06 Logistic回归与最大熵模型

从线性证据到条件概率

1 Logistic 概率分类模型

线性证据怎样变成概率与类别

概率输出保留线性证据的强弱

- **感知机的输出**
  - 线性得分的符号决定类别
  - 两个正分样本都输出正类
- **仍然缺少的信息**
  - 刚刚越过边界与远离边界的证据强弱不同
  - 离散类别不能表达这种差别
- **Logistic 回归的输出**
  - 它是直接建模 $P(y \mid x)$ 的判别式条件模型
  - 保留线性证据得分
  - 进一步给出条件类别概率

Logistic 回归的推理链

- **输入**
  - 特征向量 x 描述一个待分类样本
- **证据得分**
  - $z = w^T x + b$ 汇总各项特征证据
- **正类概率**
  - $p(x) = P(Y = 1 \mid X = x) = \sigma(z)$
- **分类结果**
  - 将 $p(x)$ 与阈值 τ 比较得到 $\hat{y}$

sigmoid 把得分映射为概率

- **概率映射**
  - $\sigma(z) = \frac{1}{1+\exp(-z)}$
  - 任意实数得分被映射到 $(0, 1)$
- **三个关键位置**
  - $z = 0$ 时，$\sigma(z) = 0.5$
  - $z > 0$ 时，正类概率大于 0.5
  - $z < 0$ 时，正类概率小于 0.5
- **职责边界**
  - sigmoid 单调保留得分顺序
  - 边界形状仍由得分函数决定

一次完整的 Logistic 推理

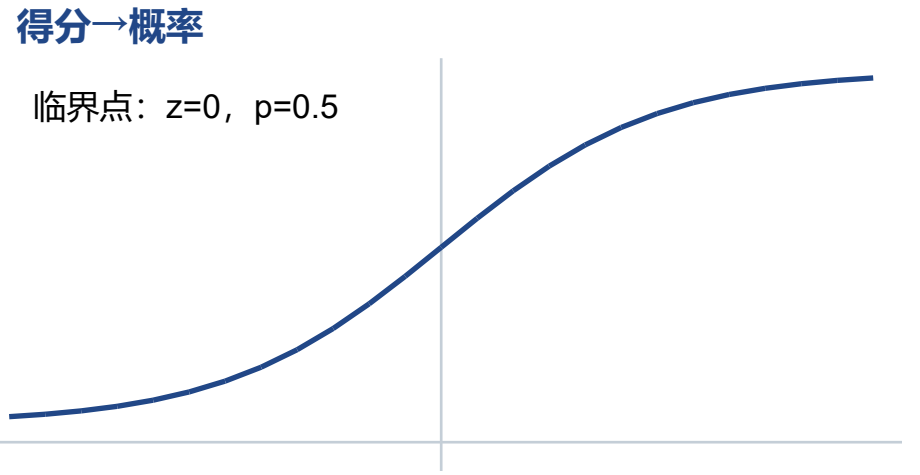
- **输入与参数**
  - $x = (2,1)$，$w = (1,-1)$，$b = 0$
- **计算线性得分**
  - $z = w^T x + b = 1 \times 2 + (-1) \times 1 = 1$
- **计算正类概率**
  - $p(x) = \sigma(1) \approx 0.731$
- **形成类别决策**
  - 取 $\tau = 0.5$，因为 $0.731 > 0.5$，所以 $\hat{y} = 1$

对数几率呈线性关系

- **几率比较两类概率**
  - 正类几率为 $p/(1-p)$
- **对数几率连接线性得分**
  - $\log\frac{p(x)}{1-p(x)} = w^T x + b$
- **模型的线性结构**
  - Logistic 回归对类别概率的对数几率作线性建模
  - 它不是对输入特征取对数，也不要求输入数据服从 Logistic 分布

概率曲线不等于决策边界

- **在得分轴上**
  - sigmoid 将 z 映射为概率
  - $z = 0$ 对应 $p = 0.5$
- **在输入空间中**
  - 当阈值为 0.5，边界满足 $w^T x + b = 0$
  - 二维输入上的基础边界是一条直线
- **两个图形承担不同作用**
  - S 形曲线描述得分到概率的映射
  - 直线描述哪些输入得到相同的临界得分



2 损失与参数学习

连续损失怎样连接真实标签与参数更新

连续损失为离散分类提供学习信号

- **分类结果只记录对错**
  - $p = 0.49$ 与 $p = 0.01$ 在阈值 0.5 下都预测错误
- **概率保留错误程度**
  - 两个预测给真实类别的概率差别很大
- **连续损失形成训练信号**
  - 损失随模型给真实类别的概率连续变化
  - 优化器才能比较参数调整前后的目标
- **职责边界**
  - 连续损失支持学习，但不保证模型一定正确或概率已经校准

熵、条件熵与交叉熵评价不同对象

- **熵评价一个完整分布**
  - $H(P) = -\sum_y P(y)\log P(y)$，表示分布 P 自身有多不确定
- **条件熵评价知道输入后的平均不确定性**
  - $H(Y \mid X) = \sum_x P(x)H(P(\cdot \mid x))$
  - 先计算每个输入下输出分布的熵，再按输入分布取平均
- **交叉熵比较真实分布与模型分布**
  - $H(Q,P) = -\sum_y Q(y)\log P(y)$
  - Q 表示真实标签分布，P 表示模型输出分布
- **监督分类中的接口**
  - 当 Q 是 one-hot 标签时，交叉熵化为 $-\log P(y_{true} \mid x)$
  - 下一页把它写成二分类形式并比较具体数值

二元交叉熵惩罚高置信错误

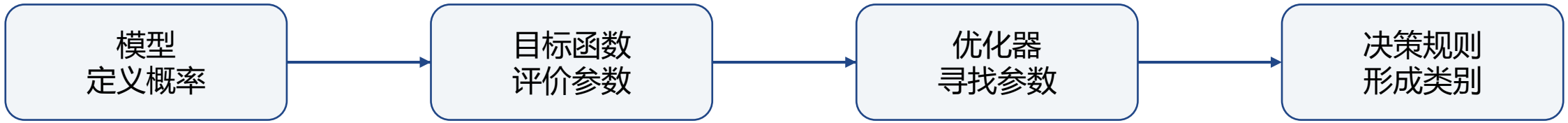
- **单样本二元交叉熵**
  - 根据真实标签，评价模型给该标签分配的概率
  - $\ell = -[y\log p + (1-y)\log(1-p)]$
- **真实标签固定为 $y = 1$**
  - 只有 $-\log p$ 分支保留
- **预测 $p = 0.9$**
  - $\ell = -\log(0.9) \approx 0.105$
  - 给真实类别较高概率，损失较小
- **预测 $p = 0.1$**
  - $\ell = -\log(0.1) \approx 2.303$
  - 给真实类别较低概率，损失较大
- **概率学习的方向**
  - 训练推动模型提高真实标签的条件概率

训练更新参数，推理固定参数

- **训练阶段**
  - 输入带标签样本 (x_i, y_i)
  - 用当前参数计算 p_i
  - 比较 p_i 与真实标签得到损失
  - 优化器更新 w, b 并重复
- **训练输出**
  - 一组固定的模型参数
- **推理阶段**
  - 输入新样本 x
  - 使用训练后的 w, b 计算得分与概率
  - 通过阈值形成类别
- **推理中不发生**
  - 不使用未知的真实标签
  - 不重新更新模型参数

模型、目标与优化器各有职责

- **模型定义概率**
  - $P(Y = 1 \mid x) = \sigma(w^T x + b)$
- **目标函数评价参数**
  - 数据损失衡量模型对训练标签的解释
  - 正则化限制过大的权重，缓解过拟合并改善稳定性
- **优化器寻找参数**
  - 根据目标函数提供的方向信息反复更新参数
- **决策规则使用概率**
  - 阈值把训练后模型的概率输出转成类别



Logistic 回归的数据集目标与梯度

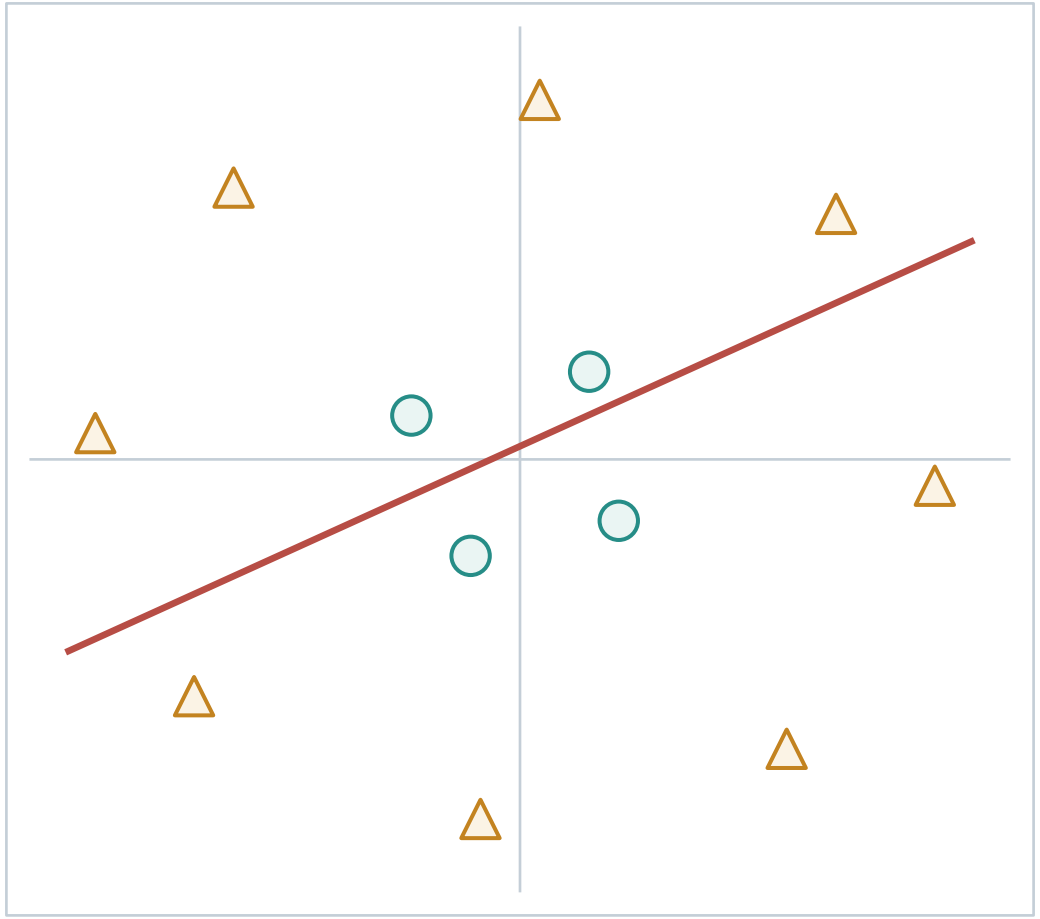
- **数据集目标**
  - $z_i = w^T x_i + b, p_i = \sigma(z_i)$
  - $J(w,b) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)] + \frac{\alpha}{2}\|w\|_2^2$
- **数值稳定实现**
  - $softplus(z) = \log(1+e^z)$
  - $\ell_i = softplus(z_i) - y_i z_i$
  - 该形式与二元交叉熵等价；实际使用框架提供的稳定 logits 损失函数，通常不正则化偏置 b
- **梯度**
  - $\nabla_w J = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)x_i + \alpha w$
  - $\frac{\partial J}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)$
- **残差怎样更新参数**
  - $p_i - y_i$ 表示预测概率与真实标签之间的残差
  - x_i 决定该残差如何作用于各个权重

3 非线性边界与特征表示

模型在特征空间线性，边界可以在原始空间呈非线性

直线无法分开中心类与外围类

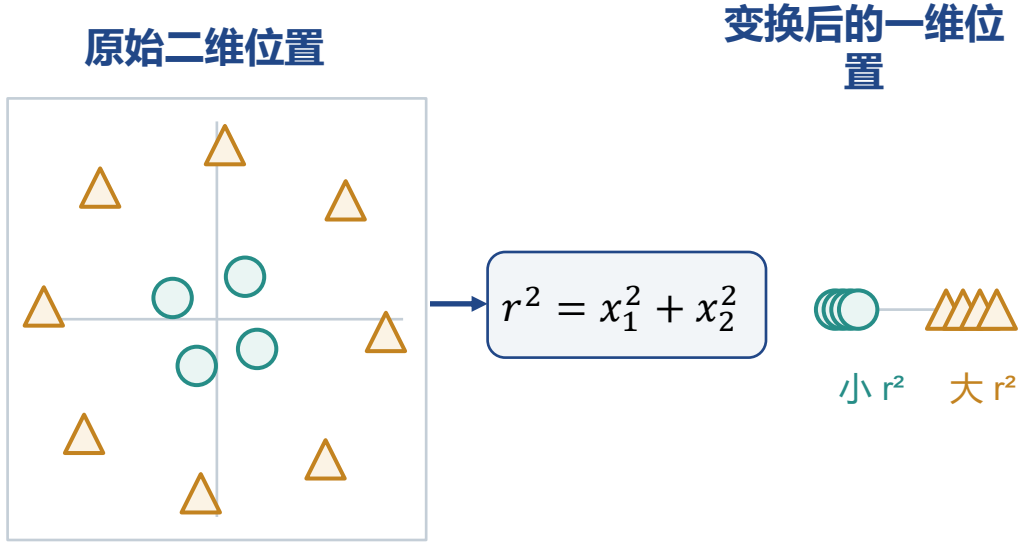
- **分类结构**
  - 一类靠近中心，另一类形成外围环带
- **原始表示**
  - 每个样本只有二维坐标 (x_1, x_2)
- **线性模型的边界**
  - $w_1 x_1 + w_2 x_2 + b = 0$ 只能形成一条直线
- **失败原因**
  - 中心—外围关系需要封闭边界
  - 移动或旋转直线仍不能包围中心类



单一直线总会把某些中心点或外围点分错

半径平方把环形分类化为阈值分类

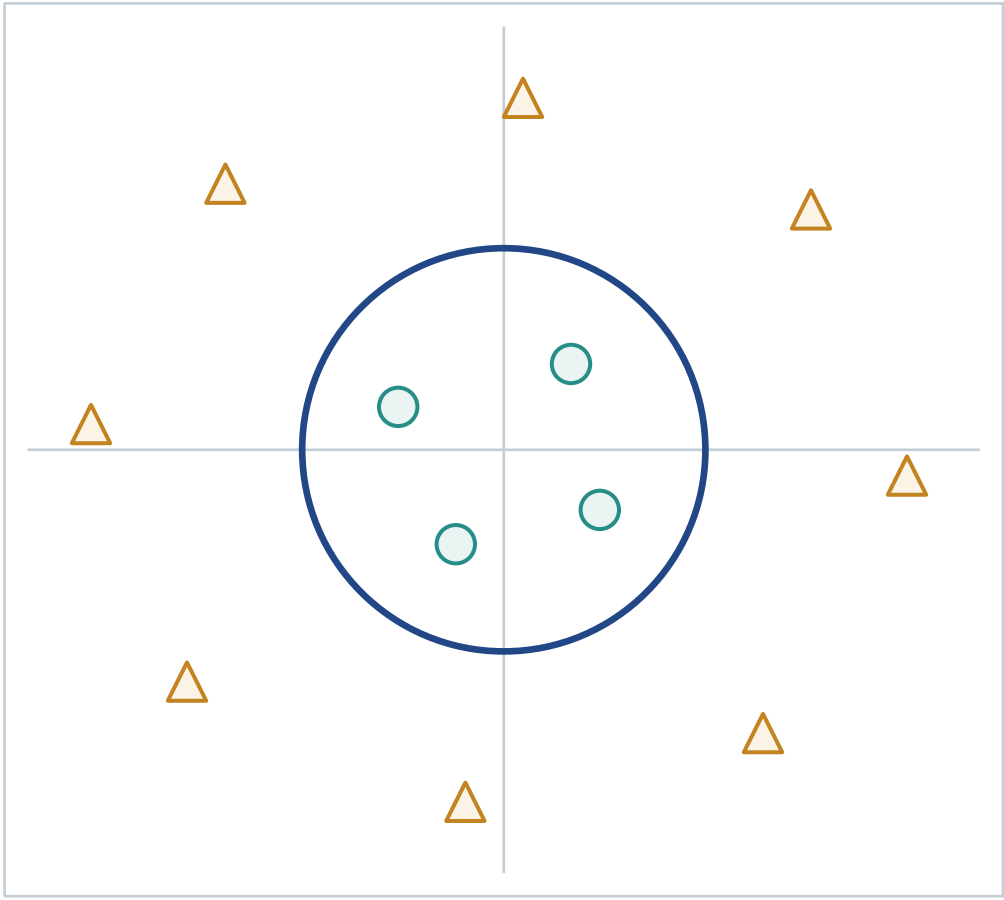
- **构造新特征**
  - $r^2 = x_1^2 + x_2^2$
- **特征的含义**
  - r^2 小表示样本靠近中心
  - r^2 大表示样本远离中心
- **变换后的模型**
  - 使用得分 $z = wr^2 + b$
  - 模型只需在一维 r^2 轴上学习一个阈值



二维的中心—外围关系，在 r² 轴上变成左右可分

线性阈值映回原空间成为圆形边界

- **变换后的特征空间**
  - 边界满足 $wr^2 + b = 0$
  - 关于模型输入 r^2 仍是线性阈值
- **原始二维空间**
  - 代回定义得到 $x_1^2 + x_2^2 = -b/w$
  - 当右侧为正数时，这是圆形边界
- **两部分各自负责**
  - 特征映射决定模型能够表达的边界形状
  - sigmoid 把最终得分映射为类别概率
- **使用边界**
  - 表示扩大能力，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 熵与最大熵模型

在满足已知约束的条件下处理剩余不确定性

条件最大熵的原始优化目标

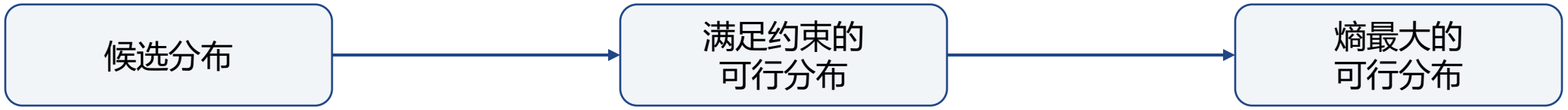
- **候选对象**
  - 对每个输入 x，候选模型给出完整分布 $P(\cdot \mid x)$
- **原始最大熵目标**
  - 在满足概率归一化和经验特征期望约束的条件下，最大化
  - $H_P(Y \mid X) = -\sum_x \tilde{P}(x) \sum_y P(y \mid x) \log P(y \mid x)$
- **与参数估计的关系**
  - 原始问题评价候选分布自身；其对偶问题产生条件对数线性模型
  - 在标准设定下，相应参数也可通过最大条件似然估计
- **解释边界**
  - 熵高低描述分布的不确定性，不直接表示分类是否正确

二分类概率越接近均匀，熵越高

- **确定分布（1，0）**
  - 一个结果的概率为 1，熵最低
- **偏斜分布（0.9，0.1）**
  - 一个结果明显占优，熵较低
- **均匀分布（0.5，0.5）**
  - 两个结果同样可能，熵最高
- **解释边界**
  - 熵比较分布的不确定性，不直接评价分类正确性

最大熵在约束内分配剩余不确定性

- **第一步：保留已知信息**
  - 概率归一化和经验事实限定可行分布集合
- **第二步：处理约束没有决定的部分**
  - 在可行集合中选择熵最大的分布
- **核心含义**
  - 模型满足选定的经验约束
  - 对证据没有区分的结果，不额外制造差别
- **特殊情形**
  - 只有在没有其他区分性约束时，均匀分布才是最大熵结果



约束 $P(A) = 0.5$ 后，最大熵均分剩余概率

- **没有额外信息**
  - 结果集合为 $\{A, B, C\}$
  - 只要求概率非负且总和为 1
  - 最大熵分布为 $(1/3, 1/3, 1/3)$
- **此时的解释**
  - 没有证据区分三个结果
- **加入约束 $P(A) = 0.5$**
  - 剩余概率满足 $P(B) + P(C) = 0.5$
  - 没有证据进一步区分 B 与 C
  - 最大熵分布为 $(0.5, 0.25, 0.25)$
- **此时的解释**
  - 已知约束优先，最大熵只处理剩余自由度

无额外约束

A, B, C 同样可能

0.33	0.33	0.33

(1/3, 1/3, 1/3)

加入约束 P(A)=0.5

A 先固定；B, C 无进一步区分证据

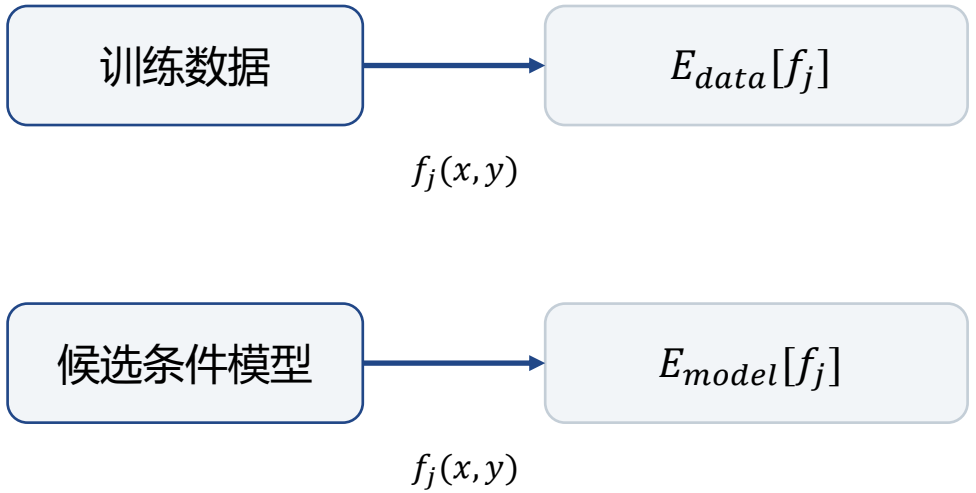
0.5	0.25	0.25

(0.5, 0.25, 0.25)

约束先固定已知部分，最大熵再分配剩余自由度

特征期望把训练事实写成模型约束

- **同一个特征函数**
  - $f_j(x, y)$ 描述输入 x 与候选类别 y 的一种关系
- **匹配总体期望**
  - 数据侧统计经验期望 $E_{data}[f_j]$
  - 模型侧计算模型期望 $E_{model}[f_j]$
  - $E_{model}[f_j] = E_{data}[f_j]$
  - 约束匹配总体期望，不逐个复制样本



$$E_{data}[f_j] = E_{model}[f_j]$$

同一特征的总体期望相等

约束保留被选中的训练事实，不复制单个样本

条件最大熵目标

- **候选对象**
  - 对每个输入 x 选择完整的类别分布 $P(y \mid x)$
- **可行条件**
  - 候选分布必须满足概率归一化和经验特征期望约束
- **在可行集合中选择**
  - $\tilde{P}(x)$ 是训练输入的经验分布，用于加权不同输入
  - $H_P(Y \mid X) = -\sum_x \tilde{P}(x) \sum_y P(y \mid x) \log P(y \mid x)$
  - 在满足上述约束的候选 P 中，最大化 $H_P(Y \mid X)$

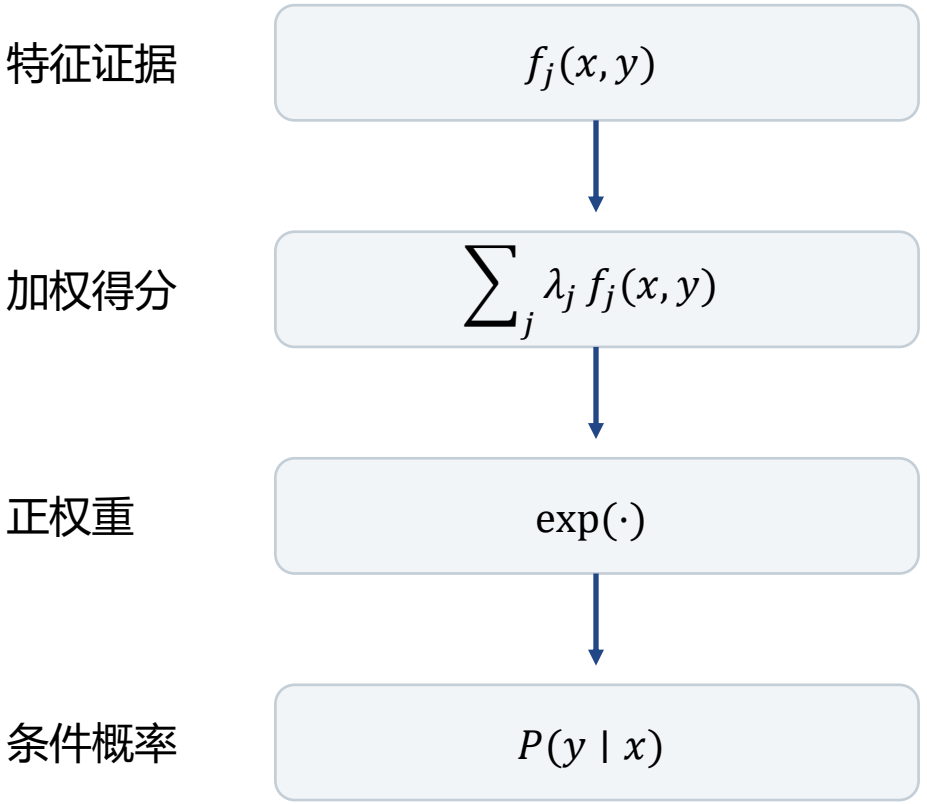
对数线性模型将加权特征归一化为概率

从约束到模型形式

- 受约束熵优化的一阶条件使每个特征约束以一个权重进入线性得分
- 候选类别的概率正比于该得分的指数；本讲采用这一结果，不展开对偶推导

模型形式

- $P_\lambda(y \mid x) = \dfrac{\exp\left(\sum_j \lambda_j f_j(x,y)\right)}{Z_\lambda(x)}$
- $Z_\lambda(x) = \sum_{y'} \exp\left(\sum_j \lambda_j f_j(x,y')\right)$



读式顺序

- 特征提供证据，权重决定方向与强度
- 指数产生正的类别权重，$Z_\lambda(x)$ 使概率和为 1

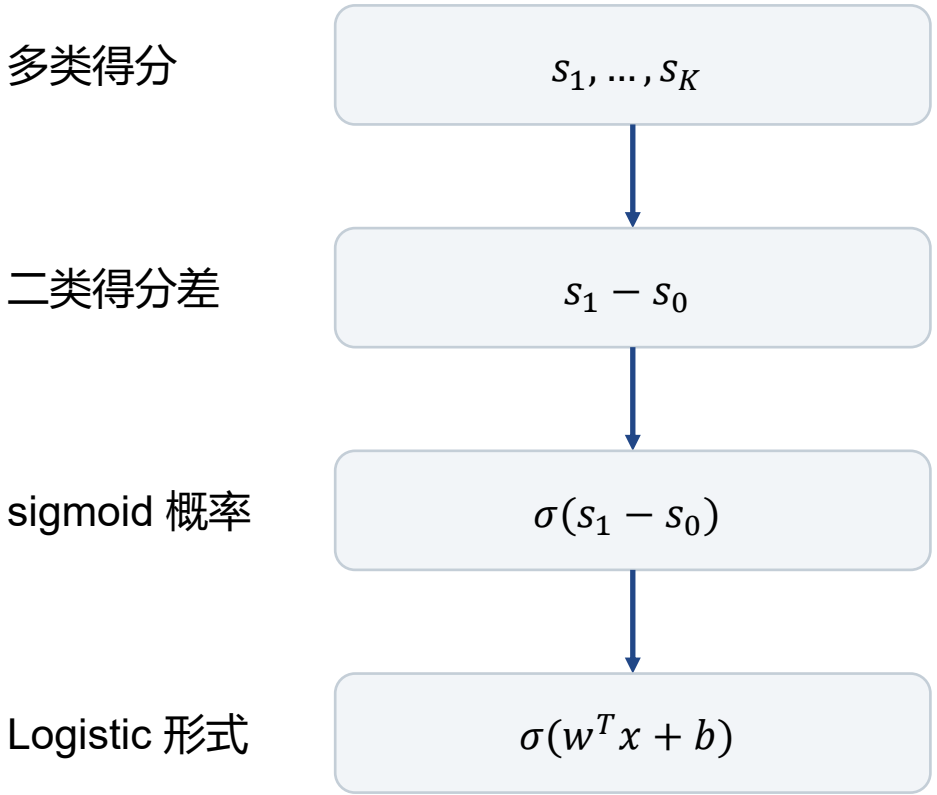
$Z_\lambda(x)$ 对同一输入的全部类别共同归一化

5 模型关系与参数优化

两条建模路径怎样汇合，参数又怎样求得

二类 softmax 化为得分差的 sigmoid

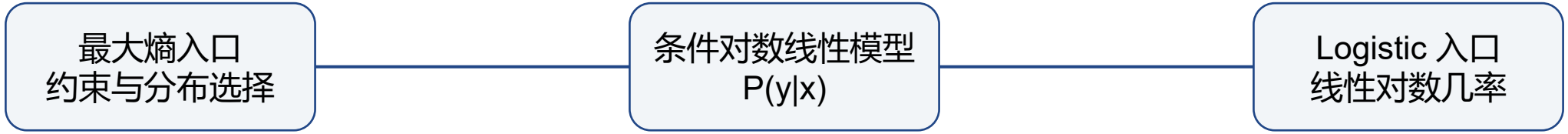
- **多类别 softmax**
  - $P(Y = k \mid x) = \frac{\exp(s_k(x))}{\sum_c \exp(s_c(x))}$
- **二分类化简**
  - $P(Y = 1 \mid x) = \frac{\exp(s_1)}{\exp(s_1)+\exp(s_0)} = \sigma(s_1 - s_0)$
  - 若 $s_1(x) - s_0(x) = w^T x + b$，便得到标准 Logistic 形式



二分类只需要类别得分之差

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



最大熵选择分布，条件似然估计参数

- **条件最大熵**
  - 在受约束的条件分布集合中提问
  - 选择满足经验特征约束且熵最大的分布
- **最大条件似然**
  - 在对数线性参数空间中提问
  - 选择使训练标签条件概率最大的参数
- **成立条件**
  - 在标准条件最大熵设定中，使用相同的特征函数、经验输入分布和无正则化目标时，最大条件熵的对偶与最大条件似然对应同一个对数线性参数估计问题
  - 正则化可解释为软化约束或引入参数先验，此时期望匹配通常不再严格成立

最大熵约束与条件似然梯度

- **条件模型与目标**
  - $P_\lambda(y \mid x) = \frac{\exp(\sum_j \lambda_j f_j(x,y))}{Z_\lambda(x)}$
  - $J(\lambda) = -\frac{1}{N}\sum_{i=1}^{N} \log P_\lambda(y_i \mid x_i)$
- **经验期望**
  - $E_{data}[f_j] = \frac{1}{N}\sum_{i=1}^{N} f_j(x_i, y_i)$
- **模型期望**
  - $E_{model}[f_j] = \frac{1}{N}\sum_{i=1}^{N} \sum_y P_\lambda(y \mid x_i) f_j(x_i, y)$
- **似然梯度**
  - $\frac{\partial J}{\partial \lambda_j} = E_{model}[f_j] - E_{data}[f_j]$
- **无正则化最优点**
  - $\frac{\partial J}{\partial \lambda_j} = 0 \Longrightarrow E_{model}[f_j] = E_{data}[f_j]$
- **加入正则化后**
  - 最优点在期望匹配和参数规模之间折中，不再要求严格相等

优化器、停止条件与解的边界

- **方法使用的信息**
  - 梯度下降使用一阶梯度；牛顿法使用梯度和 Hessian；拟牛顿法根据迭代信息近似曲率
- **凸性**
  - 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
- **停止条件**
  - 可检查目标函数下降量、梯度范数、参数更新量或最大迭代次数
- **可分数据边界**
  - 数据完全线性可分且没有正则化时，参数可能不断增大，有限极大似然解可能不存在
- **职责不变**
  - 更换优化器改变寻参路径和计算代价，不改变概率模型的定义

概率分类框架：表示、模型、目标、优化与决策

- **表示决定可表达的边界**
  - 原始或变换后的特征决定线性得分能够形成哪些边界
- **模型把证据映射为条件概率**
  - sigmoid 或 softmax 将类别得分归一化为 $P(y \mid x)$
- **参数学习形成闭环**
  - 交叉熵评价训练标签的条件概率，梯度由预测与标签的残差决定，优化器据此更新参数
- **最大熵与条件似然相连接**
  - 最大熵约束保留经验事实；标准无正则化设定下，条件似然梯度为零对应经验期望与模型期望相等
- **实现形成可执行边界**
  - 使用 logits 稳定计算损失，通过正则化控制参数规模，并根据目标变化、梯度或最大迭代次数停止；阈值或最大概率形成决策