##### QG工作室数据挖掘小组实验报告

|  |  |  |
| --- | --- | --- |
| 实习生： 李炜乐 | 导师： 张润鹏 | 日期：2019年7月20日 |

|  |
| --- |
| **实验名称：朴素贝叶斯算法** |
| **已完成内容：**   1. **二分类朴素贝叶斯** 2. **多分类朴素贝叶斯** 3. **高斯朴素贝叶斯** |
| **未完成内容：无** |
| **未完成原因：无** |
| **需要帮助：无** |

|  |  |
| --- | --- |
| **实验总结** | |
| **知识点总结：**   1. **二分类朴素贝叶斯包括数据词向量化、计算出属于不同类别的某个特征条件概率，根据朴素贝叶斯公式计算概率比较得到结果三个步骤。其中在处理朴素贝叶斯分类器的时候为了防止下溢，对计算进行取对数处理，并且得到了较好的结果。** 2. **二分类朴素贝叶斯的工作原理是基于贝叶斯决策理论，使用某个特征属性的条件概率来进行分类。** 3. **二分类朴素贝叶斯实验环境为垃圾邮箱分类，由于随机选举训练集和测试集，所以正确率并不唯一，集中于错误率基于0%-10%之间，运行时间为0.07s** 4. **高斯朴素贝叶斯包括计算数据集的均值和方差，带入高斯公式，得到高斯值，经过判断得到分类。其中在处理方差的时候采用了协方差矩阵，协方差矩阵的对角线均为各个维度上的矩阵** 5. **高斯朴素贝叶斯的工作原理就是基于高斯公式，使用某个特征的高斯概率的大小来是实现分类。** 6. **高斯朴素贝叶斯实验环境基于iris数据集，错误率处于0.1左右，运行时间为0.0148s** 7. **多分类朴素贝叶斯：多分类朴素贝叶斯居于二分类朴素贝叶斯，通过计算各个类别的各个属性的条件概，通过贝叶斯公式计算概率，计算出样本属于每个类别的概率，通过比较大小来进行分类。其中，计算先验概率和贝叶斯公式的时候，进行了拉普拉斯平滑。** 8. **多分类朴素贝叶斯实现环境四个新闻分类，由于数据集是随机分配，结果并不稳定，维持在68%-75%之间，运行时间为31.85s** | |
| **遇到问题**：在使用贝叶斯分类器进行分类的时候，需要计算多个概率的乘积，由于大多数概率趋于0导致结果并不明显 | **解决过程：采取取对数的方法解决了该问题** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **导师评价** | | | | |
| **实验分数** | **知识掌握情况** | **代码编写能力** | **建议** | **评价日期** |
|  |  |  |  |  |