# 判别模型、生成模型与朴素贝叶斯方法

JerryLead
csxulijie@gmail.com
2011 年 3 月 5 日星期六

## 1 判别模型与生成模型

上篇报告中提到的回归模型是判别模型，也就是根据特征值来求结果的概率。形式化表示为$p(y|x;\theta)$，在参数$\theta$确定的情况下，求解条件概率$p(y|x)$。通俗的解释为在给定特征后预测结果出现的概率。

比如说要确定一只羊是山羊还是绵羊，用判别模型的方法是先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。换一种思路，我们可以根据山羊的特征首先学习出一个山羊模型，然后根据绵羊的特征学习出一个绵羊模型。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。形式化表示为求$p(x|y)$（也包括$p(y)$），y 是模型结果，x 是特征。

利用贝叶斯公式发现两个模型的统一性：

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

由于我们关注的是 y 的离散值结果中哪个概率大（比如山羊概率和绵羊概率哪个大），而并不是关心具体的概率，因此上式改写为：

$$\begin{aligned}\arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y).\end{aligned}$$

其中$p(x|y)$称为后验概率，$p(y)$称为先验概率。

由$p(x|y) * p(y) = p(x,y)$，因此有时称判别模型求的是条件概率，生成模型求的是联合概率。

常见的判别模型有线性回归、对数回归、线性判别分析、支持向量机、boosting、条件随机场、神经网络等。

常见的生产模型有隐马尔科夫模型、朴素贝叶斯模型、高斯混合模型、LDA、Restricted Boltzmann Machine 等。

这篇博客较为详细地介绍了两个模型：

http://blog.sciencenet.cn/home.php?mod=space&uid=248173&do=blog&id=227964

## 2 高斯判别分析（Gaussian discriminant analysis）

1） 多值正态分布

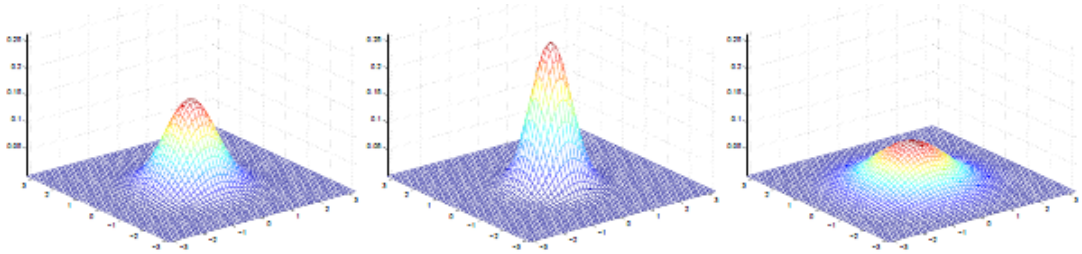
多变量正态分布描述的是 n 维随机变量的分布情况，这里的μ变成了向量，σ也变成了矩阵Σ。写作$N(\mu,\Sigma)$。假设有 n 个随机变量$X_1, X_2, \ldots, X_n$。μ的第 i 个分量是$E(X_i)$，而$\Sigma_{ii} = \mathrm{Var}(X_i)$，$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$。

概率密度函数如下：

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

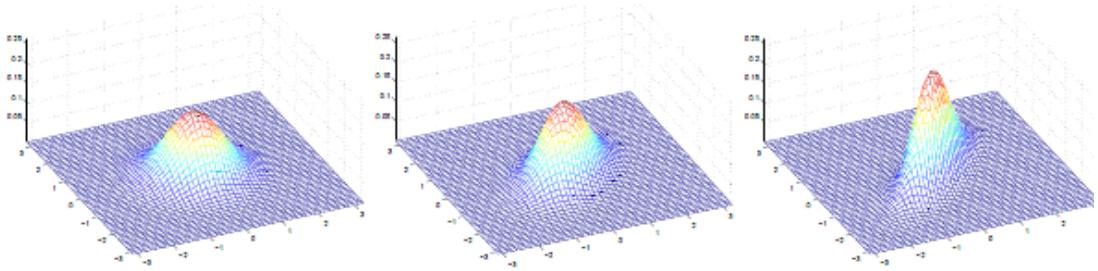
其中|Σ|是Σ的行列式，Σ是协方差矩阵，而且是对称半正定的。

当μ是二维的时候可以如下图表示：



其中μ决定中心位置，Σ决定投影椭圆的朝向和大小。

如下图：



The figures above show Gaussians with mean 0, and with covarianc matrices respectively

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

对应的Σ都不同。

2） 模型分析与应用

如果输入特征x是连续型随机变量，那么可以使用高斯判别分析模型来确定p(x|y)。

模型如下：

$$
\begin{aligned}
y &\sim \text{Bernoulli}(\phi) \\
x|y=0 &\sim \mathcal{N}(\mu_0, \Sigma) \\
x|y=1 &\sim \mathcal{N}(\mu_1, \Sigma)
\end{aligned}
$$

输出结果服从伯努利分布，在给定模型下特征符合多值高斯分布。通俗地讲，在山羊模型下，它的胡须长度，角大小，毛长度等连续型变量符合高斯分布，他们组成的特征向量符合多值高斯分布。
这样，可以给出概率密度函数：

$$
\begin{aligned}
p(y) &= \phi^y(1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)
\end{aligned}
$$

最大似然估计如下：

$$
\begin{aligned}
\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^{m} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi).
\end{aligned}
$$

注意这里的参数有两个μ，表示在不同的结果模型下，特征均值不同，但我们假设协方差相同。反映在图上就是不同模型中心位置不同，但形状相同。这样就可以用直线来进行分隔判别。

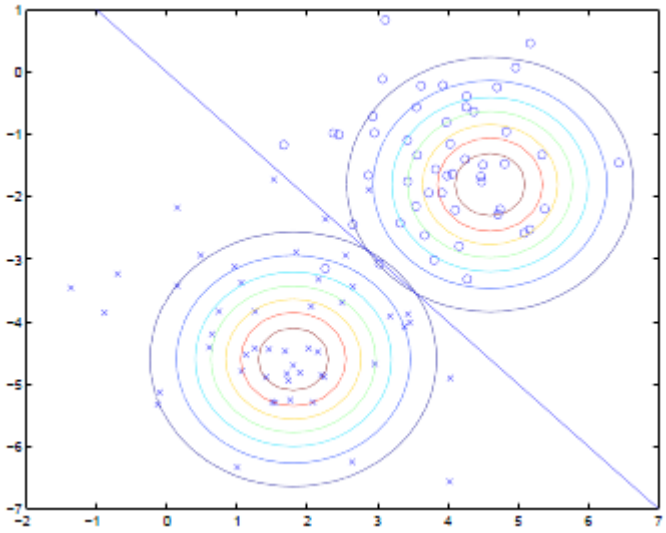
求导后，得到参数估计公式：

$$
\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)}=1\} \\
\mu_0 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}} \\
\mu_1 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m} (x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T.
\end{aligned}
$$

Φ是训练样本中结果 y=1 占有的比例。
$\mu_0$是 y=0 的样本中特征均值。

$\mu_1$是 y=1 的样本中特征均值。
Σ是样本特征方差均值。

如前面所述，在图上表示为：



直线两边的 y 值不同，但协方差矩阵相同，因此形状相同。μ不同，因此位置不同。

3） 高斯判别分析（GDA）与 logistic 回归的关系
将 GDA 用条件概率方式来表述的话，如下：

$$p(y = 1|x; \phi, \mu_0, \mu_1, \Sigma)$$

y 是 x 的函数，其中 $\phi, \mu_0, \mu_1, \Sigma$ 都是参数。
进一步推导出

$$p(y = 1|x; \phi, \Sigma, \mu_0, \mu_1) = \frac{1}{1 + \exp(-\theta^T x)},$$

这里的θ是 $\phi, \Sigma, \mu_0, \mu_1$ 的函数。
这个形式就是 logistic 回归的形式。
也就是说如果 p(x|y)符合多元高斯分布，那么 p(y|x)符合 logistic 回归模型。反之，不成立。为什么反过来不成立呢？因为 GDA 有着更强的假设条件和约束。

如果认定训练数据满足多元高斯分布，那么 GDA 能够在训练集上是最好的模型。然而，我们往往事先不知道训练数据满足什么样的分布，不能做很强的假设。Logistic 回归的条件假设要弱于 GDA，因此更多的时候采用 logistic 回归的方法。

例如，训练数据满足泊松分布，$x|y = 0 \sim \text{Poisson}(\lambda_0)$

$x|y=1 \sim \text{Poisson}(\lambda_1)$，那么 p(y|x)也是 logistic 回归的。这个时候如果采用 GDA，那么效果会比较差，因为训练数据特征的分布不是多元高斯分布，而是泊松分布。

这也是 logistic 回归用的更多的原因。

## 3 朴素贝叶斯模型

在 GDA 中，我们要求特征向量 x 是连续实数向量。如果 x 是离散值的话，可以考虑采用朴素贝叶斯的分类方法。

假如要分类垃圾邮件和正常邮件。分类邮件是文本分类的一种应用。

假设采用最简单的特征描述方法，首先找一部英语词典，将里面的单词全部列出来。然后将每封邮件表示成一个向量，向量中每一维都是字典中的一个词的 0/1 值，1 表示该词在邮件中出现，0 表示未出现。

比如一封邮件中出现了“a”和“buy”，没有出现“aardvark”、“aardwolf”和“zygmurgy”，那么可以形式化表示为：

$$x=\begin{bmatrix}1\\0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix}\begin{matrix}\text{a}\\\text{aardvark}\\\text{aardwolf}\\\vdots\\\text{buy}\\\vdots\\\text{zygmurgy}\end{matrix}$$

假设字典中总共有 5000 个词，那么 x 是 5000 维的。这时候如果要建立多项式分布模型（二项分布的扩展）。

多项式分布（multinomial distribution）

某随机实验如果有 k 个可能结局 A1，A2，…，Ak，它们的概率分布分别是 p1，p2，…，pk，那么在 N 次采样的总结果中，A1 出现 n1 次，A2 出现 n2 次，…，Ak 出现 nk 次的这种事件的出现概率 P 有下面公式：（Xi 代表出现 ni 次）

$$P(X_1=x_1,\ldots,X_k=x_k)=\begin{cases}\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k} & \text{when } \sum_{i=1}^{k}x_i=n\\0 & \text{otherwise.}\end{cases}$$

对应到上面的问题上来，把每封邮件当做一次随机试验，那么结果的可能性有$2^{5000}$种。意味着 pi 有$2^{5000}$个，参数太多，不可能用来建模。

换种思路，我们要求的是 p(y|x)，根据生成模型定义我们可以求 p(x|y)和 p(y)。假设 x 中的特征是条件独立的。这个称作朴素贝叶斯假设。如果一封邮件是垃圾邮件（y=1），且这封邮件出现词“buy”与这封邮件是否出现“price”无关，那么“buy”和“price”之间是条件独立的。

形式化表示为，（如果给定 Z 的情况下，X 和 Y 条件独立）：

$P(X|Z) = P(X|Y,Z)$

也可以表示为：

$P(X,Y|Z) = P(X|Z)P(Y|Z)$

回到问题中

$$\begin{aligned} & p(x_1,\ldots,x_{50000}|y) \\ &= p(x_1|y)p(x_2|y,x_1)p(x_3|y,x_1,x_2)\cdots p(x_{50000}|y,x_1,\ldots,x_{49999}) \\ &= p(x_1|y)p(x_2|y)p(x_3|y)\cdots p(x_{50000}|y) \\ &= \prod_{i=1}^{n} p(x_i|y) \end{aligned}$$

这个与 NLP 中的 n 元语法模型有点类似，这里相当于 unigram。

这里我们发现朴素贝叶斯假设是约束性很强的假设，“buy”从通常上讲与“price”是有关系，我们这里假设的是条件独立。（注意条件独立和独立是不一样的）

建立形式化的模型表示：

$\phi_{i|y=1} = p(x_i = 1|y = 1)$

$\phi_{i|y=0} = p(x_i = 0|y = 1)$

$\phi_{y} = p(y = 1)$

那么我们想要的是模型在训练数据上概率积能够最大，即最大似然估计如下：

$$\mathcal{L}(\phi_y, \phi_{i|y=0}, \phi_{i|y=1}) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}).$$

注意这里是联合概率分布积最大，说明朴素贝叶斯是生成模型。

求解得：

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

最后一个式子是表示 y=1 的样本数占全部样本数的比例，前两个表示在 y=1 或 0 的样本中，特征 Xj=1 的比例。

然而我们要求的是

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\
&= \frac{\left(\prod_{i=1}^{n} p(x_i|y=1)\right)p(y=1)}{\left(\prod_{i=1}^{n} p(x_i|y=1)\right)p(y=1)+\left(\prod_{i=1}^{n} p(x_i|y=0)\right)p(y=0)},
\end{aligned}
$$

实际是求出分子即可，分母对 y=1 和 y=0 都一样。

当然，朴素贝叶斯方法可以扩展到 x 和 y 都有多个离散值的情况。对于特征是连续值的情况，我们也可以采用分段的方法来将连续值转化为离散值。具体怎么转化能够最优，我们可以采用信息增益的度量方法来确定（参见 Mitchell 的《机器学习》决策树那一章）。比如房子大小可以如下划分成离散值：

| Living area (sq. feet) | $< 400$ | 400-800 | 800-1200 | 1200-1600 | $>1600$ |
|---|---|---|---|---|---|
| $x_i$ | 1 | 2 | 3 | 4 | 5 |

## 4 拉普拉斯平滑

朴素贝叶斯方法有个致命的缺点就是对数据稀疏问题过于敏感。

比如前面提到的邮件分类，现在新来了一封邮件，邮件标题是“NIPS call for papers”。我们使用更大的网络词典（词的数目由 5000 变为 35000）来分类，假设 NIPS 这个词在字典中的位置是 35000。然而 NIPS 这个词没有在训练数据中出现过，这封邮件第一次出现了 NIPS。那我们算概率的时候如下：

$$
\begin{aligned}
\phi_{35000|y=1} &= \frac{\sum_{i=1}^{m} 1\{x_{35000}^{(i)}=1 \wedge y^{(i)}=1\}}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}} = 0 \\
\phi_{35000|y=0} &= \frac{\sum_{i=1}^{m} 1\{x_{35000}^{(i)}=1 \wedge y^{(i)}=0\}}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}} = 0
\end{aligned}
$$

由于 NIPS 在以前的不管是垃圾邮件还是正常邮件都没出现过，那么结果只能是 0 了。显然最终的条件概率也是 0。

$$
\begin{aligned}
p(y=1|x) &= \frac{\prod_{i=1}^{n} p(x_i|y=1)p(y=1)}{\prod_{i=1}^{n} p(x_i|y=1)p(y=1)+\prod_{i=1}^{n} p(x_i|y=0)p(y=0)} \\
&= \frac{0}{0}.
\end{aligned}
$$

原因就是我们的特征概率条件独立，使用的是相乘的方式来得到结果。

为了解决这个问题，我们打算给未出现特征值，赋予一个“小”的值而不是 0。

具体平滑方法如下：

假设离散型随机变量 z 有{1,2,…,k}个值，我们用$\Phi_i$ = p(z = i)来表示每个值的概率。假设有 m 个训练样本中，z 的观察值是$\{z^{(1)},\ldots,z^{(m)}\}$，其中每一个观察值对应 k 个值中的一个。那么根据原来的估计方法可以得到

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}{m}.$$

说白了就是 z=j 出现的比例。

拉普拉斯平滑法将每个 k 值出现次数事先都加 1，通俗讲就是假设他们都出现过一次。

那么修改后的表达式为：

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} + 1}{m + k}.$$

每个 z=j 的分子都加 1，分母加 k。可见 $\sum_{j=1}^{k} \phi_j = 1$ 。

这个有点像 NLP 里面的加一平滑法，当然还有 n 多平滑法了，这里不再详述。

回到邮件分类的问题，修改后的公式为：

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} + 2}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} + 2}$$

## 5 文本分类的事件模型

回想一下我们刚刚使用的用于文本分类的朴素贝叶斯模型，这个模型称作多值伯努利事件模型（multi-variate Bernoulli event model）。在这个模型中，我们首先随机选定了邮件的类型（垃圾或者普通邮件，也就是 p(y)），然后一个人翻阅词典，从第一个词到最后一个词，随机决定一个词是否要在邮件中出现，出现标示为 1，否则标示为 0。然后将出现的词组成一封邮件。决定一个词是否出现依照概率 p(xi|y)。那么这封邮件的概率可以标示为 $p(y)\prod_{i=1}^{n} p(x_i|y)$。

让我们换一个思路，这次我们不先从词典入手，而是选择从邮件入手。让 i 表示邮件中的第 i 个词，xi 表示这个词在字典中的位置，那么 xi 取值范围为{1,2,…|V|}，|V|是字典中词的数目。这样一封邮件可以表示成 $(x_1, x_2, \ldots, x_n)$，n 可以变化，因为每封邮件的词的个数不同。然后我们对于每个 xi 随机从|V|个值中取一个，这样就形成了一封邮件。这相当于重复投掷|V|面的骰子，将观察值记录下来就形成了一封邮件。当然每个面的概率服从 p(xi|y)，而且每次试验条件独立。这样我们得到的邮件概率是 $p(y)\prod_{i=1}^{n} p(x_i|y)$。居然跟上面的一样，那么不同点在哪呢？注意第一个的 n 是字典中的全部的词，下面这个 n 是邮件中的词个数。上面 xi 表示一个词是否出现，只有 0 和 1 两个值，两者概率和为 1。下面的 xi 表示|V|中的一个值，|V|个 p(xi|y)相加和为 1。是多值二项分布模型。上面的 x 向量都是 0/1 值，下面的 x 的向量都是字典中的位置。

形式化表示为：

m 个训练样本表示为：$\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$

$$x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)})$$

表示第 i 个样本中，共有 ni 个词，每个词在字典中的编号为$x_j^{(i)}$。

那么我们仍然按照朴素贝叶斯的方法求得最大似然估计概率为

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

解得，

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

与以前的式子相比，分母多了个 ni，分子由 0/1 变成了 k。

举个例子：

| X1 | X2 | X3 | Y |
|---|---|---|---|
| 1 | 2 | - | 1 |
| 2 | 1 | - | 0 |
| 1 | 3 | 2 | 0 |
| 3 | 3 | 3 | 1 |

假如邮件中只有 a，b，c 这三词，他们在词典的位置分别是 1,2,3，前两封邮件都只有 2 个词，后两封有 3 个词。

Y=1 是垃圾邮件。

那么，

$$\Phi_{1|y=1} = \frac{1+0}{2+3} = \frac{1}{5}, \quad \Phi_{2|y=1} = \frac{1}{5}, \quad \Phi_{3|y=1} = \frac{3}{5}$$

$$\Phi_{1|y=0} = \frac{2+0}{2+3} = \frac{2}{5}, \quad \Phi_{2|y=0} = \frac{2}{5}, \quad \Phi_{3|y=0} = \frac{1}{5}$$

$$\Phi_{y=1} = \frac{1}{2}, \quad \Phi_{y=0} = \frac{1}{2}$$

假如新来一封邮件为 b，c 那么特征表示为{2,3}。

那么

$$P(y=1|x)=\frac{p(x,y=1)}{p(x)}=\frac{p(x=\{2,3\}|y=1)p(y=1)}{p(x=\{2,3\})}$$
$$=\frac{\Phi_{2|y=1}\Phi_{3|y=1}\Phi_{y=1}}{\Phi_{2|y=1}\Phi_{3|y=1}\Phi_{y=1}+\Phi_{2|y=0}\Phi_{3|y=0}\Phi_{y=0}}$$
$$=\frac{0.2*0.6*0.5}{0.2*0.6*0.5+0.4*0.2*0.5}=0.6$$

那么该邮件是垃圾邮件概率是 0.6。

注意这个公式与朴素贝叶斯的不同在于这里针对整体样本求的$\Phi_{k|y=1}$，而朴素贝叶斯里面针对每个特征求的$\Phi_{xj=1|y=1}$，而且这里的特征值维度是参差不齐的。

这里如果假如拉普拉斯平滑，得到公式为：

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

表示每个 k 值至少发生过一次。

另外朴素贝叶斯虽然有时候不是最好的分类方法，但它简单有效，而且速度快。