《云数据管理II -- 智能数据分析与决策支持》

2015080121 软工52 李在弦

文件目录

```
2015080121_李在弦
        / doc
                / 实验报告.pdf
        / src
                / input
                        / diabetic_data.csv
                / output
                        /（ 执行程序之后会有一些输出文件 ）
                / preprocess.py （预处理）
                / classification.py （分类）
                / kafang.py （卡方检验，单因素分析）
                / rlr.py（logistic回归，多因素分析）
                / statistic.py（统计）
```

实验环境

操作系统：Windows

IDE：PyCharm

编程语言：Python 3.6

运行设置

2015080121_李在弦/src/input/ 路径下放diabetic_data.csv

运行顺序 **preprocess.py → classification.py → kafang.py → rlr.py → statistic.py**

任务1 – preprocess.py

同一个病人仅保留第一次入院记录

先从 diabetic_data.csv 获取所有信息，并存储成字典(dict)形式。数据总共有 101766 条。有一些同一个 patient_nbr 会带着多个 encounter_id，其中 encounter_id 数字小的就是第一次入院记录。所以建立一个以 patient_nbr 为 key 的字典，然后每次发现同一个 patient_nbr 就比较它俩的 encounter_id 值，然后只保存 encounter_id 较小的数据。 这样删除的数据有30248条。剩下的有71518条数据。

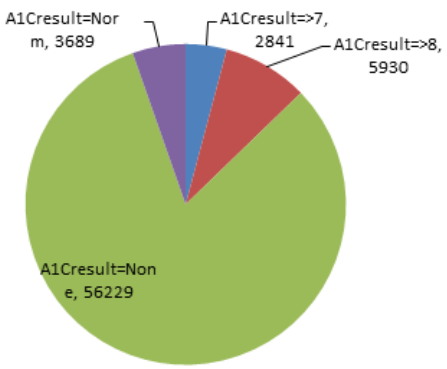
移除导致临终关怀或病人死亡的记录（discharge_disposition_id)

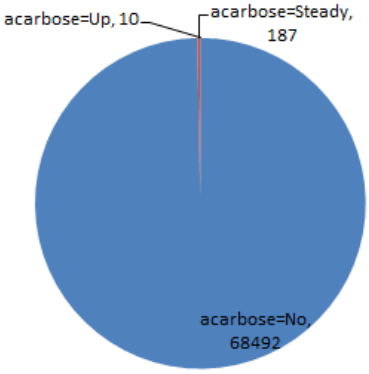
仔细看 IDs_mapping.csv 文件之后发现导致临终关怀或病人死亡的情况有
discharge_disposition_id = 11 (Expired),
discharge_disposition_id = 13 (Hospice / home),
discharge_disposition_id = 14 (Hospice / medical facility),
discharge_disposition_id = 19 (Expired at home. Medicaid only, hospice.),
discharge_disposition_id = 20 (Expired in a medical facility. Medicaid only, hospice.),
discharge_disposition_id = 21 (Expired, place unknown. Medicaid only, hospice.)
删除包含这些值的数据。剩下有69973条数据

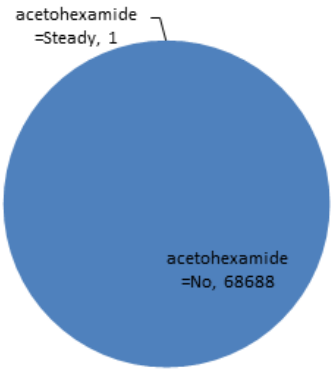
缺失值处理

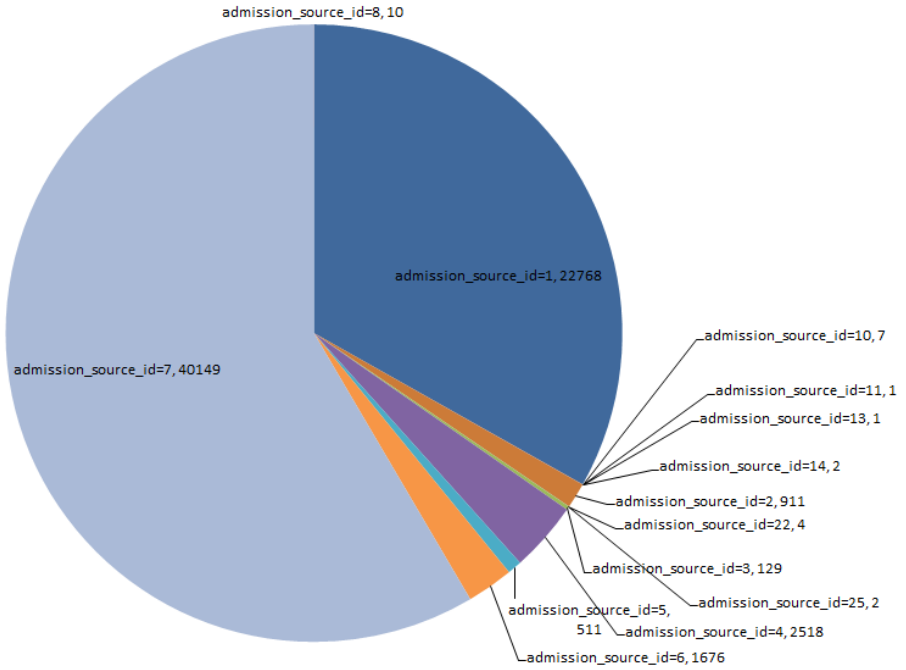
1. patient_nbr 和 encounter_id 属性对30天内再入院情况没有影响力，直接删除。
2. 'payer_code', 'medical_specialty' , 'weight' 这三个属性的缺失率太高，超过50%。直接删除
3. diag_1，diag_2，diag_3的属性值种类太多，用深度学习填充缺失值的效率太低（10%左右）。直接删除缺失这三个字段的数据（1284条数据）
4. 用深度学习处理缺失值之前，先筛选包含缺失值的数据（包含缺失值的数据有 12603 条，不包含的有56086条）
5. 为了用深度学习的方法处理缺失值，把Nominal形式的字段转换成布尔形式（比如，gender字段有 male 和 female，那么male是 [1, 0]，female是 [0, 1]）
6. 用tensorflow建立模型，对“admission_type_id” “ discharge_disposition_id” “ admission_source_id” “ race” “ gender”这五个字段做深度学习。

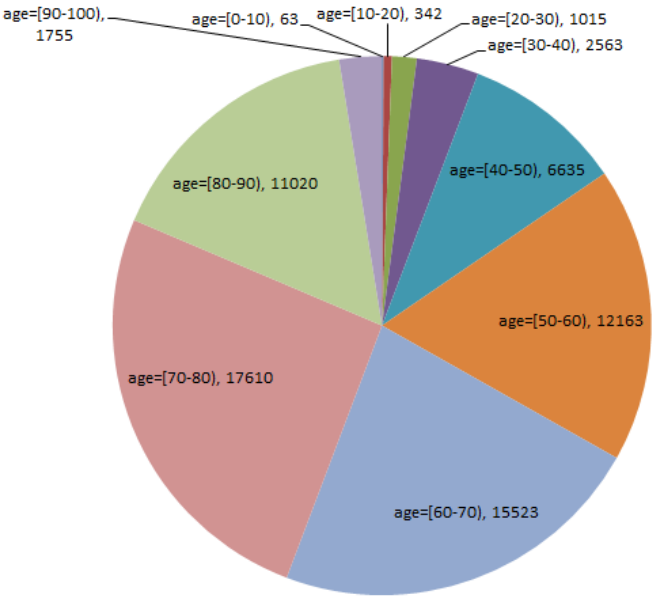
图表统计

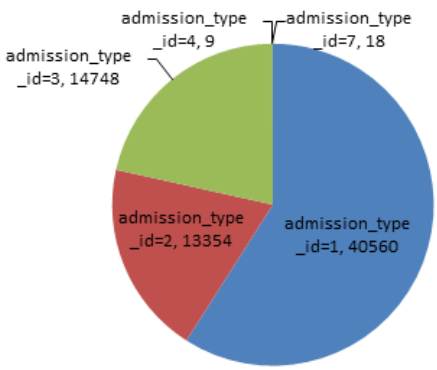

A1Cresult=Norm, 3689
A1Cresult=>7, 2841
A1Cresult=>8, 5930
A1Cresult=None, 56229

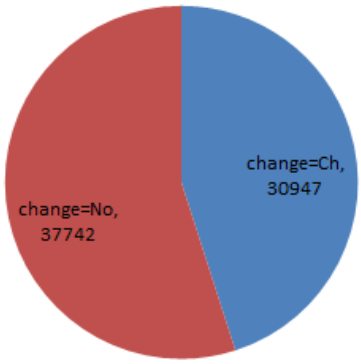

acarbose=Up, 10
acarbose=Steady, 187
acarbose=No, 68492

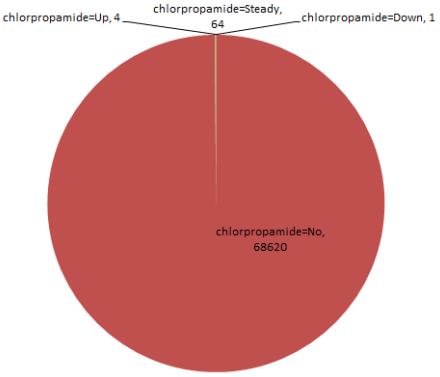

acetohexamide=Steady, 1
acetohexamide=No, 68688


admission_source_id=8, 10
admission_source_id=1, 22768
admission_source_id=7, 40149
admission_source_id=10, 7
admission_source_id=11, 1
admission_source_id=13, 1
admission_source_id=14, 2
admission_source_id=2, 911
admission_source_id=22, 4
admission_source_id=3, 129
admission_source_id=25, 2
admission_source_id=5, 511
admission_source_id=4, 2518
admission_source_id=6, 1676

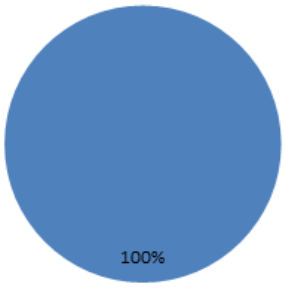

age=[90-100), 1755
age=[0-10), 63
age=[10-20), 342
age=[20-30), 1015
age=[30-40), 2563
age=[40-50), 6635
age=[50-60), 12163
age=[60-70), 15523
age=[70-80), 17610
age=[80-90), 11020


admission_type_id=4, 9
admission_type_id=7, 18
admission_type_id=3, 14748
admission_type_id=2, 13354
admission_type_id=1, 40560

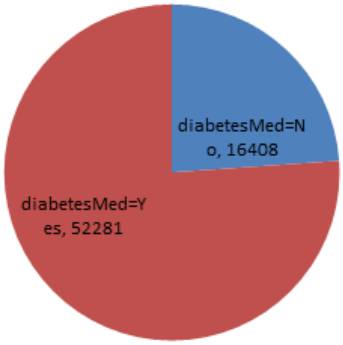


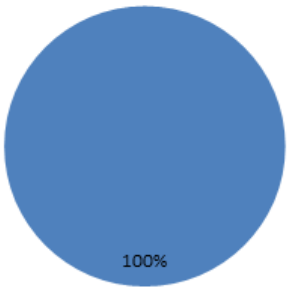


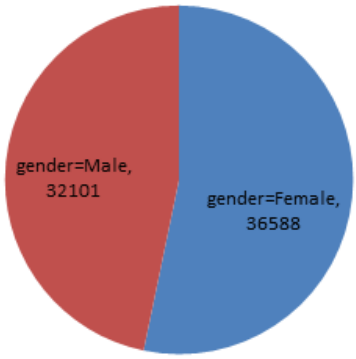
citoglipton=No

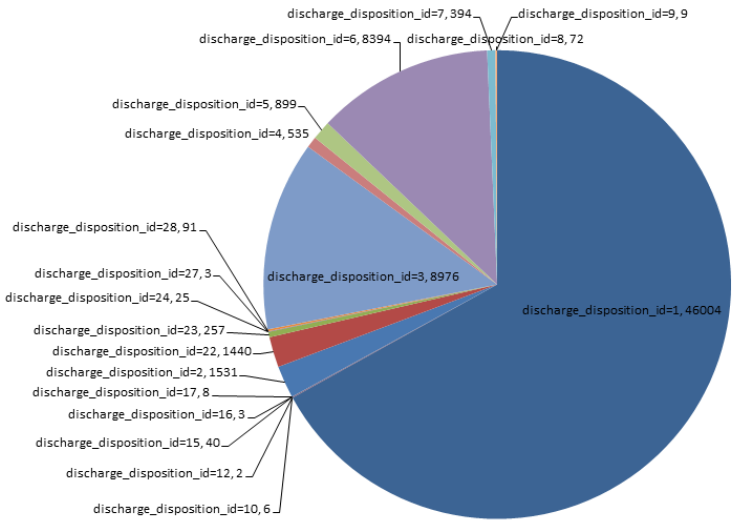


examide=No

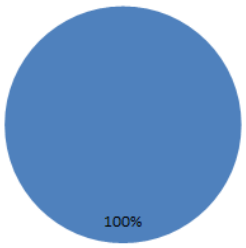


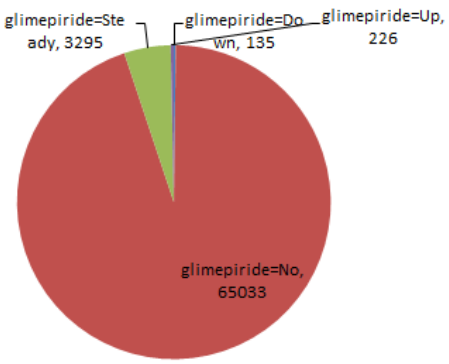


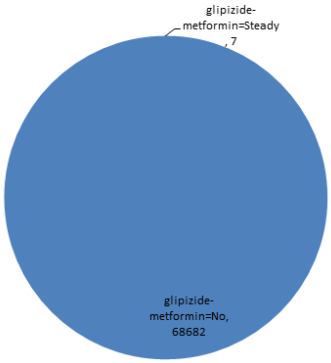


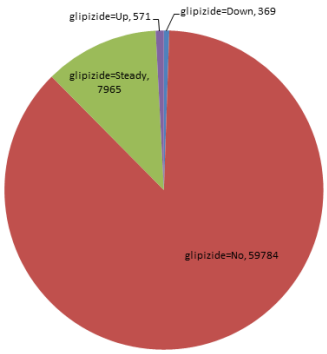


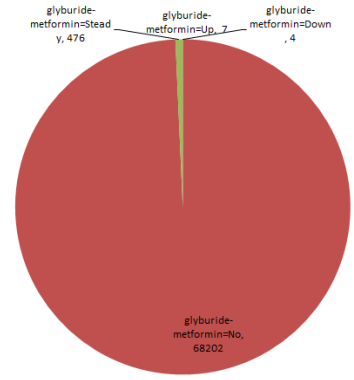
glimepiride-pioglitazone=No

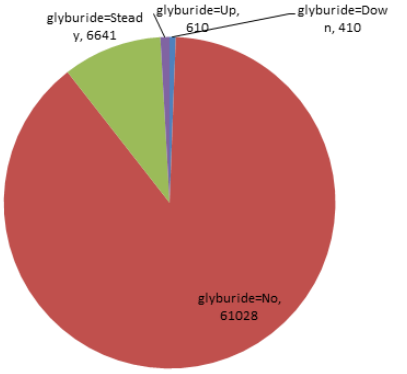


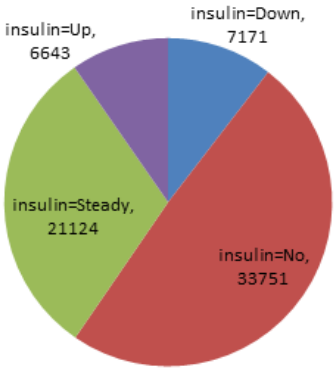


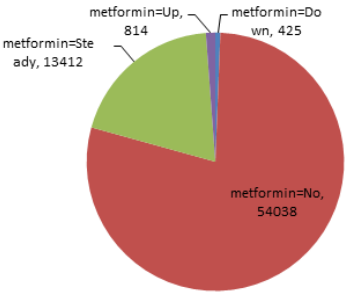


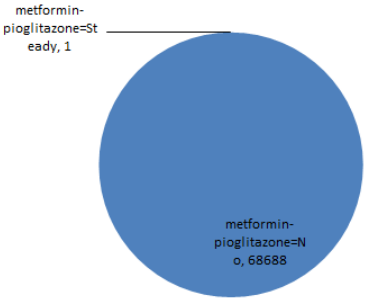


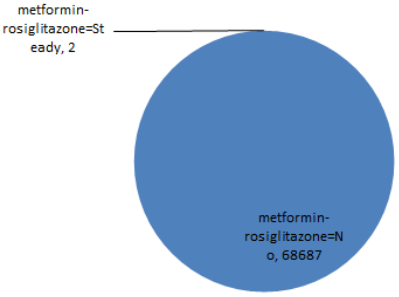

glyburide-metformin=Steady, 476
glyburide-metformin=Up, 7
glyburide-metformin=Down, 4
glyburide-metformin=No, 68202

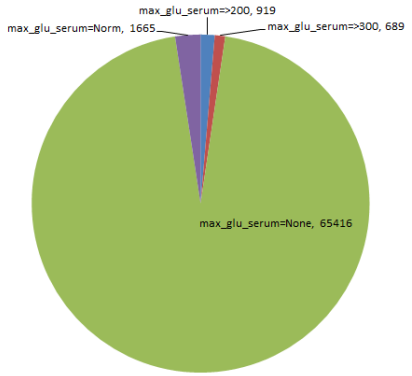

glyburide=Steady, 6641
glyburide=Up, 610
glyburide=Down, 410
glyburide=No, 61028

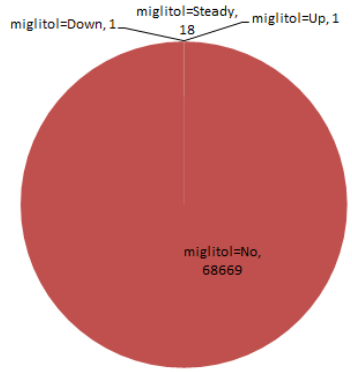

insulin=Up, 6643
insulin=Down, 7171
insulin=Steady, 21124
insulin=No, 33751

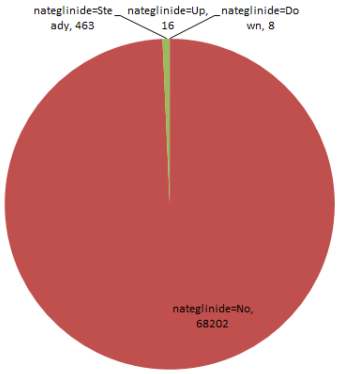

metformin=Up, 814
metformin=Down, 425
metformin=Steady, 13412
metformin=No, 54038

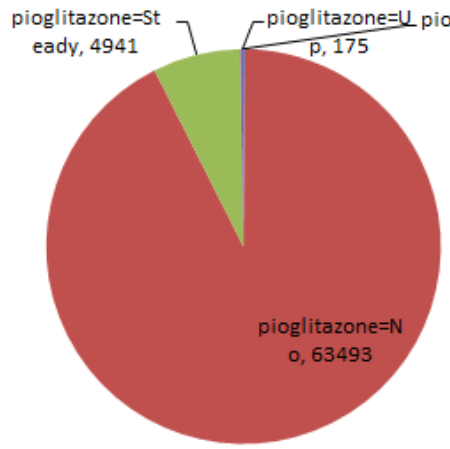

metformin-pioglitazone=Steady, 1
metformin-pioglitazone=No, 68688

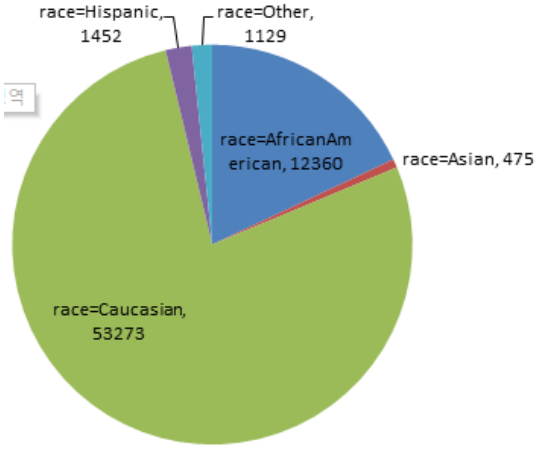

metformin-rosiglitazone=Steady, 2
metformin-rosiglitazone=No, 68687

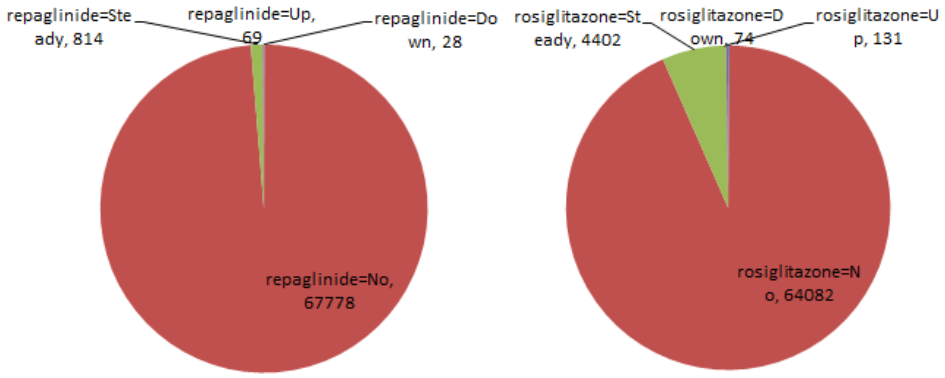

max_glu_serum=>200, 919
max_glu_serum=Norm, 1665
max_glu_serum=>300, 689
max_glu_serum=None, 65416

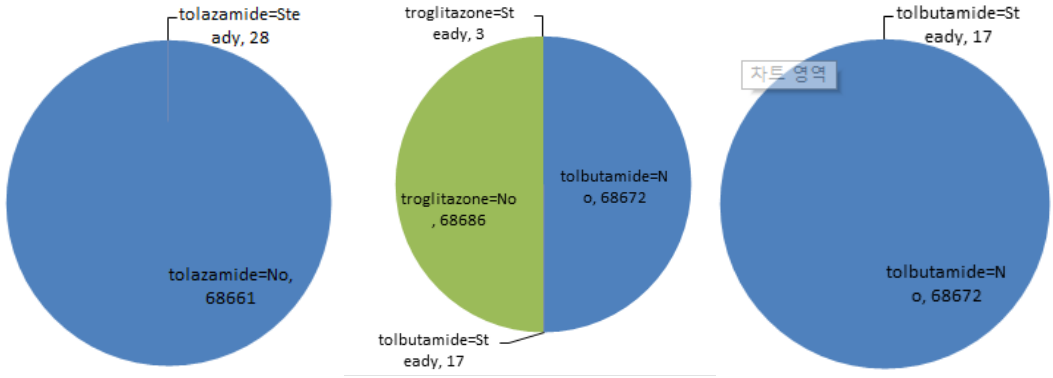

miglitol=Down, 1
miglitol=Steady, 18
miglitol=Up, 1
miglitol=No, 68669

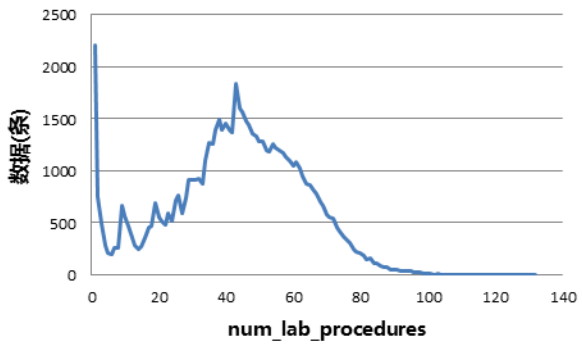

nateglinide=Steady, 463
nateglinide=Up, 16
nateglinide=Down, 8
nateglinide=No, 68202

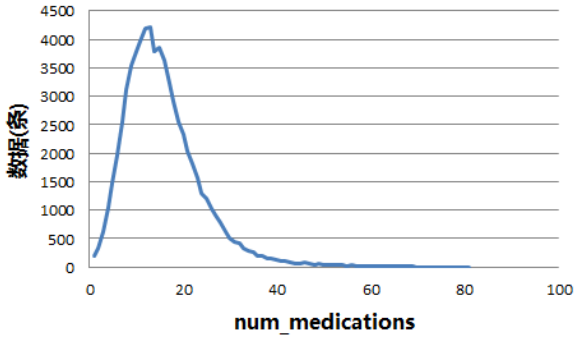

pioglitazone=Steady, 4941
pioglitazone=Up, 175
pioglitazone=Down, 80
pioglitazone=No, 63493

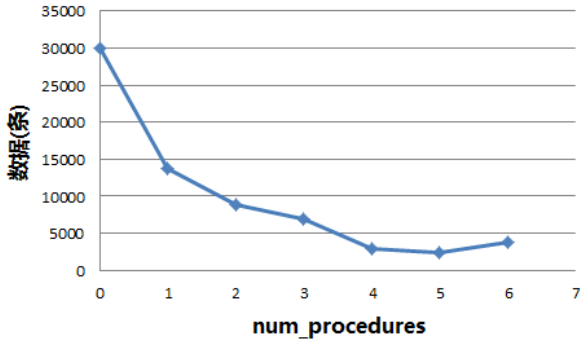

race=Hispanic, 1452
race=Other, 1129
race=AfricanAmerican, 12360
race=Asian, 475
race=Caucasian, 53273

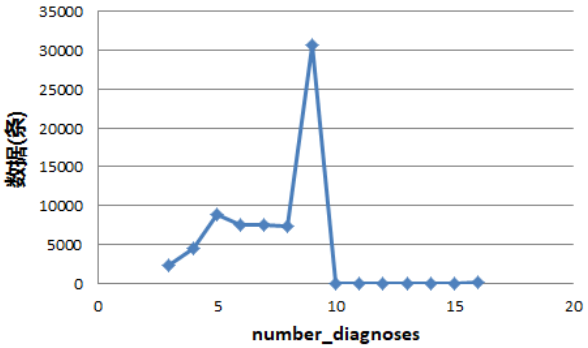

repaglinide=Steady, 814
repaglinide=Up, 69
repaglinide=Down, 28
repaglinide=No, 67778
rosiglitazone=Steady, 4402
rosiglitazone=Down, 74
rosiglitazone=Up, 131
rosiglitazone=No, 64082

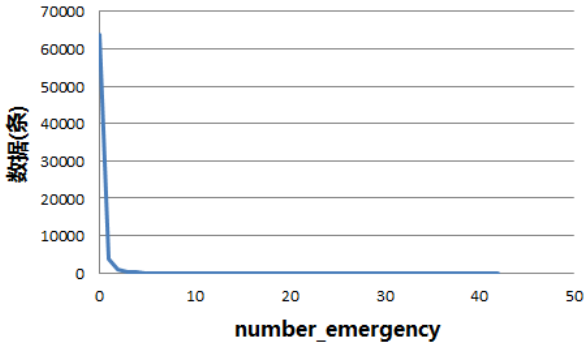

tolazamide=Steady, 28
tolazamide=No, 68661
troglitazone=Steady, 3
troglitazone=No, 68686
tolbutamide=No, 68672
tolbutamide=Steady, 17
tolbutamide=Steady, 17
tolbutamide=No, 68672

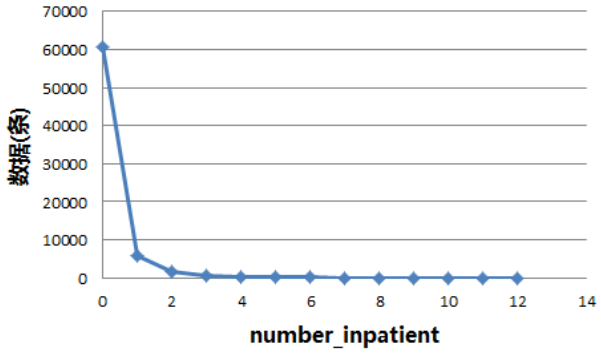

数据(条)
num_lab_procedures

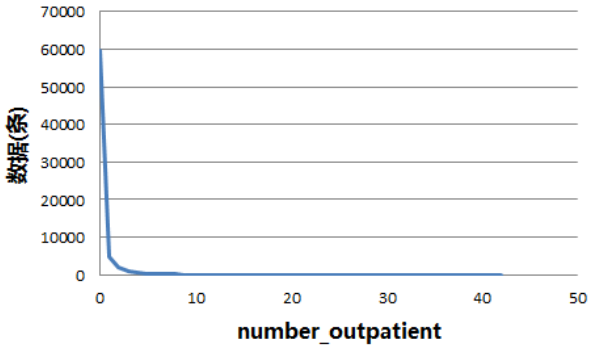

数据(条)
num_medications

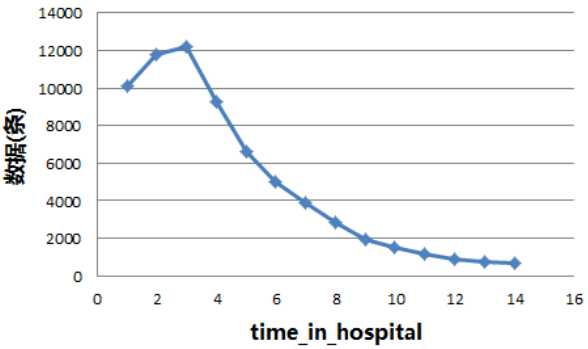

数据(条)
num_procedures

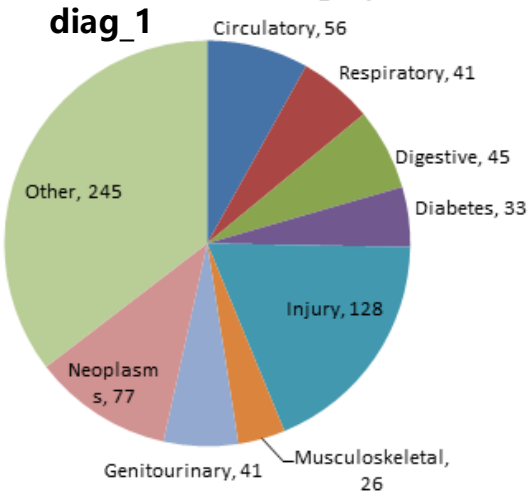

数据(条)
number_diagnoses

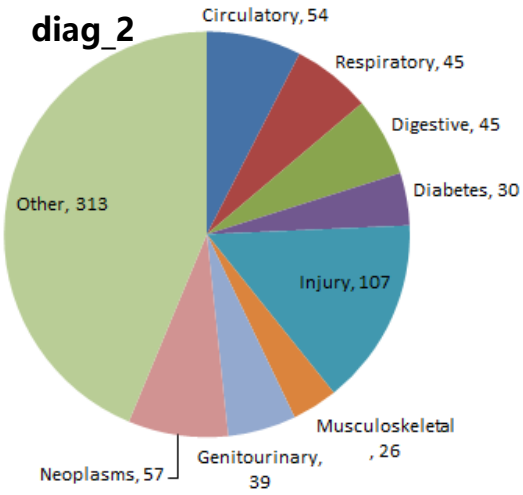

数据(条)
number_emergency

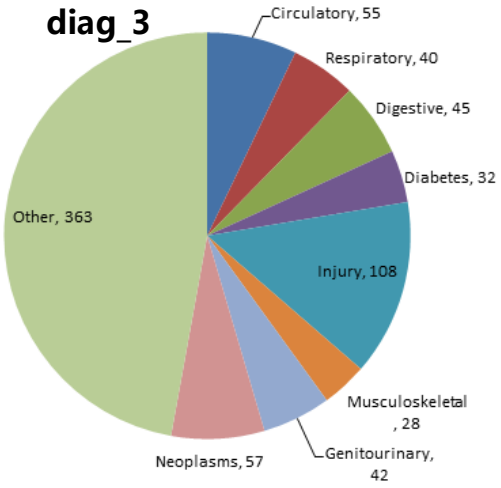

数据(条)
number_inpatient


70000
60000
50000
40000
30000
20000
10000
0
数据(条)
0
10
20
30
40
50
number_outpatient


14000
12000
10000
8000
6000
4000
2000
0
数据(条)
0
2
4
6
8
10
12
14
16
time_in_hospital


diag_1
Circulatory, 56
Respiratory, 41
Digestive, 45
Diabetes, 33
Injury, 128
Musculoskeletal, 26
Genitourinary, 41
Neoplasms, 77
Other, 245


diag_2
Circulatory, 54
Respiratory, 45
Digestive, 45
Diabetes, 30
Injury, 107
Musculoskeletal, 26
Genitourinary, 39
Neoplasms, 57
Other, 313


diag_3
Circulatory, 55
Respiratory, 40
Digestive, 45
Diabetes, 32
Injury, 108
Musculoskeletal, 28
Genitourinary, 42
Neoplasms, 57
Other, 363

任务2 – classification.py

划分训练集和测试集

为了合理划分训练集和测试集，建立了各个 race，gender，age的决策树。然后用这三个决策树来建立一个划分训练集和测试集的决策树，这个决策树的根节点是race，然后中间节点是gender，最底层节点是age。建立决策树之后在叶节点存储相应的数据。存储所有数据之后把各个叶节点分成7：3比例。其中70%的数据是训练集，30%的数据是测试集。

Bayesian分类

首先建立每个字段的决策树，然后把它们保存到C1（<30）和C2（>30，NO）相应的列表中。遍历所有训练集之后，在各个决策树的叶节点保存 P(X|Ci)值。用这个决策树列表计算 P(C1|X) 和 P(C2|X)值。比较它们两个值的大小，做Bayesian分类。

Neural Network Model分类

为了用深度学习的方法分类，把Nominal形式的字段转换成布尔形式。Numeric形式的字段直接使用数字值。为了更准确地做分类，label用三个类型， “<30” “>30” “NO”。用tensorflow建立深度学习模型。

比较分析各算法精度

Bayesian

→ **<30** accuracy： 0.056162246489859596

→ **Not <30** accuracy： 0.9757497719834755

```
check <30 accuracy: 0.056162246489859596
check no <30 accuracy: 0.9757497719834755
```

Neural Network

→ **<30** accuracy： 0.08268330733229329

→ **Not <30** accuracy： 0.9590106765384409

```
check <30 accuracy: 0.08268330733229329
check no <30 accuracy: 0.9590106765384409
```

发现Bayesian算法对‘Not <30’情况的分类效果比Neural Network算法的好一些，但是Neural Network算法对‘<30’情况的分类效果更好。Bayesian算法的分类效率比Neural Network算法稍微偏在‘Not <30’情况。所以觉得Neural Network算法的效率更好

任务3 – kafang.py，rlr.py，statistic.py

卡方检验分析 （单因素分析）

用 sklearn库的 SelectKBest 和 chi2。

单因素特征选择能够对每一个特征进行测试，衡量该特征和响应变量之间的关系，根据得分扔掉不好的特征，通过卡方检验分析获得25个最挂特征（k_score最高，k_p最小的25个因素）。

卡方分析的结果中p值越小，置信度越高。

prop	value	<30	>30	NO	k_score	k_p	r_score	k_valid	r_valid
age	[50-60)	873	3703	7587	51.42141	7.45E-13	0.495	TRUE	TRUE
age	[80-90)	1185	3835	6000	39.22408	3.78E-10	0	TRUE	FALSE
diabetesMed	No	1249	4596	10563	40.79928	1.69E-10	0.47	TRUE	TRUE
diag_1	250.7	87	189	243	37.61645	8.61E-10	0.225	TRUE	FALSE
diag_1	434	246	414	848	96.9283	7.19E-23	0.44	TRUE	TRUE
diag_1	786	155	963	1886	55.05417	1.17E-13	0.325	TRUE	TRUE
diag_1	820	120	209	455	37.40192	9.61E-10	0	TRUE	FALSE
diag_1	V58	45	32	38	126.5834	2.29E-29	0.925	TRUE	TRUE
diag_2	250	301	1153	3102	32.85922	9.91E-09	0.035	TRUE	FALSE
diag_2	342	40	39	104	36.54456	1.49E-09	0.19	TRUE	FALSE
diag_2	440	52	79	127	38.7396	4.84E-10	0.37	TRUE	TRUE
diag_3	250	647	2453	5865	36.31455	1.68E-09	0.15	TRUE	FALSE
discharge_disposition_id	1	3265	14604	28135	211.6015	6.14E-48	1	TRUE	TRUE
discharge_disposition_id	15	18	12	10	62.8828	2.19E-15	0.565	TRUE	TRUE
discharge_disposition_id	2	210	437	884	40.67222	1.80E-10	0.21	TRUE	FALSE
discharge_disposition_id	22	381	338	721	531.0696	1.65E-117	1	TRUE	TRUE
discharge_disposition_id	28	33	18	40	81.98837	1.37E-19	0.66	TRUE	TRUE
discharge_disposition_id	3	1207	2783	4986	211.7454	5.72E-48	0.66	TRUE	TRUE
discharge_disposition_id	5	185	247	467	145.4696	1.70E-33	0.87	TRUE	TRUE
num_lab_procedures	1	6211	21944	40534	625.3501	5.13E-138	0.375	TRUE	TRUE
num_medications	1	6211	21944	40534	352.0806	1.49E-78	0.1	TRUE	FALSE
number_emergency	0	6211	21944	40534	132.8575	9.71E-31	0.27	TRUE	TRUE
number_inpatient	0	6211	21944	40534	1403.438	3.76E-307	1	TRUE	TRUE
time_in_hospital	1	6211	21944	40534	415.6618	2.15E-92	0.5	TRUE	TRUE

logistic回归分析 （多因素分析）

用 sklearn库的 RandomizedLogisticRegression， LogisticRegression

使用随机特征选择方法，通过打乱设计的矩阵或者子采样的数据并，多次重新估算稀疏模型，并且统计有多少次一个特定的回归量是被选中。

通过logistic回归分析获得大概25个最挂特征（r_score大于0.25的因素）。

	prop	value	<30	>30	NO	k_score	k_p	r_score	k_valid	r_valid
2	age	[50-60)	873	3703	7587	51.42141	7.45E-13	0.495	TRUE	TRUE
3	age	[70-80)	1807	6089	9714	31.81706	1.69E-08	0.285	FALSE	TRUE
4	diabetesMed	No	1249	4596	10563	40.79928	1.69E-10	0.47	TRUE	TRUE
5	diabetesMed	Yes	4962	17348	29971	12.80455	0.000346	0.485	FALSE	TRUE
6	diag_1	428	443	1655	1772	27.21217	1.82E-07	0.355	FALSE	TRUE
7	diag_1	434	246	414	848	96.9283	7.19E-23	0.44	TRUE	TRUE
8	diag_1	786	155	963	1886	55.05417	1.17E-13	0.325	TRUE	TRUE
9	diag_1	V58	45	32	38	126.5834	2.29E-29	0.925	TRUE	TRUE
10	diag_2	440	52	79	127	38.7396	4.84E-10	0.37	TRUE	TRUE
11	diag_2	500	2	1	0	12.11214	0.000501	0.28	FALSE	TRUE
12	diag_3	250.6	95	263	299	23.44459	1.29E-06	0.3	FALSE	TRUE
13	diag_3	403	171	521	566	31.67684	1.82E-08	0.33	FALSE	TRUE
14	diag_3	V60	2	1	0	12.11214	0.000501	0.275	FALSE	TRUE
15	discharge_disposition_id	1	3265	14604	28135	211.6015	6.14E-48	1	TRUE	TRUE
16	discharge_disposition_id	15	18	12	10	62.8828	2.19E-15	0.565	TRUE	TRUE
17	discharge_disposition_id	22	381	338	721	531.0696	1.65E-117	1	TRUE	TRUE
18	discharge_disposition_id	28	33	18	40	81.98837	1.37E-19	0.66	TRUE	TRUE
19	discharge_disposition_id	3	1207	2783	4986	211.7454	5.72E-48	0.66	TRUE	TRUE
20	discharge_disposition_id	5	185	247	467	145.4696	1.70E-33	0.87	TRUE	TRUE
21	num_lab_procedures	1	6211	21944	40534	625.3501	5.13E-138	0.375	TRUE	TRUE
22	number_diagnoses	3	6211	21944	40534	57.32508	3.69E-14	0.53	TRUE	TRUE
23	number_emergency	0	6211	21944	40534	132.8575	9.71E-31	0.27	TRUE	TRUE
24	number_inpatient	0	6211	21944	40534	1403.438	3.76E-307	1	TRUE	TRUE
25	time_in_hospital	1	6211	21944	40534	415.6618	2.15E-92	0.5	TRUE	TRUE