朴素贝叶斯模型详解

李文哲

贪心学院，*NLP*训练营
wenzheli@usc.edu

Abstract

朴素贝叶斯模型是文本分析领域最为常用的模型之一，也是最为经典的模型。文本主要从教学的角度来讲解朴素贝叶斯模型以及数学原理。为了让文档具备完备性，必要的前置知识也包含在文章里。

Keywords: 朴素贝叶斯模型, 数学推导

1. 预备数学知识

1.1. 求极值问题

人工智能中最核心的数学环节是求出一个目标函数（object function）的最小值/最大值。求出一个函数最小是/最大值的方法很多，在这里我们介绍一个最经典的方法之一：直接求出极值点。这些极值点的共同特点是在这些点上的梯度为0, 如下图所示 1。这个图里面，有8个极值点，而且这些极值点中必然会存在最小值或者最大值（除去函数的左右最端点）。所以在这种方式下，我们通常x先求出函数的导数，然后设置成为0。之后从找出的极值点中选择使得结果为最小值/最大值的极值点。

例*1*：求$f(x) = x^2 - 2x - 3$的最小值. 对于这样的一个问题，其实我们都知道这个问题的答案是$x = 1$，基本上不需要计算就能看出来。接下来我们通过求导的方式来求一下最小值。首先对$f(x)$求导并把导数设置成0。

$$f'(x) = 2x - 2 = 0$$

从而得到$x = 1$, 求出来的是唯一的极值点，所以最后得出来的函数的最小值是$f(1) = 1 - 2 - 3 = -4$

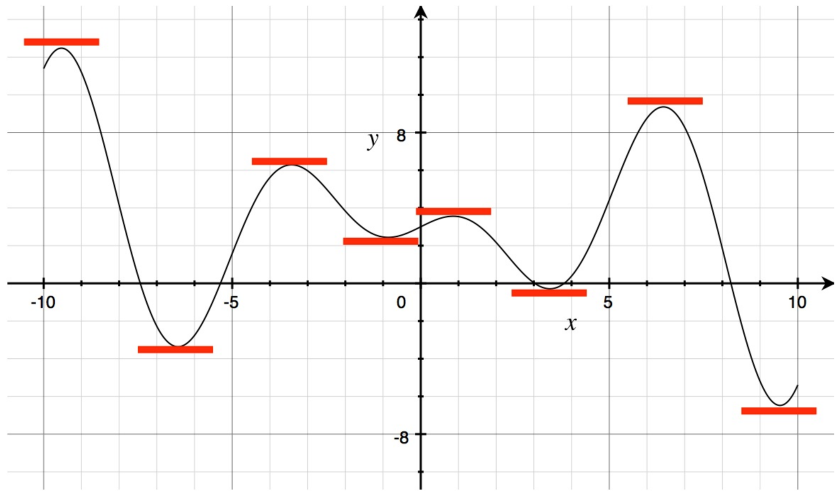

Figure 1: 函数图(包含8个极值点)

例*2*: 求$f(x)=x^4-3x^2+4$的最小值.

$$f'(x)=4x^3-6x=0$$

即可以得到$x_1=0,x_2=\sqrt{(\frac{3}{2})},x_3=-\sqrt{(\frac{3}{2})}$. 把这三个值分别带入到$f(x)$即可以得到$f(0)=4,f(\sqrt{(\frac{3}{2})})=\frac{9}{4}-\frac{9}{2}+4=\frac{7}{4},f(-\sqrt{(\frac{3}{2})})=\frac{7}{4}$。所以，$x_2,x_3$两个点都可以作为函数的最小值点，这时候函数值为$\frac{7}{4}$

*请注意：并不一定所有函数的极值都可以通过设置导数为0的方式求出。也就是说，有些问题中当我们设定导数为0时，未必能直接计算出满足导数为0的点（比如逻辑回归模型），这时候就需要利用数值计算相关的技术（最典型为梯度下降法，牛顿法..)

1.2. 拉格朗日乘法项*(Lagrangian Multiplier)*

对于某些求极值问题，函数通常带有一些条件。如下面的例子：

例*3*: 求$f(x,y)=x+y$的最大值，但有个条件是$x^2+y^2=1$. 。那这时候怎么求出最大值呢?

拉格朗日乘法项就是用来解决这类问题。我们可以把限制条件通过简单的转变加到目标函数中，这时候问题就变成了

$$maximize \quad L=x+y+\lambda(x^2+y^2-1)$$

剩下的过程就跟上面的类似了。设定导数为0，即可以得到以下三个方程：

$$f'_x(x,y,\lambda) = 1 + 2\lambda x = 0 \tag{1}$$
$$f'_y(x,y,\lambda) = 1 + 2\lambda y = 0 \tag{2}$$
$$f'_\lambda(x,y,\lambda) = x^2 + y^2 - 1 = 0 \tag{3}$$

解完之后即可以得到$\lambda_1 = \frac{\sqrt{2}}{2}, \lambda_2 = -\frac{\sqrt{2}}{2}$。针对于每一个$\lambda$我们得到的解为$(x = \frac{\sqrt{2}}{2}, y = \frac{\sqrt{2}}{2}), (x = -\frac{\sqrt{2}}{2}, y = -\frac{\sqrt{2}}{2})$。把两个解带入到原来函数里并做比较即可以确定最优解。

1.3. 最大似然估计*(Maximum Likelihood Estimation)*

最大似然估计是机器学习领域最为常见的用来构建目标函数的方法。它的核心思想是根据观测到的结果来预测其中的未知参数。我们举一个投掷硬币的例子。

假设有一枚硬币，它是不均匀的，也就是说出现正面的反面的概率是不同的。假设我们设定这枚硬币出现正面的概率为$p(H) = \theta$, 这里H指的是正面(head), 类似的还会有反面（tail）。假设我们投掷6次之后得到了以下的结果，而且我们假定每次投掷都是相互独立的事件：

$$D = \{H, T, T, H, H, H\}$$

其中D表示所观测到的所有样本。从这个结果其实谁都可以很容易说出θ，也就是出现正面的概率为$4/6$，其实我们在无意识中使用了最大似然估计法。接下来，我们从更严谨的角度来定义最大似然下的目标函数。

基于最大似然估计法，我们需要最大化观测到样本的概率，即$p(D)$。进一步可以写成：

$$p(D) = p(HTTHHH) = \theta * (1-\theta) * (1-\theta) * \theta * \theta * \theta$$

我们的目标是最大化概率值$L = \theta * (1-\theta) * (1-\theta) * \theta * \theta * \theta$。那这部分的优化即可以采用上面所提到的方法。

$$L'(\theta) = 4\theta^3(1-\theta)^2 - \theta^4 * 2 * (1-\theta) = 0$$

把这个式子整理完之后即可以得到$\theta = 2/3$, 结果跟一开始我们算出来的是一致的。

2. 朴素贝叶斯(Naive Bayes)

假设给定了一批训练数据$D = \{(x^1, y^1), ..., (x^N, y^N)\}$，其中$x^i$指的是第$i$个样本（文章），而且这个样本（文章）包含了$m_i$个单词，所以可以把$x^i$表示成$x^i = (x_1^i, x_2^i, ..., x_{m_i}^i)$. 假设$x^i$的全文内容为"今天很高兴来参加自然语言处理训练营"，那这时候x_1^i ="今天"，x_2^i="很"，x_3^i="高兴"，x_4^i="来",x_5^i="参加",x_6^i="自然语言处理",x_7^i="训练营"。所以这里$m_i = 7$。另外，y^i代表的是x^i的标签。比如在垃圾邮件应用上，它代表"垃圾邮件"，"正常邮件"。我们利用K来代表分类的个数。比如在垃圾邮件应用上$K = 2$。但在这里，我们考虑普遍的情况，很有可能$K > 2$，也称之为是多分类问题。最后，我们再定义V为词典库的大小。

朴素贝叶斯是生成模型，它的目标是要最大化概率$p(D)$，也就是$p(x, y)$。我们把前几步的推导先写一下:

$$
\begin{aligned}
p(D) &= \prod_{i=1}^{N} p(x^i, y^i) = \prod_{i=1}^{N} p(x^i|y^i)p(y^i) &(4)\\
&= \prod_{i=1}^{N} p(x_1^i, x_2^i, ..., x_{m_i}^i|y^i)p(y^i) &(5)\\
&= \prod_{i=1}^{N}\prod_{j=1}^{m_i} p(x_j^i|y^i)p(y^i) &(6)
\end{aligned}
$$

这里简单说明一下：式子(4)是利用的贝叶斯公式，式子(4)到(5)的变化是利用了一个事实：样本x^i由很多个单词来构成，这里每一个x_j^i看作是一个单词。式子(5)到(6)是利用了朴素贝叶斯的假设，也就是每个词都是相互独立的。比如给定一个句子"我们今天运动"，在给定一个标签y的情况下，概率可以写成$p($"我们今天运动"$|y) = p($我们$|y) * p($今天$|y) * p($运动$|y)$。

我们看到式子里面都是乘法项，而且多个乘法项很容易引起数据的overflow或则underflow(在这里是underflow)。所以我们一般不直接最大化$p(D)$,而是最大化$\log p(D)$, 其实这两个是等同的。因为log函数是严格递

增的函数。加上log，我们对上面式子做一些变化：

$$
\begin{aligned}
\log p(D) &= \log(\prod_{i=1}^{N}\prod_{j=1}^{m_i} p(x_j^i|y^i)p(y^i)) &(7)\\
&= \log(\prod_{i=1}^{N}\prod_{j=1}^{V} p(w_j|y^i)^{n_{ij}}p(y^i)) &(8)\\
&= \sum_{i=1}^{N}\sum_{j=1}^{V} n_{ij}\log p(w_j|y^i) + \sum_{i=1}^{N}\log p(y^i) &(9)\\
&= \sum_{k=1}^{K}\sum_{i:y^i=k}\sum_{j=1}^{V} n_{ij}\log p(w_j|y^i=k) + \sum_{k=1}^{K}\sum_{i:y^i=k}\log p(y^i=k) &(10)\\
&= \sum_{k=1}^{K}\sum_{i:y^i=k}\sum_{j=1}^{V} n_{ij}\log \Theta_{kj} + \sum_{k=1}^{K}\sum_{i:y^i=k}\log \pi_k &(11)\\
&= \sum_{k=1}^{K}\sum_{i:y^i=k}\sum_{j=1}^{V} n_{ij}\log \Theta_{kj} + \sum_{k=1}^{K} n_k\log \pi_k &(12)
\end{aligned}
$$

从式子(7)到式子(8)的转化利用了一些技巧。举例子，假设一个文章的内容为”我们 天 天 运动”，按照之前的逻辑这句话的概率为$p($”我们天天运动$|y) = p($我们$|y)p($天$|y)p($天$|y)p($运动$|y)$。但另一方面，我们也可以利用词典库的所有的词来代表这个概率。$p($我们天天运动$|y) = p($”啊”$|y)^0 p($哎$|y)^0....p($天$|y)^2$ $...p($我们$|y)^1...p($运动$|y)^1.....$ 这两者是等同的，只不过我们从词典库的维度把所有的单词都考虑了进来，并数一下每一个单词在文档i里出现了多少次，如果没有出现，相当于0次。所以从这个角度我们可以把$\prod_{j=1}^{m_i} p(x_j^i|y^i)$写成$\prod_{j=1}^{V} p(w_j|y^i)^{n_{ij}}$,其中$n_{ij}$代表单词$j$出现在文档$i$的次数。这里$w_j$代表词典库里的第$j$个单词。

从式子(8)到(9)是利用了log的性质。比如$\log x^y = y\log x, \log\prod_{i=1} p(x^i) = \sum_{i=1}\log p(x^i)$。式子(9)到式子(10)是把文档按照类别做了一个分类。也就是，一开始的时候我们是从文档1到N的顺序来循环，但现在我们先取出类别为1的文档，然后再取出类别为2的文档，以此类推。所以前面的$\sum_{i=1}^{N}$被拆分成两个sum，即$\sum_{k=1}^{K}\sum_{i:y^i=k}$。这里$i : y^i = k$代表属于类别k的所有文档。式子(10)到式子(11)是引入了两组变量。也就是我们直接设置$p(w_j|y^i = k)$为Θ_{kj},意思就是当文章分类为k的时候出现单词w_j的概率（Θ_{kj}）。另外，我们设定$p(y^i = k)$为π_k，也就是文档属于第k类的概率，这个也是朴素贝叶斯模型的先验(prior)。比如在垃圾识别应用中，

假设总共有100个垃圾邮件和1000个正常邮件，这时候一个邮件为垃圾邮件的概率为1/11，正常邮件的概率为10/11，这就是贝叶斯模型的先验，而且所有之和等于1。式子(11)到式子(12)是引入一个新的变量叫做n_k，也就是属类别k的文件个数（训练数据总统计即可，是已知的）。也就是$\sum_{i:y^i=k}\pi_k = n_k\pi_k$。

另外，有两个约束条件需要满足，分别是：

$$\sum_{u=1}^{K}\pi_u = 1 \tag{13}$$

$$\sum_{v=1}^{V}\Theta_{kv} = 1, \text{for} \quad k = 1, 2, ...K \tag{14}$$

条件(13)表示的是所有类别的概率加在一起等于1. 比如p(垃圾)$+p$(正常)$=1$。条件(14)表示的是对于任意一个分类k, 出现所有词典里的单词的总概率合为1.

所以，这个问题是有约束条件的优化问题。把约束条件和目标函数写在一起即可以得到:

$$\text{Maximize} \quad L = \sum_{k=1}^{K}\sum_{i:y^i=k}\sum_{j=1}^{V} n_{ij}\log\Theta_{kj} + \sum_{k=1}^{K} n_k \log\pi_k$$

$$s.t. \sum_{u=1}^{K}\pi_u = 1 \tag{15}$$

$$\sum_{v=1}^{V}\Theta_{kv} = 1, \text{for} \quad k = 1, 2, ...K \tag{16}$$

利用拉格朗日乘法项，我们可以把目标函数写成:

$$\text{Maximize} \quad L = \sum_{k=1}^{K}\sum_{i:y^i=k}\sum_{j=1}^{V} n_{ij}\log\Theta_{kj} + \sum_{k=1}^{K} n_k \log\pi_k + \lambda(\sum_{u=1}^{K}\pi_u - 1) + \sum_{k=1}^{K}\lambda_k(\sum_{v=1}^{V}\Theta_{kv} - 1) \tag{17}$$

2.0.1. 找出最优解π_k

我们需要设置导数为0，进而找出最优解。现在求解的是π_k，所以只要

跟π_k无关的项我们都可以不用考虑，因为它们导数为0. L对π_k的导数为：

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial (n_k \log \pi_k + \lambda \pi_k)}{\partial \pi_k} \tag{18}$$

$$= \frac{n_k}{\pi_k} + \lambda = 0 \tag{19}$$

解为$\pi_k = -\frac{1}{\lambda} n_k$。这里有个参数$\lambda$，但同时我们有一个约束条件是$\sum_{u=1}^{K} \pi_u = 1$, 我们把刚才的解带入到这里，可以得到：

$$\sum_{u=1}^{K} -\frac{1}{\lambda} n_u = 1 \tag{20}$$

则λ的值为$\lambda = -\sum_{u=1}^{K} n_u$。把$\lambda$再次带入到$\pi_k$里面，我们可以得到：

$$\pi_k = \frac{n_k}{\sum_{u=1}^{K} \pi_u} \tag{21}$$

这里面n_k为k类文档出现的个数，分母为所有文档的个数。

2.1. 找出最优解Θ_{kj}

类似的，我们需要求L对Θ_{kj}的导数为：

$$\frac{\partial L}{\partial \Theta_{kj}} = \frac{\partial (\sum_{i:y^i=k} n_{ij} \log \Theta_{kj}) + \lambda_k (\sum_{v=1}^{V} \Theta_{kv} - 1)}{\partial \Theta_{kj}} \tag{22}$$

$$= \sum_{i:y^i=k} \frac{n_{ij}}{\Theta_{kj}} + \lambda_k = 0 \tag{23}$$

解为$\Theta_{kj} = -\frac{1}{\lambda_k} \sum_{i:y^i=k} n_{ij}$。这里有个参数是$\lambda_k$，但同时我们有个约束条件是$\sum_{v=1} \Theta_{kv} = 1$，把刚才的解带入到这里，可以得到：

$$\lambda_k = -\frac{1}{\sum_{v=1}^{V} \sum_{i:y^i=k} n_{iv}} \tag{24}$$

把λ_k带入到上面的解即可以得到：

$$\Theta_{kj} = \frac{\sum_{i:y^i=k} n_{ij}}{\sum_{v=1}^{V} \sum_{i:y^i=k} n_{iv}} \tag{25}$$

这个式子中分子代表的是在所有类别为k的文档里出现了多少次w_j,也就是词典库里的j个单词。分母代表的是在类别为k的所有文档里包含了总共多少个单词。

到此为止，模型的参数已经得到。如有发现错误，请联系作者。感谢阅读。