2023 深圳赛(东三省)A 题第四小问
代码说明

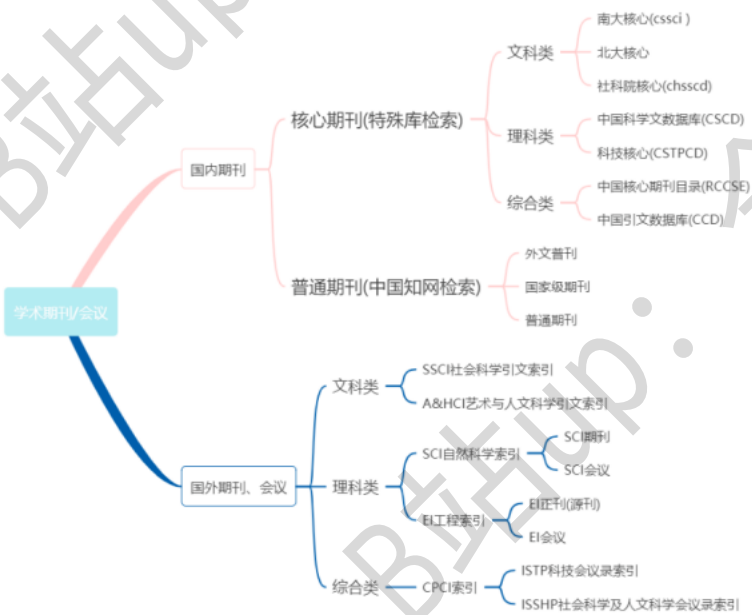
(本文档由 B 站 UP：全糖奶茶屋提供)

特别提示：本次东三省的 ABCD 题在赛后，均可转为 EI 国际会议，一份文章两份成果. 8 月即可录用!!!

只需要把您的文章交给我们，剩下的修改翻译，由我们全部负责，所有价格共计 5499(含一切版面费)，正规公司，合同保障，不能发表全额退款.

含金量：SCI源刊 > SCI会议 > EI源刊 > EI会议（权威会议）> 中文核心期刊 = 南大核心 > EI会议（一般型） > 国家级期刊 ≥ 外文普刊 > 省级期刊 > 一般普刊

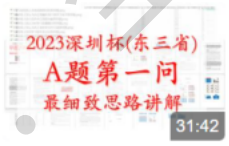
大家在选择期刊的时候一定要确定是可以被哪个库检索到的!!!



2023 深圳赛(东三省)A 题第四小问
代码说明

所有代码相关运行过程，都在 B 站有视频，请参照每一小问的视频理解模型，并且根据视频运行代码!!!

本小问的详细讲解，代码如何运行，软件如何安装视频已经上传，请参看 B 站视频：



2023深圳杯(东三省)数模竞赛A题第一问详细讲解

2023年07月26日 12:05:22

73 0 1 0 1 0 2

1. 缺失值的处理

为了将居民进行合理分类，本文档由 B 站 up：全糖奶茶屋提供本文主要居民的身体指标为居民的特征进行数据处理. 居民的身体指标主要有:年龄，身高，体重等 14 项特征.

对居民的身体指标进行缺失值统计，如图 1 所示.

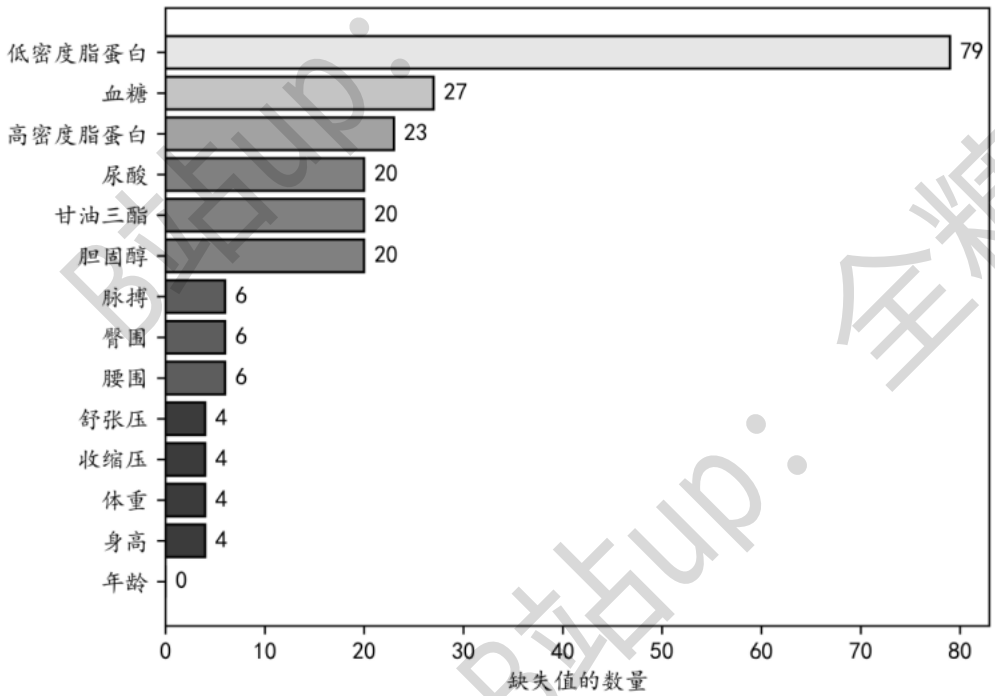


图 1:居民各项特征的缺失值数量

如图 1,缺失值最多的身体指标为低密度脂蛋白,有 79 个居民在该数据上存在缺失,其他特征缺失数量均不足 30.本文档由 B 站 up：全糖奶茶屋提供由于本文的总样本量为 17709,存在缺失值的数据仅占 0.5%,故本文对存在缺失值的样本进行剔除,不影响总体的样本分布.处理后的数据如表 1 所示.

表 1:去除缺失值后的数据

ID	年龄	身高	体重	腰围	臀围	收缩压	舒张压	脉搏	胆固醇	血糖	高密度脂蛋白	低密度脂蛋白	甘油三酯	尿酸
10001	40	161.5	50.5	68	92	98	62	70	3.79	4.61	1.59	1.88	0.91	210.4
10002	29	166	48.5	60	88	104	70	80	4.24	4.59	1.59	2.17	0.81	270.6
10003	47	183	64	77	95	104	70	80	4.84	4.13	1.33	2.92	1.59	298.2
10004	39	177.2	78.5	89.1	106	126	82	72	5.2	4.68	0.92	3.6	1.84	327.9
10005	55	173.6	73	86.8	97	106	70	70	5.1	4.41	1.37	3.15	1.87	350.9
10006	54	174.2	56	67.2	88.4	114	78	66	4.67	4.76	1.38	2.85	0.42	360.3
10027	60	152.4	49	65	83	100	66	88	5.21	4.72	1.88	2.92	0.84	261.2
……														
17709	37	171.4	53	71	90	100	68	70	4.85	3.94	1.27	2.8	0.9	341.8

为了使得聚类结果不受数据本身的量纲影响,本文对去除缺失值后的数据进行归一化处理.

2. 对去除缺失值后的数据建立 kmeans 模型进行聚类

本文对问题一的数据进行空间和时间的划分，本文档由 B 站 up：全糖奶茶屋提供从而精确描述共享汽车使用的分布情况。对空间划分采用 k 均值聚类算法(k-means Clustering Algorithm)。该算法是一种迭代求解的算法，对数据集中各个样本点的进行相似度比较，即根据各个采样点的经纬度数据计算距离。k-meas 模型的核心思想是设置样本中心，将样本中心附近的样本点划分为同一簇，即可以对数据集中的样本进行空间上的划分。

表 2：kmeans 流程

步 数	具 体 操 作
1	随机产生 k 个区域中心
2	对每个样本点计算区域中心的距离，并分配到较近的区域
3	如果分配结果不改变，则结束，否则继续下一步
4	计算新的区域中心
5	跳转到第 2 步

STEP 1: 利用肘部法确定聚类数量 k

其中 k-means 的代价函数为误差平方和公式，即公式（1）

$$SSE = \sum_{i=1}^{m} [dist(x_i, u_{C_i})]^2 \qquad (1)$$

其中 $dist(\bullet)$ 为欧式距离计算公式，u_{C_i} 为样本 x_i 所在簇的中心。

簇中心的计算公式为

$$u_i^{(j)} = \frac{1}{|C_i|} \sum_{x \in C_i} x^{(j)} \qquad (2)$$

其中 $u_i^{(j)}$ 簇 C_i 中心的第 j 个特征，$|C_i|$ 为簇 C_i 中样本数量。

将 k 值依次取 1-9 代入 k-means 算法中计算得到 SSE，如下图。

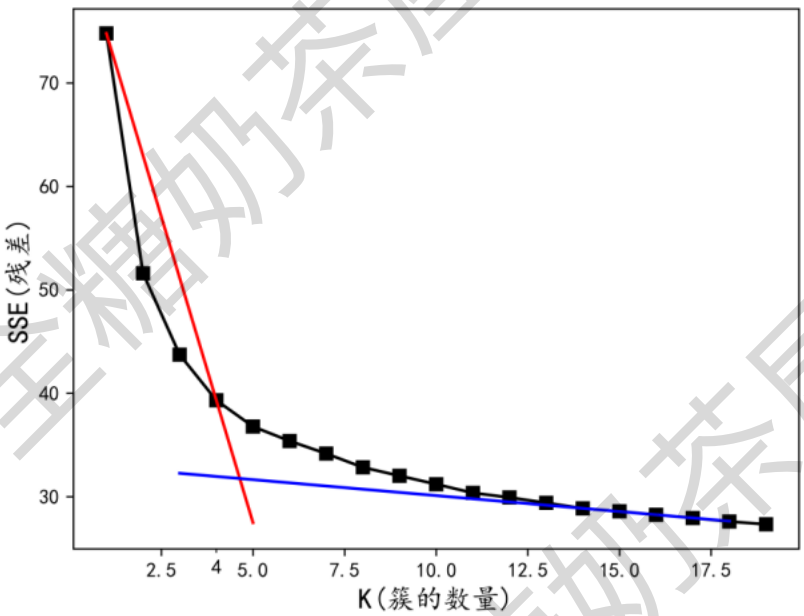


图 2:SSE 随 k 变化折线图

如图 2 所示，当 k<5 时 SSE 随聚类的数量增加而迅速下降，当 k>5 时 SSE 随聚类数量增加下降速度减缓. 因此 k=5 时可以认为簇的结构相对稳定，故对居民分类的数量为 5 类.

STEP 2: 将 k=5 代入模型进行聚类

将聚类结果进行向二维映射,并进行二维可视化.

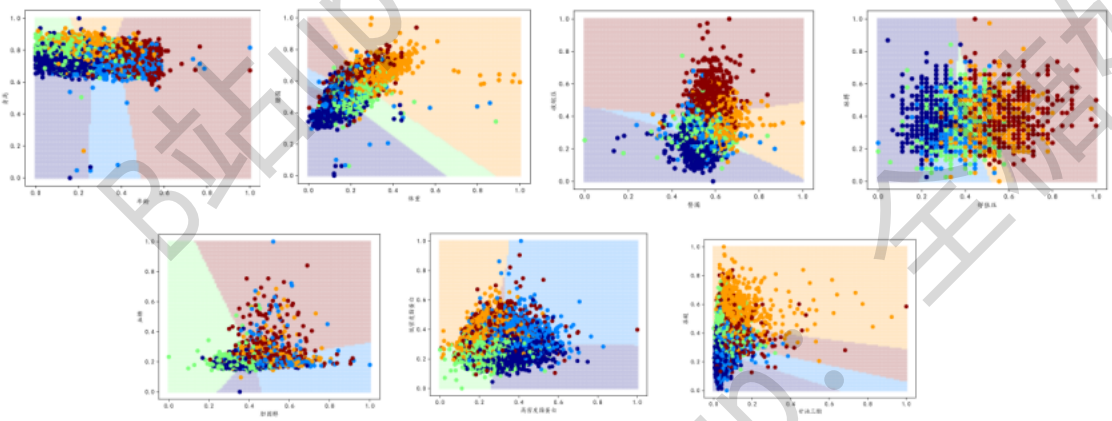


图 3: 聚类结果的二维可视化

如图 3 所示,不同颜色即为对象在二维平面上映射的分类结果,本文档由 B 站 up: 全糖奶茶屋提供背景区块为决策边距在二维平面上的映射区域.其中 5 个不同类别的聚类中心如表 3 所示.

表 3: 各类别聚类中心

类别	年龄	身高	体重	腰围	臀围	收缩压	舒张压

0	0.1755	0.7337	0.1325	0.438	0.5302	0.2235	0.294
1	0.3467	0.73	0.1782	0.5123	0.5582	0.2946	0.3679
2	0.1863	0.7808	0.2106	0.516	0.5605	0.2934	0.3757
3	0.2473	0.7971	0.3014	0.617	0.6049	0.3454	0.4599
4	0.4338	0.7415	0.2215	0.577	0.5803	0.4803	0.5355

类别	**脉搏**	**胆固醇**	**血糖**	**高密度脂蛋白**	**低密度脂蛋白**	**甘油三酯**	**尿酸**
0	0.3649	0.3781	0.1894	0.3387	0.2574	0.0319	0.2455
1	0.3344	0.4745	0.2065	0.3413	0.3614	0.0496	0.2864
2	0.3586	0.3713	0.1941	0.2499	0.2696	0.0484	0.3849
3	0.3613	0.473	0.2076	0.2342	0.3725	0.0967	0.5117
4	0.3872	0.4695	0.2302	0.2804	0.3646	0.0733	0.3696

将聚类中心的数据逆归一化恢复为正常数据,如表 4 所示.

表 4: 聚类中心逆归一化数据

类别	**年龄**	**身高**	**体重**	**腰围**	**臀围**	**收缩压**	**舒张压**
0	45.49	158.86	51.68	70.93	89.33	101.53	66.46
1	61.59	158.32	58.13	79.4	93.48	112.19	73.11
2	46.51	165.82	62.7	79.82	93.82	112	73.82
3	52.24	168.23	75.5	91.33	100.4	119.82	81.39
4	69.78	160.02	64.24	86.78	96.76	140.05	88.2

类别	**脉搏**	**胆固醇**	**血糖**	**高密度脂蛋白**	**低密度脂蛋白**	**甘油三酯**	**尿酸**
0	73.73	4.34	4.77	1.36	2.52	0.91	238.18
1	71.41	5.44	5.21	1.37	3.4	1.39	263.09
2	73.26	4.26	4.89	1.1	2.62	1.36	323.11
3	73.46	5.42	5.23	1.06	3.49	2.67	400.42
4	75.43	5.38	5.8	1.19	3.42	2.03	313.83

根据表 4 的数据可以看出不同类别的人群在年龄本文档由 B 站 up：全糖奶茶屋提供、身高、体重、腰围、臀围、血压、胆固醇、血糖等方面有些差异：

- 类别 0(可能是正常人群)：
  年龄较年轻, 体重指数正常, 血压正常, 胆固醇正常, 血糖正常。
- 类别 1(可能有超重/肥胖症)：
  年龄较大, 体重、腰围和臀围偏大, 血压偏高, 胆固醇偏高, 血糖正常。
- 类别 2(可能有超重症)：
  年龄适中, 体重、腰围偏大, 血压偏高, 胆固醇正常, 血糖正常。
- 类别 3(可能有肥胖症)：
  年龄适中, 体重非常之大, 腰围很大, 血压偏高, 胆固醇偏高, 血糖正常。
- 类别 4(可能有高血压)：
  年龄较大, 体重正常, 血压很高, 胆固醇正常, 血糖偏高。

所以不同类别之间在体重、血压、血脂方面有明显的差异, 这可能跟年龄、生活方式、运动、饮食等因素有关。

针对不同类别的人群，本文给出以下健康建议：

- 对于类别 0 的人群：保持良好的生活习惯，本文档由 B 站 up：全糖奶茶屋提供坚持适量运动，保持健康的体重，注意饮食结构。
- 对于类别 1 的人群：适当控制饮食卡路里，多吃高蛋白低脂肪食物，适量运动以降低体重。定期监测血压和胆固醇，必要时服用降压药或降胆固醇药。
- 对于类别 2 的人群：控制饮食卡路里摄入，适当增强有氧运动量，缩小腰围。定期监测血压。
- 对于类别 3 的人群：积极改变生活方式，控制卡路里摄入量，同时适量进行有氧运动。可能需要药物控制血压和胆固醇。
- 对于类别 4 的人群：适当控制饮食中的碳水化合物摄入，增强有氧运动量。定期监测血糖，根据医生建议服用降糖药物。