宿舍成员分配算法

算法的输入

性别相同的两组学生，按地域划分为省内组和省外组，每个人由一组已知的信息表示，比如学号、性别、起居时间段、兴趣爱好等等。

对于每个宿舍成员划分，主要要求

1. 尽量均匀来自两个组，比如保证省内2名、省外2名，最好不要出现全都是省内/外的情况；
2. 性格动静结合，但尽量不要出现3个安静1个活跃型这种情况；
3. 家庭人均收入尽量不要悬殊，比如一个500以下，三个3000以上；
4. 其它条件尽量相似。

思路分析

这个问题如果暴力实现的话，需要枚举的情况数为：

$$C_4^4 * C_8^4 * \ldots * C_n^4$$

也就是说，这是一个组合爆炸问题，求最优解是很耗时的一件事（n较大的时候基本是不可能做到的），所以我转而求近似解，通过贪心法来组合配对。

算法流程

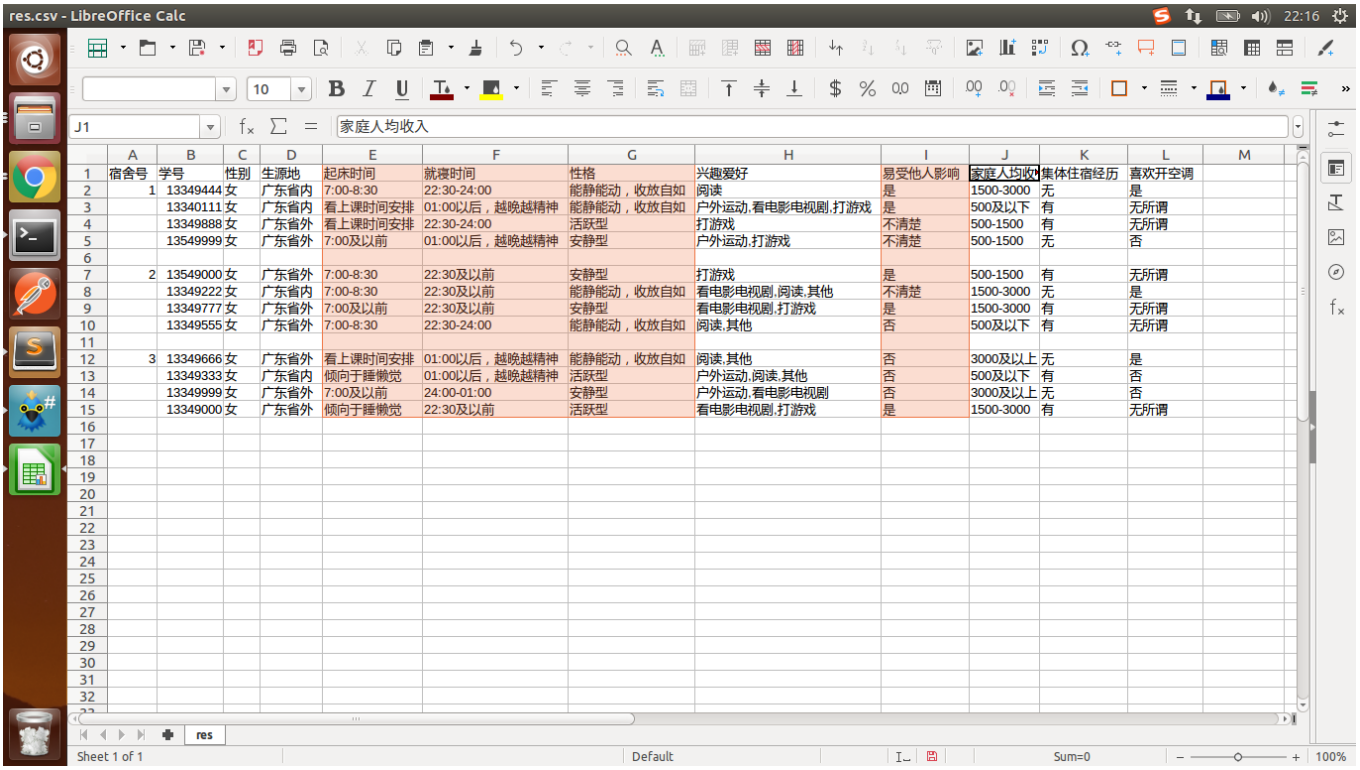
1. 均分两个组的人数，抽取人数较多的一组的部分人到另外一组，使得两组的人数相差不超过1；
2. 组内两两配对，根据相似度来配对，两人的特征值平均得到一个新的特征来代表这两个人的组合；
3. 将两个人组合后的特征在两个组之间配对，即1个GroupA的组合加1个GroupB的组合为一个宿舍，也是根据相似度最相近来组合。

细节

1. **从人数较多的组抽取多少人到另外一个组呢？**
   假设两个组的人数为x和y，并且有$x \leq y$，那么只需要抽取$\lfloor \frac{y-x}{2} \rfloor$个人给x就可以了，容易证明有$y - x \leq 1$。
2. **从人数较多的组抽取哪些人放到另外一组呢？**
   我的想法是，这些人要尽量不相同，因为抽取后它们就成为同一组的人了，如果是他们比较相似的话，很容易在“组内两两配对”阶段就组合到了一起。
   比如一般省外的同学比较多，那么组合后就变成了两个省外的同学，如果这样的话，那么在“组与组之间配对”的时候，很容易形成4个都是省外的同学的宿舍，不满足第一条约束。
   所以说，抽取出来的人要尽量不同。那怎么做到这一点呢？我就使用了kmeans，比如要抽取K个人，那么就人数较多的一组聚类分成K类，每类里面抽取那个离类中心最近的点（离簇中心最近，最能够表示那一类的特点，反过来也是与其它簇最不相同的点）。
   从另外的角度来看，不同的簇代表不同个性、志趣的人群，为了更多元化地进行group combination，这样子抽取也是比较合理的。
3. **kmeans会形成空簇，怎么解决这个问题？**
   这个问题有很多解决方法，我选取了容易实现的方法：增量更新簇的中心。
   因为当一个簇减少为只有一个点的时候，那么这个点就是这个簇的中心，保证了每个簇至少有一个点，绝对不可能形成空簇！
4. **人数参差不齐的时候怎么办？**
   比如省内有5个人，省外有12个人，那么省外分了(12-5)/2=3个人到省内组，变成8:9，省外组多了一个人。
   总体的想法就是——把落单的人凑到同一个宿舍里，不过具体得分类讨论，上面说到均分后有$y - x \leq 1$，那么总的情况可以分为：
   （1）当$x = y$时，xy要么都是奇数，要么都是偶数，偶数的话，假设二者都是2k人，那么总共4k，刚好完整；奇数的话，假设二者都是2k+1人，那么“组内两两配对”之后，每组刚好都剩下一个人，将其作为special单独一间宿舍。
   （2）当$y = x + 1$时，很明显此时xy必定是一奇一偶，此时经过“组内两两配对”之后，必定是人数为奇数的那个组剩下一个人没能配对（直接放到special宿舍里），而偶数组的配对数可能：
   a）跟奇数组的一样（比如人数分别为8和9，配对数都是4）；
   b）比奇数组多一个（比如人数分别为9和10，配对数分别是4和5）。
   配对数不一样也无所谓，总是用人数少的那一组来从人数多的那一组里面找配对就可以了，而最后如果还有剩下（情况b最后人数为10的那一组肯定是剩下两个人没有参与最终的配对），那么就将起放到special宿舍里。
5. **将落单的人分配到同一个宿舍真的好吗？**
   我觉得是没什么大的问题的，可以从“他们为什么会落单”的原因来看，配对的时候用的是”贪心法“，总是找当前群体里最相似的个体来配对。之所以落单，主要是因为他们在群体中与其他人不是特别相似，要么就是比较平均的特征，要么就是比较极端的特征，那么将他们分配到同一个宿舍的话，相似的可能性也比较大，再说了，还有最后的人工检查分配结果的阶段，如果发现确实不合适，可以略做调整或重新分过。

分配结果

目前的数据是我自己随便造的，没有真实的数据来测试，目前的分类结果如下（由于kmeans的初始化是随机选择中心点的，所以每次运行得到的结果不一定一样）：



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	宿舍号	学号	性别	生源地	起床时间	就寝时间	性格	兴趣爱好	易受他人影响	家庭人均收	集体住宿经历	喜欢开空调	
2	1	13349444	女	广东省内	7:00-8:30	22:30-24:00	能静能动，收放自如	阅读	是	1500-3000	无	是	
3		13340111	女	广东省内	看上课时间安排	01:00以后，越晚越精神	能静能动，收放自如	户外运动,看电影电视剧,打游戏	是	500及以下	有	无所谓	
4		13349888	女	广东省外	看上课时间安排	22:30-24:00	活跃型	打游戏	不清楚	500-1500	有	无所谓	
5		13549999	女	广东省外	7:00及以前	01:00以后，越晚越精神	安静型	户外运动,打游戏	不清楚	500-1500	无	否	
6													
7	2	13549000	女	广东省外	7:00-8:30	22:30及以前	安静型	打游戏	是	500-1500	有	无所谓	
8		13349222	女	广东省内	7:00-8:30	22:30及以前	能静能动，收放自如	看电影电视剧,阅读,其他	不清楚	1500-3000	无	是	
9		13349777	女	广东省外	7:00及以前	22:30及以前	安静型	看电影电视剧,打游戏	是	1500-3000	有	无所谓	
10		13349555	女	广东省外	7:00-8:30	22:30-24:00	能静能动，收放自如	阅读,其他	否	500及以下	有	无所谓	
11													
12	3	13349666	女	广东省外	看上课时间安排	01:00以后，越晚越精神	能静能动，收放自如	阅读,其他	否	3000及以上	无	是	
13		13349333	女	广东省内	倾向于睡懒觉	01:00以后，越晚越精神	活跃型	户外运动,阅读,其他	否	500及以下	有	否	
14		13349999	女	广东省外	7:00及以前	24:00-01:00	安静型	户外运动,看电影电视剧	否	3000及以上	无	否	
15		13349000	女	广东省外	倾向于睡懒觉	22:30及以前	活跃型	看电影电视剧,打游戏	是	1500-3000	有	无所谓	

比如上图选中的列划分还算不错，不过其它列可能是因为比重比较小以及这个算法还有较大的改善的空间等原因，划分得不够好。

后续工作

其实我一直在考虑如何衡量分配结果的好坏，因为要求是多个维度上的相似，且各个维度对总体的比重大相径庭，用肉眼来看的话，分在同一宿舍里的人，有的维度上挺好的，但有的维度上又不太好，目前有一个想法就是用kmeans的SSE定义，就是同一宿舍内距离起“中心点“的距离的平方和，将所有宿舍的SSE加起来就的到了总体的误差是怎样的。

但是关键是，有了衡量标准，**怎么根据这个标准使得算法自动去做得更好呢**？有待后续完善。

目前还存在的问题就是，由于用的是最简单的kmeans，它聚类的结果**严重受限于初始化**，而初始化又是**随机的**，所以划分的结果**比较不稳定**（也可能是当前的测试数据量太少了，只有12个），这也是后续改善算法的一个方向——如何较好地初始化kmeans的簇中心。