山东大学 软件 学院

模式识别技术 课程实验报告

|  |  |  |
| --- | --- | --- |
| 学号：201500301125 | 姓名： 马嫣然 | 班级： 15级1班 |
| 实验题目：贝叶斯实验 | | |
| 实验目的：  A.实验任务题目：  1. 利用贝叶斯算法进行数据分类操作，并统计其预测正确率。  2. 随机产生10000组正样本和20000负样本高斯分布的数据集合（维数设为二维），要求正样本：均值为[1;3]，方差为[2 0;0 2]；负样本：均值为[10;20]，方差为[10 0;0 10]。先验概率按样本量设定为1/3和2/3.分别利用最小错误概率贝叶斯分类器和最小风险概率贝叶斯分类器对其分类。（假设风险程度正样本分错风险系数为0.6，负样本分错风险为0.4，该设定仅用于最小风险分析）  B.实验完成要求：  1.仔细阅读并了解实验数据集；  2.使用任何一种熟悉的计算机语言（比如 C，Java或者MATLAB）实现朴素贝叶斯算法；  3.利用朴素贝叶斯算法在训练数据上学习分类器，训练数据的大小分别设置为：前100个数据，前200个数据，前500个数据，前700个数据。前1000个数据，前1350个数据；  4.利用测试数据对学习的分类器进行性能评估；  5.演示实验，提交代码，统计分析实验结果并上交实验报告； | | |
| 实验步骤与内容：   1. 代码来源   代码是在参考开源代码的基础上做出的改进  <http://www.docin.com/p-1914267304.html> 贝叶斯实验报告  来自湖南大学的人工智能实验报告，由智能与科学1302班的某位同学匿名上传至该网站   1. 算法设计说明   实验环境：  硬件环境 个人笔记本电脑  软件环境 JDK Eclipse  所用语言： Java  实验数据分析：  汽车评价数据集  共1728个数据，每个数据特征为6维，分为4类，类别标记为unacc，acc，good，V-good  四个类别标记分别表示汽车性价比等级（由低到高）  unacc：1210个  acc：384个  good：69个  V-good：65个  6个特征分别为：（6个属性）   1. buying （取值：v-high、high、med、low） 表示购买价格 2. maint （取值： v-high、high、med、low） 表示维修价格 3. door （取值：2、3、4、5-more） 车门数量 4. Persons （取值：2、4、more） 可容纳人数 5. Lug\_boot （取值：small、med、big） 行李箱大小   Safety （取值：low、med、high） 安全系数  下载链接：<http://archive.ics.uci.edu/ml/datasets/Car+Evaluation>  算法设计：  1.数据要求  训练数据（Test Data）：用于模型构建  测试数据（Test Data）：用于检测模型构建，此数据只在模型检验时使用，用于评估模型的准确率。绝对不允许用于模型构建过程，否则会导致过度拟合。   1. 思路   来自CSDN博客——朴素贝叶斯分类  <http://blog.csdn.net/chinesesword/article/details/10108747>  朴素贝叶斯分类的正式定义如下：  1、设x={a1,a2......am}为一个待分类项，而每个a为x的一个特征属性。  2、有类别集合C={y1,y2......yn}。  3、计算P(y1|x),P(y2|x)......P(yn|x)。  4、如果P(yk|x)=max{P(y1|x),P(y2|x)......P(yn|x)}，则x属于yk。  那么现在的关键就是如何计算第3步中的各个条件概率。我们可以这么做：  1、找到一个已知分类的待分类项集合，这个集合叫做训练样本集。  2、统计得到在各类别下各个特征属性的条件概率估计。  即P(a1|y1),P(a2|y1)......P(am|y1);  P(a1|y2),P(a2|y2)......P(am|y2);  ................................  P(a1|yn),P(a2|yn)......P(am|yn);  3、如果各个特征属性是条件独立的，则根据贝叶斯定理有如下推导：  P(yi|x)=(P(x|yi)\*P(yi))/P(x)  因为分母对于所有类别为常数，因为我们只要将分子最大化即可，即求P(x|yi)P(yi)最大。又因为各特征属性是条件独立的，所以有：P(x|yi)P(yi)=P(a1|yi)\*P(a2|yi)......P(am|yi)\*P(yi)  从中找出最大值，类别i即为最可能分类  再计算准确率   1. 具体实现   **(1)声明变量，存储各类统计**  11    **(2)汽车属性类**  12  **(3)获取训练样本**  声明并实现**public** **void** ReadFile(String filename)方法，将数据集的前特定行数读取，并将每一行的数据属性按（，）分割开，各自统计个数  15  **(4)计算概率**  **public** **void** Calculated\_probability()方法  计算训练样本中各属性在各类中出现的概率，作为P（am|yi)    13  **(5)导入测试样本数据**  原理同上文读取训练样本数据  **(6)对测试样本分类**  先对测试样本每一条进行处理，对每一个数据都计算它的每一个属性在unacc,acc,good,vgood下的概率，即P(x|yi)=P(a1|yi)\*P(a2|yi).....\*P(am|yi)  因为P(yi|x)=(P(x|yi)\*P(yi))/P(x)，所以从P（x|yi）中选出最大的值,类别i即为分类  判断预测结果和开始给定的结果是否相等并记录  14    **(7)计算分类器的准确率**  分类准确数/总测试数   1. 改进：   a.所参考的代码中，对训练样本的读取方式是一整个文件读到底，即**while** (temp !=**null** )循环，  测试数据的录入也是如此。而实验数据网站中只给了一个汽车数据集的文件，所以要将它们分成训练数据和测试数据，修改文件读取方式。  实验完成要求3中说“利用朴素贝叶斯算法在训练数据上学习分类器，训练数据的大小分别设置为：前100个数据，前200个数据，前500个数据，前700个数据。前1000个数据，前1350个数据”，所以需要将ReadFile方法改成可控制大小的读取文件，就可以用数据集文件充当训练样本了。  而数据集从第1350到结尾的数据都没有被训练样本要求过，所以可以充当测试样本。  b.如果将样本集的大小设置为1200行以内，就会出现下图的错误  1  对数据进行打印排查之后发现  2  可见是ClassValue[2]即“good”这一类出了问题  看数据集可知，前1199行没有出现过“good”类，只有其它三类  同理如300行以内没有vgood与good类，所以上图中三四的ClassValue都为0  最终没有能够优化好算法中这个问题，只得修改数据集，将部分数据剪切至前几行  如此不再出现错误警告  实验结果：  测试数据选择数据集文件第1351行至结尾  训练数据按照要求选择前100、200、500、700、1000、1350行  前100个数据  分类器的准确率为:58.201054%  前200个数据  分类器的准确率为:58.201054%  前500个数据  分类器的准确率为:58.201054%  前700个数据  分类器的准确率为:58.201054%  前1000个数据  分类器的准确率为:58.201054%  前1350个数据  分类器的准确率为:68.783066%  实验结果分析：  由上文的实验结果可见，训练样本越大，分类器的准确率就越高  但前四次测试结果不变，且接近于随机分类，可能是由于训练样本的分布有问题，导致各属性并不能很好地独立  3）实验结果截图  前100、200、500、700、1000个训练样本  4  1350个训练样本  3 | | |