# 04 贝叶斯分类器

从概率更新到分类决策

# 学习目标

- 解释概率表示与证据更新
  - 说明训练频数怎样用于概率估计，以及概率怎样表达当前信息下的不确定性并支持预测和决策
  - 解释类别先验、类条件概率、证据与后验的关系，完成一次单一证据更新
- 解释并运用朴素贝叶斯分类模型
  - 解释生成式分类的学习对象与条件独立假设，完成离散朴素贝叶斯训练、推理与归一化
  - 说明零概率的来源、加一平滑的作用，以及有限数据与模型假设的局限
- 区分类别推断、参数学习与分类决策
  - 区分类别先验与参数先验，比较 MLE、MAP 和完整贝叶斯学习的结果与预测方式
  - 说明最大后验分类与 0–1 损失的联系，以及错误代价改变时的决策边界

# 1 概率与贝叶斯更新

区分两个条件方向，用各类别对同一证据的贡献建立贝叶斯更新关系，并完成一次邮件概率计算

# 类条件概率与类别后验的区别

判断含“优惠”的邮件属于垃圾类的概率。课程构造的 10 封邮件中，4 封为垃圾类 $G$，6 封为正常类 $N$；两类中含“优惠”的邮件分别为 4 封和 1 封。

记 $O = 1$ 表示含“优惠”，条件概率竖线右侧是已知条件，左侧是待判断事件。

| 条件方向 | 统计范围 | 所回答的问题 |
| --- | --- | --- |
| $P(O = 1 \mid G)$ | 4 封垃圾邮件 | 垃圾邮件中多大比例含优惠? |
| $P(G \mid O = 1)$ | 含优惠的邮件 | 看到优惠后，多大可能是垃圾邮件? |

类内比例 $4/4 = 1$ 不能直接作为观察后的类别概率，因为正常邮件也会出现相同特征。

**类别判断须结合类别基线与当前特征在各类别下的出现概率。**

概率可描述重复现象的规律，也可量化当前信息下的不确定性，为类别比较与决策提供依据。
训练频数用于概率估计；有限样本比例不等于总体真实概率。

# 先验、类条件概率、证据与后验

$Y$ 是类别变量，$X$ 是观测特征变量，$y, x$ 是具体取值；$P(y \mid x)$ 简写 $P(Y = y \mid X = x)$。单一证据例子取 $X = O, x = 1$。

| 对象 | 含义 | 对类别判断的作用 |
| --- | --- | --- |
| **类别先验** $P(y)$ | 尚未观察当前样本特征时的类别概率 | 提供类别基线 |
| **类条件概率** $P(x \mid y)$ | 给定类别后出现当前观测的概率 | 表示该类别对证据的支持 |
| **证据** $P(x)$ | 不限定类别时出现当前观测的总概率 | 汇总所有类别贡献并归一化 |
| **类别后验** $P(y \mid x)$ | 观察当前特征后的类别概率 | 分类所需的概率判断 |

联合概率 $P(x, y)$ 描述特征取值为 $x$ 且类别为 $y$ 的同一事件，可连接两个条件方向。

# 贝叶斯公式：由联合概率到类别后验

- **同一联合事件有两种分解**
  - $P(x, y) = P(x \mid y)P(y) = P(y \mid x)P(x)$.
  - $P(x \mid y)P(y)$ 是类别 $y$ 对当前观测的概率贡献
- 当 $P(x) > 0$ 时，贝叶斯关系给出类别后验
  - $P(y \mid x) = \frac{P(x|y)P(y)}{P(x)}$.
- 互斥且完备的类别贡献共同构成证据概率
  - $P(x) = \sum_c P(x \mid c)P(c)$.
  - 分母汇总各类贡献，使类别后验之和为 1

类别后验同时取决于先验与类条件概率，不能仅凭类条件概率选择类别。

# 单一证据下的邮件概率更新

复用同一组邮件计数，只观察 $O = 1$。频数模型给出

$$P(G) = 4/10 = 0.4, P(N) = 6/10 = 0.6, P(O = 1 \mid G) = 4/4 = 1, P(O = 1 \mid N) = 1/6.$$

| 类别 | 对优惠证据的贡献 |
| --- | --- |
| 垃圾 $G$ | $1 \times 0.4 = 0.4$ |
| 正常 $N$ | $(1/6) \times 0.6 = 0.1$ |

$$P(O = 1) = 0.4 + 0.1 = 0.5, P(G \mid O = 1) = \frac{0.4}{0.5} = 0.8.$$

观察“优惠”后，垃圾类后验高于先验，正常类后验仍为 0.2。两类贡献共同决定后验，因此类条件概率为 1 时，类别后验仍可能小于 1。

# 2 朴素贝叶斯分类

用条件独立假设建立朴素贝叶斯模型，从邮件计数估计概率，再完成分类与平滑重算；区分概率推断与损失条件下的类别选择

# 生成式分类的训练与推理

前面只使用“优惠”这一证据。要同时利用“会议”和“附件”等特征，需要建立可从数据学习的多特征概率模型。

记 $X = (O, M, A)$ 为三个二值特征，依次表示优惠、会议、附件；1 表示有，0 表示无。$x$ 是当前邮件的具体特征取值向量。

先学习类别先验与类条件分布，再根据观测推断类别，这一路线称为**生成式分类**。概率模型描述各类别下的数据分布。

- **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 $P(x \mid y)$
- **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 $P(y \mid x)$

判别方法则直接学习 $P(y \mid x)$ 或分类边界。

# 条件独立假设与类条件概率分解

三个二值特征有 $2\times2\times2=8$ 种组合，一般 $d$ 个二值特征有 $2^d$ 种组合。直接估计 $P(x \mid y)$ 需要在每个类别内统计组合，有限数据可能没有覆盖大量组合。

**用结构假设减少待估计的概率，让有限样本集中支持单特征估计。**代价是放弃对类内特征依赖的表示。

- 给定同一类别 $Y=y$ 后，假定特征之间相互独立
  - 在该假设下，$P(x \mid y)=\prod_{j=1}^{d} P(x_j \mid y)$
  - $x_j$ 是第 $j$ 个特征的观测取值
  - 本例的特征顺序为优惠、会议、附件

采用这一条件独立结构的生成式分类模型称为**朴素贝叶斯分类器。**

# 条件独立模型的参数负担与边界

自由参数是可以独立指定的概率数值。对同一类别、同样的 $d$ 个二值特征：

| 类条件模型 | 估计对象 | 每类自由参数数目 |
| --- | --- | --- |
| 联合分布 | 所有特征组合的概率 | $2^d - 1$：$2^d$ 个概率之和为 1 |
| 条件独立模型 | 各单特征取值的概率 | $d$：每个特征只估计取值为 1 的概率，取值为 0 的概率由补概率得到 |

**参数负担由指数增长降为线性增长；模型同时舍弃类内特征依赖。**有限样本下更易估计，来自这一建模取舍，独立性仍须由任务与数据检验。

# 统计训练邮件的类别与特征

在给定类别后的特征条件独立假设下，从邮件计数估计类别先验与各特征的类条件概率，再计算待分类邮件的类别得分。

| 类别 | 样本数 | 优惠=1 | 会议=1 | 附件=1 |
|---|---|---|---|---|
| 垃圾 $G$ | 4 | 4 | 1 | 0 |
| 正常 $N$ | 6 | 1 | 4 | 3 |

- 本例沿用课程构造的 10 封邮件，垃圾类记为 $G$，正常类记为 $N$；表中统计各类样本数及类内单特征计数
- 二值特征 $O, M, A$ 分别表示优惠、会议、附件；1 表示有，0 表示无
- 取值为 0 的计数由类内总数减去取值为 1 的计数得到，如垃圾邮件中会议=0 的计数为 $4 - 1 = 3$

# 类别先验与类条件概率的频数估计

本例采用**极大似然估计**（MLE）：选择使已观察训练数据在模型下出现的可能性最大的参数。对当前离散模型，估计结果为相应的样本频率；$\hat{P}$ 表示估计得到的概率。

- **类别先验：在全部训练邮件中统计**
  - $\hat{P}(y) = N_y / N$
  - $N$ 是训练样本总数，$N_y$ 是类别 $y$ 的样本数；用该类样本数除以总样本数，要求 $N > 0$
- **类条件概率：在指定类别内统计**
  - $\hat{P}(X_j = a \mid y) = N_{j,a,y} / N_y$
  - $N_{j,a,y}$ 是类别 $y$ 中特征 $j$ 取值为 $a$ 的样本数；分母只计该类样本，要求 $N_y > 0$

# 按待分类邮件的特征读取模型概率

待分类邮件含“优惠”、不含“会议”、带有附件，记为 $x=(O=1,M=0,A=1)$。按当前邮件的特征取值，从已训练模型中读取以下概率。

类别先验为 $\hat{P}(G)=4/10$、$\hat{P}(N)=6/10$。

| 当前邮件的特征取值 | 垃圾类 $G$ 的类条件概率 | 正常类 $N$ 的类条件概率 |
| --- | --- | --- |
| 优惠 $O=1$ | $\hat{P}(O=1 \mid G)=4/4$ | $\hat{P}(O=1 \mid N)=1/6$ |
| 会议 $M=0$ | $\hat{P}(M=0 \mid G)=3/4$ | $\hat{P}(M=0 \mid N)=2/6$ |
| 附件 $A=1$ | $\hat{P}(A=1 \mid G)=0/4$ | $\hat{P}(A=1 \mid N)=3/6$ |

# 朴素贝叶斯的类别得分与比较

对同一封邮件 $x = (O = 1, M = 0, A = 1)$，按给定类别后的条件独立假设组合上一页的模型概率

- **得分规则：类别先验 × 各特征的类条件概率**
  - $s_y(x) = \hat{P}(y) \prod_j \hat{P}(x_j \mid y)$
- **代入当前邮件所需的概率**
  - 垃圾类：$s_G(x) = \frac{4}{10} \times \frac{4}{4} \times \frac{3}{4} \times \frac{0}{4}$
  - 正常类：$s_N(x) = \frac{6}{10} \times \frac{1}{6} \times \frac{2}{6} \times \frac{3}{6}$
  - 两条计算链的因子顺序均为类别先验、优惠、会议、附件
- **按 0–1 损失规则选择类别**
  - 0–1 损失中，预测正确为 0，预测错误为 1
  - 各类别的误判代价相同
  - 同一邮件的各类别后验具有共同分母，得分排序就是后验排序；取 $\hat{y} = \mathrm{argmax}_y s_y(x)$，即选择得分最大的类别
  - 得分尚未除以共同分母，不能直接作为归一化后的类别后验概率报告

# 一个零计数使整类得分归零

- **得分比较与当前预测**
  - $s_G(x) = 0$，$s_N(x) = \frac{1}{60}$
  - 归一化后，当前模型给出 $P(G \mid x) = 0$、$P(N \mid x) = 1$；这些概率来自频数模型，不能保证真实类别判断正确
  - 在最大后验类别规则下，正常类得分较大，当前模型预测为正常
- **垃圾类得分归零的原因**
  - 4 封垃圾邮件中没有观察到带附件的邮件
  - 频数 MLE 给出 $\hat{P}(A = 1 \mid G) = 0/4 = 0$；该零因子使垃圾类的整个得分乘积为零
- **有限样本下的估计边界**
  - “训练集中没有观察到”不能证明“真实世界中不可能发生”
  - 处理零计数对概率估计的影响后，再比较同一封邮件的类别得分

# 二值特征的加一平滑

训练邮件中没有垃圾类带附件的样本，频数估计因而给出零概率，使垃圾类得分归零。未观察到不代表不可能发生；加一平滑调整估计规则，为未出现的特征取值保留非零概率。

| 类条件概率估计 | 规则 | 零计数的处理 |
| --- | --- | --- |
| 频数 MLE | $\hat{P}(X_j = a \mid y) = N_{j,a,y}/N_y$ | 零计数直接产生零概率 |
| 二值特征加一平滑 | $\tilde{P}(X_j = a \mid y) = (N_{j,a,y} + 1)/(N_y + 2)$ | 每个取值的计数加 1；二值特征的分母加 2 |

- **本例只改变类条件概率的估计规则及其数值**
  - 训练数据、查询邮件和条件独立假设保持不变
  - 类别先验仍为 $4/10$ 与 $6/10$，不作平滑

# 平滑后的邮件分类计算

沿用同一封待分类邮件 $x = (O = 1, M = 0, A = 1)$，使用加一平滑后的类条件概率重新计算。

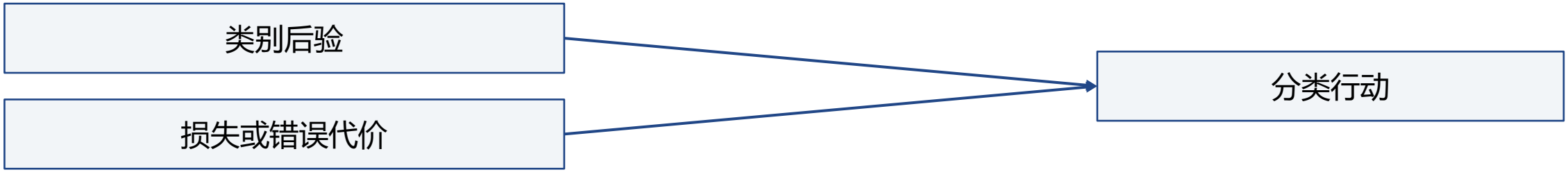
- **两个未归一化得分**
  - $s_G(x) = \frac{4}{10} \times \frac{5}{6} \times \frac{4}{6} \times \frac{1}{6} = \frac{1}{27}$
  - $s_N(x) = \frac{6}{10} \times \frac{2}{8} \times \frac{3}{8} \times \frac{4}{8} = \frac{9}{320}$
- **归一化后验**
  - 两类得分的和是共同归一化分母
  - $P(G \mid x) = \frac{s_G(x)}{s_G(x)+s_N(x)} = \frac{1/27}{1/27+9/320} \approx 0.568$
  - $P(N \mid x) = 1 - P(G \mid x) \approx 0.432$
- **预测与边界**
  - 在本讲 0–1 损失下采用最大后验类别规则，预测由平滑前的正常邮件变为垃圾邮件
  - 本例中平滑改变了预测，但不保证平滑总会改善分类

# 0–1 损失下，最大后验分类使错误概率最小

前面的得分比较已执行最大后验选类。其最优性取决于错误代价：

- **0–1 损失：正确为 0，错误为 1，各类误判代价相同**
  - 给定观测 $x$，选择类别 $a$ 的条件错误概率为 $1 - P(Y = a \mid x)$
  - 最小化错误概率 ⇔ 最大化所选类别的后验概率

**错误代价改变时，同一后验分布可能对应不同的最优行动。**



# 训练、推理与决策：输入、输出及评价

| 阶段 | 输入 | 输出 |
| --- | --- | --- |
| 训练 | 带标签数据和估计规则 | 概率模型：类别先验与类条件概率 |
| 推理 | 当前观测和固定模型 | 各类别得分，必要时归一化为类别后验 |
| 决策 | 类别后验和错误代价 | 一个预测类别或行动 |

- **分类正确性**：预测类别是否与真实标签一致；本例无真实标签，不能据预测改变判断准确率提高
- **概率校准**：预测概率是否与相应样本中该类的实际比例相符；后验归一化不保证校准

# 3 参数学习与总结

区分类别推断与参数学习，用同一伯努利数据比较 MLE、MAP 和完整贝叶斯的学习结果与预测方式，最后总结本讲

# 类别推断与参数学习的对象区别

前面用频数 MLE训练概率模型，再固定模型推断邮件类别。现在考察的未知对象是参数：学习可返回一个参数点，也可保留参数后验，预测时综合这种不确定性。

参数是控制模型分布的量。例如 $\theta = P(A = 1 \mid G)$ 是垃圾类带附件的概率；有限计数估计这个量，参数后验描述它仍有多少不确定性。

| 未知对象 | 推断问题 | 所用信息 | 结果 |
| --- | --- | --- | --- |
| 类别 $Y$ | 当前样本属于哪一类? | 给定模型和观测 $x$ | 类别后验 $P(y \mid x)$ |
| 参数 $\theta$ | 估计参数值及其不确定性 | 训练数据 $D$ | 参数点，或参数后验 $p(\theta \mid D)$ |

- 类别先验 $P(y)$ 描述类别基线；参数先验 $p(\theta)$ 描述纳入训练数据前对参数的不确定性

朴素贝叶斯规定条件独立的模型结构，参数可用频数 MLE 估计；使用贝叶斯公式分类，并不要求保留整个参数后验。

# 参数似然、先验与后验更新

- 参数似然 $p(D \mid \theta)$ 是关于参数的函数
  - 固定已观察训练数据 $D$，比较不同参数 $\theta$ 对数据的解释
  - 类别推断中的 $P(x \mid y)$ 则比较各类别对当前观测的支持
- 参数先验与似然形成参数后验
  - $p(\theta \mid D) \propto p(D \mid \theta)p(\theta)$
  - $\propto$ 表示相差一个与 $\theta$ 无关的比例常数
  - 连续参数用密度表示不确定性：区间内曲线下的面积是概率，密度高度本身不是单点概率

# 伯努利数据与 Beta 参数后验

另取独立的伯努利示例，与邮件计数分开。每次只有成功与失败，成功概率为 $\theta \in [0,1]$；给定参数后4次试验独立同分布，数据 $D$ 有3次成功、1次失败：

$$p(D \mid \theta) \propto \theta^3(1-\theta).$$

Beta 分布在 $[0,1]$ 上描述概率参数。对 $a,b>0$，其密度正比于 $\theta^{a-1}(1-\theta)^{b-1}$。本例给定 $\mathrm{Beta}(2,2)$ 先验，密度正比于 $\theta(1-\theta)$：

$$p(\theta \mid D) \propto \theta^3(1-\theta)\theta(1-\theta) = \theta^4(1-\theta)^2.$$

因此，参数后验为 $\mathrm{Beta}(5,3)$。它描述未知参数的不确定性，与当前样本的类别概率或单个参数点不同。

# MLE、MAP 与完整贝叶斯：学习结果与预测

在同一数据和上述示例先验下：

| 方法与依据 | 学习结果 | 下次成功概率 |
| --- | --- | --- |
| 极大似然估计（MLE）：数据似然，不用参数先验 | 参数点 $\hat{\theta}_{\mathrm{MLE}} = 3/4$ | 代入该点：$3/4$ |
| **最大后验估计**（MAP）：数据与参数先验，取后验众数 | 参数点 $\hat{\theta}_{\mathrm{MAP}} = 4/6 = 2/3$ | 代入该点：$2/3$ |
| 完整贝叶斯：数据与参数先验 | 整个 $\mathrm{Beta}(5,3)$ 参数后验 | 按后验平均：$5/8$ |

**同一数据下，预测差异来自对参数的处理：选一个点，或保留分布并按后验平均。**

# 伯努利示例的参数点与后验平均计算

- **参数点：由指数求最大值位置**
  - $\theta^r(1-\theta)^s$ 在 $r, s > 0$ 时最大值位于 $r/(r+s)$
  - MLE：似然指数为 3、1，$3/4 = 0.750$
  - MAP：后验指数为 4、2，$4/6 = 2/3$
- **后验预测：对参数的不确定性求平均**
  - $P(Z = 1 \mid D) = \mathbb{E}[\theta \mid D]$
  - $Z = 1$ 为下一次成功；$\mathbb{E}[\theta \mid D]$ 是后验均值
  - $\mathrm{Beta}(a, b)$ 均值为 $a/(a+b)$
  - 本例：$5/8 = 0.625$，按整个后验平均

**接回加一平滑：**若另取 $\mathrm{Beta}(1,1)$ 先验，$n$ 次观察有 $k$ 次成功，下一次成功的后验预测为 $(k+1)/(n+2)$。这正是二值加一平滑；与本例 $\mathrm{Beta}(2,2)$ 下的 $5/8$ 区分。

# 本讲总结

- **模型建立：以结构假设降低估计负担**
  - 条件独立把联合概率拆成单特征概率；更易估计的代价是舍弃类内特征依赖
- **类别推断与决策：证据更新认识，损失决定行动**
  - 先验与类条件概率形成后验；0–1 损失下选最大后验类别，错误代价改变时行动可能改变
- **参数学习：估计值背后仍有不确定性**
  - MLE、MAP 选参数点，完整贝叶斯保留参数后验并平均预测；二值加一平滑可由 $\mathrm{Beta}(1,1)$ 后验预测解释