信息内容安全实验课第五周

Python数据分析

chapter 5 使用朴素贝叶斯算法分类垃圾邮件

文本分类

文本分类就是在给定的分类体系下,让计算机根据给定文本的内容，将其判别为事先确定的若干个文本类别中的某一类或某几类的过程。

一般来说，文本分类可以分为一下过程：

（1） 预处理：将原始语料格式化为同一格式，便于后续的统一处理；

（2） 索引：将文档分解为基本处理单元，同时降低后续处理的开销；

（3） 统计：词频统计，项（单词、概念）与分类的相关概率；

（4） 特征抽取：从文档中抽取出反映文档主题的特征；

（5） 分类器：分类器的训练；

（6） 评价：分类器的测试结果分析。

典型的分类算法包括Rocchio算法、朴素贝叶斯分类算法、K-近邻算法、决策树算法、神经网络算法和支持向量机算法等。

朴素贝叶斯分类算法

算法基本思想：贝叶斯定理

$$p(c_i/x,y)=\frac{p(x,y/c_i)\,p(c_i)}{p(x,y)}$$

条件概率公式：$p(A/B)=\frac{p(AB)}{p(B)}$

朴素贝叶斯分类算法

算法基本思想：利用先验概率和条件概率估算后验概率

后验概率　　条件概率　　先验概率

$$p(c_i/x,y)=\frac{p(x,y/c_i)\,p(c_i)}{p(x,y)}$$

先验概率与后验概率

事情还没有发生，那么这件事情发生的可能性的大小，是先验概率。

事情已经发生，那么这件事情发生的原因是由某个因素引起的可能性的大小，是后验概率。

朴素贝叶斯分类算法—后验概率的简化

$$p(c_i/x,y)=\frac{p(x,y/c_i)\,p(c_i)}{p(x,y)}$$

以垃圾邮件分类为例：

假设c_0是正常邮件，c_1是垃圾邮件，x与y分别是邮件的两个特征，那么，当邮件有x和y两个特征时，

其为正常邮件的概率为$p(c_0/x,y)=\frac{p(x,y/c_0)\,p(c_0)}{p(x,y)}$

其为垃圾邮件的概率为$p(c_1/x,y)=\frac{p(x,y/c_1)\,p(c_1)}{p(x,y)}$

比较上述两个概率的大小，选择较大的概率作为结果。

因为两个概率的分母完全一样，因此在比较两者大小时可忽略分母。

朴素贝叶斯分类算法—计算先验概率

忽略分母后，后验概率变成了$p(c_i/x,y) = p(x,y/c_i)\,p(c_i)$

那么只需要计算$p(x,y/c_i)$和$p(c_i)$两项即可

$p(c_i)$的计算十分简单。 $p(c_0)$表示在邮件数据集中，正常邮件的概率，$p(c_1)$则表示在邮件数据集中，垃圾邮件的概率。因为数据集是有标注的，所以计数之后除以邮件总数即可。

朴素贝叶斯分类算法—计算条件概率

$$p(c_i/x,y) = p(x,y/c_i)\,p(c_i)$$

为了简化运算，引入条件独立性假设，即x和y的条件概率相互独立。这也是朴素二字的由来。

那么，$p(x,y/c_i) = p(x/c_i)\,p(y/c_i)$

同样地，$p(x/c_i)$和$p(y/c_i)$可以对数据进行计数而直接得出。

朴素贝叶斯分类算法步骤

1、计算先验概率$p(C = c_k),\ k = 1,2,\dots,K$

2、计算独立条件概率$p\left(X^i = a^i / C = c_k\right),\ i = 1,2,\dots,n$

3、计算总条件概率$\prod_{i=1}^{n} p\left(X^i = a^i / C = c_k\right),\ i = 1,2,\dots,n$

4、计算后验概率$p(C = c_k)\prod_{i=1}^{n} p\left(X^i = a^i / C = c_k\right)$

5、选取最大后验概率确定x的类别

朴素贝叶斯分类算法举例

由下表的训练数据学习一个朴素贝叶斯分类器并判断样本$(2,S)$的类标记Y，表中$X^{(1)},X^{(2)}$为特征，取值的集合分别为$A_1=\{1,2,3\},A_2=\{S,M,L\},Y$为类标记，$Y\in C=\{1,-1\}$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

朴素贝叶斯分类算法举例

1、计算先验概率：$P(Y=1)=\frac{9}{15}, P(Y=-1)=\frac{6}{15}$

2、计算独立条件概率：

$$P\left(X^{(1)}=1/Y=1\right)=\frac{2}{9} \quad P\left(X^{(1)}=2/Y=1\right)=\frac{3}{9} \quad P\left(X^{(1)}=3/Y=1\right)=\frac{4}{9}$$

$$P\left(X^{(2)}=S/Y=1\right)=\frac{1}{9} \quad P\left(X^{(2)}=M/Y=1\right)=\frac{4}{9} \quad P\left(X^{(2)}=L/Y=1\right)=\frac{4}{9}$$

$$P\left(X^{(1)}=1/Y=-1\right)=\frac{3}{6} \quad P\left(X^{(1)}=2/Y=-1\right)=\frac{2}{6} \quad P\left(X^{(1)}=3/Y=-1\right)=\frac{1}{6}$$

$$P\left(X^{(2)}=S/Y=-1\right)=\frac{3}{6} \quad P\left(X^{(2)}=M/Y=-1\right)=\frac{2}{6} \quad P\left(X^{(2)}=L/Y=-1\right)=\frac{1}{6}$$

朴素贝叶斯分类算法举例

3、计算样本$(2,S)$的后验概率

$$P(Y=1)P\left(X^{(1)}=2/Y=1\right)P\left(X^{(2)}=S/Y=1\right)=\frac{9}{15}\times\frac{3}{9}\times\frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P\left(X^{(1)}=2/Y=-1\right)P\left(X^{(2)}=S/Y=-1\right)=\frac{6}{15}\times\frac{2}{6}\times\frac{3}{6}=\frac{1}{15}$$

后者概率更大，因此样本$(2,S)$属于$y=-1$类

两个后验概率相加不等于1是因为我们忽略了分母

朴素贝叶斯分类算法优缺点

主要优点有：

1）朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

2）对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。

3）对缺失数据不太敏感，算法也比较简单，常用于文本分类。

主要缺点有：

1） 理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型给定输出类别的情况下，假设属性之间相互独立，这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

2）需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。

3）由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

4）对输入数据的表达形式很敏感。

分类算法评价

泛化:机器学习模型学习到的概念在它处于学习的过程中时模型没有遇见过的样本时候的表现。

好的机器学习模型的模板目标是从问题领域内的训练数据到任意的数据上泛化性能良好。这让我们可以在未来对模型没有见过的数据进行预测。

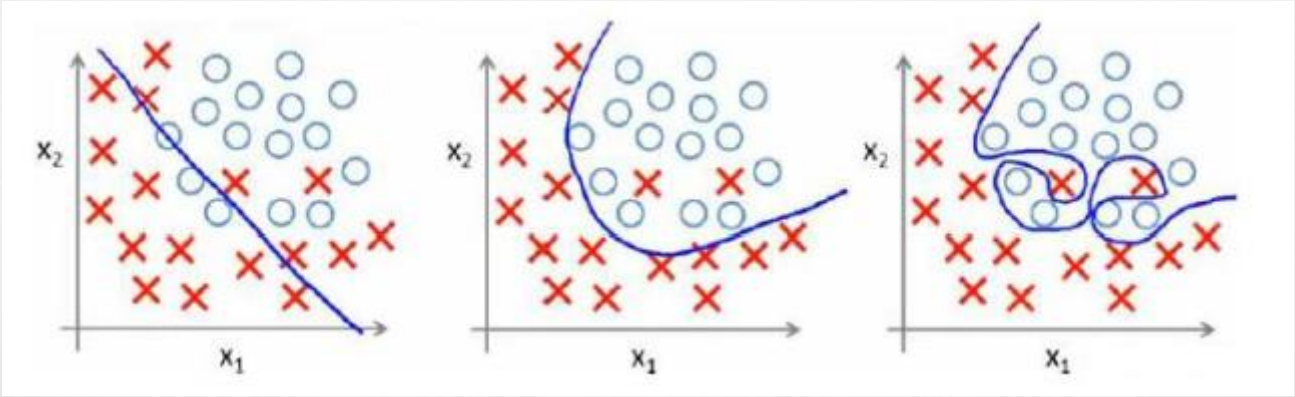
讨论一个机器学习模型学习和泛化的好坏时，我们通常使用术语：过拟合、欠拟合

分类算法评价

欠拟合: 欠拟合指的是模型在训练和预测时表现都不好的情况。

过拟合: 指的是模型对于训练数据拟合程度过当的情况。通常表现为在训练数据上准确率很高，在测试数据上准确率大幅下降。

常用的解决欠拟合及过拟合的方法有：归一化、PCA、增加交叉验证集等



分类算法评价

查准率（Precision）、查全率（召回率）（Recall）、准确率(Accuracy)

我们将算法预测的结果分成四种情况：

1. **正确肯定**（True Positive,TP）：预测为真，实际为真
2. **正确否定**（True Negative,TN）：预测为假，实际为真
3. **错误肯定**（False Positive,FP）：预测为真，实际为假
4. **错误否定**（False Negative,FN）：预测为假，实际为假

则：

查准率P=TP/（TP+FP）

查全率R=TP/（TP+FN）

准确率ACC=(TP+FN) / (TP+TN+FP+FN)

除此之外，还有精确率和召回率的调和均值F，**2/F=1/P+1/R　　F=2*P*R / (P+R)**

F= 2TP / (2TP+FP+FN)

分类算法评价

某池塘有1400条鲤鱼，300只虾，300只鳖。现在以捕鲤鱼为目的。撒一大网，逮着了700条鲤鱼，200只虾，100只鳖。那么，这些指标分别如下：

查准率= 700 / (700 + 200 + 100) = 70%

召回率= 700 / 1400 = 50%

F值 = 70% * 50% * 2 / (70% + 50%) = 58.3%

如果把池子里的所有的鲤鱼、虾和鳖都一网打尽，这些指标又有何变化：

查准率 = 1400 / (1400 + 300 + 300) = 70%

召回率 = 1400 / 1400 = 100%

F值 = 70% * 100% * 2 / (70% + 100%) = 82.35%

由此可见，查准率是评估捕获的成果中目标成果所占得比例；召回率，顾名思义，就是从关注领域中，召回目标类别的比例；而F值，则是综合这二者指标的评估指标，用于综合反映整体的指标。

示例：朴素贝叶斯分类算法分类垃圾邮件

1、使用中文邮件数据集，正常邮件和垃圾邮件各100个；

2、使用jieba分词，对200个邮件分词，并去除重复词与符号，作为词汇表使用；

3、对每个邮件进行单独分词操作，并产生一个词汇表长度的向量，该向量唯一标识该邮件。向量的形式为(0,0,0,1,0,0)，即一个词只要出现（不论次数），便将向量中的该词在词汇表索引位置的0变为1。例如：

词汇表：（中文、邮件、垃圾、推广、新闻）

内容：垃圾邮件　向量表示为(0, 1, 1, 0, 0)

示例：朴素贝叶斯分类算法分类垃圾邮件

4、对上述200个向量计算每个词的条件概率

直接采用向量相加的形式计算，例如：

邮件1：(0, 1, 1, 0, 0)

邮件2：(1, 1, 1, 0, 1)

邮件3：(0, 1, 0, 0, 1)

相加得：(1, 3, 2, 0, 2)

三封邮件中，一共出现了1+3+2+0+2=8个词

则每个词的条件概率为$(\frac{1}{8},\frac{3}{8},\frac{2}{8},\frac{0}{8},\frac{2}{8})$

两个改进：

1、由于条件独立性假设，需要对条件概率进行乘法运算，若某个样本不出现，即概率为0，则最后结果也为0。所以假设每个样本至少出现一次。

2、在实际运算中，条件概率可能会很小，即接近于0，那么在乘法运算中很可能会有下溢出的问题。所以将全部乘法运算改为log运算。

示例：朴素贝叶斯分类算法分类垃圾邮件

```
def trainNB(trainMatrix, trainCategory):  #trainMatrix为所有邮件的矩阵表示，trainCategory为表示邮件类别的向量
    numTrainDocs = len(trainMatrix)  #邮件总数量
    numWords = len(trainMatrix[0])  #词典长度
    pSpam = sum(trainCategory) / float(numTrainDocs)  #统计垃圾邮件的总个数，然后除以总文档个数（先验概率）
    p0Num = np.ones(numWords)  #将向量初始化为1，表示每个词至少出现1次
    p1Num = np.ones(numWords)  #同上
    p0Denom = 2.0; p1Denom = 2.0  #分母初始化为2
    for i in range(numTrainDocs):
        if trainCategory[i] == 1:  #如果是垃圾邮件
            p1Num += trainMatrix[i]  #把属于同一类的文本向量相加，实质是统计某个词条在该类文本中出现频率
            p1Denom += sum(trainMatrix[i])  #把垃圾邮件向量的所有元素加起来，表示垃圾邮件中的所有词汇
        else:
            p0Num += trainMatrix[i]
            p0Denom += sum(trainMatrix[i])
    p1 = np.log(p1Num / p1Denom)  #统计词典中所有词条在垃圾邮件中出现的概率
    p0 = np.log(p0Num / p0Denom)  #统计词典中所有词条在正常文邮件中出现的概率
    return pSpam, p1, p0
```

实验报告（五）

截止时间
2017年5月6日
18:00

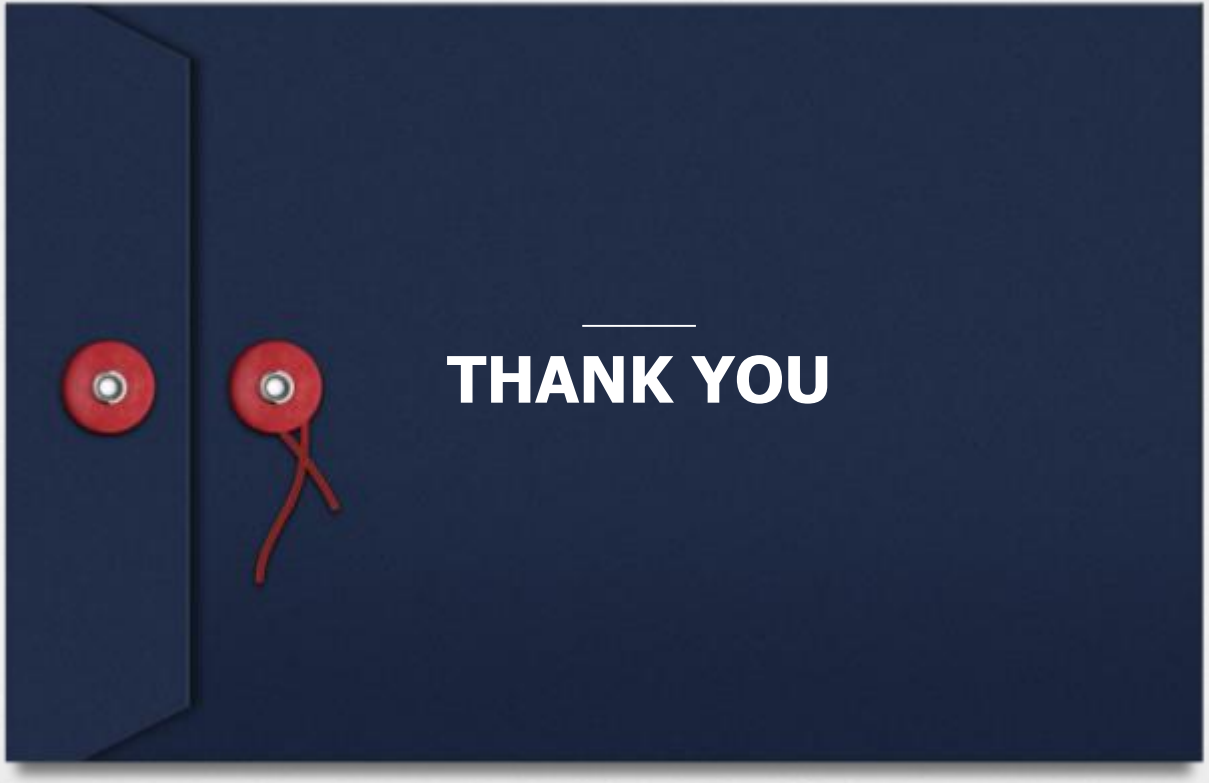
请各位同学在截止时间之前将实验报告发送给课程学习委员

实验五 内容：

1. 使用python语言，完善朴素贝叶斯分类垃圾邮件代码，并编写测试代码，统计该模型的分类正确率。

实验要求：

1. 独立完成，请勿抄袭，自行选择模块内容请勿与其他同学完全雷同。
2. 关键步骤请截图，并保存在实验报告文档中，**截图总数应大于10张**。
3. 实验报告文档命名格式：学号-姓名-实验X
4. 如在实验基础上有创新与拓展、可获得额外分数奖励。


THANK YOU