

04 贝叶斯分类器

从概率更新到分类决策

1 概率与贝叶斯更新

概率怎样描述规律、表达不确定性，并随证据更新

■ 重复现象的长期规律

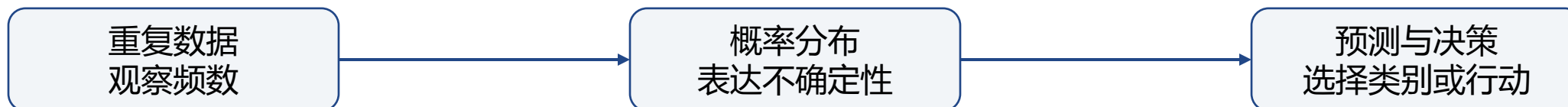
- 同类现象反复出现时，各种结果会呈现相对稳定的比例
- 训练频数为概率估计提供数据依据

■ 当前信息下的确信程度

- 概率分布表达对未知类别或参数的不确定性
- 新证据到来后，概率分布可以更新

■ 预测与决策的量化输入

- 模型比较各类别的概率并形成预测
- 决策规则结合概率与错误代价选择行动



■ 垃圾邮件中出现优惠

- 已知类别是垃圾邮件，求出现“优惠”的概率
- 记 $Y = G$ 表示垃圾邮件， $O = 1$ 表示出现“优惠”
- $P(O = 1 | Y = G)$

■ 出现优惠后是垃圾邮件

- 已知邮件出现“优惠”，求类别是垃圾邮件的概率
- $P(Y = G | O = 1)$

■ 条件决定问题方向

- 两个概率通常不同，需要用贝叶斯公式建立联系

贝叶斯公式中的四个概率对象

■ 四个概率对象

- 类别先验 $P(y)$ 与类条件概率 $P(x | y)$ 描述观察前的基线及类别对观测的解释
- 证据 $P(x)$ 用于归一化；类别后验 $P(y | x)$ 是更新结果

■ 同一联合事件的两种分解

- $P(x, y) = P(x | y)P(y) = P(y | x)P(x)$

■ 由分解得到更新公式

- $P(y | x) = \frac{P(x|y)P(y)}{P(x)}$
- 分子是类别 y 对当前观测 x 的概率贡献

■ 垃圾邮件

- 共有 4 封
- 4 封出现 “优惠”

■ 正常邮件

- 共有 6 封
- 1 封出现 “优惠”

■ 待求后验

- 已知一封邮件出现 “优惠” , 求它是垃圾邮件的概率
- $P(Y = G \mid O = 1)$

- 垃圾邮件对证据的贡献

- $P(O = 1 | G)P(G) = 1 \times 0.4 = 0.4$

- 正常邮件对证据的贡献

- $P(O = 1 | N)P(N) = \frac{1}{6} \times 0.6 = 0.1$

- 证据概率

- $P(O = 1) = 0.4 + 0.1 = 0.5$

- 更新后的类别概率

- $P(G | O = 1) = \frac{0.4}{0.5} = 0.8$

■ 观察证据以前

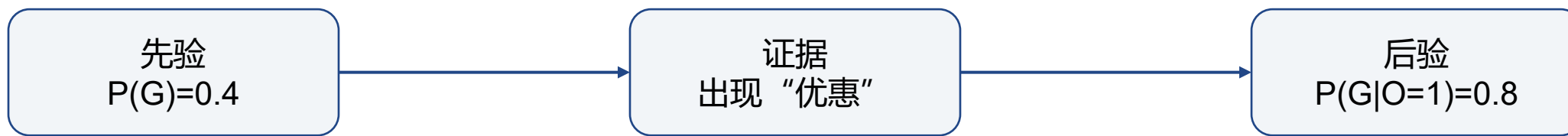
- 垃圾邮件的类别先验为 0.4

■ 当前证据的解释能力

- 垃圾邮件一定出现“优惠”，正常邮件也有 1/6 的概率出现“优惠”

■ 观察证据以后

- 垃圾邮件后验提高到 0.8
- 证据增强了垃圾邮件判断，但仍保留 0.2 的正常邮件可能性



2 朴素贝叶斯分类

从训练数据估计概率，再组合多个特征完成预测

■ 训练阶段

- 从带标签数据估计类别先验 $P(y)$
- 估计各类别下的特征分布 $P(x | y)$

■ 推理阶段

- 输入新样本的观测特征 x
- 用贝叶斯公式计算类别后验 $P(y | x)$

■ 与判别路线的区别

- 判别方法直接学习 $P(y | x)$ 或分类边界
- 本讲只用该区别定位学习对象

■ 直接估计联合概率

- 需要统计每一种特征组合在各类别中的出现频率
- 特征增加后，有限数据很难覆盖全部组合

■ 给定类别后的条件独立

- 假定给定类别 y 后，各特征条件独立
- $P(x | y) = \prod_j P(x_j | y)$

■ 收益与边界

- 模型只需估计多个单特征条件概率
- 真实特征之间可能仍存在条件依赖，朴素独立假设无法表示这些关系

■ 垃圾邮件：4 封

- 优惠 = 1: 4 封
- 会议 = 1: 1 封
- 附件 = 1: 0 封

■ 正常邮件：6 封

- 优惠 = 1: 1 封
- 会议 = 1: 4 封
- 附件 = 1: 3 封

■ 类别先验

- $P(G) = 4/10$; $P(N) = 6/10$

■ 查询取值的类条件概率

- 优惠: $P(O = 1 | G) = 4/4$; $P(O = 1 | N) = 1/6$
- 会议: $P(M = 0 | G) = 3/4$; $P(M = 0 | N) = 2/6$
- 附件: $P(A = 1 | G) = 0/4$; $P(A = 1 | N) = 3/6$

■ 查询邮件

- O 、 M 、 A 分别表示优惠、会议、附件特征
- $x = (O = 1, M = 0, A = 1)$

■ 垃圾邮件得分

- 记垃圾邮件类别为 G , $s_G(x) = \frac{4}{10} \times \frac{4}{4} \times \frac{3}{4} \times \frac{0}{4}$

■ 正常邮件得分

- 记正常邮件类别为 N , $s_N(x) = \frac{6}{10} \times \frac{1}{6} \times \frac{2}{6} \times \frac{3}{6}$

■ 得分的含义

- 一般形式为 $s_y(x) = P(y) \prod_j P(x_j | y)$
- $s_y(x)$ 与对应类别后验成正比, 比较类别时取 $\hat{y} = \operatorname{argmax}_y s_y(x)$
- 当前得分尚未归一化, 不能直接作为后验概率报告

■ MLE 计算结果

- $s_G(x) = 0$
- $s_N(x) = \frac{1}{60}$

■ 零因子的来源

- 4 封垃圾邮件中没有观察到带附件的邮件
- 频数 MLE 给出 $P(A = 1 | G) = 0/4 = 0$

■ 有限数据带来的问题

- 一个零因子使该类别的整个乘积为零
- “训练集中没有观察到” 不能证明 “真实世界中不可能发生”

■ 频数 MLE

- $\hat{P}(X_j = a | y) = \frac{N_{j,a,y}}{N_y}$
- 零计数直接产生零概率

■ 二值特征的加一平滑

- $\tilde{P}(X_j = a | y) = \frac{N_{j,a,y} + 1}{N_y + 2}$
- 每个取值加 1; 二值特征的分母加 2

■ 只改变估计规则

- 训练数据、查询邮件和条件独立假设保持不变
- 类条件概率的估计规则及其数值发生改变

■ 两个未归一化得分

- $s_G(x) = \frac{4}{10} \times \frac{5}{6} \times \frac{4}{6} \times \frac{1}{6} = \frac{1}{27}$
- $s_N(x) = \frac{6}{10} \times \frac{2}{8} \times \frac{3}{8} \times \frac{4}{8} = \frac{9}{320}$

■ 归一化后验

- 两类得分的和是共同归一化分母
- $P(G | x) = \frac{s_G(x)}{s_G(x) + s_N(x)} = \frac{1/27}{1/27 + 9/320} \approx 0.568$
- $P(N | x) = 1 - P(G | x) \approx 0.432$

■ 预测与边界

- 本例在最大后验规则下预测为垃圾邮件
- 该结果只说明本例中平滑改变了预测，不保证平滑总会改善分类

■ 训练

- 输入带标签数据和估计规则
- 输出概率模型：类别先验与类条件概率

■ 推理

- 输入当前观测和固定模型
- 输出各类别得分，必要时归一化为类别后验

■ 决策

- 输入类别后验和错误代价
- 输出一个预测类别或行动

3 参数学习与决策

区分类别概率更新、参数学习结果与最终行动

类别先验与参数先验作用于不同对象

■ 未知对象不同

- 类别推断：当前样本的类别 Y ；参数学习：控制数据分布的参数 θ

■ 先验作用于各自的未知对象

- 类别先验 $P(y)$ ；参数先验 $p(\theta)$

■ 数据项与后验也分别对应

- 类条件概率 $P(x | y)$ 对应类别后验 $P(y | x)$
- 参数似然 $p(D | \theta)$ 对应参数后验 $p(\theta | D)$

■ 共同名称，不同对象

- 每次使用“先验”“似然”或“后验”都需要说明作用对象

- **MLE 保留似然最大的参数点**

- 使用数据 D 得到 $\hat{\theta}_{MLE}$; 预测时代入该点

- **MAP 保留参数后验最大的点**

- 使用数据 D 与参数先验 $p(\theta)$ 得到 $\hat{\theta}_{MAP}$; 预测时代入该点

- **完整贝叶斯学习保留参数后验分布**

- 使用数据 D 与参数先验 $p(\theta)$ 得到 $p(\theta | D)$; 预测时按参数后验综合不同参数下的结果

■ 参数与数据

- θ 表示一次试验成功的概率
- 独立观察 4 次, 其中 3 次成功、1 次失败
- 参数似然为 $p(D | \theta) \propto \theta^3(1 - \theta)$

■ 参数先验

- 设 $\theta \sim \text{Beta}(2, 2)$
 - 该先验只为形成便于手算的课程例子
- 本例比较 MLE、MAP 与完整贝叶斯学习怎样处理同一组数据。

■ MLE 输出一个点

- $\hat{\theta}_{MLE} = 3/4$
- 只使用观察数据，选择似然最大的参数

■ MAP 输出一个点

- 参数后验为 $\theta \mid D \sim \text{Beta}(5,3)$
- 先验中的成功、失败参数分别加上观测计数，得到 $(2 + 3, 2 + 1)$
- $\hat{\theta}_{MAP} = (5 - 1)/(5 + 3 - 2) = 4/6$

■ 完整贝叶斯保留分布

- 学习结果是整个参数后验 $\text{Beta}(5,3)$
- 下一次成功的后验预测概率为后验均值 $5/8$

■ 概率推断

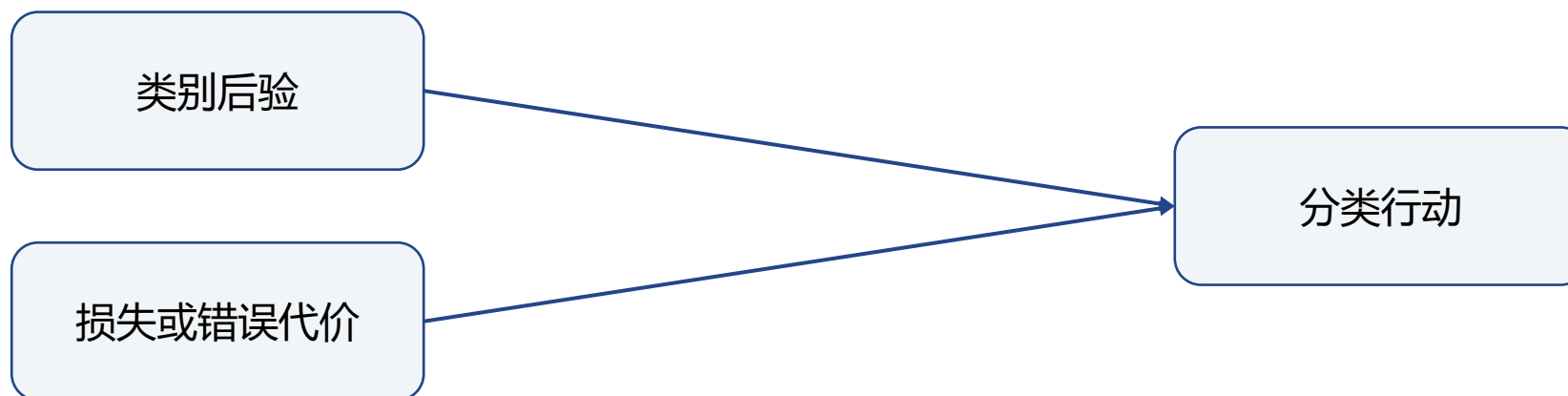
- 模型根据当前观测得到类别后验 $P(y | x)$, 表示各类别在当前信息下的概率

■ 分类决策

- 决策规则结合类别后验与错误代价选择行动; 在各类误判代价相同的 0-1 损失下, 选择后验最大的类别

■ 成立边界

- 错误代价不同时, 同一个后验分布可能产生不同决策



■ 数据与估计规则形成概率模型

- 有限频数提供估计依据；MLE、MAP 与完整贝叶斯学习产生不同的参数表示

■ 模型假设组织概率关系

- 条件独立降低联合估计负担，平滑限制零计数的影响，但二者都不保证模型正确

■ 贝叶斯更新形成后验

- 类别后验由类别先验与当前观测在各类别下的解释能力共同形成

■ 损失规则形成决策

- 后验结合损失产生行动；错误代价改变时，最优分类也可能改变