JerryLead

All things are difficult before they are easy.

判别模型、生成模型与朴素贝叶斯方法

转载时请注明来源：http://www.cnblogs.com/jerrylead

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θ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X1,X2,…,Xn。μ的第i个分量是E(Xi)，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$|\Sigma|$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：

公告

Contact me via

csxulijie@Gmail.com

Save page as PDF

昵称：JerryLead
园龄：4年9个月
粉丝：1502
关注：4
+加关注

导航

博客园
首页
新随笔
联系
订阅XML
管理

< 2011年3月 >

日	一	二	三	四	五	六
27	28	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

统计

随笔 - 27
文章 - 0
评论 - 354
引用 - 0

搜索

找找看

谷歌搜索

常用链接

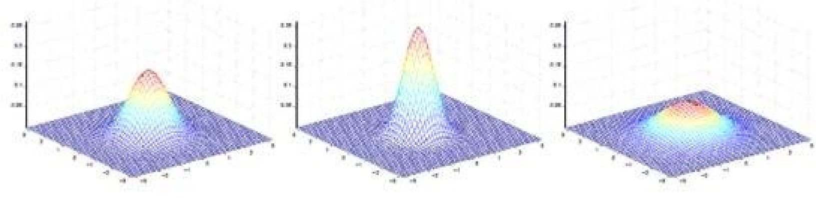
我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签

Machine Learning(22)
Big Data(3)
Maths(1)

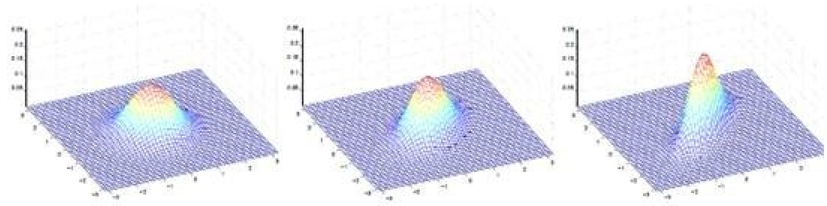
随笔档案(27)

2013年4月 (1)
2012年8月 (2)
2012年5月 (1)
2011年8月 (1)
2011年6月 (1)
2011年5月 (2)
2011年4月 (10)



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：

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

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：

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

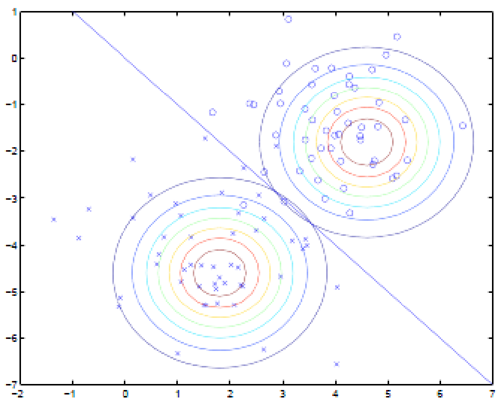
Φ是训练样本中结果y=1占有的比例。

μ_0是y=0的样本中特征均值。

μ_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：

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

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：

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

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

$\phi_y = p(y = 1)$

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。

Technorati 标签: Machine Learning

回到邮件分类的问题，修改后的公式为：

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

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
1	2	-
2	1	-
1	3	2
3	3	3

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，

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

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

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

假如新来一封邮件为b，c那么特征表示为{2,3}。

那么

$$
\begin{aligned}
\mathrm{P}(y=1|x) = \frac{p(x,y=1)}{p(x)} &= \frac{p(x=\{2,3\}|y=1)p(y=1)}{p(x=\{2,3\})} \\
&= \frac{\Phi_{2|y=1}\Phi_{3|y=1}\Phi_{y=1}}{\Phi_{2|y=1}\Phi_{3|y=1}\Phi_{y=1} + \Phi_{2|y=0}\Phi_{3|y=0}\Phi_{y=0}}
\end{aligned}
$$

$$= \frac{0.2*0.6*0.5}{0.2*0.6*0.5+0.4*0.2*0.5} = 0.6$$

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：

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

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。

标签：Machine Learning

« 上一篇：对线性回归，logistic回归和一般回归的认识
» 下一篇：小谈导数、梯度和极值

评论

#1楼 2011-03-08 20:25 schindlerlee

赞!
好!

支持(0) 反对(0)

#2楼 2011-09-27 13:24 第二视点[未注册用户]

很好，了解了一下拉普拉斯平滑

#3楼 2012-03-17 14:38 mz31098

后验概率那个公式是不是贴错了？
后面那个例子，数据太不完整了"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。"

后两封有3个词，怎么也没看出来，而且没有关于y的信息。

支持(0) 反对(0)

#4楼 2012-08-12 11:15 lvpengdlut

请问博主插图是从哪得到的？是不是有什么讲义啊

支持(0) 反对(0)

#5楼 2012-08-15 10:45 duanchw37

@lvpengdlut
有Andrew Ng的课件，网易公开课上也有相应的视频，我正在学习之！！

支持(0) 反对(0)

#6楼 2012-08-30 01:27 victor0535

"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 "

这段啥意思啊？？？ 怎么感觉跟上面介绍的朴素贝叶斯不一样？？

我觉得你理解误，"随机决定一个词是否要在邮件中出现"，怎么能随机呢， 而是根据用训练时选取的"词汇向量" 比对该邮件，出现的了词 相应的分量记为1 不出现记为0 最后计算概率时， 不能只乘出现的概率，而不出现的也要乘 即（1－p） p为训练时取得的出现该词的概率 ......

如果有问题 望继续探讨~~~~~

支持(1) 反对(0)

#7楼 2014-03-27 10:47 CorrectAdv

写的好不错

支持(0) 反对(0)

#8楼 2014-04-17 15:30 yazifei

楼主，您好 。关于最后那个例子有点疑问
1.y的标签在你的表格中没体现出来，所以有点没看懂
2.文本事件模型中，比如词A在文档里出现了很多次，选其位置作为特征的时候是随机选其中的一个位置？，那么是不是有很多种可能，然后会生成不同特征，这对分类结果是不是有影响？

支持(0) 反对(0)

#9楼 2014-04-21 20:31 紫梦lan

我想问下：为什么GDA中多模型的协方差矩阵是一样的？

支持(0) 反对(0)

#10楼 2014-08-19 09:36 keyalone

@紫梦lan
你也可以不使用一样的啊！只不过是为了拟合方便简单点而已。guassian判别法要求数据量比较大。

支持(0) 反对(0)

#11楼 2014-10-29 17:17 程序员阿力

拉普拉斯平滑中，最后一个公式，分母为什么+2
按照每种可能都加1的话，分母应该+k啊

支持(0) 反对(0)

#12楼 **2015-03-16 17:12** **rushshi**

@keyalone
也可以理解为对于X来说，不管Y是多少，输入特征X的相关性认为一致的。

支持(0) 反对(0)

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。

最新**IT**新闻:
· 作为员工，如何识别初创企业健康状况
· 如何招到靠谱的产品经理？
· 创业跟风者的15项特征
· 在网上没人知道你是一条狗的时候，你会怎么做？
· 看完豆瓣读书这份年度榜单，才知道今年错过了多少好书
» 更多新闻...

最新知识库文章:
· Git协作流程
· 企业计算的终结
· 软件开发的核心
· Linux概念架构的理解
· 从涂鸦到发布——理解API的设计过程
» 更多知识库文章...