IDA 课程作业实验报告

软件 51　2015013190　安彦哲

2018.11.22

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	$[75-100)$	$[75-100)$	$[75-100)$	$[50-75)$	$[50-75)$

并用平均值填补缺漏值。

3) 因为 $Unknown/Invalid$ 记录条数较少（3 条，两条 $[70-80)$ ，一条 $[60-70)$ ），所以选择手动处理（根据这两个年龄段的 $Weight$ 平均值，填补为 $[75-100)$）。

2

3