贝叶斯决策论

前言

本章主要探讨贝叶斯相关的概率框架下的决策。我们会探讨最基本的贝叶斯公式，在贝叶斯的公式基础上，我们介绍朴素贝叶斯和半朴素贝叶斯分类器，以及贝叶斯网络、期望最大化算法。半朴素贝叶斯和贝叶斯网络将会是难点，理解EM算法和参数更新迭代的过程。这就是本章的主要内容。

本人对贝叶斯没有什么深入研究，这是我第二遍过贝叶斯部分，所以会比较生...深入浅出这部分内容是有必要的。

同样，本章将参考西瓜书和统计学习方法两本书。

早在很早之前，人们就一件事情发生的概率是事先明确的还是经过不断地实验迭代而来的吵得不可开交。其中，前者坚信世间万物都已确定，即频率派，我们可以统计样本空间事件的个数来确定发生的概率；而后者则坚信，事件的概率由不断的连续的样本反复迭代而来，即贝叶斯派，事件的概率由原来事件的概率+修正得来。

https://www.bilibili.com/video/BV1kcW5zHEYk?vd_source=e52d78598c9e345db52a38406feeeae0

概率的本质是什么？频率派 vs 贝叶斯派视角：两种应对不确定性的哲学_哔哩哔哩_bilibili

概率的本质是什么？是客观存在的规律，还是主观信念的反映?从统计学的两大流派——频率派和贝叶斯派出发，我们将一起看看:他们是如何理解“概率”?又是如何用完全不同的统计学工具，来应对这个世界的不确定性。, 视频播放量 287152、弹幕量 587、…

两党之争，向来如此。但是如果你仔细思考就会发现，实际上二者是相互融合的。

贝叶斯派不得不使用先验概率来估计后验概率，而先验概率哪里来的？是由统计得来。

而频率派所说的概率表示的是事件发生频率的极限值。当重复试验的次数趋近无穷大时，事件发生的频率会收敛到真实的概率之上。这不就是贝叶斯派的迭代思想么？

所以本章大部分内容实际上不会过多解释那个属于什么派，而是总体来看。

贝叶斯决策论

贝叶斯公式

关于贝叶斯公式：

$$P(A|B)^{①} = \frac{P(B|A)^{②}P(A)^{③}}{P(B)^{④}}$$

有必要的说明：

1. 后验概率①：在 B 样本集的基础上，A 发生的概率

2. 似然度②：在事件 A 发生的前提下，观察 D 发生的概率。可以看作是先验概率修正的动态权重，或者高中的说法叫条件概率
3. 先验概率③：发生 A 事件了，其在 B 样本集基础上的概率
4. 证据概率④：B 样本集发生的概率

期望损失

贝叶斯决策理论是统计模式识别和机器学习领域的基石，它为我们提供了一种在不确定性下做出最优决策的数学框架。其核心思想非常直观：***在不确定性的情况下，最佳的决策是选择那个具有最大成功概率（或最小风险）的方案。***它巧妙地将我们已有的“先验知识”和从数据中观察到的“新证据”结合起来，通过贝叶斯公式，得出“后验概率”，并基于此进行决策...

所以，我们需要首先需要表示最大成功率（最小风险）。假设有 N 种可能的类别标记，即 $\mathcal{Y}=\{c_1,c_2,\dots,c_N\}$，$\lambda_{ij}$ 是将一个真实标记为 c_j 的样本分类为 c_i 所产生的损失。基于后验概率 $P(c_i|x)$ 可获得将样本 x 分类为 c_i 所产生的期望损失(expected loss)和，$R(\cdot|\cdot)$ 期望损失函数是样本 x 的“条件风险”(conditional risk)，也就是对于单个样本而言：

$$R(c_i|x)=\sum_{j=1}^{N}\lambda_{ij}P(c_j|x),\quad (7.1)$$

我们的任务是寻找一个判定准则 $h:\mathcal{X}\longrightarrow\mathcal{Y}$ 以最小化总体风险，即期望（平均）后，即总体风险：

$$R(h)=\mathbb{E}_x[\ R(h(x)\mid x\ )\ ],\quad (7.2)$$

对于每个样本而 x 言，若能使 h 能最小化条件风险 $R(h(x)|x)$，则总体风险 $R(h)$ 也将被最小化。这就产生了贝叶斯判定准则(Bayes decision rule)：最小化总体风险，只需要在每个样本上选择哪个能使条件风险 $R(c|x)$ 最小的类别标记，即

$$h^*(x)=\arg\min_{c\in\mathcal{Y}}\ R(c|x),\quad (7.3)$$

此时，h^* 被称为贝叶斯最优分类器(Bayes optimal classifier)，与之对应的总体风险 $R(h^*)$ 被称为贝叶斯风险(Bayes risk)。$1-R(h^*)$ 反映了分类器所能到达的最好性能，即通过机器学习所能产生的模型理论精度上限。

具体来说，若目标是最小分类错误率，则误判损失 λ_{ij} 可以写为（对应之前在SVM的0/1损失函数）：

$$\lambda_{ij}=\begin{cases}0, & if\ i=j;\\ 1, & otherwise;\end{cases},\quad (7.4)$$

此时条件风险

$$R(c|x)=1-P(c|x),\quad (7.5)$$

求 $P(c|x)$ 最小值，就是求 $R(c|x)$ 最大值。于是，最小化分类错误率的贝叶斯最优分类器为

$$h^*(x)=\arg\max_{c\in\mathcal{Y}}\ P(c|x),\quad (7.6)$$

即对于每个样本 x ，选择能使后验概率 $P(c|x)$ 最大的类别标记。

后验概率的构建差异

不过，要使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c|x)$ ，但是在现实任务中，后验概率通常很难获取。而机器学习所要实现的是基于有限的训练样本集尽可能准确地估计出后验概率 $P(c|x)$ 。也就是两种策略：

给定 x ，直接构建 $P(c|x)$ 来预测，这样得到的是“判别式模型”(discriminative models)

或者对联合概率分布 $P(c|x)$ 建模，然后再由此获得 $P(c|x)$ ，这样得到的是“生成式模型”(gennerative models)。这在之前的线性判别分析中是没有介绍的。

这也和刚开始介绍的频率派与贝叶斯派有关。生成式模型的数学框架与贝叶斯思想有天然的亲和性，而判别式模型的实用主义风格与频率派哲学有共通之处。

再谈贝叶斯公式

对于生成式模型来说，必然需要考虑贝叶斯公式：

$$
\begin{aligned}
P(c|x) &= \frac{P(x,c)}{P(x)}, \quad (7.7) \\
&= \frac{P(c)P(x|c)}{P(x)}, \quad (7.8)
\end{aligned}
$$

对于给定样本 x ，其证据因子与类标记无关，所以估计 $P(c|x)$ 的问题就转化为了如何基于训练数据 D 来估计先验 $P(c)$ 和似然 $P(x|c)$ 。

类先验概率 $P(c)$ 表达了样本空间中各类样本所占的比例，根据大数定律，当训练集包含充足的独立同分布样本时， $P(c)$ 可通过各类样本出现的频率来进行估计。

对类条件概率 $P(x|c)$ 来说，由于它涉及关于 x 所有属性的联合概率，直接根据样本出现的频率来估计将会遇到严重的困难。例如，假设样本的 d 个属性都是二值的，则样本空间将有 2^d 种可能的取值，在实际应用中，这个值往往远大于训练样本数 m ，也就是说，很多样本取值在训练集中可能就没出现过，直接用频率来估计 $P(x|c)$ 显然不可行，因为存在未被观察的数据。