机器学习笔记–朴素贝叶斯和多项分布模型

空修菜

1 朴素贝叶斯方法（Naive Bayes）

1.1 一些理论

我们设 $\phi_{j|y=0} = p(x_j = 1|y = 0), \phi_{j|y=1} = p(x_j = 1|y = 1), \phi_y = p(y = 1)$. 由这个表达式可以知道 ϕ_y 是一个常数，因为概率为 1，所以 $p(y = 0) = 1 - \phi_y$, 由于 $\phi_{j|y=0} = p(x_j = 1|y = 0)$ 是 $y = 0$ 时，x 的第 j 个元素为 1 的概率, 所以 $p(x_j = 0|y = 0) = 1 - \phi_{j|y=0}$, 同理, $\phi_{j|y=1} = p(x_j = 1|y = 1)$ 是 $y = 1$ 时，x 的第 j 个元素为 1 的概率, 所以 $p(x_j = 0|y = 1) = 1 - \phi_{j|y=1}$。

由上面的叙述，在 $y = 1$ 时，x_j 取 0 或者取 1 的概率就可以表示为

$$\phi_{j|y=1}^{x_j}(1 - \phi_{j|y=1})^{1-x_j}.$$

在知道 $p(Y), p(Y|X)$ 后，利用 Bayes 的先验方法得到的后验概率就可以表示为

$$p(Y|X) = \frac{p(Y)p(X|Y)}{p(X)}.$$

所以 X, Y 的联合分布可以写为

$$L(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \prod_{i=1}^{m} \left\{ (1 - \phi_y)\phi_{j|y=0}^{x_j^{(i)}} \Big(1 - \phi_{j|y=0}\Big)^{1-x_j^{(i)}} \right\}^{1-y^{(i)}} \left\{ \phi_y \phi_{j|y=1}^{x_j^{(i)}} \Big(1 - \phi_{j|y=1}\Big)^{1-x_j^{(i)}} \right\}^{y^{(i)}}.$$

基于上面的分布表达式，我们就得到似然函数 ℓ

$$\begin{aligned}\ell(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}) = \log L = &\sum_{i=1}^{m}(1 - y^{(i)}) \log(1 - \phi_y) + \sum_{i=1}^{m}(1 - y^{(i)})x_j^{(i)} \log \phi_{j|y=0} \\ &+ \sum_{i=1}^{m}(1 - y^{(i)})(1 - x_j^{(i)}) \log(1 - \phi_{j|y=0}) + \sum_{i=1}^{m} y^{(i)} \log \phi_y \\ &+ \sum_{i=1}^{m} y^{(i)} x_j^{(i)} \log \phi_{j|y=1} + \sum_{i=1}^{m} y^{(i)}(1 - x_j^{(i)}) \log(1 - \phi_{j|y=1})\end{aligned}$$

由上面的表达式，利用极大似然估计可以得到

$$0 = \frac{\partial \ell}{\partial \phi_y} = \frac{1}{\phi_y} \sum_{i=1}^{m} y^{(i)} - \frac{1}{1-\phi_y} \sum_{i=1}^{m} (1 - y^{(i)}),$$

整理上式可得

$$\phi_y = \frac{1}{m} \sum_{i=1}^{m} y^{(i)}.$$

利用同样的方法

$$0 = \frac{\partial \ell}{\partial \phi_{j|y=0}} = \frac{1}{\phi_{j|y=0}} \sum_{i=1}^{m} (1 - y^{(i)}) x_j^{(i)} - \frac{1}{1-\phi_{j|y=0}} \sum_{i=1}^{m} (1 - y^{(i)})(1 - x_j^{(i)}),$$

由上式整理可得

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} (1 - y^{(i)}) x_j^{(i)}}{\sum_{i=1}^{m} (1 - y^{(i)})}.$$

用同样的方法

$$0 = \frac{\partial \ell}{\partial \phi_{j|y=1}} = \frac{1}{\phi_{j|y=1}} \sum_{i=1}^{m} y^{(i)} x_j^{(i)} - \frac{1}{1-\phi_{j|y=1}} \sum_{i=1}^{m} y^{(i)} (1 - x_j^{(i)}),$$

由上式整理可得

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} y^{(i)} x_j^{(i)}}{\sum_{i=1}^{m} y^{(i)}}.$$

利用 indicate function 我们就可以将所得到的结果表示为

$$\begin{aligned}
\phi_y &= \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)} = 1\} \\
\phi_{j|y=0} &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0 \wedge x_j^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \\
\phi_{j|y=1} &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1 \wedge x_j^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}
\end{aligned}$$

在完成以上参数的估计以后，对一给定的 x 我们就可以进行预测了，主

要是利用以下的 Bayes rule，

$$\begin{aligned} p(y=1|x) &= \frac{p(y)p(x|y)}{p(x)} \\ &= \frac{\prod_{j=1}^{n} p(x_j|y=1)p(y=1)}{\prod_{j=1}^{n} p(x_j|y=1)p(y=1) + \prod_{j=1}^{n} p(x_j|y=0)p(y=0)} \\ &= \frac{\prod_{j=1}^{n} \phi_y \phi_{j|y=1}^{x_j}(1-\phi_{j|y=1})^{1-x_j}}{\prod_{j=1}^{n} \phi_y \phi_{j|y=1}^{x_j}(1-\phi_{j|y=1})^{1-x_j} + \prod_{j=1}^{n}(1-\phi_y)\phi_{j|y=0}^{x_j}(1-\phi_{j|y=0})^{1-x_j}} \end{aligned}$$

1.2 拉普拉斯光滑（Laplace smoth）

模型总结

1. 生成适合于训练的训练集。S 是原始的要检查的邮件。在朴素贝叶斯方法中，假设有一本词典存在，每个样本 x 的每个特征按照词典中词的顺序排列，x 初始化为一个 0 向量。若 x 的特征出现在 S 中，则将 x 中对应特征的值标为 1；

2. 朴素贝叶斯的一个强假设是 x 的每个分量 x_i, x_j 之间是条件独立的，即在同一个类别中分量之间的互不影响：

$$p(x_i \mid y=1) = p(x_i \mid y=1; x_j).$$

3. ϕ_y 是垃圾邮件的概率，所以不是垃圾邮件的概率是 $1-\phi_y$；

4. $\phi_{j|y=1} = p(x_j = 1 \mid y = 1)$ 计算的是：当邮件是垃圾邮件时，x 的第 j 个词出现的概率。因此，当邮件是垃圾邮件时，x 的第 j 个词不出现的概率是 $1-\phi_{j|y=1}$；

5. 在预测时，需要计算 $p(x_i \mid y=1)$，此时，如果 $x_i = 1$，那么就直接取值 $\phi_{i|y=1}$，否则取值 $1-\phi_{i|y=1}$。同理，可计算 $p(x_j \mid y=0)$；

6. **问题**。上面的模型存在的问题是：如果用于预测的 x 中的特征所代表的单词从没有在训练时出现过，那么概率 $p(y=1 \mid x)$ 的值不确定。假设 x_j 在训练时从没出现过，即在训练集中，无论是 $y=1$，还是 $y=0$，都有 $x_j = 0$，所以 $\phi_{j|y=0} = 0$，$\phi_{j|y=1} = 0$，因此，

$$\begin{aligned} p(x_j = 1 \mid y = 1) &= 1 \times 0 = 0, \\ p(x_j = 1 \mid y = 0) &= 1 \times 0 = 0, \end{aligned}$$

再由 $p(y=1 \mid x)$ 的计算公式以及乘法原理可知

$$p(y=1 \mid x) = 0/0.$$

7. 处理上述问题的方法称为 Laplace 光滑：k 是 y 的类别数，

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0 \wedge x_j^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} + k}$$
$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1 \wedge x_j^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} + k},$$

对于二分类问题 $k = 2$。

2 多项分布模型（Event models for text classification）

2.1 一些理论

当我们不再用 0 和 1 来表示一个以文本为元素的向量，而采用这些字在字典里的代号数字，我们就得到了这样的表示

$$x = \begin{pmatrix} x1 = 234 \\ x_2 = 1234 \\ \vdots \\ x_n = 86320 \end{pmatrix}$$

此时，我们引入记号 $\phi_{k|y=0} = p(x_j = k|y = 0)$, $\phi_{k|y=1} = p(x_j = k|y = 1)$,$\phi_y$ 与之前的朴素贝叶斯方法中的参数相同，即 $\phi_y = p(y = 1)$。在这里，我们允许每一个 x 的维数不一样，比如 $x^{(1)}$ 有 10 维数，而 $x^{(7)}$ 有 100 维，所以我们知道维数是依赖于 i 的，所以我们将每一个 $x^{(i)}$ 的维数记为 n_i。有了以上的符号，我们就可以写出分布函数 $p(x^{(i)}, y^{(i)})$

$$\begin{aligned} p(x^{(i)}, y^{(i)}) = p(y^{(i)})p(x^{(i)}|y^{(i)}) &= p(y^{(i)}; \phi_y)p(x^{(i)}|y^{(i)}; y) \\ &= p(y^{(i)}; \phi_y) \prod_{j=1}^{n_i} p(x_j^{(i)}|y^{(i)}; y) \end{aligned}$$

注意到 $p(y^{(i)})$ 中的 $y^{(i)}$ 既可能取 0 也可能取 1，所以 $p(y^{(i)})$ 有两个可能的值；同时，在 $y^{(i)}$ 已经取完定值 1 的时候，$p(x^{(i)}|y^{(i)} = 1)$ 由于 $x^{(i)}$ 等于或者不等 k, 所以 $p(x^{(i)}|y^{(i)} = 1)$ 也有两种可能的值。这样我们就可以把 $p(x^{(i)}|y^{(i)})$ 写成如下形式

$$\begin{aligned} p(x_j^{(i)}|y^{(i)}; y) &= p(x_j^{(i)}|y^{(i)} = 1)^{y^{(i)}} p(x_j^{(i)}|y^{(i)} = 0)^{1-y^{(i)}} \\ &= \underbrace{p(x_j^{(i)}|y = 1)^{y^{(i)}}}_{A} \underbrace{p(x_j^{(i)}|y = 0)^{1-y^{(i)}}}_{B} \end{aligned}$$

$$A = \left\{ \Big(p(x_j^{(i)} = k|y = 1)\Big)^{1\{x_j^{(i)}=k\}} \Big(1 - p(x_j^{(i)} = k|y = 1)\Big)^{1\{x_j^{(i)}\neq k\}} \right\}^{y^{(i)}}$$

注意到

$$1\{x_j^{(i)} = k\} = 1 - 1\{x_j^{(i)} \neq k\},$$

利用前面的记号，A 可以写为

$$\begin{aligned} A &= \left\{ \Big(p(x_j^{(i)} = k|y = 1)\Big)^{1\{x_j^{(i)}=k\}} \Big(1 - p(x_j^{(i)} = k|y = 1)\Big)^{1-1\{x_j^{(i)}=k\}} \right\}^{y^{(i)}} \\ &= \left\{ \phi_{k|y=1}^{1\{x_j^{(i)}=k\}} \Big(1 - \phi_{k|y=1}\Big)^{1-1\{x_j^{(i)}=k\}} \right\}^{y^{(i)}} \end{aligned}$$

同理，B 可以写为

$$\begin{aligned} B &= \left\{ \Big(p(x_j^{(i)} = k|y = 0)\Big)^{1\{x_j^{(i)}=k\}} \Big(1 - p(x_j^{(i)} = k|y = 0)\Big)^{1-1\{x_j^{(i)}=k\}} \right\}^{1-y^{(i)}} \\ &= \left\{ \phi_{k|y=0}^{1\{x_j^{(i)}=k\}} \Big(1 - \phi_{k|y=0}\Big)^{1-1\{x_j^{(i)}=k\}} \right\}^{1-y^{(i)}} \end{aligned}$$

我们再利用 ϕ_y 来表示 $p(y^{(i)}; \phi_y)$

$$p(y^{(i)}; \phi_y) = \phi_y^{y^{(i)}} (1 - \phi_y)^{1-y^{(i)}}.$$

基于上面的讨论我们就得到了似然函数 $L(\phi_y, \phi_{k|y=0}, \phi_{k|y=1})$：

$$\begin{aligned} L(\phi_y, \phi_{k|y=0}, \phi_{k|y=1}) &= \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) = \prod_{i=1}^{m} \Big(p(y^{(i)}; \phi_y) \prod_{j=1}^{n_i} p(x_j^{(i)}|y^{(i)}; y)\Big) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left\{ \phi_y \phi_{k|y=1}^{1\{x_j^{(i)}=k\}} \Big(1 - \phi_{k|y=1}\Big)^{1-1\{x_j^{(i)}=k\}} \right\}^{y^{(i)}} \left\{ (1 - \phi_y) \phi_{k|y=0}^{1\{x_j^{(i)}=k\}} \Big(1 - \phi_{k|y=0}\Big)^{1-1\{x_j^{(i)}=k\}} \right\}^{1-y^{(i)}} \end{aligned}$$

令 $\ell(\phi_y, \phi_{k|y=0}, \phi_{k|y=1}) = \log L$, 所以

$$\begin{aligned}\ell = \log L = &\sum_{i=1}^{m} y^{(i)} \log \phi_y + \sum_{i=1}^{m} (1 - y^{(i)}) \log(1 - \phi_y) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} y^{(i)} 1\{x_j^{(i)} = k\} \log \phi_{k|y=1} \\ &+ \sum_{i=1}^{m} \sum_{j=1}^{n_i} y^{(i)} (1 - 1\{x_j^{(i)} = k\}) \log(1 - \phi_{k|y=1}) \\ &+ \sum_{i=1}^{m} \sum_{j=1}^{n_i} (1 - y^{(i)}) 1\{x_j^{(i)} = k\} \log \phi_{k|y=0} \\ &+ \sum_{i=1}^{m} \sum_{j=1}^{n_i} (1 - y^{(i)}) (1 - 1\{x_j^{(i)} = k\}) \log(1 - \phi_{k|y=0}).\end{aligned}$$

我们关于似然函数 ℓ 分别对 $\phi_y, \phi_{k|y=0}, \phi_{k|y=1}$ 求导，并令其导数为 0 就可以得到三个参数的估计表达式。

$$0 = \frac{\partial \ell}{\partial \phi_y} = \frac{1}{\phi_y} \sum_{i=1}^{m} y^{(i)} - \frac{1}{1 - \phi_y} \sum_{i=1}^{m} (1 - y^{(i)}),$$

整理上式可得

$$\phi_y = \frac{1}{m} \sum_{i=1}^{m} y^{(i)},$$

同理，

$$0 = \frac{\partial \ell}{\partial \phi_{k|y=1}} = \frac{1}{\phi_{k|y=1}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} y^{(i)} 1\{x_j^{(i)} = k\} - \frac{1}{1 - \phi_{k|y=1}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} y^{(i)} (1 - 1\{x_j^{(i)} = k\}),$$

整理上式可得

$$\phi_{k|y=1} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} y^{(i)} 1\{x_j^{(i)} = k\}}{\sum_{i=1}^{m} n_i y^{(i)}},$$

再由 $\phi_{k|y=1}$ 与 $\phi_{k|y=0}$ 的对称性，立即可得

$$\phi_{k|y=0} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (1 - y^{(i)}) 1\{x_j^{(i)} = k\}}{\sum_{i=1}^{m} n_i (1 - y^{(i)})}.$$

关于上面的结果利用 indicate function 可以写为

$$\begin{aligned}
\phi_y &= \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)}=1\} \\
\phi_{k|y=0} &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{y^{(i)}=0\}1\{x_j^{(i)}=k\}}{\sum_{i=1}^{m} n_i 1\{y^{(i)}=0\}} \\
\phi_{k|y=1} &= \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{y^{(i)}=1\}1\{x_j^{(i)}=k\}}{\sum_{i=1}^{m} n_i 1\{y^{(i)}=1\}}.
\end{aligned}$$

REMARK 1. 注意到在 $\phi_{k|y=1}$ 中的 k 就是给定一个向量 x 时,x 的每一个分量所对应的数字。从本节开头例子来看，$k = 234, 1234, ..., 86320$。

对于给定的向量 x，现在我们可以来计算它的后验概率 $p(y=1|x)$, 即进行预测

$$\begin{aligned}
p(y=1|x) &= \frac{p(y=1)p(x|y=1)}{p(x)} \\
&= \frac{p(y=1)p(x|y=1)}{p(y=1)p(x|y=1)+p(y=0)p(x|y=0)} \\
&= \frac{p(y=1)\prod_{s=1}^{n_s} p(x_s|y=1)}{p(y=1)\prod_{s=1}^{n_s} p(x_s|y=1)+p(y=0)\prod_{s=1}^{n_s} p(x_s|y=0)} \\
&= \frac{\phi_y \prod_{s=1}^{n_s} \phi_{x_s|y=1}}{\phi_y \prod_{s=1}^{n_s} \phi_{x_s|y=1}+(1-\phi_y)\prod_{s=1}^{n_s} \phi_{x_s|y=0}}
\end{aligned}$$

2.2 Summary

1. 多项分布模型一般用于文本分类，它与朴素贝叶斯模型的区别是该模型考虑了一个文本中语词出现的频率，朴素贝叶斯模型只关注语词是否出现（出现 1，不出现 0）。

2. 最终的模型中的两个参数 $\phi_{k|y=0}, \phi_{k|y=1}$ 都需要进行 Laplace 光滑。