实验三 朴素贝叶斯分类器

1. **实验目的**
2. 在真实数据集上实现朴素贝叶斯分类器，并验证其分类效果
3. 了解如何在测试数据集上实现一个机器学习算法
4. 了解如何评价分类效果
5. 了解如果分析实验结果
6. **算法原理**

假设各条件相互间独立，即：

|  |  |
| --- | --- |
|  | (1) |

在训练时训练 *P*(*y*)以及*P*(*xi*|*y*)。

测试时输出：

|  |  |
| --- | --- |
|  | (2) |

1. **数据集处理**
2. **数据集描述**

实验给定了Adult数据集，其中adult.data为训练集（32561条数据），adult.test为测试集(16281条数据)，每行数据代表一个人，共有15个维度的特征，最后一个特征为该人的收入是否超过了50K。

数据集中部分特征是连续数据，部分数据可能未知(用?表示)

1. **分类器的性能评价指标：**

本次实验中，采用精确率（Precision）和召回率（Recall）作为朴素贝叶斯分类器性能的评价指标，计算方法为：

|  |  |
| --- | --- |
|  | (3) |

|  |  |
| --- | --- |
|  | (4) |

1. **数据集的变量及其含义**

| ****变量名**** | ****意义**** | ****数据类型**** | ****元数据**** |
| --- | --- | --- | --- |
| age | 年龄 | 连续数据 |  |
| work\_class | 职业类型 | 离散数据 | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| fnlwgt | 最终重量(?) | 连续数据 |  |
| education | 学历等级 | 离散数据 | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| education\_num | 学历的数字等级 | 连续数据 |  |
| marital-status | 婚姻状况 | 离散数据 | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| occupation | 职业 | 离散数据 | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces |
| relationship | 家庭关系 | 离散数据 | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| race | 人种 | 离散数据 | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| sex | 性别 | 离散数据 | Female, Male |
| capital\_gain | 资本收益 | 连续数据 |  |
| capital\_loss | 资本损失 | 连续数据 |  |
| hours\_per\_week | 每周工作时长 | 连续数据 |  |
| native-country | 出生国 | 离散 | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands |
| income | 收入 | 离散 | >50K, <=50K. |

1. 数据探索
2. 未知数据？的处理

你打算如何处理未知数据？

提示：

（1）统计未知数据的数量在总数据集中的比例，如果不高，可以先尝试直接去掉未知数据后的训练模型的性能。

（2）再做一个尝试，将未知数据通过某种方式补全后，再对比训练模型的性能。

1. 数据合并

在这些数据项中，是否有数据项可以合并？

提示：Never-worked和Without-pay可以合并为Without-pay字段。

1. 数据的离散化

数据集中有多个数据项是连续数据类型，对于这些数据项，可以有两种处理方式：不处理，或者离散化。请考虑如果要将以下数据项离散化，该如何进行：

1. Age：如何分割？分割粒度为多少？
2. 资本收益：观察资本收益的数据范围和变化范围，思考直接将资本收益离散化是否合适？提示：考虑到投资和收入之间存在一定的关系，而且可以很明显的看出收益为0的占了所有数据的大部分，所以将其分为三个类别：没有资本收益，资本收益较少，资本收益较大。除去数据中为0的值之后，求其平均值、中位数、方差，然后观察以平均值还是中位数作为界点？
3. 资本损失：与资本收益做同样的处理。
4. 每周工作时间：如果分割粒度设置为5，该如何分割？
5. 训练模型
6. 使用adult.data为训练集（32561条数据），adult.test为测试集(16281条数据)，编写python程序训练朴素贝叶斯模型。http://mlr.cs.umass.edu/ml/datasets/Adult
7. 朴素贝叶斯模型可以参考：<https://dataminingguide.books.yourtion.com/chapter-6/chapter-6-6.html>
8. 注意0概率的处理
9. 当测试集中某个数据的某条特征值取了某个值*xi*，但训练集中该特征值并没有取过该值*xi*，则在训练时*P*(*xi*|*y*)=0，由此在做测试时计算其概率

|  |  |
| --- | --- |
|  | (5) |

时会得到概率为0。

解决方法：通常我们会进行拉普拉斯平滑处理：

|  |  |
| --- | --- |
|  | (6) |

1. 实验结果分析
2. 训练集规模的影响

训练集的规模对分类效果有什么影响？

分别选取5%， 50%， 100%的训练集数据训练分类模型，然后比较精准率，填写如下表格：

| ****比例**** | ****训练集数目**** | ****精准率**** | ****召回率**** |
| --- | --- | --- | --- |
| 5% | 1494 |  |  |
| 50% | 15075 |  |  |
| 100% | 30162 |  |  |

分析结论：训练集大小对分类器的影响是什么？

1. 特征集的影响

选取不同的特征组合对分类结果有什么影响？

分别选取8-10个各种特征组合训练分类模型，然后比较精准率，填写如下表格：

| ****特征集**** | ****精准率**** | ****召回率**** |
| --- | --- | --- |
| 年龄 工作类别 学历等级 婚姻状况 家庭关系 投资利得 投资损失 收入 |  |  |
| 年龄 工作类别 学历等级 婚姻状况 职业 投资利得 投资损失 收入 |  |  |
| 。。。。。。 |  |  |

分析结论：选取哪个特征集可以获得最好的分类效果？

1. 连续特征以及未知特征的处理
2. 各连续特征不离散化与离散化的分类效果对比，分别得到如下表格：

| ****特征**** | ****分割粒度**** | ****精准率**** | ****召回率**** |
| --- | --- | --- | --- |
| age | 不分割 |  |  |
| age | 3 |  |  |
| age | 5 |  |  |
| age | 10 |  |  |

| ****特征**** | ****分割粒度**** | ****精准率**** | ****召回率**** |
| --- | --- | --- | --- |
| hour | 不分割 |  |  |
| hour | 3 |  |  |
| hour | 5 |  |  |
| hour | 10 |  |  |

| ****特征**** | ****分割粒度**** | ****精准率**** | ****召回率**** |
| --- | --- | --- | --- |
| 投资收益与损失 | 不分割 |  |  |
| 投资收益与损失 | 1000 |  |  |
| 投资收益与损失 | 无，低，高 |  |  |

请讨论连续特征离散化处理的效果。

1. 未知数据如何处理

对于未知数据的处理，做两种情况的对比，一种是直接忽略掉这些数据，另一种是将‘？’补全，并得到如下表格：

| ****处理方案**** | ****精准率**** | ****召回率**** |
| --- | --- | --- |
| 忽略未知数据 |  |  |
| 补全数据 |  |  |