IDA 课程作业实验报告

软件 51　2015013190　安彦哲

2018.12.16

1　数据预处理

1.1　遗漏数据处理

1.1.1　*Race* (2%)

1) 经过统计，所给数据表中共有 101766 条数据，其中 $AfricanAmerican$ 有 19210 条，*Asian* 有 641 条，*Caucasian* 有 76099 条，*Hispanic* 有 2037 条，*Other* 有 1506 条，遗漏 2273 条。

2) 处理方案：将遗漏处均填补为 *Caucasian* （众数）。

1.1.2　*Weight* (97%)

1) 统计不同 *Age* ，不同 *Gender* ，不同 *Weight* 的数据条数，得到的结果如下：

Male/Female \ Age / Weight	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)
[0-25)	3/0	0/0	0/1	1/0	5/2
[25-50)	1/1	1/3	0/1	1/3	3/4
[50-75)	0/0	4/10	16/23	12/12	18/20
[75-100)	0/0	0/0	7/10	11/23	44/36
[100-125)	0/0	0/0	2/7	6/4	30/36
[125-150)	0/0	0/0	0/1	6/3	15/10
[150-175)	0/0	0/0	1/0	0/2	2/2
[175-200)	0/0	0/0	0/0	0/0	3/1
>200	0/0	0/0	0/0	0/1	0/0

Male/Female \ Age / Weight	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
[0-25)	3/1	5/7	7/7	2/4	0/0
[25-50)	6/7	4/9	6/9	2/27	2/7
[50-75)	37/59	67/88	71/170	66/165	11/48
[75-100)	88/100	150/145	252/215	136/95	15/9
[100-125)	78/60	123/70	119/55	25/10	0/0
[125-150)	25/20	22/14	13/9	4/3	0/0
[150-175)	4/7	9/4	1/3	0/0	0/0
[175-200)	2/3	0/1	1/0	0/0	0/0
>200	0/0	1/1	0/0	0/0	0/0

2) 根据以上数据计算出不同 *Age* ，不同 *Gender* 的平均 $Weight$ ：

Weight \ Age / Gender	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)
Male	$[0-25)$	$[50-75)$	$[75-100)$	$[75-100)$	$[75-100)$
Female	$[25-50)$	$[50-75)$	$[75-100)$	$[75-100)$	$[75-100)$
Weight \ Age / Gender	[50-60)	[60-70)	[70-80)	[80-90)	[90-100)
Male	$[75-100)$	$[75-100)$	$[75-100)$	$[75-100)$	$[50-75)$
Female	$[75-100)$	$[75-100)$	$[75-100)$	$[50-75)$	$[50-75)$

3) 从上述统计结果中可看出，一是 $Weight$ 的缺失率较高，二是 $Weight$ 取值的区间长度较大导致绝大多数分组的平均值都落在了 $[75-100)$ 范围内，因此最终决定将 $Weight$ 字段删除。

1.1.3 *Diagnosis* (1%)

根据题目要求，删除 *Diagnosis* 2 和 *Diagnosis* 3 字段，仅保留 *Diagnosis* 1 字段。

1.1.4 *Payer Code* (52%)

该字段与其他字段相关性不大，无法通过其他字段推测该字段的取值，且缺失率较高，因此将该字段删除。

1.1.5 *Medical Specialty* (53%)

该字段缺失率较高，故将其删除。

1.2 移除无关记录

根据题目要求，移除 *Discharge_disposition_id* 为 11(*Expired*)，13(*Hospice /home*)，14(*Hospice/medical facility*)，19(*Expired at home. Medicaid only, hospice.*)，20(*Expired in a medical facility. Medicaid only, hospice.*)，21(*Expired, place unknown. Medicaid only, hospice.*) 的相关记录。

2 数据分类与预测

2.1 朴素贝叶斯分类模型

2.1.1 算法流程

（具体代码见 */src/classification/bayesian_classifier.py*）

(1) 读取 *csv* 文件（*read_csv*，在代码中对应的函数名，下同）;

(2) 划分训练集和测试集（*split_dataset*）;

(3) 根据 *readmitted* 将训练集分为三类（*seperate_trainset*）;

(4) 统计不同种类、不同属性的数据的数目（*standardize_attributes*）;

(5) 根据上一步的结果，计算条件概率（*calculate_probabilities*）;

(6) 做出预测，计算正确率（*predict_testset*）。

2.1.2 遇到的问题及解决方案

在应用朴素贝叶斯分类模型进行分类的过程中，最需要解决的一个问题是：大部分属性为离散值，而有一部分属性为连续值。

根据课程内容及所查资料，离散值的条件概率即为该属性值出现次数和数据总数之比，而计算连续值的条件概率时，只需要认为其服从高斯分布即可。

2.2 KNN 分类模型

2.2.1 算法流程

（具体代码见 */src/classification/knn_classifier.py*）

(1) 读取 *csv* 文件（*read_csv*，在代码中对应的函数名，下同）;

(2) 划分训练集和测试集（*split_dataset*）;

(3) 计算出值为连续型的属性的最大值和最小值（*calculate_max_and_min*）;

(4) 做出预测，计算正确率（*knn*）。

2.2.2 遇到的问题及解决方案

遇到的问题和上面方法一样，处理方法为：计算距离时，离散值若相同则记为 0，不同记为 1；连续值记为值与该属性取值区间长度之比。

3 *logistic* 回归模型

实现该模型调用了 *sklearn* 的 *LogisticRegression*，只需要将离散值处理为具体数值（用自然数来做标记，比如该属性有 3 个取值，便将这三类分别标记为 1，2，3），再调用该库中的相关构造函数即可完成 *logistic* 回归验证。

4 分类模型测试结果对比

（*Ratio* 指训练集和测试集大小之比；以下结果均为多次测试取平均值所得）

Accuracy / Ratio / Model	0.67	0.75
朴素贝叶斯分类模型	51.73%	52.19%
KNN 分类模型	47.24%	47.75%

根据测试结果，朴素贝叶斯分类模型的准确率普遍要比 *KNN* 分类模型高，用时也更少；同时，*KNN* 分类模型的测试结果还相当不稳定，多次出现了 30% 左右的准确率。

logistic 回归模型所测得的准确率为 64.63%。