机器学习讲义(L17-A)：生成式贝叶斯分类器

授课教师：王贝伦 / 助教：张嘉琦，黄旭，谈笑，徐浩卿

1 贝叶斯分类器（Bayes Classifiers）

1.1 生成模型与判别模型

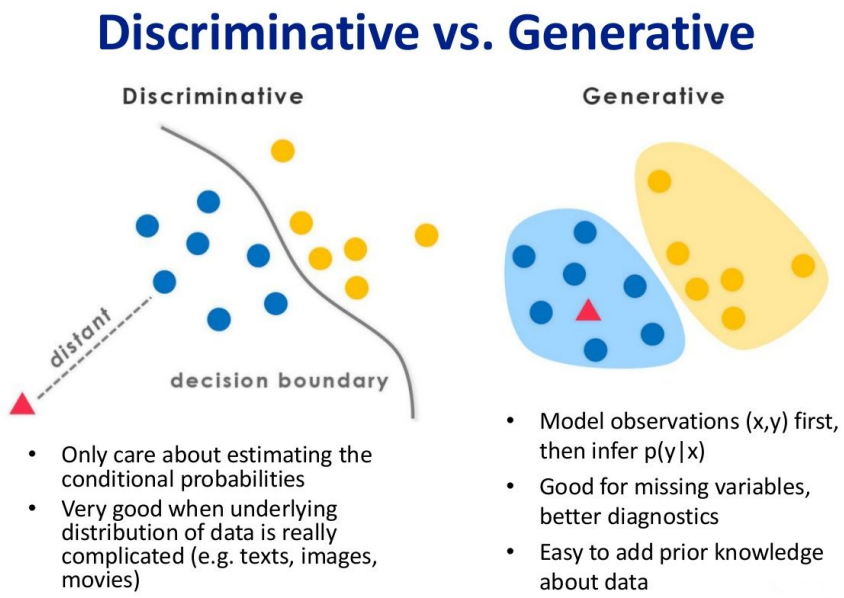

Figure 1: 生成模型和判别模型的对比

其实分类机器学习模型的任务是从属性 x 预测标记 c，即求概率 $P(c|x)$；

对于**判别式模型**来说，对于未知实例 x，根据 $P(c|x)$ 可以求得标记 c，即可以直接判别出来，如上图的左边所示，实际就是直接得到了判别边界，所以传统的、耳熟能详的机器学习算法如线性回归模型、支持向量机 SVM 等都是判别式模型，这些模型的特点都是输入属性 x 可以直接得到 c。比如对于二分类任务来说，实际得到一个 score，当 score 大于阈值（threshold）时则为正类，否则为反类。

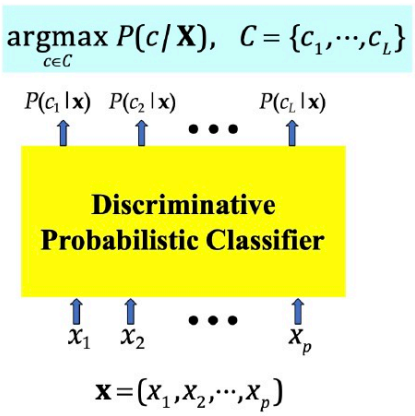

Figure 2: 判别式模型

而对**生成式模型**来说，对于未知实例 x，我们需要求出 x 与所有标签之间的联合概率分布，再通过贝叶斯规则求 $P(c|x)$，即

$$\begin{aligned} P(C=c_i|\boldsymbol{X}=\boldsymbol{x}) &= \frac{P(\boldsymbol{X}=\boldsymbol{x}|C=c_i)P(C=c_i)}{P(\boldsymbol{X}=\boldsymbol{x})} \\ &\propto P(\boldsymbol{X}=\boldsymbol{x}|C=c_i)P(C=c_i) \\ &for\ i=1,2,...,L \end{aligned} \tag{1}$$

对于所有的 $P(C=c_i|\boldsymbol{X}=\boldsymbol{x})$ 来说，分母 $P(\boldsymbol{X}=\boldsymbol{x})$ 都是相等的，而分子实质上就是联合概率分布 $P(\boldsymbol{X}=\boldsymbol{x},C=c_i)$，这时候我们的目标就变成选取使联合概率分布最大的那一个标签作为最后的结果。如图 1 右边所示，并没有什么边界存在，对于未知实例（红三角），

求两个联合概率分布（有两个类），比较一下，取那个大的（注意：这里是由于我们的联合概率分布和后验概率分布成正比，所以通常取后验概率最大，可以等同于取联合概率分布最大）。机器学习中朴素贝叶斯模型、隐马尔可夫模型 HMM 等都是生成式模型。下图展示了生成式模型的分类器生成的概率为 $P(x|c)$，并以此为基础生成联合分布概率进行分类评估。

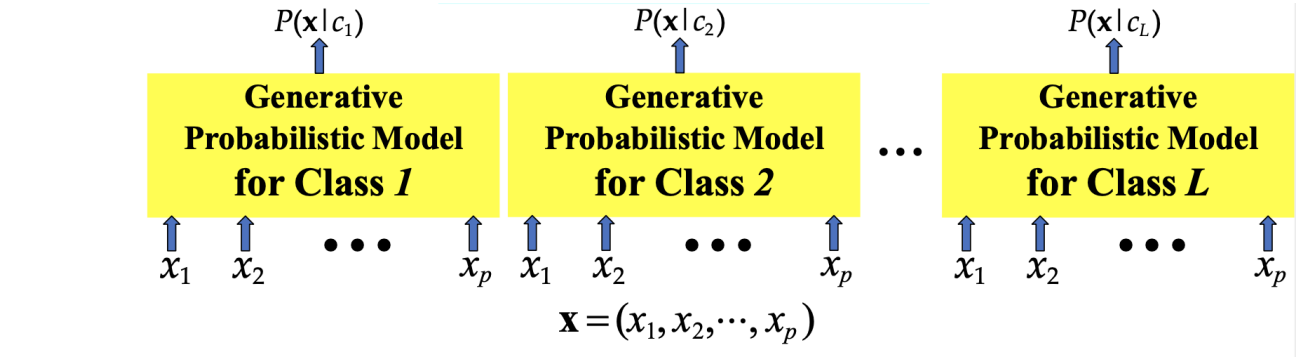

Figure 3: 生成式模型

而所谓的 MAP，全称为 Maximum a posterior estimation，即最大后验概率估计，求使 $P(X|\theta)P(\theta)$ 最大的参数 θ，

$$\begin{aligned}\operatorname{argmax} P(x|\theta) &= \operatorname{argmax} \frac{P(x|\theta)P(\theta)}{P(x)} \\ &= \operatorname{argmax} P(x|\theta)P(\theta) = \operatorname{argmax} \prod_i P(x^{(i)}|\theta)P(\theta)\end{aligned}$$

然后跟极大似然估计一样，对其进行对数处理，对目标函数求导使其等于 0。这里是通过对参数的调整使得 $P(C = c_k|\boldsymbol{X} = \boldsymbol{x})$ 最大（这里的 c_k 指正确的类别）。

判别式模型举例：要确定一个羊是山羊还是绵羊，用判别模型的方法是从历史数据中学习到模型，然后通过提取这只羊的特征来预测出这只羊是山羊的概率，是绵羊的概率。

生成式模型举例：利用生成模型是根据山羊的特征首先学习出一个山羊的模型，然后根据绵羊的特征学习出一个绵羊的模型，然后从这只羊中提取特征，放到山羊模型中看概率是多少，再放到绵羊模型中看概率是多少，哪个大就是哪个。

总结一下，生成模型基于输入 x 和标签 c 的联合概率 $P(x, c)$，并通过使用贝叶斯规则计算 $P(c|x)$ 进行预测，然后选择最可能的标签 c 。判别式分类器直接对后验概率 $P(c|x)$ 建模，或者学习从输入 x 到类标签的直接映射。

1.2 生成式贝叶斯分类器（Generative Bayes Classifiers）

这里我们直接用一个例子进行说明，该例子由 14 个样本组成，数据包括四个特征，第一个为天气（outlook），有三个可能值：Sunny，Overcast，Rain；第二个特征为温度（temperature），有三个可能值：Hot，Mild，Cool；第三个特征为湿度（humidity），有两个可能值：High，Normal；第四个特征为风力（wind），有两个可能值：Weak，Strong。表格的最后一列是类别，表示这种外界环境是否适合打网球，有两个可能值：Yes，No。

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
	X_1	X_2	X_3	X_4	C

Figure 4: 例子

由 (1) 式可知，只有考虑所有的参数，才能保证对于实例所有的可能性都考虑在内，所以我们需要考虑所有参数取值组合的联合概率。对于上图，我们以 $|X_i|$ 表示该特征的可取值个数，有 $|X_1|=3, |X_2|=3, |X_3|=2, |X_4|=2$，而类别的可取值个数，$|C|=2$，那么该模型一共有 $3\times3\times2\times2\times2=72$ 个参数。如下图，列出所有的情形

Outlook (3 values)	Temperature (3 values)	Humidity (2 values)	Wind (2 values)	Play=Yes	Play=No
sunny	hot	high	weak	0/9	1/5
sunny	hot	high	strong	.../9	.../5
sunny	hot	normal	weak	.../9	.../5
sunny	hot	normal	strong	.../9	.../5
....	....	....	....	....	....
....	....	....	....	....	....
....	....	....	....	....	....
....	....	....	....	....	....

Figure 5: 所有参数

此外，由于生成模型求的是联合概率，所以我们还需要求得先验概率

$$P(Play=Yes)=\frac{9}{14}\quad P(Play=No)=\frac{5}{14}$$

当一个新的实例出现时，其必为图 5 这些参数中的一种，我们只需要计算最后的联合概率分布就可以了。

那么这些参数怎么求呢？在这里我们使用极大似然估计，听起来很复杂，其实很简单，就是以数据出现的频率代替概率 [2]。对于先验概率来说，我们想求 Play=Yes 的先验概率就是统计 Play=Yes 的个数然后除以总数，如图 4，我们可以数得标红的条目共有 9 个，所以其先验概率为 $\frac{9}{14}$；而对于条件概率来说其实也是相同的办法，我们以图 4 为例，我们想知道在 Play=Yes 的情况下，Outlook=Overcast，Temperature=Hot，Humidity=High，Wind=Weak 的概率是多少，那么就是在这标红的 9 条条目中寻找符合的情况，共找到一条，那么它的概率就是 $\frac{1}{9}$。

我们可以看出来，这种方法过于依赖训练数据，一旦数据过少或者失真，预测值将严重偏离真实值，而同时，仅是上述这种简单的表格，其参数就达到 72 之多，其运算效率极其低下。

2 朴素贝叶斯分类器（Naïve Bayes Classifier）

为了减少运算，我们通常会对样本特征做出一些假设。我们先介绍一个最简单的生成式贝叶斯分类器，**朴素贝叶斯分类器**。同样的，对于样本数据 $X=(x_1,x_2,\cdots,x_p)$，朴素贝叶斯分类器的目的也是预测

$$P(C\mid X=(x_1,x_2,\cdots,x_p))=\frac{P(X=(x_1,x_2,\cdots,x_p)\mid C)P(C)}{P(X=(x_1,x_2,\cdots,x_p))} \tag{2}$$

从之前对于生成式贝叶斯分类器的介绍中我们可以看到，如果每个属性 x_i 的可选择值比较多的话，计算 $P(X=(x_1,x_2,\cdots,x_p)\mid C)$ 就比较复杂。以我们之前的网球的例子来说，我们一共有 4 个属性，这些属性去之不同的组合有 $3\times3\times2\times2=36$ 种，因此我们需要计算并保存这 36 种组合的概率 $P(X\mid C)$。对于复杂的问题来说，这部分的开销会很大。为了降低计算成本，朴素贝叶斯方法为模型加上了一个比较强的假设，即**每个属性 $\mathbf{x_i}$ 之间是条件独立的，满足 $\mathbf{x_i}\perp\mathbf{x_i}\mid\mathbf{C},\ \forall\mathbf{i}\neq\mathbf{j}$**。有了这个条件独立的假设，条件概率 $P(X\mid C)$ 就可以写成

$$P(X=(x_1,\cdots,x_p)\mid C)=P(x_1\mid C)\times P(x_2\mid C)\times\cdots\times P(x_p\mid C) \tag{3}$$

因此，对于朴素贝叶斯分类器来说，我们只需要计算针对于每一个属性的条件概率 $P(x_i\mid C)$ 而非联合概率。这样我们的计算就会大大减少。仍旧以网球的问题为例，在朴素贝叶斯分类器中，我们只需要计算 $3+3+2+2=10$ 个概率。

2.1 训练朴素贝叶斯分类器

朴素贝叶斯分类器的训练重点仍然在与计算多个条件概率 $P(X\mid C)$ 以及先验概率 $P(C)$。我们依旧用数据集中样本出现的频率来代替概率进行计算。具体来说，我们有

$$\begin{aligned}P(x_i\mid c_j)&=\frac{N(X_i=x_i,C=c_j)}{N(C=c_j)}\\ P(c_j)&=\frac{N(C=c_j)}{N}\end{aligned} \tag{4}$$

其中 N 是样本的数量，而 $N(a)$ 是满足条件 a 的样本数量。上面的公式（4）用形象一点的形式来解释就是

$$\begin{aligned}P(x_i\mid c_j)&=\frac{\text{第 i 个属性为 } x_i \text{ 且类别为 } c_j \text{ 的样本数量}}{\text{类别为 } c_j \text{ 的样本数量}}\\ P(c_j)&=\frac{\text{类别为 } c_j \text{ 的样本数量}}{\text{样本总量}}\end{aligned} \tag{5}$$

2.2 利用朴素贝叶斯分类器预测

计算得到了我们需要的所有概率值之后，我们如何对一个新的样本进行分类？这就需要用到贝叶斯法则了。假设我们有两个属性 $x_1 = \{a_1, a_2\}, x_2 = \{b_1, b_2\}$ 以及两个类别 $C = \{\text{Yes, No}\}$，并且我们已经通过数据计算并保存了所有需要的概率值。现在对于一个新的样本 $X = (x_1 = a_1, x_2 = b_1)$。注意到对于朴素贝叶斯分类器，我们有

$$\begin{aligned} P(C \mid X) &= \frac{P(X \mid C)P(C)}{P(X)} \\ &= \frac{P(x_1 \mid C)P(x_2 \mid C)P(C)}{P(X)} \\ &\propto P(x_1 \mid C)P(x_2 \mid C)P(C) \end{aligned} \tag{6}$$

因此我们只需要计算出 $P(x_1 \mid \text{Yes})P(x_2 \mid \text{Yes})P(\text{Yes})$ 与 $P(x_1 \mid \text{No})P(x_2 \mid \text{No})P(\text{No})$，对比哪个数值更大，则这个样本就属于哪个类别。因此对于这个样本 $X = (x_1 = a_1, x_2 = b_1)$，我们需要比较 $P(x_1 = a_1 \mid \text{Yes})P(x_2 = b_1 \mid \text{Yes})P(\text{Yes})$ 与 $P(x_1 = a_1 \mid \text{No})P(x_2 = b_1 \mid \text{No})P(\text{No})$。

2.3 朴素贝叶斯分类器示例

我们依然以之前的打网球的例子来具体说明在朴素贝叶斯分类器来完成分类。假设我们的样本数据如图6所示。可以看到 $P(x_1 = \text{Rain} \mid C = \text{Yes}) = \frac{3}{9}$。

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes ←
D4	[Rain]	Mild	High	Weak	Yes ←
D5	[Rain]	Cool	Normal	Weak	Yes ←
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes ←
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes ←
D10	[Rain]	Mild	Normal	Weak	Yes ←
D11	Sunny	Mild	Normal	Strong	Yes ←
D12	Overcast	Mild	High	Strong	Yes ←
D13	Overcast	Hot	Normal	Weak	Yes ←
D14	Rain	Mild	High	Strong	No

Figure 6: 样本数据。

同理我们可以计算出其他所有条件概率值，如图7, 8, 9以及10所示。并且我们有先验概率 $P(C = \text{Yes}) = \frac{9}{14}$ 以及 $P(C = \text{No}) = \frac{5}{14}$

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Figure 7: outlook 数据的条件概率。

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Figure 8: temperature 数据的条件概率。

Humidity	Play=*Yes*	Play=*No*
High	3/9	4/5
Normal	6/9	1/5

Figure 9: humidity 数据的条件概率。

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

Figure 10: wind 数据的条件概率。

现在，假设我们有一个新的数据 $X = (\text{Outlook} = \text{Sunny}, \text{Temperature} = \text{Cool}, \text{Humidity} = \text{High}, \text{Wind} = textStrong)$，我们对于两个不同的类别 Yes, No 分别计算概率为

$$\begin{aligned} P(\text{Yes} \mid X) &= 0.0053 \\ P(\text{No} \mid X) &= 0.0206 \end{aligned} \tag{7}$$

所以我们将这个样本 X 的类别分为 No。我们这里省略了计算过程，具体的计算很简单，只是用到公式（4）以及贝叶斯定理。

2.4 朴素贝叶斯分类器总结

朴素贝叶斯分类器是一个非常简单的模型，由于它的简单性和不差的预测准确度，朴素贝叶斯分类器也经常被使用。朴素贝叶斯分类的基本思想就在于假设所有属性之间是条件独立的，虽然这是一个比较强的假设并且这个假设不符合很多问题的实际情况，但这个假设为后续计算节省下了大量的成本。

在使用朴素贝叶斯分类器时，因为我们是以收集到的数据为基础来计算的，由于数据的局限性，我们会遇到一些难以处理的情况。比如如果一些情况的数据在训练的时候从未出现过，那么对于新出现的有这种属性的数据，我们就无法进行分类。仍然以打网球为例，假设我们的数据中从未出现过 outlook $=$ sunny 并且类别是 Yes 的数据，那么我们就无法计算 $P(\text{outlook} = \text{sunny} \mid \text{Yes})$ 的概率。对于一个新的数据，如果这个新数据的属性 outlook $=$ sunny，我们就无法计算出它属于 Yes 的概率。

还存在的一种问题就是零概率 $P(x \mid c) = 0$ 的问题。假设我们知道 outlook 可以取到 sunny 这个值，但是在我们收集到的数据中，没有哪一个数据的 outlook $=$ sunny，所以我们计算得到 $P(\text{outlook} = \text{sunny} \mid \text{Yes}) = P(\text{outlook} = \text{sunny} \mid \text{No}) = 0$。显然，这和我们的直觉是不符的，零概率表示一件事永远不会发生，然而 outlook $=$ sunny 并且类别为 Yes 这件事是完全有可能发生的。出现这种错误的原因在于我们收集的数据并不全面。为了处理这类问题，我们需要对计算的条件概率做一些平滑处理，即公式（4）中对于条件概率的计算应该改为

$$P(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + K_i} \tag{8}$$

其中 K_i 表示属性 x_i 可能取到的值的数量。对于打网球问题的 outlook 属性来说，它能取到三个值 sunny，rain 以及 overcast，因此我们有 $K_{\text{outlook}} = 3$。通过这样的平滑，我们的计算中就永远不会出现零概率的值。

引用

[1] https://www.zhihu.com/question/20446337

[2]https://www.cnblogs.com/LuffysMan/p/9748820.html

[3] Naïve Bayes Classifier: https://en.wikipedia.org/wiki/Naive_Bayes_classifier