逻辑回归与朴素贝叶斯的战争

原创 夕小瑶 夕小瑶的卖萌屋 2017-04-13

首先，小夕带领大家回顾一下文章《逻辑回归》、《Sigmoid与Softmax》、《朴素贝叶斯》中的几点内容，这几点内容也是本文的前置知识：

1. 逻辑回归模型的表达式(假设函数)：$h_\theta(x) = sigmoid(\theta \cdot x)$，其中$sigmoid(z) = \frac{1}{1+e^{-z}}$。
2. 逻辑回归模型$h_\theta(x)$本质上是二类分类问题中其中一个类别的后验概率$P(y|x)$。
3. 用于二类分类的sigmoid函数只是用于多类分类的softmax函数的一个特例。
4. 朴素贝叶斯模型本质上计算并比较的是某样本x与某类别y的联合概率$P(x,y)$。

如果对上述前置知识有疑问，小夕强烈建议再参考那三篇文章理解一下哦。

好了，上面的知识在本文中已默认为常识，不再额外解释啦~

在朴素贝叶斯模型中，P(x,y)是基于贝叶斯定理和独立性假设来近似得到的，而不是像回归模型计算P(y|x)那样直接计算出来。那么有没有一种表示来直接得到P(x,y)的表达式呢？

还记得我们在《sigmoid与softmax》中定义的，小夕将$e^{w1 \cdot w2}$定义为向量w1与w2的“亲密度”，而某个样本x属于某个类别的后验概率P(y|x)就可以解读为“**类别y与样本x的亲密度占所有类别与样本x的亲密度之和的比例**”，用数学语言(softmax)描述就是这样子的（K为类别数，wj是小夕解读过的描述类别j的向量，同时也是大众理解的模型参数）：

$$P(y=j|x=i) = \frac{e^{x_i w_j}}{\sum_{k=1}^{K} e^{x_i w_k}}$$

仔细观察一下小夕用亲密度解释后验概率的这句话，有没有发现这句话非常生动的描绘了P(y|x)呢？(不是自夸啦\(//∇//)\，下文要用到...)

---> 固定住x，因此计算亲密度时忽略其他样本的存在（准确讲，**忽略P(x)**的分布情况），只关心当前的样本x。

那如果我们要描绘P(x,y)呢？描绘x与y的联合概率分布的话，肯定既要描绘出全部的y的情况，**又要描绘出全部的x的情况**，机智的你或许已经想到了，那我们不固定x了，而是考虑全部的x不就行啦。所以，某样本x与某类别y的

联合概率P(x,y)就是“类别y与样本x的亲密度占所有类别与所有样本的亲密度之和的比例”，也就是只需要让分母照顾到所有样本就行啦~所以：

$$P(x=i,y=j)=\frac{e^{x_i w_j}}{\sum_{m=1}^{M}\sum_{k=1}^{K}e^{x_m w_k}}$$

没错，这就是朴素贝叶斯模型背后的东西，它本质上就是额外考虑了样本x自身分布情况的逻辑回归(多类时的softmax回归)。所以本质上，逻辑回归模型与朴素贝叶斯模型之间隔着的墙就是这个p(x)。一个优雅的数学公式总结一下：

$$P(x,y)=P(y|x)*P(x)$$

于是机器学习模型基本上兵分两路：像朴素贝叶斯这种，通过计算样本x与类别y的联合分布来进行分类的机器学习模型被称为生成式模型；像逻辑回归这种，在固定住特定样本x的情况下，计算该样本x与类别y的条件分布来进行分类的机器学习模型被称为判别式模型。

有了这两个定义以后，战争爆发了。

2 战争

战争焦点：以朴素贝叶斯模型为代表的判别式模型与以逻辑回归为代表的生成式模型哪个更好呢？

理论上说，生成式模型不仅考虑(计算中包含)了后验概率，又包含了样本x自身的分布情况，因此比判别式模型涵盖更多的信息量，所以应该更准确才是。但是实际上，从历史战况来看，除了文本分类等个别任务外，判别式模型的代表，逻辑回归模型，往往比代表生成式模型的朴素贝叶斯模型表现更佳。

这是为什么呢？

从上文的朴素贝叶斯的公式可以看出，想要基于全部信息，来计算完整的p(x,y)其实是很困难的，因此需要像朴素贝叶斯一样做一些独立性假设才能近似计算p(x,y)。然而，这些假设又过度简化了p(x)，使得它的估计很不准确，导致哪怕在朴素贝叶斯模型表现优异的场景下，它对各个p(x,y)的计算实际上都是很不准的。

在此，有一个小实验大家可以做一下：

用朴素贝叶斯分类器完成某个分类任务，记下分类器对每个预测结果的把握(即每个P(x,y))。然后把每个样本的每一维度的特征复制成两个。即让X=[x1,x2,x3...]变成X=[x1,x1,x2,x2,x3,x3...]，然后再训练，然后看看对预测结果的把握有没有增大或减小。

我们知道，这样肯定不会带来任何额外的信息量，也不会改变p(x)的分布，然而，这样却会导致朴素贝叶斯增大对预测结果的把握度，也就是增大了对p(x,y)的估计值，这显然是大大的误差。

而判别式模型，由于固定了x值，所以不会考虑p(x)的问题，也就是说对p(x)的分布呈中立态度，自然不会因此引入额外的误差。而对于分类问题，去考虑和计算p(x)的分布情况本就是多此一举的，因此，反倒是判别式模型往往要优于生成式模型的。

所以，暂且就认为朴素贝叶斯模型战败了。

3 战后悄悄话

然而，就像朴素贝叶斯与逻辑回归这个**生成式-判别式对**，同样的战争蔓延到了分类之外的战场上...

欲知后事如何，且听小夕下回，也可能下下回，或者下下下回，或者...