

- 判别式：直接针对要的东西进行建模
- 生成式：联合分布，再转化过去（但两次计算可能有两次误差）。更全面，拎出一部分都有用

生成式构建出模型后，可输入噪声产生新样本

朴素贝叶斯，HMM，MRF，DBN

贝叶斯分类器！=贝叶斯学习

LDA判别函数

$$\begin{aligned}
 \arg \max_y p(y | \mathbf{x}) &= \arg \max_y \ln p(y)p(\mathbf{x} | y) \\
 &= \arg \max_y \ln \pi_y + \ln p(\mathbf{x} | y) \\
 &= \arg \max_y \ln \pi_y - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) \\
 &= \arg \max_y \ln \pi_y + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y - \frac{1}{2} \boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y \\
 &= \arg \max_n \delta_n(\mathbf{x}).
 \end{aligned}$$

可以看出，当已知类别先验 π_n 以及各类别条件（高斯）分布的参数 $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 时，可以通过计算 $\{\delta_n(\mathbf{x})\}_{n=1}^N$ ，其中最大 $\delta_n(\mathbf{x})$ 所对应的类别即为 LDA 预测的类别。

可以看出， $\delta_n(\mathbf{x})$ 一定程度反映了样例 \mathbf{x} 预测为 n 类的置信度，置信度越高，则 $\delta_n(\mathbf{x})$ 越大。且 $\delta_n(\mathbf{x})$ 可以分解为

$$\underbrace{\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y}_{w_n} + \underbrace{\left(\ln \pi_y - \frac{1}{2} \boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y \right)}_{b_n}. \quad (37)$$

因此，LDA 中类别置信度的计算是关于示例 \mathbf{x} 的线性函数，和线性分类器的置信度计算方法类似。

本题中 LDA 的判别方法和贝叶斯最优决策论相关，详细内容将在第 7 章讨论。可以看出，LDA 通过对类别的先验、似然建模，推导出后验概率。当假设不同类别的类别条件概率均为高斯分布，且类别之间共享协方差矩阵，则以最大后验概率为目标推导出的类别预测函数 δ_n 是关于示例 \mathbf{x} 的线性函数，而 LDA 中“线性”的含义即源于此。

一、贝叶斯决策论

考虑相关概率和误判损失已知的理想情况，如何选择最优类别标记

需要知道真实后验概率才能算，但我们只能得到近似后验概率

所以达不到最优风险，只能逼近，方式有判别式和生成式

- 先验：还没发生就根据常识开始猜，猜各个结果的可能性分布

- 后验：根据结果猜是哪个原因方式引起的
- 似然：原因（参数）固定，猜结果可能性

贝叶斯最优分类器

- 后验概率： $P(c_i | \mathbf{x})$ ：将样本 \mathbf{x} 误分类成 c_i 的概率
- 条件损失（风险）： \mathbf{x} 分类为 c_i 的整体损失

$$R(c_i | \mathbf{x}) = \sum_{j=1}^N \lambda_{ij} P(c_j | \mathbf{x})$$

- 最优分类器 h^*

选择使每个样本条件风险最小的类别标记

$$h^*(\mathbf{x}) = \arg \min_{c \in \mathcal{Y}} R(c | \mathbf{x})$$

- 总体风险

寻找一个判定准则 $h: \mathcal{X} \mapsto \mathcal{Y}$ 以最小化总体风险

$$R(h) = \mathbb{E}_{\mathbf{x}} [R(h(\mathbf{x}) | \mathbf{x})] .$$

贝叶斯公式：后验=先验*似然

$$P(c | \mathbf{x}) = \frac{P(c) P(\mathbf{x} | c)}{P(\mathbf{x})}$$

- $P(\mathbf{x})$ ：证据因子，对所有类别标记相同（这个是固定的结果分布情况）
- $P(c)$ ：先验概率，样本空间中各类样本所占比例，可通过各类样本出现的频率来估计

先验是人的认识，比如硬币正反面都是1/2

频率学派不考虑先验，只从实际样本来判定

但各个 $P(c)$ 完全相同时，先验与后验相同

- $P(\mathbf{x} | c)$ ：似然，样本相对于类标记 c 的类条件概率

似然是数据观察的结果，看到的数据越多，似然就越大

$P(x_i | c)$ 表示 c 类中属性 x 的取值是 x_i 的概率

二、极大似然估计MLE

求解概率模型基本方法，概率问题解参数一定要

概率模型训练过程就是参数估计过程

基于生成式，试图在 θ_c 所有可能取值中，找使数据出现"可能性"最大的值

理解

思想：概率模型的参数应当使当前观测到的样本是最有可能被观测到的，即当前数据似然最大

人话：估计的分布应该是使得当前抽样结果发生可能性最大那个，真相可能有很多种，我们选择最大概率出现目前现象的那个作为真相

通常情况下，一枚硬币是均匀的，投出正反的概率均为 0.5。假设事件 H 为投出正面，事件 T 为投出反面， p_H 为投出正面的概率。

(1) 假设此时硬币正常，已知 $p_H = 0.5$ 。求三次试验的结果为 HHT 的概率。

(2) 假设此时硬币不知道正常与否， p_H 未知，但是已知经过三次试验结果为 HHT ，试估计出 p_H 。

解：

(1) $P(HHT | p_H = 0.5) = 0.5 \times 0.5 \times 0.5 = 0.125$

(2) $L(p_H | HHT) = P(HHT; p_H) = p_H \times p_H \times (1 - p_H) = p_H^2(1 - p_H)$

我们希望能找到一个 p_H 使得发生 HHT 的概率最大，也即找出 $p_H^2(1 - p_H)$ 的最大值，

计算得到 $p_H = \frac{2}{3}$ 取最大值，故估计量 $p_H = \frac{2}{3}$

知乎 @霖霖霖

两学派

- 频率主义学派：参数虽然未知但固定，可通过优化似然函数来确定参数
- 贝叶斯学派：参数随机变量有分布，假定参数的先验分布，再基于观测数据计算后验分布

步骤

假定 $P(x | c)$ 具有确定的概率分布形式，且被参数 θ_c 唯一确定，则任务就是利用训练集 D 来估计参数 θ_c

θ_c 对于训练集 D 中第 c 类样本组成的集合 D_c 的似然(likelihood)为

$$P(D_c | \theta_c) = \prod_{x \in D_c} P(x | \theta_c)$$

连乘易造成下溢，因此通常使用对数似然 (log-likelihood)

$$LL(\theta_c) = \log P(D_c | \theta_c) = \sum_{x \in D_c} \log P(x | \theta_c)$$

于是， θ_c 的极大似然估计为 $\hat{\theta}_c = \arg \max_{\theta_c} LL(\theta_c)$

- 假设 $P(x|c)$ 有确定形式并被参数向量 θ_c 唯一确定
- 用高斯分布建模概率分布(两个最大似然参数)

如果实际分布远不满足高斯，则效果不好

$$\hat{\mu}_c = \frac{1}{|D_c|} \sum_{x \in D_c} x,$$

$$\hat{\sigma}_c^2 = \frac{1}{|D_c|} \sum_{x \in D_c} (x - \hat{\mu}_c)(x - \hat{\mu}_c)^T.$$

评价

- 结果准确性严重依赖假设分布与实际的接近程度

最大后验估计MAP

- 贝叶斯视角，考虑先验
- 只是后面加入一项

$$\hat{\theta}_c = \arg \max_{\theta_c} \sum_{x \in D_c} \log P(x | \theta_c) + \log P(\theta_c)$$

三、朴素贝叶斯分类器

主要障碍：

- 似然是所有属性联合概率，难以从有限训练样本中获得

比如在数据集中统计各属性组合的出现次数作为 $P(x|c)$ 的离散时估计：
组合太多，样本稀疏，而且有的组合根本没出现

基本思路：条件独立性假设

因此可分解单个属性组合的条件概率 $P(x|c)$

分解后每个单个属性的条件概率 $P(x_i|c)$ 大大增加
 d 是属性数

不是独立同分布（所有样本从同一分布中独立的抽样出来），只是不同属性间独立

$$P(c | x) = \frac{P(c) P(x | c)}{P(x)} = \frac{P(c)}{P(x)} \prod_{i=1}^d P(x_i | c)$$

$$h_{nb}(x) = \arg \max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^d P(x_i | c)$$

两个估计目标

- 先验概率 $P(c)$ 估计：直接通过训练集标记获得
- 条件概率 $P(x_i|c)$ 估计：分离散和连续讨论

$P(x_i|c)$ 表示 c 类中属性 x 的取值是 x_i 的概率

- 对离散属性，令 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集合，则

$$P(x_i | c) = \frac{|D_{c,x_i}|}{|D_c|}$$

- 对连续属性，考虑概率密度函数，假定 $p(x_i | c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$

$$p(x_i | c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

例：朴素贝叶斯

- 所有 x_i 已知

好瓜里色泽青绿概率

后, 为每个属性估计条件概率 $P(x_i | c)$:

$$P_{\text{青绿}|\text{是}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375 ,$$

$$P_{\text{青绿}|\text{否}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333 ,$$

$$P_{\text{蜷缩}|\text{是}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{是}) = \frac{5}{8} = 0.375 ,$$

$$P_{\text{蜷缩}|\text{否}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333 ,$$

$$P_{\text{浊响}|\text{是}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750 ,$$

$$P(\text{好瓜} = \text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \\ \times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \approx 0.038 ,$$

$$P(\text{好瓜} = \text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \\ \times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \approx 6.80 \times 10^{-5} .$$

比较好瓜和坏瓜的后验概率大小来分类（实际上未必全要算出来，只需要大小关系）

实际计算把连乘替换为连加避免数值下溢

拉普拉斯修正

- 应对缺点：某一条件概率为0时，连乘会出事

$$P_{\text{清脆}|\text{是}} = P(\text{敲声} = \text{清脆} | \text{好瓜} = \text{是}) = \frac{0}{8} = 0$$

- 把所有连乘中的都平滑处理

训练集变大时，修正引入的先验 (prior) 影响逐渐变小

令 N 表示训练集 D 中可能的类别数， N_i 表示第 i 个属性可能的取值数

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N}, \quad \hat{P}(x_i | c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

$$\hat{P}(\text{好瓜} = \text{是}) = \frac{8 + 1}{17 + 2} \approx 0.474, \quad \hat{P}(\text{好瓜} = \text{否}) = \frac{9 + 1}{17 + 2} \approx 0.526.$$

类似地， $P_{\text{青绿}|\text{是}}$ 和 $P_{\text{青绿}|\text{否}}$ 可估计为

$$\hat{P}_{\text{青绿}|\text{是}} = \hat{P}(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3 + 1}{8 + 3} \approx 0.364,$$

朴素贝叶斯用处

- 对预测速度需求高时：先算好所有概率估值，用时查表
- 数据更替频繁时：懒惰学习：不训练，需要预测时再估值
- 数据不断增加时：增量学习，基于现有估值修正新样本涉及的概率估值

四、半朴素贝叶斯

某些属性不独立

独依赖估计ODE

- 每个属性在类别之外最多依赖于一个其他属性，即父属性

贝叶斯网