记第一次建模比赛经历

（比赛完了之后，其实没啥可写的，毕竟感觉自己做的挺差劲的，现在写的话是因为自己不知不觉间马上要参加第二场建模了，对自己的第一次比赛有了点感想。

唠叨一下

什么叫数学建模比赛，就是给你几个题，各题目涉及不同背景下对实际问题的求解，就我参加的华中杯而言，**一个A 题 马赛克瓷砖选色问题，另一个B 题 技术问答社区重复问题**，队伍三人都是新手，然后选题发出来，哦豁，有点傻眼，把关键词摘取了到知网上搜了搜，发现第一题是比较传统的建模题，图论方向。（虽然后来看群里有人讨论拓扑结构，还听到别的队有讨论最短路的，可以说图论不一定是误判哦）。第二题比较新颖，搜不到论文，自己大致猜测了一下，算数据处理，机器学习方面吧，由于自己专业方向，和对数据库，后端有点好奇，还是比较想做B题的（本人没做这个题，网上搜了搜)

有一说一，关键词搜索完了，看了看，咦，好像没啥用，怎么做还是不会，于是发题的那个晚上，我们队在沉默中度过了。凸(>皿<)凸，然后第二天，网上有关赛题的思路出来了（比赛在五一假期，所以估计人比较闲），大致看了看，哎，B题思路都看不懂，自然语言处理啥的，电信专业不是很合适，并且编程语言方面不是特别强，想想就放弃了。（结果，这次比赛A题：894，B题：238，好像是这样，并且有关B题，主办方给的题目有很多可讨论的点，看看那个解答，就觉得很麻烦。）

A题

看了看，网上这样归纳。

第一题可以采用遗传算法，模拟退货算法等找出最优解。第二题可以使用评价模型，建立评价标准。第三题可以采用评价模型决定最优瓷砖。

A应该不需要机器学习，但是要启发式计算，启发式算法是个大概念。在数学建模里面，通常分为进化计算与群智计算，进化计算就是遗传算法与他的改良，群智就包括蚁群蜂群鱼群粒子群退火等算法，我想这题用遗传算法比较好，因为颜色本质上可以看作二进制串，也可以看作染色体，蒙特卡洛模拟？也是个牛批的办法 ，我用模拟退火算的还没有我蒙特卡洛暴力解的好，这个A要是有人用元胞自动机那可就有意思了，A题 我觉得用聚类算法比较合适

第一题

附件 2 是图像 1 中的 216 种颜色，附件 3 是图像 2 中的 200 种颜色，请找出与每种颜色最接近的瓷砖颜色，将选出的瓷砖颜色的编号按照附件 4 的要求输出至结果文件。

怎么搞?

- 我们组的方法：

首先，先把两个附件的数据导入呗，RGB的话，可以用向量表示，要求颜色相近，

判断不同的颜色与 22 种瓷砖颜色间的相似度，我们采用两种方法组合讨论：

1：分别计算一种颜色与 22 种瓷砖的两个向量的欧几里得距离

2：计算两向量的余弦相似度

最后设置一个综合相似度，把两个结果处理一下，争取结果的准确性

- 其他：

针对问题一，寻找算法准则，为两组需求数据自动对应颜色，换句话说根据RGB信息自动匹配相近的颜色。假设把RGB看成空间的坐标，我们可以定义不同的距离来衡量需求数据中的每一个点和已有22组颜色的距离。寻找距离最短的记为最相似。最后可视化出自动匹配的效果。常用的距离有马氏距离、欧氏距离、以及相似余弦等；

简而言之，遍历找最小距离。

然后把数据优化一下：

怎么搞？

- 距离求了，会发现结果有近似的，可以在做一个方差计算，方差越小，距离越近。
- RGB转换为HSV，两种表示方法，分别计算，对结果分析。

有一些算法，我们没有想到但是可以去猜测一种做法：某个队根据几个值得出了一个非线性公式并有多个限制条件（非线性规划）然后配合模拟退火等智能优化算法得出更优解

第二题

2）如果该厂技术革新，计划研发新颜色的瓷砖。那么，不考虑研发难度，只考虑到拼接图像的表现力，应该优先增加哪些颜色的瓷砖？当同时增加 1 种颜色、同时增加 2 种颜色、……、同时增加 10 种颜色时，分别给出对应颜色的 RGB 编码值。

- 对于第二问，这个表现力就是说能够近似体现多种颜色，优先增加与多个颜色差别较小，不管增加几种颜色，但本质上还是选近似颜色，只不过同时多增加几种颜色就要考虑新聚类中心的分布问题，FCM思想比较适合解决该问题，目标函数可以就以FCM的来。

  该算法思想就是其他点到聚类中心的距离的倒数之和最小，但是别直接套用该算法程序，其中的距离公式需要更改。第二问说是从技术革新的角度，那么本问被聚类的点应当为256*256*256个点了，并不是附件2和附件3的点来做聚类分析，当然在选出新聚类中心颜色后，可以再去算一附件2和附件3的J函数值，对于题目提到的表现力，颜色越近似就说明表现力越好，表现力函数公式可以直接是J函数的倒数。

  本文的程序设计思路可以做个参考：

  1.两个自变量，除了固定的22个颜色外额外增加的颜色数、聚类范围(切记不是半径为R的圆形)

  2.(优化算法)随机产生n个个体，每个个体拥有m个新聚类中心(m个RGB值，看做是三维坐标)和1个聚类范围，被聚类的点就是256*256*256个点，按第一问距离公式计算各聚类中心距离。

  3.设立目标函数J，如果想将表现力设为目标函数，这里目标函数就设为1/J。（第一问两个公式加权聚合为一个公式，权重自行设立，效果不好就调试下），求J函数最小化或者1/J函数最大化。

  4.迭代多次后，输出新增最佳m个颜色RGB。

- 先考虑加一个瓷砖的情况，也就是现在有 23 块瓷砖，通过游历其他可能的瓷砖，带入到附件二三我们便能找到，使综合相似度最大的那一个瓷砖参数，便可得加一个瓷砖时的结果。我们接下来便可将求出的这块瓷砖当作已知的，求第二块瓷砖，也就是在原来 23 块瓷砖的基础上再添加一块瓷砖使总相似度最大，而后面的三块，四块…… 便可以此类推。

  先以加一种瓷砖为例，所有的瓷砖颜色为原本 22 种，还有新的瓷砖（X，Y，Z）共

  有 23 个颜色。此时可以计算图像与这 23 个瓷砖的相似程度函数。通过遗传算法来计算最好的 X，Y，Z 取值。

  产生随机群体 (50)，然后进行迭代 (运用交叉和变异)，生成进一步多的后代群体，计算后代群体的适应度，比对分析出最大值，在一次次进化中使算法一步步趋近最优解。

  后续可以继续添加瓷砖，逐步添加样本个数，运行遗传算法，一步步求出最优解。

- 可以很容易想到，把22种颜色分类，去找1600万个点中距离这些类的“重心”最远的那个点一定是第一个增加的

  这里所谓的“重心”就是建立合适的坐标系求得的中心

  我们队建立的坐标系是：RGB坐标系->HSV坐标系->圆锥坐标系下HSV坐标系（有理由的）

  其实关键在于如何分类。开始想到了神经网络，奈何我的水平太差写不出来，后来发现kmeans算法我可以很快的实现，就用kmeans去分类，分好类之后我们可以对于每一个类去二分球的半径，不在所有球内的就是答案

  至于如何判断是否在求内，其实O(n)跑一遍所有颜色到22种颜色的重心排序后的结果就好

  step1.把22种瓷砖颜色的RGB值转化为圆锥坐标系下坐标

  step2.枚举类数K∈[1,22]，用Kmeans算法求得重心坐标

  step3.计算22种已有颜色到重心的距离和，并确定合适的K值

  step4.枚举增加一个其他颜色的RGB值，用Kmeans算法求得重心坐标

  step5.计算重心偏移量，降序排序并把偏移量最大颜色设为已有（必选）

  step6.二分球的半径并判断其他颜色是否在已有颜色画出的球内

  step7.若不在则设为已有（必选），重复操作6，直到球的半径在误差精度内

第三题

- 如果研发一种新颜色瓷砖的成本是相同的，与颜色本身无关，那么，综合考虑成本和表现效果，你们建议新增哪几种颜色，说明理由并给出对应的RGB 编码值。 这一问是对第一问和第二问的应用，所以不需要新的模型。这里只给出简单的思路，不在给出实际代码了。 对于增强图型表现力的单目标问题来说肯定是颜色越丰富表现力越强。但是因为时间管理成本、开发成本不能够这么干，所以我们应该提高颜色的利用率。建议定义如下指标：

  1、已有颜色的对整个颜色空间的`覆盖率`，每增加一个，覆盖率越高，但是覆盖率的增加速率会变慢。所有个数临界值可以选择每增加一个颜色颜色覆盖率变慢时的颜色个数。

  2、第一问中我们针对每一个颜色都选择距离最近的颜色编号，同时我认为所有颜色数据和对应编号之前的`距离`（绝对或者相对距离都可）也是一个不错的指标。

- 我们在第二问的基础上给出了11种颜色，并（基本可以）确定第三问的答案一定在这11种中。（至于为什么不是第二问题目问的10种，也是有原因的哦）

  然后用二进制去表示选还是不选，一共有2047种状态（即2^11-1，全是0的状态也不可），美其名曰0-1规划罢了

  然后把这些状态跑一遍，用权重和除以1的个数即为该状态下综合权值（综合考虑成本和表达效果）

  具体解释上面这句话：

  权重和：权重哪里来的？我们用了比较主观的层次分析法（为什么用主观点的也需要解释理由哦）

  用层次分析法分析H、S、V三者在决定颜色中的权重，并对于每一个因素又分析了11种颜色的权重，并分别进行一致性检验

  最后给出每一种颜色的权重

  值得一提的是，第一问我们直接用的RGB，后面两问我们采取了HSV（为什么这样选其实就蕴含在二者的对比之中）

  而权重和就是每个状态二进制是1的颜色权重相加

  因为每种颜色的成本一样，那么总状态的成本可以用颜色的个数，即1的个数来表示

  这里查到一个类似的定义“投资回报率”，“收入/成本”可以相当于我们的“权重和/1的个数”

然后取出最后总权重最大的那个状态对应的颜色就好了

B题

有一说一，专业性很有点强，适合NLP（自然语言处理的大佬搞

题到现在也没啥思路，直接cv吧

- 请根据附件给出的问题文本数据及问题配对信息，建立一个能判断问题是否重复的分类模型，并解决：

  1）输出样本问题组为重复问题的概率；

  2）从附件问题列表中，给出与目标问题重复概率最大的前 10 个问题的编号；

  对于每个问题的预测结果采用 top K 列表对其进行评估；

可以总结为一个问题，第二个问题只是在第一个问题的基础上的应用而已。而解决第1个问题的关键就在于构建一个机器学习模型，从而能够根据两个非结构化的文本来输出，两者之间是否重复。说到底就是一个监督学习的问题。

要解决这个监督学习问题，首先就要将非结构化的文本转换为结构化的，类似于表格或者向量的数据。为了解决这个问题，本文将英语文本进行拆分，停用词过滤，提取词根，在采用词袋模型配合TF-IDF方法，最终将非结构化的英语文本转换为一个向量。

之后将附件2，与处理过后的附件一进行合并，从而获得用以机器学习的数据集。考虑到直接合并产生的数据及占用的空间非常大，所以本文在进行合并操作的时候进行了适当的筛选。

然后由于重复的数据比起非重复的数据，数量上差距太大，如果直接用于监督学习，会由于类别不均衡问题，导致模型的效果太差。所以在训练模型之前，本文采用过采样和欠采样的方法，来解决数据的类别不均衡问题。

再考虑到数据的特征非常多，容易造成维度灾难，因此本文将采用卡方检验来排除那些不相关的特征。

最后本文将采用逻辑回归模型，根据上述处理过后的数据，即按 7:3 的比例拆分成训练集和测试集，在训练集中训练模型，在测试集中评价模型评价指标为 F1 值。

到机器学习模型之后，要解决第2问就比较容易了。只需要将目标问题与其他问题的两两配对，输入到逻辑回归模型之中，就可以得出相应的概率，然后依据概率最大，提取出前10个问题即可。

逻辑回归模型当然不仅可以输出概率，也可以输出类别。我们只要应用逻辑回归模型对目标问题和这前10个问题进行预测，得到标签，结合原数据标签，就可以得出可以K值。

不过，按照我的那个模型，算出的 **K 值比较低**... 原因可能是因为在训练模型的时候，数据没有全部投入的缘故，又或者是解决类别不均衡时，引入了太多的误差。因此，这里就仅仅实现了一部分了

若要解决这个问题，可以：

1. 加大训练数据，最好将所有的数据都投进去。
2. 给相似数据赋予一个较大的权重，使得少数类数据能够被加强学习

上述B题

另一个：

先说说这个题目吧，专业技术问答平台(国内CSDN，国外StackOverflow)上面有很多提问，问题当中存在冗余，这些冗余会降低检索系统的召回质量。题目希望使用NLP技术做一个标重模型，把重复的问题折叠起来，提升检索质量。之后看了看数据，给了7295条标注数据，包含英语原文和中文翻译。那个中文翻译简直惨不忍睹，new运算符都给翻译出来了，果断用原文。

。。。看不懂了

NLP做B题

逛论坛发现有个大佬整理了一下有关这次学习的收获，链接放这里

大一新生参加华中杯

杂谈

再写个后记吧，建模比赛，你就比如编程手，python,matlab,lingo,spss,一堆要学的，然后要做啥，就是根据建模提的方案，你去码代码，把结果求出来，然后画几个图，分析一下数据。（excel！最强图表分析工具，{就是不会用}，，然后这次的话python里面的pandas+mysql ，听说蛮好的，可以代替excel

建模手，提前了解一定量的模型很重要

写作，latex真的赞，不用担心排版，就是latex画表格真的值得吐槽，然后表格的话，后期是再excel里面搞了个插件**Excel2Latex**，直接转化为latex的，就是你转换后还要微调一下，然后公式工具

latex在线公式编辑器

真的香，然后一些图片插入，先把图片放进文件夹，然后再引用，蛮简单的，就是第一次用，总有些报错，也别急，慢慢来，搜一下，改一下，也还好，然后画图软件，visio,spss,ppt,都可以了解一下。

今天把论文再看了看，发现，，里面的内容有很多小细节需要改，所以最后最好一起看一下论文。

、