数据分析课程教学中存在问题和浅析

李志刚

武汉软件工程职业学院　湖北武汉　430079

【摘　要】数据分析存在的问题有：python基础没有学好，三维数组置换难以理解，随机数掌握不好，不会对数据进行可视化，针对这些问题笔者给出了解决方案。

【关键词】转置、随机数、可视化

Existing Problems and Brief Analysis in the Curriculum Teaching of Data Analysis

Zhigang Li

Wuhan Software Engineering Vocational College，Wuhan City, Hubei Province 430079

Abstract: The problems existing in data analysis are: the python foundation is not learned well, the 3 d array replacement is difficult to understand, the random number is not good, the data will not be visualized, the author gives a solution to these problems.

Key words: transpose, random number, visualization

人工智能专业的学生学习数据分析会有如下问题：

第一：python程序设计学得不好，对循环、列表、元组等知识掌握很差，特别是需要重点掌握的切片。

第二：数组的转置掌握较差，特别是三维数组转置，因为空间想像力不够，不知道三维是如何变化的。

第三：对随机数的使用不理解，不知道随机数在生活中有什么应用，不知道什么时候用rand()，什么时候用randn()，什么时候用randint()。

第四：对异常值的处理方法不理解，不知道如何找到异常值，找到异常值方法的适用范围，异常值一定是不合理的值吗?

第五：对数据可视化不理解，不明白为什么要划分子图，不知道什么时候用直方图、条形图、水平条形图、折线图、散点图、堆积区域图、饼状图和雷达图，也弄不懂各种图中每个参数的意义。

第六：不明白什么是重采样，也不知道重采样的作用，更不能理解滑动窗口有什么作用。

第七：对文本数据分析的过程不理解，不知道中文和英文的分析有什么区别。

第八：对综合项目案例不知道该如何去分析

针对以上问题,有以下解决方案：

第一：在学习数据分析之前，花一周的时间给学生复习python程序设计的知识：包括顺序，选择，循环三种结构，特别是多重循环；理解元组，列表之间的区别，对于一维数组和多维数组的切片要重点复习。

第二：对于二维数组的转换比较好理解，就是行列进行转换，比如三行四列的数组进行置换就可以得到四行三列的数组。对于三维数组学生很难理解，可以这样举例：对于一个2*3*4的三维数组，数组里面的元素为1到24，比如数字2，原来的位置的维度是0,0,1；如果进行转换后维度变成了3*4*2，那么2这个数字转换后的位置的维度就变为0,1,0。对于空间想像能力不够的同学来说，这是一个比较容易理解的方法。还可以这样思考，三维转换可以考虑为坐标辆的转换，比如是转换是的维度为(1,2,0)，意思是原来的Y坐标值转化为X坐标，原来的Z坐标转化为Y坐标，原来的X坐标转换为Z坐标。

第三：随机数要根据实际情况来帮助学生理解。rand()一般用于产生0–1的随机数，比如在游戏中怪物的移动有四个方向，左上，右上，左下，右下。Rand()可以产生0–1的随机数，如果产生的随机数在0–0.25之间，怪物就向左上运动；如果随机数在0.25–0.5之间，怪物就向右上运动；如果随机数在0.5到0.75之间，怪物就向左下运动；如果随机数在0.75–1之间，怪物就向右下运动。randn()是指产生平均数为0，方差为1的随机数，它产的生的数据不会局限在在0–1的范围内。randint()可以产生随机整数，可以用于验证码和斗地主发牌。对于验证码，可以先把十个数字，26个大写字母和26个小写字母都放在一个数组中，然后随机生成0–51的随机数，随机数作为数组的下标，取到相应下标对应的数组元素，循环四次就可以得到一个验证码。如果想在验证码中加入汉字，可以考虑在数组中加入你想要的汉字即可。对于斗地主发牌也是类似的，随机产生0–53的随机数，然后根据点数取

相应的牌，但存在一个问题，就是如何出现重复的牌该如何处理？有三种方法，第一种就是每次产生的随机数放入一个数组中，新产生的随机数和数组中的依次比较，如果有一个相同就重新产生随机数，但这种方法在最后几个随机数的时候会浪费大量时间，影响程度的效率，不太推荐这种解决方法。第三种方法每次产生一个随机数，在数组中取出这个随机数所对应的牌，然后把这张牌从数组中删掉，下次就在新数组中重新随机，这样就能保持每次的牌不重复。当然还有第三种方法：第一次是产生0–53的随机数，取出相应数组的元素，与第一个元素进行交换；第二次产生的是1–52的随机数，取出对应元素与第二元素交换；第三次产生2–53的随机数，依次类推，就可以产生不重复的随机数。

第四：异常值的检测主要包括基于拉依达原则和基于箱形图检测，前者适合于符合正态分布的数据，比如说学生的期中和期末考试成绩，正态分布的第一个参数是平均值，第二个参数是方差，如果数据落在平均值减去3倍的方差和平均值加上3倍的方差这个范围内，那么它就是正常值，否则就是异常值。而后者适合于所有数据，比如说餐饮的菜价，每个区的经济数据。检测可以找出异常值，但异常值不一定是不合理的，要根据实际情况来讨论。如某个区是经济特区，经济数据远超其它区，从箱形图来看它是异常数据，但也是合理的。

第五：数据可视化主要是为了将数据以图表化的形式呈现出来，在一个图表中有时候会需要包括多个图，如果把多个图都画在一起会显示得很杂乱，这时候就需要用到子图，经如可以分成3*3的子图，就可以选择合适的位置画图形，如果对划分的子图不满意，后面可以把子图修改为3*1。常见图表包括折线图、直方图、水平柱状图、堆积区域图、饼状图和雷达图。直方图可以表示不同数据的差异，比如同一个班的数学成绩分布，落在各个分数段的人数可以画一个直方图，可以很清楚的看出在各个分线段分数的人数的差值，它的第一个参数代表数据源，第二个参数bins代表产生的条柱个数，比如高中数学按照十分一个分数段，可以分为15个条柱。折线图是用折线将各个数据点连在一起形成的，可以表示同一数据不同时间的变化，比如可以表示一个学生在不同学期的平均分的变化，它有两个参数，第一个参数代表横坐标，第二个参数代表纵坐标，也可以只设置一个参数代表纵坐标。饼图是把数据以扇形的方形显示，可以表示一类数据占总体的比例，比如网民中各个年龄段的人数百分比就可以考虑饼图，饼状图的第一个参数是数据源，第二个参数explode是饼状图的每一块离圆心的距离，这个参数需要用数组的形式表示，如果饼状图包括8个扇形，那么这个数组中就需要定义8个值，第三个参数labels扇形对应的文本，第四个参数autopct是百分比的格式，一般是设置保留小数点后几位，第四个参数pctdistance是代表扇形对应的数值距圆心的距离和半径的比值，它是一个相对值，第五个参数shadow:表示是否显示阴影，第六个参数labeldistance标签文本绘制位置和半径的比值，第七个参数startangle表示开始绘制的角度，角度以逆时针来计算的，第八个参数表示radius半径。对于一个合理的饼状图来说，其中数据源，labels和autopct这三个参数是必须的，不然就算能画出饼状图也没有任何意义。条形图和直方图有点类似，不过它可以表示多个数据的对比，比如旅游的时候可以对不同旅游景点的面积和旅游的人数生成一个条形图。水平条形图主要有两个参数，第一个参数表示纵坐标，第二个参数代表横坐标，这和一般的条形图不一样，注意区别。在水平条形图中有时候需要把每一块的数值都要显示出来，这时候要使用plt.text方法，这个方法的第一个参数是数值的横坐标，第二个参数是数据的纵坐标，第三个参数是显示的数值。散点图一般用于查看数据的分布情况，第一个参数是横坐标，第二个参数是纵坐标。堆积区域图一般有两个参数，第一个参数代表横坐标，一般用一维数组，第二个参数代表纵坐标，一般用二维数组；当然也可以第一个参数代表横坐标，后面二三四个参数也分别用一维数组代表纵坐标。画雷达图首先要对区域进行划分，比如需要显示六科的成绩，就需要六块，然后画雷达图，然后对雷达图进行填充；雷达图常见错误是数据和划分的区域不一致，比如数据有五行六列，划一共划分为六块，这时候就需要对数据进行转置然后就可以画出雷达图。

第六：重采样包括升采样和降采样，降采样主要是对时间颗粒变大，原来是每一天统计数据，现在改为每一周或每一个月统计数据，比如知道一年中每一天的汽车销售，现在要统计每一个月的销售量就需要用到升采样，重采样在生活中很常见，因为很多时候去统计每一天的销售量会发现波动太大，也没有办法得到有用的信息，而看每一个月的信息差距就会非常明显。而升采样是指时间颗粒变小，原来是一周统计一次数据，现在改为一天统计一次数据，因为升采样面临着数据量不够的问题，所以一般升采样是将一周的数据复制下来再分配到一周的每一天，通过升采样得到的数据因为只是简单

的复制所以会和实际的数据有一定的差距。滑动窗口是为了统计一个时间段的数据，比如一个产品在打广告前统计一个月的销售量，打完广告后也分别统计一个月的销售量，可以看出一个打广告对产生销售量产生的影响是否达到预期要求。

第七：文本数据分析首先要进行分词，因为英文是写完一个单词空一格的，所以英文分词很简单，直接按空格来进行分词的；而中文分词分为全模式和精确模式，其中全模式会把所有可能的词全部输出，可能会出现有的字重复使用，一个字既和前面的字组成词，也和后面的字组成词，而精确模式不会出现重复的字，所以一般是采用精确模式进行分词。然后是词性标注，对每一个词按照名词，动词，形容词等进行标记。因为在英文中有的词有多种表示方法，比如动词有一般现在时，一般过去时，现在进行时三种不同的状态，所以需要进行词形归一化；在进行词性归一化也有可能会出现问题，因为有的词既有可能是某个动词的过去时，也有可能是一个名词，所以有时候如果词形归一化没有还原，需要再加一个表示词性的参数。做完词形归一后需要进行删除停用词操作，在英文中会有the,this,that一类的词，在中文中会有这个和那个的词，这类词删掉一般对分析的文本没有任何影响；对停用词进行删除需要用到停用词表，英文停用词表比较简单，可以直接使用，但中文停用词表一般都不太完美，需要在使用的时候再添加一些用得比较多的停用词。然后我们可以对文本的情感进行分析，比如评价一件衣服，一台电脑，一本书的好坏；在评价的时候首先要找到代码情感的词，比如说好，坏，烂一类的，还要找到一些程序副词，比如相当，非常，比较一类的，然后对每个程度赋值一个权重，如相当赋值为4，非常赋值为3,比较赋值为1；对正向评价赋值为1，负向评价赋值为-1，然后用评价乘以程度副词再进行相加，如果值为正数，则代表为正面评价，如果值为负数，则代表负面评价，但进行文本情感判断的时候遇到陌生词汇计算机又不认识，具有一定的局限性。接着可以进行文本相似度的评价，首先要进行分词和词形归一化的操作，然后把所有的词的词频按从高到底显示出来，可以显示频率出现最高的100个单词。然后生成两个文本的词频向量，意思就是这两个文本中是否出现这100个单词，如果出现了在对应的位置显示1，如果没有就显示0，最后两个计算词频向量之间的夹角，如果两个向量的夹角比较小，则说明两个文本相似度比较高，如果夹角比较大，则说明相似度比较低。最后还可以对文本进行分类，比如说可以统计一个学生的名字的最后一个汉字和学生性别之间的关系，如最后一个轩字，则他是男生是概率是多少，是女生的概率又是多少。首先要进行训练，看哪些字最能代表一个人的性别，一般来说可以考虑用1000个来训练，再用1000个来测试，看准确率多少。

第八：对于综合项目分析，首先需要读取数据，因为有些数据表的数据项是英文的，需要转换为中文的，然后进行重复项的检测，如果有重复项，就把重复项删掉，再进行空值的检测，空值的处理视情况而定，可以删掉，也可以用平均值来填充空值。接着我们需要对数据进行分组，最好通过不同的字段来进行分组。最后我们要对分组的数据进行图表化，为了让数据的对比更加强烈，我们需要用到折线图，柱状图，饼状图。

参考文献：

[1] 黑马程序员编著. Python数据分析及应用. 北京：中国铁道出版社，2019:216-218
[2] 黑马程序员. Python数据预处理.北京：人民邮电出版社，2021:48-50
[3] 黑马程序员. Python数据可视化.北京：人民邮电出版社，2021:45-46