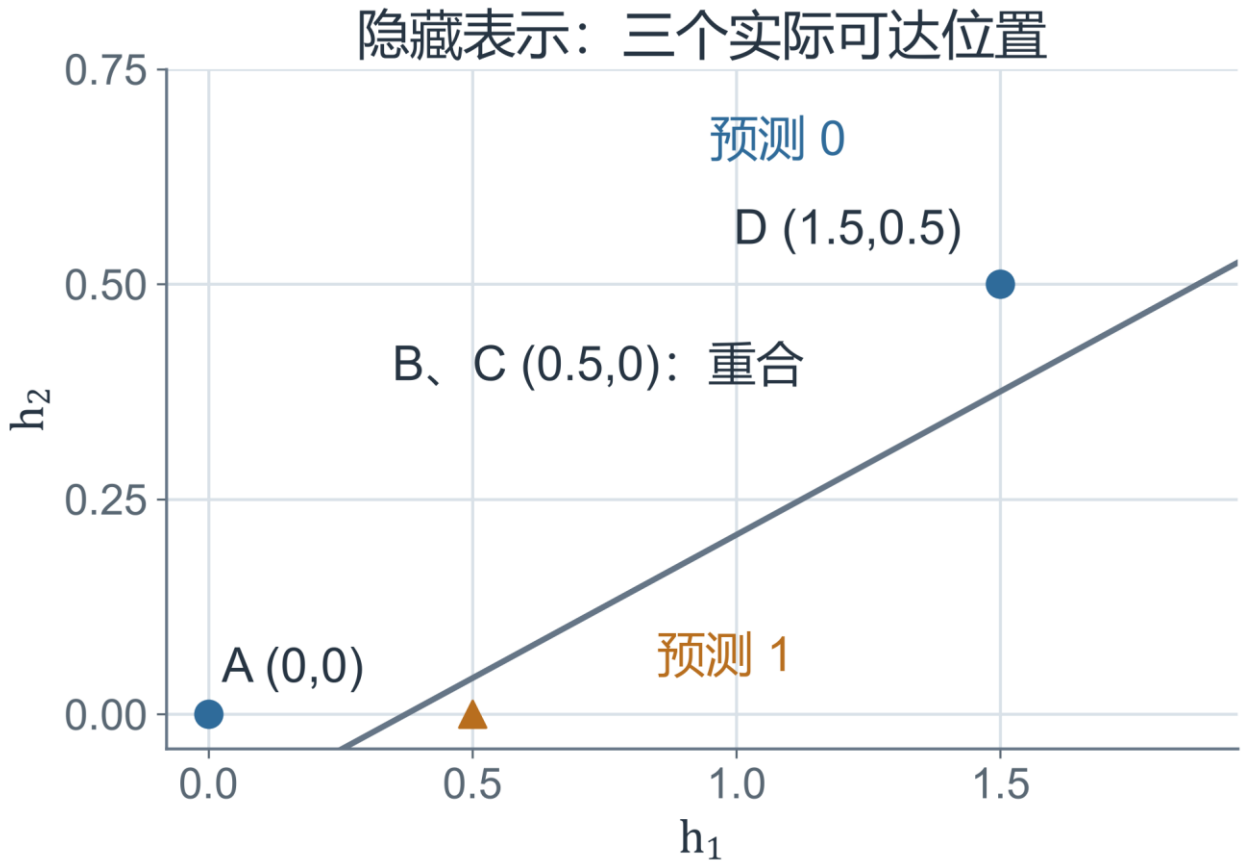
XOR 的隐藏特征构造与验证

- 分类要求：输入和决定 XOR 标签
  - $t = x_1 + x_2 \in \{0,1,2\}$
  - $t = 1$ 为正类；$t \in \{0,2\}$ 为负类
- 人工构造：两个 ReLU 单元响应不同阈值
  - $h_1 = \max(0, t - 0.5)$
  - $h_2 = \max(0, t - 1.5)$
  - 输出：$o = 4h_1 - 12h_2 - 1.5$
  - 核对：$t = 0,1,2$ 对应 $o = -1.5, 0.5, -1.5$；按 $o \geq 0$ 判为 1，得到标签 0、1、0
- 第一单元使中间状态得分变正，第二单元抵消末端的响应
  - 下一页用相同 A、B、C、D 观察隐藏空间中的分离

隐藏表示中的 XOR 分类关系

- B、C 映射为同一隐藏表示 $(0.5,0)^{\mathrm{T}}$，仍同属正类
- 边界 $4h_1 - 12h_2 - 1.5 = 0$ 将 B、C 与 A、D 分开





表达能力、训练过程与泛化表现

- 表达：是否存在能够实现任务关系的参数
  - 三角形响应展示一种可表达函数；选讲 XOR 也是人工构造验证
- 训练：能否根据数据找到合适的参数
  - 前向求值使用人工参数，尚未执行参数更新
- 泛化：学得模型在目标分布的未见样本上表现如何
  - 需要相应的数据与评价，当前计算例没有这类证据
- 隐藏参数改变表示，输出参数改变表示的使用方式
  - 宽度和深度可改变表达范围，训练与泛化仍需分别评价

本讲总结

- 非线性隐藏表示使网络能够表达超出仿射复合的关系
- 单隐藏层网络由隐藏变换与输出变换组成
  - 输入 → 预激活 → 隐藏表示 → 输出得分 → 任务输出
- 预测规则从输出得到类别，损失结合真实标签评价输出
- 参数同时影响隐藏表示和输出，平均损失规定学习目标
- 给定参数的前向结果与训练、泛化结果分别判断