06 Logistic回归与最大熵模型

从线性证据到条件概率

本讲学习目标

- 解释 Logistic 回归从特征、得分、概率到类别的推理，以及对数几率的线性结构
- 区分模型、监督目标、梯度与优化器，读懂训练—推理闭环
- 用特征变换解释原始空间的非线性边界，区分它与 sigmoid 的作用
- 解释熵、约束与条件最大熵，读懂特征加权、指数和归一化的模型形式
- 说明最大熵、条件似然、Logistic 与 softmax 的对应条件，认识正则化与求解边界

1 Logistic 概率分类模型

从线性得分建立类别概率，用给定参数完成预测，并解释对数几率与决策边界的线性结构。

从线性得分到类别概率

感知机按线性得分的符号输出类别。两个不同的正得分都判为正类，类别输出未保留证据强弱的差别。概率分类保留这种差别，供学习与决策使用。

▪ 把线性证据写成概率关系

- 得分可取任意实数，两个类别的概率必须非负且总和为 1
- Logistic 回归让两类概率之比的自然对数——**对数几率**——随特征线性变化
- 同一模型可写成 sigmoid 概率映射，再通过阈值判断类别

▪ 概率输出支持学习与决策

- 同样预测为正类的样本，模型给出的正类概率可以不同
- 训练评价模型给真实类别的概率；预测按概率与阈值判断类别

Logistic 回归直接建模 $P(y \mid x)$，属于判别式条件模型。线性对数几率是一种模型选择，概率的取值要求并不能唯一确定 sigmoid。

Logistic 回归的预测过程

输入与参数

- 特征向量 x 表示一个样本；二分类标签为 $y \in \{0,1\}$，正类为 1
- 使用给定的权重 w 和偏置 b 完成预测

线性得分

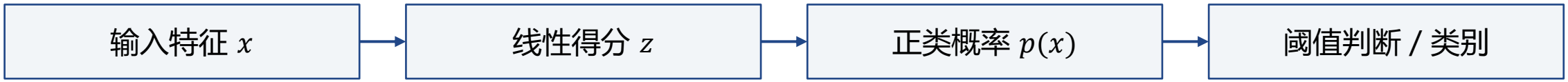
- $z = w^T x + b$，其中 $w^T x = \sum_j w_j x_j$

分类决策

- 以 τ 为阈值：$p(x) \geq \tau$ 预测为 1，否则预测为 0；本讲取 $\tau = 0.5$

条件类别概率

- $p(x) = P(Y = 1 \mid X = x) = \sigma(z) = \frac{1}{1+\exp(-z)}$
- 正类概率位于 $(0, 1)$，负类概率为 $1 - p(x)$
- **sigmoid** 单调递增：得分越大，正类概率越高
- $z = 0$ 时 $p = 0.5$，$z > 0$ 时 $p > 0.5$，$z < 0$ 时 $p < 0.5$



Logistic 回归的预测示例

- **输入与参数**
  - $x = (2,1)$，$w = (1,-1)$，$b = 0$
- **计算线性得分**
  - $z = w^T x + b = 1\times2 + (-1)\times1 = 1$
- **计算正类概率**
  - $p(x) = \sigma(1) \approx 0.731$
- **判断类别**
  - 取 $\tau = 0.5$，因为 $0.731 > 0.5$，所以 $\hat{y} = 1$
- **结果含义**：得分是特征的加权结果，概率是模型的条件预测，类别由阈值决定

本例参数为教学构造，该样本的真实类别未知。

对数几率与线性得分

sigmoid 形式用于计算类别概率。对数几率形式表达同一个模型，说明它的线性结构及概率阈值对应的决策边界。

▪ 对数几率连接线性得分

- 正类几率 $p/(1-p)$ 是正类概率与负类概率之比
- 对数几率是几率的自然对数
- 由 $p=\sigma(z)$ 得到 $p/(1-p)=e^z$，因此

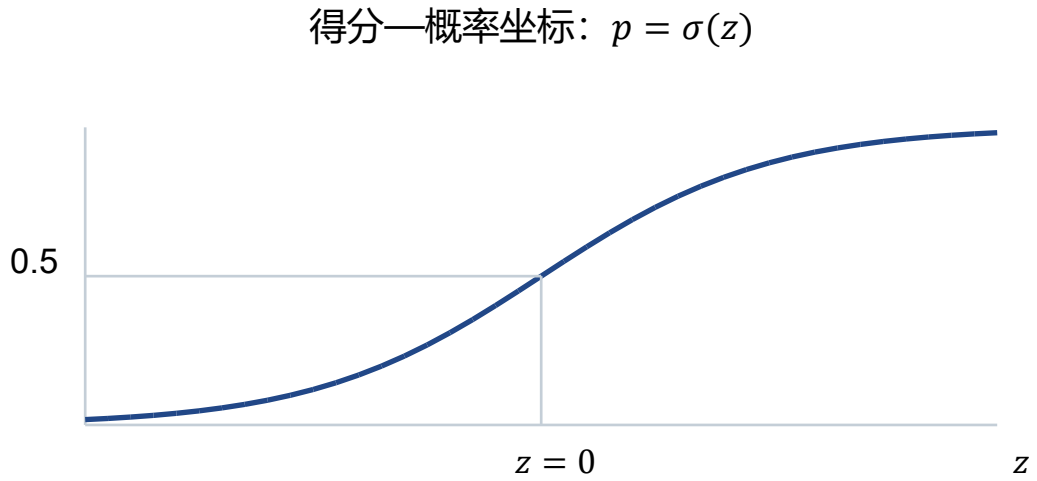
$$\log\frac{p(x)}{1-p(x)}=z=w^Tx+b$$

- 对数作用于两类概率之比；输入特征通过加权和与偏置决定对数几率，因此对数几率关于输入特征呈线性关系

概率阈值与线性决策边界

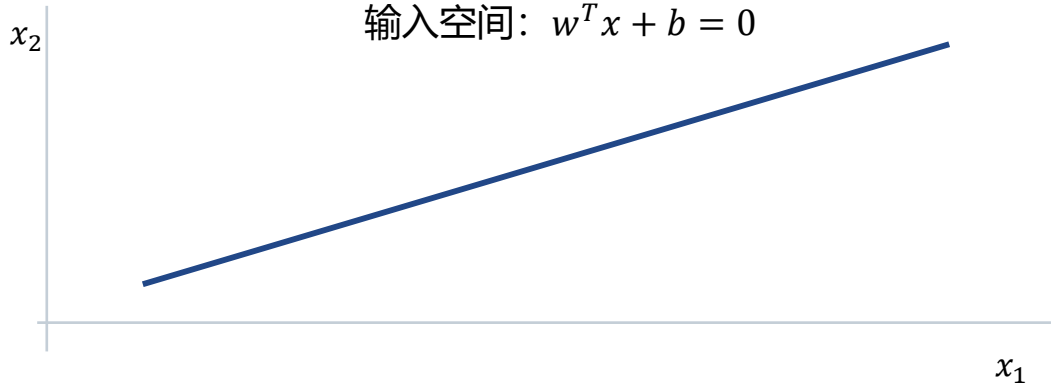
概率阈值确定决策边界

- 阈值为 0.5 时，临界概率 $p = 0.5$ 对应得分 $z = 0$
- 因此，输入空间中的边界满足 $w^T x + b = 0$
- 当 $w \neq 0$ 时，二维输入上的边界是一条直线



概率曲线与决策边界

- S 形曲线位于得分—概率坐标中，说明得分怎样转成概率
- 直线位于输入空间中，表示得分为零的输入；边界形状由得分函数决定



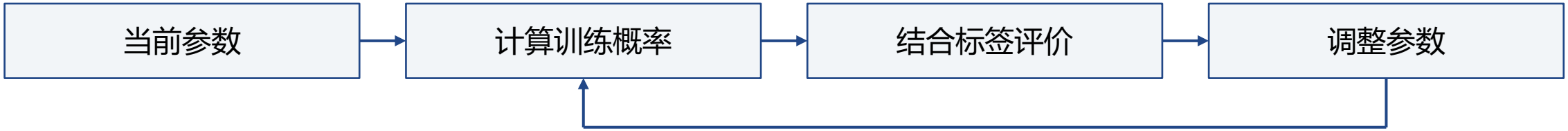
2 损失与参数学习

用真实标签评价预测概率，将损失汇成训练目标，再利用梯度调整参数，区分训练与预测的职责。

Logistic 回归参数学习的基本思路

上一例用给定参数完成预测。训练数据提供真实标签，可用于评价模型分配的类别概率，并指导参数调整；这种标签反馈形成训练目标。

- **计算概率**：使用当前的 w, b，计算训练样本的类别概率
- **评价参数**：根据预测概率与真实标签计算损失，将样本损失取平均并加入正则化，形成目标函数
- **调整参数**：优化器利用目标的变化信息调整参数，再重新计算概率和目标值



用交叉熵损失评价真实类别的预测概率

真实标签固定为 $y = 1$ 时，$p = 0.49$ 与 $p = 0.01$ 在阈值 0.5 下都预测错误，但模型给真实类别的概率不同。**学习应提高模型给真实类别的概率，而不只统计类别是否判错。**

- **先评价已观察类别的概率**
  - 取真实类别的模型概率，损失为 $-\log P(y_{true} \mid x)$
  - 真实类别概率越低，**负对数损失**越大；二分类按真实标签选取正类或负类概率
- **用交叉熵精确表达同一评价**
  - Q 为标签分布：已观察类别取 1，其余取 0，称为 **one-hot**；P 为模型输出分布
  - **交叉熵**：$H(Q,P) = -\sum_y Q(y)\log P(y)$
  - one-hot 使求和只保留已观察类别，得到上面的真实类概率负对数
- **符号与约定**：y 遍历类别；log 为自然对数

二元交叉熵损失的计算示例

▪ 单样本二元交叉熵

- 根据真实标签，评价模型给该标签分配的概率
- $\ell = -[y\log p + (1-y)\log(1-p)]$
- 当 $y=1$ 时保留 $-\log p$；当 $y=0$ 时保留 $-\log(1-p)$

另取一组独立计算例子：真实标签固定为 $y=1$，比较两种预测概率。

正类概率	损失计算	解释
$p=0.9$	$\ell = -\log(0.9) \approx 0.105$	给真实类别较高概率，损失较小
$p=0.1$	$\ell = -\log(0.1) \approx 2.303$	给真实类别较低概率，损失较大

训练推动模型提高真实标签的条件概率。 两种标签分支遵循同一评价原则。

Logistic 回归的训练目标

参数改变预测概率，进而影响目标值。训练最小化 $J(w,b)$。

- **平均监督目标**：训练集有 N 个样本 (x_i, y_i)，$y_i \in \{0,1\}$
  - $z_i = w^T x_i + b, p_i = \sigma(z_i)$
  - $J(w,b) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)] + \frac{\alpha}{2}\|w\|_2^2$
  - $\alpha \geq 0$ 是**正则化**强度，$\|w\|_2^2$ 为权重平方和；本讲不惩罚偏置 b
  - 正则化限制过大的权重，缓解过拟合并改善稳定性
  - 单样本损失为 $\ell_i = -\log P(Y = y_i \mid X = x_i)$；按独立样本建模，训练标签的条件概率相乘形成条件似然
  - 取负对数使乘积变为损失求和；不加正则化时，求和或平均的最小化都对应最大条件似然

Logistic 回归的残差与梯度

前页已定义训练目标。梯度用概率与标签的残差，描述参数的局部变化对目标的影响。

- **梯度**与残差：梯度表示目标随参数的局部变化
  - $\nabla_w J = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)x_i + \alpha w, \frac{\partial J}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)$
  - $p_i - y_i$ 是概率与 0/1 标签的残差：正类标签时为 $p_i - 1 < 0$，负类标签时为 $p_i > 0$。x_i 将残差转换为各权重的样本贡献，汇总后用于计算同一训练目标的梯度
  - 某个参数的偏导数为正时，小幅增大该参数会使目标局部增大；偏导数为负时，目标局部减小。优化器据此寻找降低目标的参数

Logistic 回归的训练与预测对比

训练与预测共用概率模型；输入信息、参数处理和输出各有不同。

比较项	训练	预测
输入信息	带标签样本 (x_i, y_i)，真实标签用于计算损失	新样本 x，真实标签未知
参数处理	计算概率与损失，优化器更新 w, b 并重复	固定训练后的 w, b，计算得分与概率
输出结果	学到的一组模型参数	样本的类别概率，以及通过阈值得到的预测类别

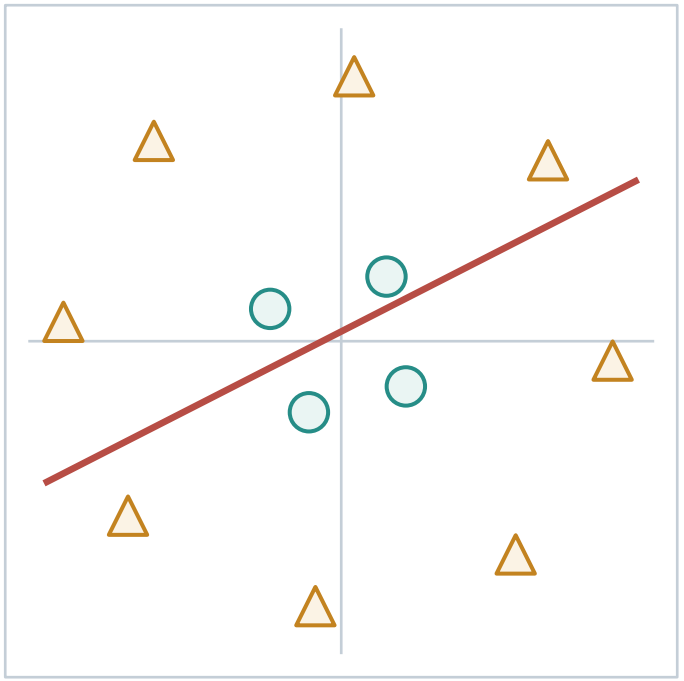
3 Logistic 回归的特征表示与非线性边界

以中心类与外围类说明原始表示的直线限制，用半径平方特征建立阈值，再映回原空间的圆形边界。

直线无法分开本例的中心类与外围类

原始二维坐标下，Logistic 回归调整参数仍只能移动或旋转直线边界。

- **原坐标的表示限制**
  - 特征为 (x_1, x_2)，边界满足 $w_1x_1 + w_2x_2 + b = 0$
  - 当 w_1, w_2 不同时为零时，边界是一条直线
  - 本例中心类靠近原点，外围类分布在环带
  - 直线无法完全分开两类，封闭边界可以围住中心区域

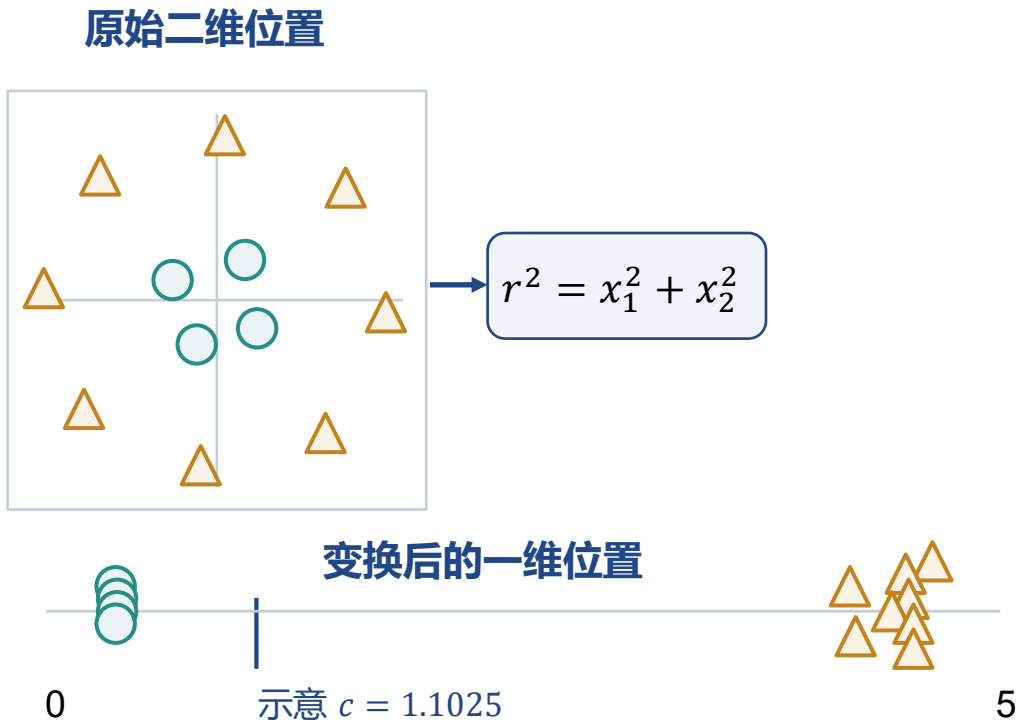


两类样本按离原点远近区分。将这种几何关系作为特征，才能改变当前直线表示的限制。

半径平方特征与一维阈值分类

二维坐标映射为半径平方特征。

- **半径平方特征**
  - $r^2 = x_1^2 + x_2^2$
  - 数值越小越靠近原点，越大越远离原点
- **同一批样本在新特征轴上的位置**
  - 中心点：0.365 至 0.3725；外围点：4.2125 至 4.645

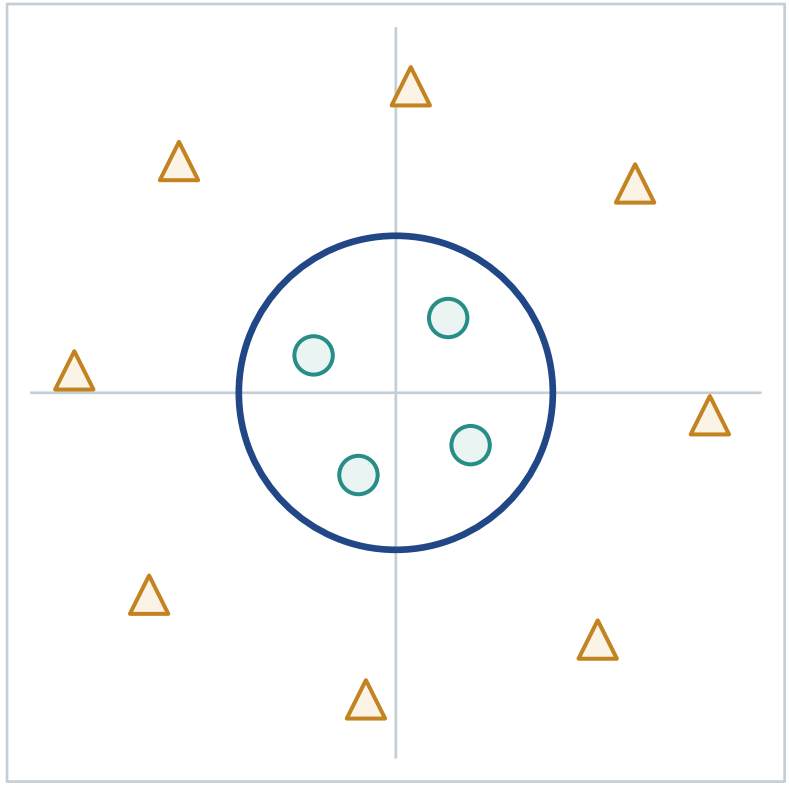


- **使用新特征的 Logistic 回归**
  - 得分 $z = wr^2 + b$ 关于新特征线性，sigmoid 将得分转为类别概率
  - 本例提供半径平方特征，模型学习权重 w 与偏置 b，形成轴上阈值

线性阈值在原始空间中形成圆形边界

同一边界在半径平方轴上是阈值，在原始二维空间中是圆。

- **特征轴上的线性阈值**
  - 概率阈值为 0.5，边界满足 $wr^2 + b = 0$
- **原空间中的圆形边界**
  - 代入 $r^2 = x_1^2 + x_2^2$，得到 $x_1^2 + x_2^2 = -b/w$
  - $w \neq 0$ 且 $-b/w > 0$ 时，边界是以原点为圆心的圆
- **表示与概率映射的分工**
  - 非线性边界来自特征变换；sigmoid 仍负责得分到概率的映射
  - 特征变换改变表达能力，也可能增加过拟合风险



特征空间：$wr^2 + b = 0$　　原始空间：$x_1^2 + x_2^2 = c$

4 条件最大熵的建模与参数学习

从数据约束未唯一确定分布的问题进入最大熵原则，再把特征统计写成约束，连接条件模型与参数学习。

类别概率建模的两条路径

共同任务是利用带标签的数据，建立输入 x 到类别概率 $P(y \mid x)$ 的模型。

▪ 前面的 Logistic 路径

- 先定义从特征得分到类别概率的模型
- 再通过训练目标学习参数；特征表示决定模型能表达怎样的边界

▪ 现在的最大熵路径

- 先把训练数据中的输入与类别关系写成特征约束
- 再选择满足这些约束的条件类别分布

▪ 为什么需要选择原则

- 多个分布可能满足相同约束，数据提供的信息尚未唯一确定答案
- 已知事实应保留；对事实尚未区分的部分，需要避免无依据的额外偏好
- 最大熵把这一选择思路写为约束内的熵最大化

熵衡量概率分布的不确定性

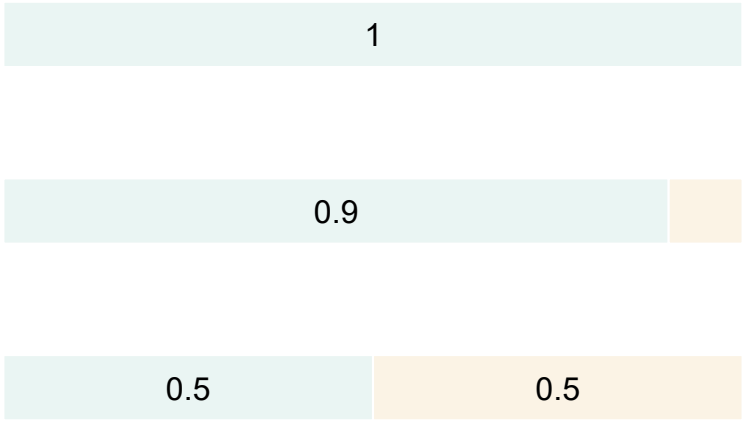
当多个分布满足已知约束时，熵提供比较分布自身不确定性的尺度。

$H(P) = -\sum_y P(y)\log P(y)$

P 为结果分布，y 遍历结果，log 为自然对数，约定 $0\log 0 = 0$。

熵将各结果的信息量按概率加权，衡量分布自身的不确定性。

分布	结果的不确定性	熵比较
确定分布（1，0）	一个结果的概率为 1，结果确定	最低
偏斜分布（0.9，0.1）	一个结果明显占优，仍有不确定性	较低
均匀分布（0.5，0.5）	两个结果同样可能	最高



熵评价分布自身；监督交叉熵用已观察标签评价模型预测。 熵高不代表分类更准确。

最大熵原则：在约束内选择分布

最大熵原则先保留已知信息，再在满足这些信息的分布中选择熵最大的一个。

- **约束确定可行集合**
  - 候选分布的概率非负且总和为 1，并满足选定的经验约束
  - 满足全部约束的分布构成可行集合
- **熵用于选择分布**
  - 在可行集合内最大化熵，处理约束尚未确定的部分
  - 对约束未区分的结果，避免加入额外偏好
- **均匀分布的适用情形**
  - 对有限结果集合，只有非负与归一化要求、没有其他区分性约束时，最大熵结果是均匀分布

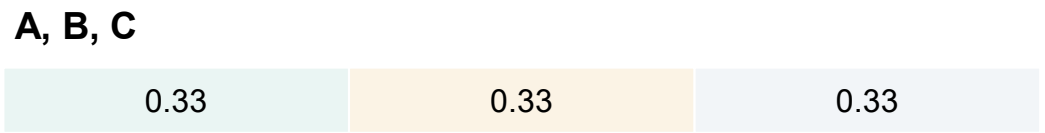


最大熵分布选择示例

结果为 $\{A, B, C\}$：比较无额外约束与 $P(A) = 0.5$ 两种情形。

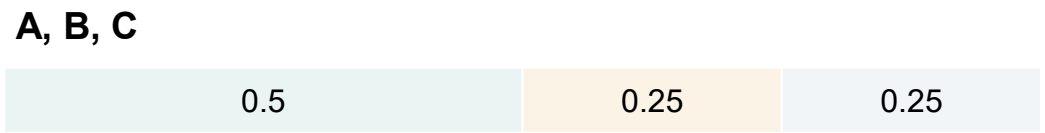
▪ 只有概率非负与归一化要求

- 没有信息区分三个结果，最大熵分布为 $(1/3, 1/3, 1/3)$



▪ 加入约束 $P(A) = 0.5$

- 剩余概率满足 $P(B) + P(C) = 0.5$；$(0.5, 0.5, 0)$ 与 $(0.5, 0.25, 0.25)$ 均可行
- 将剩余质量均分给未被区分的 B, C，熵较大，选 $(0.5, 0.25, 0.25)$



▪ 两种结果的共同原则

- 先满足已知约束，再在约束未确定的部分最大化熵
- 给定 $P(A) = 0.5$ 后，三类均分已经不满足约束

经验平均与模型平均描述同一特征关系

训练中观察到的关系，怎样约束概率模型？以下四个构造样本中，固定输入，比较真实标签计数与模型预计计数。

输入	真实标签	样本数
$x = 1$	1,1,0	3
$x = 0$	0	1

- **特征函数** $f_1(x, y)$
  - $x = 1$ 表示输入证据出现，$y = 1$ 表示正类
  - $f_1(x, y)$ 在 $x = 1, y = 1$ 时取 1，否则取 0
- **经验平均：数据实际出现多少**
  - 该特征出现两次，$E_{data}[f_1] = 2/4 = 0.5$
- **模型平均：模型认为会出现多少**
  - 令 $q = P(Y = 1 \mid x = 1)$，前三个输入各贡献 q，最后一个贡献 0，因此 $E_{model}[f_1] = 3q/4$

两种平均使用同一特征，约束要求模型平均匹配经验平均。

将训练数据中的特征关系写成约束

特征函数 $f_j(x, y)$ 描述输入—候选类别关系，区别于输入特征 x。本例的 0/1 函数表示“证据出现且为正类”，其经验平均提供统计约束。

▪ 经验期望：观察到的统计事实

- $E_{data}[f_j] = \frac{1}{N}\sum_{i=1}^{N} f_j(x_i, y_i)$
- 按真实标签求特征平均；N 为样本数，i 为样本索引，j 为特征编号

▪ 模型期望：预测同一统计量

- $E_{model}[f_j] = \frac{1}{N}\sum_{i=1}^{N} \sum_y P(y \mid x_i) f_j(x_i, y)$
- 对每个训练输入按类别概率加权，再对同一批输入取平均

▪ 统计约束：匹配整体平均

- $E_{model}[f_j] = E_{data}[f_j]$
- 四样本中，$3q/4 = 0.5$ 得到 $q = 2/3$；$x = 0$ 下的概率仍未确定

模型不要求逐样本复制标签，而要求所选统计关系在整体上匹配。

条件最大熵在可行分布中作选择

$x = 0$ 下正类概率可取 0、$1/2$ 或 1，都满足同一特征约束。

- **已知信息：特征期望约束**
  - 模型需满足 $E_{model}[f_j] = E_{data}[f_j]$
- **可行集合：满足约束的条件分布**
  - 对每个 x，要求 $P(y \mid x) \geq 0$、$\sum_y P(y \mid x) = 1$，并满足全部选定的特征约束
- **本例中的剩余自由度**
  - $P(Y = 1 \mid x = 1) = 2/3$ 已固定；输入权重为 $3/4$ 与 $1/4$
  - 最大熵只选择尚未确定的 $x = 0$ 分布，得到 $P(Y = 1 \mid x = 0) = P(Y = 0 \mid x = 0) = 1/2$
- **选择原则：在可行集合内最大化条件熵**
  - **条件熵**按固定经验输入频率 $\tilde{P}(x)$ 加权各输入下的类别熵
  - $H_P(Y \mid X) = -\sum_x \tilde{P}(x) \sum_y P(y \mid x) \log P(y \mid x)$

条件最大熵的解具有对数线性形式

统计约束为特征配置权重，熵的导数产生对数概率，按类别归一化得到概率模型；下面给出这三项作用的最短数学联系。

▪ 熵与约束构成拉格朗日函数

- 拉格朗日乘子 λ_j 对应特征期望约束；同时加入归一化约束

▪ 驻点：令概率偏导为零

- 熵项导数：对 $p > 0$，$\frac{d}{dp}(-p\log p) = -(1 + \log p)$
- $\log P(y \mid x) = \sum_j \lambda_j f_j(x, y) + c(x)$

▪ 指数化并按类别归一化

- 定义特征加权得分 $s_\lambda(x, y) = \sum_j \lambda_j f_j(x, y)$
- $P_\lambda(y \mid x) = \frac{\exp(s_\lambda(x,y))}{Z_\lambda(x)}$
- $Z_\lambda(x) = \sum_{y'} \exp(s_\lambda(x, y'))$
- y' 遍历全部类别；与类别无关的项 $c(x) = -\log Z_\lambda(x)$ 由归一化确定

适用条件：只覆盖正经验输入频率下的正概率内点；强迫零概率时不能用有限权重表示。

对数线性模型重现同一约束分布

- **同一例子的解**
  - 仅有 $f_1(x, y)$，取 $\lambda_1 = \log 2$
  - $x = 1$ 时两类的非归一化权重为 2 和 1，正类概率为 $2/(2+1) = 2/3$
  - $x = 0$ 时两类得分为 0，非归一化权重均为 1，两类概率各为 $1/2$

同一权重同时解释两种输入：有证据时偏向正类，无证据时两类均分。它重现了前面的约束与最大熵选择；一般权重可通过条件似然估计。

条件最大熵与最大条件似然

条件最大熵在分布空间选择分布，最大条件似然在参数空间估计权重。

分布空间视角：条件最大熵

- 给定经验特征约束
- 在满足约束的条件分布中选择条件熵最大的分布

参数空间视角：最大条件似然

- 给定前页的对数线性模型 $P_\lambda(y \mid x)$
- 最大化训练标签的条件似然，等价于最小化平均负条件对数似然
- $J(\lambda) = -\frac{1}{N}\sum_{i=1}^{N} \log P_\lambda(y_i \mid x_i)$

成立条件

- 标准条件最大熵设定，相同特征函数、相同经验输入分布、无正则化，并满足前页有限权重解的适用条件

标准条件下，两种入口对应同一对数线性参数估计问题。

负对数似然梯度与特征期望差

当前模型低估特征平均

- 同一四样本的经验平均为 $E_{data}[f_1] = 1/2$
- 当前取 $\lambda_1 = 0$，已定义的正类概率为 $q = 1/2$
- 模型平均为 $E_{model}[f_1] = 3q/4 = 3/8$，低于经验平均

最优点的期望匹配与边界

- 无正则化且存在有限可微最优点时，零梯度对应两种期望相等
- 加入正则化后，最优点在期望匹配与参数规模之间折中

期望差给出学习方向

- $\frac{\partial J}{\partial \lambda_j} = E_{model}[f_j] - E_{data}[f_j]$
- 本例梯度 $3/8 - 1/2 = -1/8$，增大 λ_1 使目标局部下降
- 增大权重提高 q，朝匹配值 $2/3$ 调整
- 两种期望对同一批输入取平均；J 沿用前页目标

5 共同模型与求解边界

在匹配条件下比较最大熵、softmax 与 Logistic 的共同形式，随后说明数值计算和参数求解的边界，回收全讲认识。

二类 softmax 化为得分差的 sigmoid

多类 **softmax** 按指数得分归一化；二类形式可以改写为两个类别得分之差的 sigmoid。

多类别 softmax

- $P(Y=k \mid x)=\frac{\exp(s_k(x))}{\sum_c \exp(s_c(x))}$
- $s_k(x)$ 是类别 k 的得分，c 遍历全部候选类别

二分类化简

- $P(Y=1 \mid x)=\frac{\exp(s_1)}{\exp(s_1)+\exp(s_0)}=\sigma(s_1-s_0)$
- 若 $s_1(x)-s_0(x)=w^T x+b$，便得到标准 Logistic 形式

得分差决定二类概率

- 两个得分共同加同一常数不改变概率，二类概率只依赖得分差

条件最大熵与 Logistic 共享对数线性形式

最大熵入口

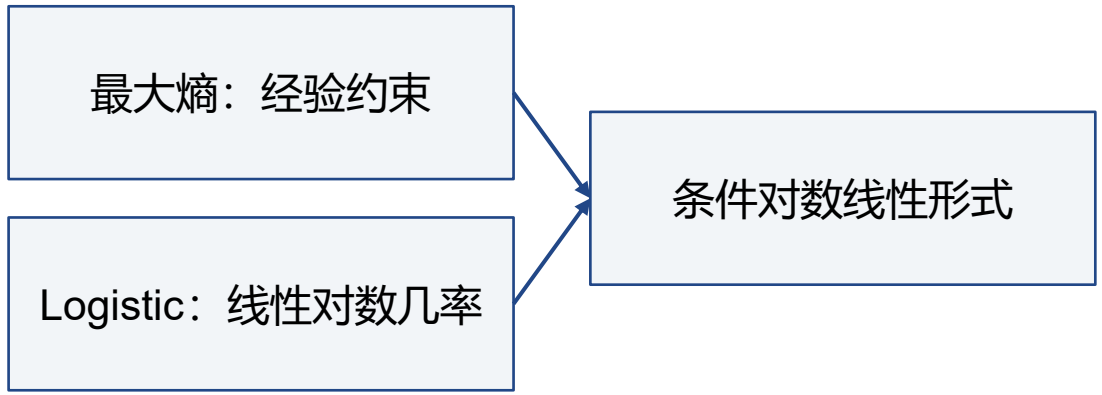
- 从经验特征约束和最大条件熵选择分布

共同模型形式

- 特征证据加权形成类别得分
- 指数化并按类别归一化得到 $P(y \mid x)$

Logistic 入口

- 从线性对数几率直接定义条件概率模型



成立条件

- 使用相匹配的条件模型、特征表示和参数化
- 多类形式对应 softmax，二类基准参数化对应 Logistic 回归

同一例子的参数对应

- 前述四样本示例中 $x \in \{0,1\}$，两类得分为 $\lambda_1 x$ 和 0
- 因此 $P(Y=1 \mid x)=\sigma(\lambda_1 x)$，对应 Logistic 的 $w=\lambda_1=\log 2, b=0$
- $x=1$ 时为 $2/3$，$x=0$ 时为 $1/2$；此对应固定 $b=0$，另学自由偏置会改变模型与约束

二分类 Logistic 回归的稳定损失计算

交叉熵可等价写成得分形式，用于稳定计算，训练目标的含义不变。

- **已有监督目标**：二分类单样本交叉熵为 $\ell_i = -[y_i \log p_i + (1-y_i)\log(1-p_i)]$，其中 $y_i \in \{0,1\}$、$p_i = \sigma(z_i)$
- **等价的得分形式**
  - $softplus(z) = \log(1+e^z)$，$\ell_i = softplus(z_i) - y_i z_i$
  - logits 指概率的对数几率，在当前模型中即得分 $z_i = w^T x_i + b$
- **稳定计算的用途**
  - 实际计算使用库提供的稳定 logits 损失；直接计算指数仍可能溢出
  - 等价计算保持已有数据集监督目标、正则化约定和学习含义

优化方法、有限解与停止条件

训练目标已确定，优化方法负责搜索参数；有限解的存在性与迭代的停止条件需分别判断。

优化方法使用的信息

- 梯度下降使用一阶梯度
- 牛顿法使用梯度与 Hessian，Hessian 表示目标的二阶变化
- 拟牛顿法根据迭代信息近似曲率

优化问题的理论边界

- 二分类线性 Logistic 回归的负对数似然是凸函数，加入 L_2 正则化后仍为凸优化问题
- 数据完全线性可分且无正则化时，参数可能不断增大，有限极大似然解可能不存在
- 凸性不保证有限最优参数存在

工程停止条件

- 目标下降量、梯度范数、参数更新量或最大迭代次数，可作为停止依据
- 满足停止条件表示迭代结束，不自动证明已得到最优参数

优化器决定如何搜索参数，不改变概率模型本身。

本讲总结：Logistic 回归与最大熵模型

- **Logistic 回归的预测**：线性得分经 sigmoid 转为概率，再按阈值判类；对数几率线性
- **监督目标与参数学习**：平均交叉熵评价真实类概率，可加正则化；优化器利用梯度的局部变化信息迭代寻参
- **特征变换与非线性边界**：特征变换改变原空间的边界形状；sigmoid 负责得分到概率的映射
- **条件最大熵的建模与学习**：保留经验特征约束并最大化条件熵；对数线性模型通过条件似然学习参数
  - 适用条件：标准模型、无正则化、有限可微最优点
  - 在该最优点，梯度为零对应模型与经验特征期望相等
- **两类模型的形式联系**：条件模型、特征与参数化相匹配时，共享对数线性形式；多类对应 softmax，二类化为 sigmoid

求解边界：稳定计算、正则化与停止条件限定求解；模型、数据与目标决定有限解是否存在。