机器学习实验-朴素贝叶斯分类器

计 41 张盛豪 2014011450

2017-03-18 23:01:28

Contents

1 提示：个人联系方式及代码简要说明、运行方式在 README.md 文件中 1

2 实验目的 1

3 算法实现原理 2

4 数据集处理 2
4.1 数据集描述 . . . 2
4.2 分类器的性能评价指标： . . . 2
4.3 数据集的变量及其含义 . . . 2
4.4 特殊数据的处理方式 . . . 3
4.4.1 未知数据? 的处理 . . . 3
4.4.2 数据合并 . . . 3
4.4.3 数据的离散化 . . . 3
4.4.4 各连续变量的范围及对应的分割粒度如下： . . . 5

5 实验结果的分析 5
5.1 训练集规模的影响 . . . 5
5.1.1 选取 5%，50%，100% 的训练集数据训练分类模型 . . . 5
5.1.2 重复随机抽取样本实验（5 次），记录最小，最大，平均准确率 . . . 5
5.2 0 概率的处理 . . . 6
5.3 连续特征以及未知特征的处理 . . . 6
5.3.1 连续数据如何进行处理 . . . 6
5.3.2 未知数据如何处理 . . . 7
5.4 选取的特征值对比 . . . 7
5.5 交叉验证 . . . 7

6 实验总结及结论 8

7 参考资料 8

1 提示：个人联系方式及代码简要说明、运行方式在 README.md 文件中

2 实验目的

- 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
- 了解如何在测试数据集上实现一个机器学习算法
- 了解如何评价分类效果
- 了解如果分析实验结果

3 算法实现原理

假设各条件相互间独立，即

$$P(y|x_1,\cdots,x_n)P(y) \propto \prod_{i=1}^{n} P(x_i|y)$$

在训练时训练 $P(y)$ 以及 $P(x_i|y)$

测试时输出

$$\hat{y} = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

4 数据集处理

4.1 数据集描述

实验给定了 Adult 数据集，其中 adult.train 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)

4.2 分类器的性能评价指标:

本次实验中，我采用了准确率作为朴素贝叶斯分类器性能的评价指标，计算方法为:

$$Accuracy = \frac{number\ of\ correctly\ classified\ records}{number\ test\ records}$$

4.3 数据集的变量及其含义

变量名	意义	数据特征	处理方式
age	年龄	连续数据	分段离散
work_class	职业类型	离散数据	
fnlwgt	最终重量 (?)	连续数据	无意义、忽略
education	学历等级	离散数据	
education_num	学历的数字等级	连续数据	重复、忽略
marital-status	婚姻状况	离散数据	
occupation	职业	离散数据	
relationship	家庭关系	离散数据	
race	人种	离散数据	
sex	性别	离散数据	
capital_gain	资本利得	连续数据	离散化
capital_loss	资本损失	连续数据	离散化
hours_per_week	每周工作时长	连续数据	离散化
native-country	出生国	离散	
income	收入	离散	

4.4 特殊数据的处理方式

4.4.1 未知数据? 的处理

数据集中有未知的数据 (?)，我的处理方式是将这类数据直接忽略掉

4.4.2 数据合并

使用 R 语言对数据进行统计，发现，Never-worked 和 Without-pay 可以合并为 Without-pay 字段。

4.4.3 数据的离散化

4.4.3.1 数据规律探索

考虑到 R 语言对数据处理的优越性，因此采用 R 语言对数据规律进行探索，利用 R 语言读入测试集和训练集

```
# 读取测试集，已清除?
test = read.csv("after.test",
                  sep=",", header=F, col.names=c("age", "work_class", "fnlwgt", "education",
                      "education_num",
                                                 "marital-status", "occupation", "relationship",
                                                     "race", "sex",
                                                 "capital_gain", "capital_loss",
                                                     "hourr_per_week","native-country", "income"),
                  fill = FALSE, strip.white = T)
```

然后对连续数据做统计之后得到如下结果

4.4.3.1.1 Age 的规律

```
> table(train$age)

 17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41
     42  43  44
328 447 594 629 621 674 824 752 799 745 789 808 774 813 851 789 837 836 828 852 828 791 786 765 769
    741 743 704
 45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69
     70  71  72
706 711 683 523 555 575 571 455 448 394 386 343 337 344 332 276 259 213 186 173 136 110 111  90  80
    64  54  40
 73  74  75  76  77  78  79  80  81  82  83  84  85  86  88  90
 49  38  34  29  20  14  15  16  13   7   5   8   3   1   3  35
```

可知年龄范围为 17~90，我们设置其分割粒度为 5

4.4.3.2 资本收益

可以从统计结果看出，投资收益相互间差别很大，直接以固定颗粒度分割并不适合，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，我们将其分为三个类别，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，求得其平均值，中位数，方差如下:

```
> # 资本收益
> table(train$capital_gain)

    0   114   401   594   914   991  1055  1086  1151  1173  1409  1424  1455  1471  1506  1639  1797  1831
27624     6     1    28     8     3    21     1     8     1     3     3     1     7    14     1     5     7
 1848  2009  2036  2050  2062  2105  2174  2176  2202  2228  2290  2329  2346  2354  2387  2407  2414  2463
    5     2     4     3     2     8    46    20    15     3     5     4     4    10     1    18     6    11
 2538  2580  2597  2635  2653  2829  2885  2907  2936  2961  2964  2977  2993  3103  3137  3273  3325  3411
    1    12    20    10     4    30    22    10     3     2     6     8     1    94    36     6    53    21
 3418  3432  3456  3464  3471  3674  3781  3818  3887  3908  3942  4064  4101  4386  4416  4508  4650  4687
    3     2     2    20     7    13    10     5     6    31    12    40    19    67    11    11    40     3
 4787  4865  4931  4934  5013  5060  5178  5455  5556  5721  6097  6360  6418  6497  6514  6723  6767  6849
   22    16     1     7    69     1    91    11     4     3     1     3     7    11     4     2     3    26
 7298  7430  7443  7688  7896  7978  8614  9386  9562 10520 10566 10605 11678 13550 14084 14344 15020 15024
  240     8     5   270     2     1    52    16     4    43     6     9     2    25    39    26     5   337
15831 18481 20051 22040 25124 25236 27828 34095 41310 99999
    6     2    33     1     2    11    32     3     2   148
```

Figure 1: 资本收益

```
> # 资本收益平均值
> mean(train$capital_gain[train$capital_gain!=0])
[1] 12977.6
> # 资本收益中位数
> median(train$capital_gain[train$capital_gain!=0])
[1] 7298
> # 资本收益方差
> sd(train$capital_gain[train$capital_gain!=0])
[1] 22311.91
```

可以看出其方差较大，以平均值作为界点不合适，我们取其中位数作为资本收益高低的界点（训练集与测试集这个数据差别不大）

同理，资本损失的数据也有这样的规律，我们也用这样的方法对其进行处理。

```
> # 资本损失平均值
> mean(train$capital_loss[train$capital_loss!=0])
[1] 1867.898
> # 资本损失中位数
> median(train$capital_loss[train$capital_loss!=0])
[1] 1887
> # 资本损失收益方差
> sd(train$capital_loss[train$capital_loss!=0])
[1] 361.8574
```

资本损失的差异值不大，我们直接取中位数作为分界点

4.4.3.3 每周工作时间

```
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30
  5   9  11  20  22  19  13  50   7 110   7  64   4  12 173  73  13  49   5 548  18  17  14  87 246   9  11  51   7 478
 31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  7 133  22  20 592 106  91 225  25 7107  22 113  73  92 849  47  33 239  10 1376   7  65  12  20 348  43   2  10   2 680
 61  62  63  64  65  66  67  68  69  70  72  73  74  75  76  77  78  79  80  84  85  86  88  89  90  92  96  98  99
  2   5   5   7 108   6   2   4   1 142  35   1   2  37   1   3   5   1  69  25   4   2   2   1  13   2   4   3  45
```

Figure 2: 每周工作时间

工作时间为 1~99，分割粒度设置为 5

4.4.4 各连续变量的范围及对应的分割粒度如下：

变量名	范围	分割粒度
age	17~90	5
capital_gain	0 ~ 99999	0，0~7298，7298+
capital_loss	0 ~ 4356	0，0~1887，1887+
hours_per_week	1~99	5

5 实验结果的分析

5.1 训练集规模的影响

问题：训练集的规模对分类效果有什么影响?

5.1.1 选取 5%，50%，100% 的训练集数据训练分类模型

测试数据的选择，经过多方面的对比测试，发现选取的特征值为

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

时，训练的模型分类效果最好。以此为基础，分别选取 5%，50%，100% 的训练集数据训练分类模型时的训练结果对比如下：

比例	训练集数目	准确率
5%	1494	83.71%
50%	15075	83.99%
100%	30162	84.12%

由这个表格可以看出，随着数据集的增加，分类器的分类效果越来越好。而实际上，即使只取了 5%（1494 条训练数据），训练出的分类器分类效果仍然挺好的，可以由此体会到贝叶斯分类的高效性和实用性。

5.1.2 重复随机抽取样本实验（5 次），记录最小，最大，平均准确率

随机比例	训练集数目	准确率
70.13%	21131	83.99%
47.92%	14458	84.01%
94.40%	28472	84.04%
58.65%	17687	84.14%
11.83%	3566	83.95%

最小值	最大值	平均值
83.95%	84.14%	84.03%

综合两个测试结果可以得出如下结论：数据集的规模会对分类效果产生一定影响，但这种影响并不是绝对的，当抽取的训练集具有随机性时，小训练集也有可能会有特别好的分类效果，不过整体来看，训练集规模越大，分

类效果越好。

5.2 0 概率的处理

当测试集中某个数据的某条特征值取了某个值 x_i，但训练集中该特征值并没有取过该值 x_i，则在训练时 $P(x_i|y)=0$，由此在做测试时计算其概率

$$\hat{y} = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

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，即在计算条件概率时对每个 x_i 做 $+\lambda$ 处理，对应的总数也需要做 $+M\lambda$，经过测试可以发现，在未做拉普拉斯平滑时，训练集取 50% 时，分类准确率为 83.95%，加上拉普拉斯平滑处理之后，分类准确率为 83.99%，准确率有一定提升。

5.3 连续特征以及未知特征的处理

5.3.1 连续数据如何进行处理

说明：**【数据的分析及离散化方案】**见前面【3.4.3 数据的离散化】小节。由于后期经过测试发现特征选取年龄 工作类别 学历等级 婚姻状况 职业 投资利得 投资损失 收入这 8 个特征时分类效果最好，故前面测试样例的特征选取均选择了这 8 个，为了测试连续数据分割方案对分类效果的影响，将每周工作时间这一连续特征加入分类的特征中。

测试时，训练集取 100%，拉普拉斯平滑处理的 $\lambda=1$

各连续特征值分割粒度与其分类效果对比

特征	分割粒度	准确率
age	不分割	83.816%
age	3	83.831%
age	5	83.751%
age	10	83.784%

特征	分割粒度	准确率
hour	不分割	83.845%
hour	3	83.845%
hour	5	83.752%
hour	10	83.804%

特征	分割粒度	准确率
投资收益与损失	不分割	【85.20%】
投资收益与损失	1000	83.62%
投资收益与损失	无，低，高	83.75%

（**注意**：测试某一特征时，其余特征分割情况默认为：年龄分割粒度 5，每周工作时间分割粒度为 5，投资收益和损失分割为无， 收益/损失低，收益/损失高）

结果分析：这里发现了一些很尴尬的结果，对这三个连续特征值进行不同粒度离散之后发现，【不进行离散】，直接以每一个数据单独作为一个类别进行分类时准确率反而比对其进行不同间距离散之后分类【准确率高】，特别是【投资收益与损失】的两个特征值，不进行离散时其准确率甚至高达 【85.2%】，而进行了等间距离散或者以无，低、高，结合前面对各个特征的分布情况的统计可以进行如下猜测：

> 结合前面 age，work hour，capital_gain，capital_loss 几个特征值的取值特征统计结果，可以发现，age 和 work hour 在各个取值中虽然较为分散，但是也有小范围集中，简单的等间距离散，对其分类的优化效果不大，甚至于有可能因为粒度过大而使分类效果显著下降，当分割粒度恰好使得集中数据分在了一个 category 时，其准确率会略高一点，而如果恰好使之分散开，可能会对分类准确率有反作用，如 work hour 特征分割粒度为 5 和 10 的对比。
>
> 对于投资收益和损失这两个特征值，分析其数据特征可以发现，有超过 70% 的数据是 0，而其他数据就较为零散并且差异值极大，简单地等间距分割，或者以中位数，平均数作为临界点进行分割都是不太合理的，因此这种情况下不进行离散化分类效果反而会更好，（不知道高斯分布处理会不会优化其分类效果，由于能力和时间限制，没来得及进行测试）

5.3.2 未知数据如何处理

对于未知数据的处理，我做了两种情况的对比，一种是直接忽略掉这些数据，另一种是将 ‘? ’ 也视为一种数据和特征

处理方案	准确率
忽略未知数据	84.12%
视为新类型	【84.45%】

可以看出，将未知数据视为一种特殊的新类型时其分类准确率有较大提高，可知这些数据某种程度上也能够反映出其收入的高低

5.4 选取的特征值对比

选取的特征值	准确率
年龄工作类别重量学历等级教育年限婚姻状况职业家庭关系人种性别投资利得投资损失每周工作时间出生国收入	79.77%
年龄工作类别学历等级婚姻状况职业家庭关系人种性别投资利得投资损失每周工作时间出生国收入	81.85%
年龄工作类别学历等级婚姻状况职业家庭关系性别投资利得投资损失每周工作时间出生国收入	81.83%
年龄工作类别学历等级婚姻状况职业人种投资利得投资损失每周工作时间收入	83.74%
年龄工作类别学历等级婚姻状况职业投资利得投资损失每周工作时间收入	83.75%
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	84.12%
年龄工作类别教育年限婚姻状况职业收入	81.79%

由这个表的对比分析可以明显地感受到特征值的选取对分类效果的影响，【年龄，工作类别，学历等级，婚姻状况，职业，投资利得，投资损失】这几个特征值能够很大程度上反应出其收入的高低，特别是投资收入和损失，这个特征与收入有较大的相关性，这也是为什么在【5.3.1 连续数据如何进行处理】一节中提到对这两个特征不进行分割时其分类效果甚至可以高达 85.20% 的一个原因。

这也提醒我们在选取训练数据时注意对特征值的选取，有代表性的特征对其分类准确率有促进作用，而一些无关的特征则会对分类效果有负作用。

5.5 交叉验证

选取不同比例的测试集数据用作训练，观察其对分类效果的影响

测试集比例	训练集 + 测试集	准确率
0%	15075 + 0	83.99%
5%	15075 + 751	84.02%
50%	15075 + 7525	84.23%
100%	15075 + 15060	84.26%

[注] 训练集选取的是 50% 训练集数据

由这个数据对比可以看出，训练模型时加入部分测试集数据，对分类效果有提升作用。

6 实验总结及结论

本次实验实现了贝叶斯分类算法，并探讨了数据规模、特征值的选择对分类效果的影响，以及连续值，未知值的不同离散方案和处理方案对分类效果的影响，并探讨了拉普拉斯平滑对分类效果的影响。

通过多方面的对比，主要得出了以下一些结论：

- 特征值的选取对分类效果影响较大，可以根据特征值与类别的相关性大小判断其对分类是有帮助的还是有干扰的。
- 特征值个数也对分类效果有一定影响，选择合适的，足够的特征作为分类依据是较好的，特征值的选择和个数的选择可通过参数的调整进行尝试后得出最优方案
- 数据规模对分类效果有一定影响，但是只要特征值选取较好，训练数据较少情况下训练出的模型其分类效果也较好。
- 连续值的处理方式对分类效果影响较大，对于一些本身比较离散并且该特征值对类型影响较大情况下，分割粒度越小，分类效果会更好，尤其是本实验中投资收入和投资损失两个特征值，不进行分割时其准确率甚至高达 85.2%，可明显体会到数据离散粒度对分类效果的影响。
- 未知数据的处理，未知数据视为一种特殊类型也是一种比较有效的处理方式，背后原因可能是这些没统计到数据的人群可能具有某方面的共性，其收入也会受到这方面的影响。

本次实验中自己收获很大，特别是学会了从不同的方面去评价一个算法的性能的方式的分析方法。此外，对编程也有一些新的启发，在编程时需要注意面向对象编程，考虑到各种可能的变化，本次实验中的贝叶斯分类器的实现方法我参考了【参考资料 1 使用 python 编写朴素贝叶斯分类器】一文的实现方式，这种实现可以达到一种自适应的状态，不论特征值有多少个，是什么类型的数据，只要给出参数，均可对其进行训练和分类，这就方便了我们进行不同的特征值对分类效果的影响分析时的测试，只需要对测试文件和训练文件做对应修改即可。此外，在实现过程中设置了很多参数，可以方便进行训练数据比例，是否随机选取，特征值分割粒度，是否进行拉普拉斯平滑处理等进行设置，极大地方便了测试和分析。

7 参考资料

- 使用 Python 编写朴素贝叶斯分类器 https://datamininguide.books.yourtion.com/chapter-6/chapter-6-6.html
- Dataset: Adult (R) http://scg.sdsu.edu/dataset-adult_r/