贝叶斯决策论与朴素贝叶斯算法

徐迪深

July 1, 2021

1 介绍

1.1 贝叶斯决策论

贝叶斯决策论（Bayesian decision theory）是运用概率实施决策的基本方法之一。当你的目的是预测未来所有可能发生事件发生的概率，而你掌握了历史上与你所要预测的事件相关信息的概率时，贝叶斯决策论就基于贝叶斯公式提供了一种误判损失最小的预测方法。

例 1 假设你想买到甜而且皮薄的西瓜，消费这种西瓜的效用对你来说是最大的。之前你吃过的西瓜中你注意到西瓜的甜度和皮的厚薄在色泽，纹理，根蒂这三个特征上有差异，那么你根据之前吃的瓜发现当色泽更深，纹理更加清晰，根蒂蜷缩的瓜更像甜而且皮薄的西瓜。所以你在挑选的瓜就会尽量挑选根上述三个特征接近的瓜。因为你认为这样最有可能选到你想要的瓜。

1.2 用数学的语言来解释贝叶斯决策论

在例 1 中我们通过挑选西瓜的方式对贝叶斯决策论的思路有一个感性的理解。那么数学上贝叶斯决策论是如何推导的?

因为贝叶斯决策论的思路是在概率的框架下找到误判损失最小的条件，所以我们需要用条件概率来表示误判风险，然后去寻找一个能够使得这个风险最小的映射。

以例 1 作为背景，假设 x_1 表示色泽，x_2 表示纹理，x_3 表示根蒂，这样我们有三个可以参考的特征 $X = (x_1, x_2, x_3)$，假设 y_1 表示甜，皮薄的瓜，y_2 表示不甜，皮薄的瓜，y_3 表示甜，皮厚的瓜，y_4 表示不甜，皮厚的瓜，那么我们未来可能发生的事情就是 $Y = (y_1, y_2, y_3, y_4)$。那么 $P(y_i|X), i = 1, 2, 3, 4$ 就代表拿到特征为 X 的是上述四种情况的概率。那么对于一个瓜的一组特征 $\bar{X}$, 假设这个瓜实际是一个甜而且皮薄的瓜，假设你错误判断成剩下三种你不想要的西瓜而没有买所导致的效用损失是 $\lambda_i = U_1 - U_i$，$i = 2, 3, 4$ 表示你误判成剩下三个西瓜的种类，那么你猜错的条件风险（conditional risk）就是

$$\sum_{i=2}^{4} \lambda_i P(y_i|\bar{X}) \qquad (1)$$

假设有一个判断准则是 $h : X \to Y$，那么基于 (1) 式，就有

$$(1) = \sum_{i=2}^{4} \lambda_i P(h(\bar{X})|\bar{X}) \qquad (2)$$

很显然，在 (2) 中，λ_i 对每一个 i 来说是常数。所以我们希望找到一个分类法则 $h^\star : X \to Y$ 使得 $P(y_i|\bar{X}), i = 2, 3, 4$ 最小，一次来达到我们的条件风险最小，而这实际上等价为将 $P(y_1|\bar{X})$。这样的 $h\star$ 被称为贝叶斯最优分类器 (Bayes optimal classifier)，而将贝叶斯最优分类器带入 (2) 就得到了贝叶斯风险 (Bayes risk)。如果要衡量贝叶斯最优分类器的具体精度，在例 1 下，用 $P(y_1|\bar{X})$ 的大小即可。

1.2.1 如何最小化误判的概率

刚才我们得到了结论是如果想要最准确地预测，那么就要最小化误判的概率也就是最大化正确判断的概率。我们仍然以例 1 继续说明如何最大化正确判断的概率 $P(y_1|\bar{X})$。

由条件概率公式我们有

$$P(y_1|\bar{X}) = \frac{P(\bar{X}, y_1)}{P(\bar{X})} \qquad (3)$$

显然 $P(\bar{X})$ 对于总体是固定的，那么原来最大化 $P(y_1|\bar{X})$ 的问题就转化成了 $P(\bar{X}, y_1)$ 最大化的问题。由全概率公式可得

$$P(\bar{X}, y_1) = P(\bar{X}|y_1)P(y_1) \qquad (4)$$

将 (4) 和 (3) 联立，得

$$P(y_1|\bar{X}) = \frac{P(\bar{X}|y_1)P(y_1)}{P(\bar{X})} \qquad (5)$$

这也是贝叶斯公式。在 (5) 中，$P(y_i)$ 是类别得先验概率 (prior probability)，$P(\bar{X}|y_1)$ 是样本特征相对于类别 y_1 的条件概率（即在 y_1 类别的一堆瓜之中挑出一个瓜，有多大可能可以拿到特征为 $\bar{X}$ 的瓜），这样的条件概率也被称为似然概率 (likelihood)。显然，先验概率是固定的，当训练样本足够的情况下由大数定理可知概率可以用频率估计，那么这个时候我们的目标再次转变为最大化似然概率。

1.2.2 极大似然估计

当我们想要最大化似然概率时，我们最常用的方法就是极大似然估计。

背景 极大似然估计的背景是频率学派 (Frequentist) 对于总体概率分布的认知：总体的概率分布的参数是未知且无法精准预测的，但是客观存在而且是一个固定的值，所以可以通过对总体抽样，利用样本估计。与之相对，贝叶斯学派 (Bayesian) 则认为参数本身是一个随机变量，本身也服从某一个分布，然后从观测到的数据来计算参数得后验分布。

而极大似然估计就是频率学派用于估计概率分布参数的经典方法。

极大似然估计的步骤 首先先假设我们要估计的概率分布。这个假设并不是无依据的。比如在样本容量足够大的时候根据中心极限定理，我们就可以我们估计的概率分布是一个未知的正态分布。

在例 1 的背景下，假设 D_0 是指第一类（甜而且皮薄的瓜）的集合，那么这些瓜在色泽，纹理和根蒂上就是一个矩阵:

$$D_0 = \begin{pmatrix} x_{1i} & x_{2i} & x_{3i} \end{pmatrix}$$

i 指第一类样本中的第 i 个样本。那么对于每一个样本，它们是独立同分布的。我们将每一个样本服从的概率分布的参数的条件概率求积，数学表达如下:

$$P(D_0|\theta_1) = \prod_{D_i \in D_0} P(D_i|\theta_1) \qquad (6)$$

对 θ_1 进行极大似然估计就是想找到能使 $P(D_0|\theta_1)$ 最大的 $\hat{\theta_1}$，我们通过对 (6) 求导就可以找到这个数字。

1.3 朴素贝叶斯分类器

1.3.1 贝叶斯分类器

在贝叶斯决策论下，我们想要找到一组最符合我们预期的样本所包含的特征。贝叶斯分类器则是利用贝叶斯决策论的思想，反其道而行之：当我们在学习完历史数据之后，现在给我们一个未知样本，我们根据这个样本的特征则可以知道这个样本最大可能是什么类别。这样的算法我们称为贝叶斯分类器。

1.3.2 朴素的假设

之前我们得出了如果想要最大化我们的后验概率，那么就要最大化我们的似然概率。根据条件概率公式

$$P(X|y_i y_i) = p(X,Y)P(Y) \qquad (7)$$

$X = \{x_1, x_2, ..., x_n\}$ 是所有特征的集合，$y_i \in Y = \{y_1, y_2, ..., y_m\}$ 则是所有类别的集合。不难看出，我们想要求得联合分布（$P(X,Y)$）非常困难的。所以我们为了计算方便会加入一个各个特征之间互相独立的假设，即朴素的假设。此时，有

$$(7) = P(X|y_i y_i) = \prod_{x_i \in X} P(x_i|y_i) \qquad (8)$$

将 (8) 和 (5) 联立，有

$$P(y_1|\bar{X}) = \frac{\prod_{x_i \in \bar{X}} P(x_i|y_1)P(y_1)}{P(\bar{X})} \qquad (9)$$

在 (9) 中，分母 $P(\bar{X})$ 在总体和样本都是一个确定的值，所以真正能够影响到后验概率的，是分母。这样我们就给出了贝叶斯分类器的判断式

$$h_{best\ classification}(x) = argmax_{y_i \in Y} \prod_{x_i \in X} P(x_i|y_i)P(y_i)$$

2 程序实现：如何通过财务指标判断来对上市公司是否是 ST 公司

Edward Altman(1968) 提出了运用财务数据构造特征来预测公司的违约风险，建立了如下模型:

$$Z = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

X_1 是 (流动资产-流动负债)/总资产，X_2 是留存收益/总资产，X_3 是息税前利润/总资产，X_4 是股票市值/总债务账面价值，X_5 是销售收入/总资产。

Z 值越小，公司的违约概率越大。当 Z 值小于 1.81 时认为该公司大概率会发生违约，而 Z 大于 2.99 时认为公司的违约事件属于小概率事件，然而当 Z 处于两个值的临界时无法可靠判断。

Z 指标使用的模型是典型的线性模型，而我们可以尝试运用同样的财务指标观察在中国的 A 股市场贝叶斯风险是否低于 z 指标的条件风险（即贝叶斯分类器是否在预测方面有更优秀的表现）