【NLP】经典分类模型朴素贝叶斯解读

原创 小Dream哥 有三AI 2019-07-02

> 贝叶斯分类器在早期的自然语言处理任务中有着较多实际的应用，例如大部分的垃圾邮件处理都是用的贝叶斯分类器。贝叶斯分类器的理论对于理解后续的NLP模型有很大的进益，感兴趣的小伙伴一定要好好看看，本文会详细的讲述贝叶斯分类器的原理。
>
> 本文会是我们NLP基础系列最后一篇机器学习模型的讲解，后面会进入深度学习相关的内容。
>
> 作者&编辑 | 小Dream哥

1 贝叶斯决策论

贝叶斯决策论是在统计概率框架下进行分类决策的基本方法。对于分类任务来说，在**所有相关概率**都已知的情况下，贝叶斯决策论考虑如何基于这些概率和误判损失来预测分类。

假设在一个分类任务中，有N种可能的分类，y={c1,c2,c3,...,cN}。我们会这样定义将一个样本预测为ci的期望损失，又叫“条件风险”：

$$R(c_i \mid x) = \sum_{j=1}^{N} \lambda_{i,j} P(c_j \mid x)$$

```
1、其中lambda_i_j，是将一个第j类样本预测为i类的损失
2、P(c_j|x)表示为将样本x预测为j类的概率
```

那么学习的任务是什么呢？

学习任务是寻找一个判定准则，利用该判定准则（分类器）进行分类预测，能够最小化条件风险：

$$R(h) = E_x[R(h(x) \mid x)]$$

如果我们针对每个样本x都最小化其条件风险，那么整体的风险也会最小。

这就是所谓的贝叶斯判定准则：为最小化总体风险，只需在每个样本上选择那个能使条件风险最小的类别标记，即

$$h^*(x) = \arg\min_{c \in y} R(c \mid x)$$

h*称为贝叶斯最优分类器。
讲了这些理论，估计大家更是云里雾里，那我们不妨来看看实际的朴素贝叶斯分类器是怎么构建的。

我们先假设lambda_i_j有这样的形式：

$$\lambda_{i,j} = \begin{cases} 0, & \text{if}, i = j \\ 1, & otherwise \end{cases}$$

那么

$$R(c_i \mid x) = \sum_{j=1}^{N} \lambda_{i,j} P(c_i \mid x) = 1 - P(c_i \mid x)$$

这样的话，最小化分类错误率的贝叶斯最优分类器为：

$$h^*(x) = \arg\max_{c \in y} P(c \mid x)$$

怎么理解呢？小Dream哥的理解是，**根据贝叶斯判定准则，我们要预测一个样本属于哪一个类别，计算所有的后验概率P(c|x)，概率最大的那一个类别的后验概率就是预测到的类别了。**

那么该如何去计算后验概率P(c|x)呢？

贝叶斯模型是一种生成模型，先计算联合概率P(c,x)，再通过联合概率计算后验概率，也就是利用如下的贝叶斯公式：

$$P(c|x)=\frac{P(c,x)}{P(x)}=\frac{P(c)P(x|c)}{P(x)}$$

OK，那联合概率和先验概率该怎么计算呢？**朴素贝叶斯模型**就该登场了。

2 朴素贝叶斯分类器

我们再来仔细的分析贝叶斯公式，在有一个训练集的情况下：

> 1、P(c)为样本为某个类别的概率，给定样本及其label后容易计算
> 2、P(x)为某个样本（所有属性相同）出现的概率，给定样本后，容易得到

比较难计算的是P(x|c):

$$P(x|c)=P(x_1,x_2,\ldots,x_m|c)$$

其中m为样本属性的个数，例如预测西瓜是不是甜的模型，如果基于西瓜的花纹是否清晰、敲起来响声是否清亮这两个属性来判断的话，属性个数为2，也就是m=2。

在朴素贝叶斯模型中，有一个样本属性条件独立性假设，即：

$$P(x_1,x_2,\ldots,x_m|c)=\prod_{i=1}^{m}P(x_i|c)$$

这样贝叶斯公式就变成了：

$$P(c|x)=\frac{P(c)}{P(x)}\prod_{i=1}^{m}P(x_i|c)$$

那么，朴素贝叶斯模型得公式就调整为：

$$h_{nb}^{*}(x) = \arg\max_{c \in y} \frac{P(c)}{P(x)} \prod_{i=1}^{m} P(x_i \mid c)$$

对于所有类别来说，P(x)相同，所以上式可以简化为：

$$h_{nb}^{*}(x) = \arg\max_{c \in y} P(c) \prod_{i=1}^{m} P(x_i \mid c)$$

好了，这就是朴素贝叶斯模型基础理论的所有内容了。

到这里，反应快的同学就会说："你说了这么多原理和公式，那么这个模型到底是怎么训练和预测的呢？"下面我们就来讨论这个问题。

3 朴素贝叶斯模型的训练和预测

我们好好看看朴素贝叶斯模型最后的表达式，带计算的参数有P(c)，P(x_i|c)。**训练的过程，其实就是计算所有的P(c)，P(x_i|c)的过程。**

假设数据集为D，Dc表示数据集中C类样本组成得集合。|D|表示数据集中样本的个数，|Dc|表示C类样本的个数。

那么P(c)可以如下表示：

$$P(c) = \frac{|D_c|}{|D|}$$

P(x_i|c)可以用下式表示：

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{|D_c|}$$

|Dc,x_i|表示样本属于c类，第i个属性为x_i的样本的数目。在已知数据集的情况下，上面两个式字都很容易计算，得到所有P(c)和P(x_i|c)后，就完成了学习的过程。

那么，当来了一个新样本，该如何对该样本的类别进行预测呢?

假设新样本X(x_1,x_2,_x_3,....x_m)，总共有n个类别。根据最终的贝叶斯公式

$$\mathrm{h_{nb}}^{*}(x)=\arg\max_{c\in y} P(c)\prod_{i=1}^{m}P(x_i\mid c)$$

预测步骤如下:

(1)根据训练获得的概率值矩阵，第1个类别的P(c_1)和P(x_1|c_1),P(x_2|c_1),...P(x_m|c_1)，并计算他们的乘积，得到属于第一个类别的概率
(2)同上，计算样本属于其他类别的概率
(3)取概率最大的类别为预测样本的类别

这里总结一下:
朴素贝叶斯模型在训练过程，利用数据集D，计算P(c)，P(x_i|c)。在预测时，输入样本，利用贝叶斯公式，计算n个类别的概率，最后输出概率最大的那个类别，作为预测的类别。

$$P(c_j)\prod_{i=1}^{n}P(x_i\mid c_j)$$

总结

整个看下来，朴素贝叶斯模型的本质是针对样本属性的统计概率模型。要想朴素贝叶斯模型的效果好，前期的特征工程和数据清洗是非常重要的工作。早期的机器学习分类模型，特征选择是至关重要的工作，直接决定了模型的效果，这点与现在的深度学模型有很大的差别。神经网络中，通常是在模型内进行特征提取与学习，这就大大减少了特征工程方面的工作。

这是NLP基础理论系列文章中最后一篇机器学习方面的文章了，后面开始介绍深度学习相关的内容了。其他经典的模型，例如SVM，决策树，EM等，如有需要，大家可以留言，小Dream哥视情况，要不要再补上。

下期预告：递归神经网络RNN