贝叶斯文本分类器

1、**贝叶斯公式**

P（C│X）=P(X│C)/P(X)

1. **分词**

　训练集是有限的，而句子的可能性则是无限的。所以覆盖所有句子可能性的训练集是不存在的。我们不能拿句子作为特征，需要拿句子里面的词语作为特征去考虑。

也就是把一整句话拆分成更细粒度的词语来进行表示。另外，分词之后去除标点符号、数字甚至无关成分(停用词)是特征预处理中的一项技术。

　　观察（“我”,“司”,“可”,“办理”,“正规发票”,“保真”,“增值税”,“发票”,“点数”,“优惠”)，这可以理解成一个向量：向量的每一维度都表示着该特征词在文本中的特定位置存在。

1. ****贝叶斯文本分类器****

贝叶斯文本分类的任务就是将表示成向量的待分类文本X(x1,x2,....xn)归类到与其关联最紧密的类别C（C1,C2,......,Cn）中去。其中X(x1,x2,....xn)为待分类文本的特征向量，C（C1,C2,......,Cn)为给定的类别体系。也就是求解向量X(x1,x2,....xn)属于给定类别C1，C2，......,Cn的概率值，其中Pj为X(x1,x2,....xn)属于Cj的概率，则max(P1,P2,......,Pn)所对应的类别就是文本X所属的类别。

因此，此类问题可以描述为求解(1)式最大值

P(cj│x1,x2,....,xn)=

P(x1,x2,....,xn│cj)P(cj)/P(c1,c2,......,cn) (1)

其中：

1. P(cj)-----文本属于类别cj的概率
2. P(x1,x2,....,xn│cj)------如果待分类文本属于类别cj，则类别cj中包含向量(x1,x2,......,xn)的概率
3. P(c1,c2,......,cn)-------给定的所有类别的联合概率

对于给定的类别，P(c1,c2,......,cn)是一个常数，所以求解（1）式的最大值转换为求解(2)式的最大值

P(x1,x2,....,xn│cj)P(cj) (2)

又根据贝叶斯假设，文本特征向量属性x1,x2,....xn，独立同分布，其联合概率分布等于各个属性特征概率分布的乘积

P(x1,x2,....,xn│cj)=

所以变成求(3)式的最大值：

P(cj) (3)

为了计算分类函数的最大值，（3）式中的先验概率值分别估计如下：

1. P(cj)=M(C=cj)/N

其中，M(C=cj)----训练文本集中属于cj类别的文本数量；

N----训练文本集总数量。

1. P(xj|cj)=N(Xi=xi，C=cj)/N(C=cj)

其中，N(Xi=xi，C=cj)-----类别cj中包含属性xi的训练文本数量

N(C=cj)------类别中的训练文本数量