一些奇妙的期末考试模拟题

#MLorDS

本卷是北京理工大学机器学习考试题，考试时间 120 分钟，总分 100 分

适合北京理工大学 AI 系

祝考试愉快！

题目 1：灌铅的骰子（10 分）

一些桌面角色扮演（TRPG）游戏中，会掷骰子决定游戏的进程。小雨同学正在玩某种 TRPG 游戏，该游戏使用一个 100 面骰子完成技能检定，掷骰子得到的数值越低，技能检定就越容易通过，因此，小雨准备在一颗骰子中灌铅以提高成功率。她一共准备了两枚骰子，其中只有一枚骰子是灌铅的，第 i 颗骰子扔出 1-50 点的概率都相同，为 θ_i；掷出 51-100 点的概率也相同，为 $0.02-\theta_i$。她进行了三次实验，在每次实验中，她将**同一枚骰子**投掷 5 次，实验结果是：

- 第一次：23，25，46，57，18
- 第二次：03，45，72，34，49
- 第三次：45，78，39，72，55
  可惜的是，小雨自己也忘记哪枚骰子是灌铅的了。请你帮她找出在第几次实验中使用的骰子更可能是灌铅的？θ_1 和 θ_2 分别是多少？

题目 2：篮球游戏（15 分）

在上题中我们已经说到，小雨非常喜欢玩游戏，于是现在小雨又构造了一个篮球游戏。她叫来 A, B, C 三个小伙伴，在一定距离外对着篮筐投球。其中，A 的命中率始终为 0.6，B 的命中率始终为 0.4，C 的命中率始终为 0.8；当轮到某个人投球时，如果他投中了，那么他继续投球；如果他没投中，则球按照 A->B->C->A 的顺序循环地传递给下一个人。小雨在一旁观看 A, B, C 投球，如果轮到某人投球时，他投中了，那么小雨有 50%的概率认为该人进行了一次“优美的投球”，此时，小雨将举起一朵小花花。初始时，A, B, C 三个小伙伴通过公平的抽签决定谁先开始投球。作为局外人的你无法看到谁在投球，只能看见小雨是否举起了小花花。

（1）三人共投了三次球，你发现小雨分别{举起，没有举起，举起}小花花。请问三人投球顺序为{A, A, B}的概率是多大？

（2）使用不同于 （1）中的计算方法验证你在（1）中得到的结果是否正确。

（2）三人共投了三次球，你发现小雨分别{没有举起，没有举起，举起}小花花。请问三人概率最大的投球顺序是什么？

题目 3：异常用户（15 分）

作为一名管理建设地球发动机的工程师，你最近发现有一些奇怪的用户在各大社交平台上宣传“信数字生命，得永生”。你认为这种思想是错误的，而且将严重阻碍地球发动机的建设进程，这样你们就来不及在氦闪来临前将地球开出太阳系了！你认为这些人都是被 MOSS 控制了，于是你准备用魔法对抗魔法，使用简单的程序分辨一个用户是不是一位被“数字生命”洗脑的异常用户。你发现，一个异常用户有如下特征：

- 在社交媒体上的发言中过于频繁地提及“数字生命”等敏感词
- 24 小时不休息，不停地在社交媒体发言
  于是，你收集了几个用户的信息如下：

编号	发言频率	每条发言中敏感词的数量	标签

编号	发言频率	每条发言中敏感词的数量	标签
1	12	8	异常用户
2	4	3	正常用户
3	2	1	正常用户

（1）使用硬间隔 SVM 求解一个划分两类用户的超平面
（2）一个用户的发言频率是 25，每条发言中敏感词的数量是 1。你认为他属于哪一类用户？
（3）现在有一核函数 $\phi(x)$，与之对应的内积变换函数为 $f(x,y)$

$$f(x,y)=\exp(-0.002\,x^Ty)$$

利用这个核函数，求分类超曲面 $w^T\phi(x)+b=0$ 中 w 和 b 的值

题目 4：找啊找啊找朋友（15 分）

迈尔斯-布里格斯类型指标（Myers–Briggs Type Indicator）是由美国作家伊莎贝尔·布里格斯·迈尔斯和她的母亲凯瑟琳·库克·布里格斯共同制定的一种人格类型理论模型。该指标通过四个维度上的划分（E/I：外倾/内倾；S/N：实感/直觉；T/F：理智/情感；J/P：判断/理解）来将人格划分为 16 个类型。作为一名充满正义感的银河棒球侠，你认为星间列车上最近的气氛有些沉闷，于是你决定找找列车上哪些人的性格比较相近，以便让他们成为更好的朋友。你收集了五位列车乘员的四项指标得分如下：

名字缩写	I/内倾得分	N/直觉得分	T/理智得分	P/理解得分
TR	3.23	4.21	1.19	0.45
M	-5.42	-0.45	-2.29	0.13
W	2.13	1.03	-2.77	-0.45
D	3.09	0.21	1.35	-0.96
H	-1.46	0.71	-0.23	-0.43

（1）请你使用**经过 Z-Score 标准化后的**数据之间的**哈密顿距离**作为数据之间的距离度量，写出成对距离矩阵
（2）请你使用**数据点间的平均距离**作为类间距离的度量依据，利用 （1）中求得的成对距离矩阵，完成**凝聚的**系统聚类，画出聚类谱系图
（3）根据聚类结果，你认为哪些人更容易成为朋友？

题目 5：悄悄话（10 分）

众所周知，Minecraft 中的村民只能发出几个简单的音节，玩家完全听不懂他们在讲些什么。但是，由于你今天不慎把头撞到了电线杆子上，你发现你竟然可以神奇地听懂村民之间的悄悄话了。经过仔细理解，你认为他们的语言中只有 AAA、BBB、CCC 和 DDD 这四个单词，他们谈话的主题也只包括三个，分别是保卫村庄（G）、与玩家交易（T）和日常事务（R）。你收集了他们的三份谈话内容，并决定使用隐狄利克雷分配算法为三份谈话中的每个词分配主题。三份谈话中的单词与你给每个单词分配的初始主题如下所示：

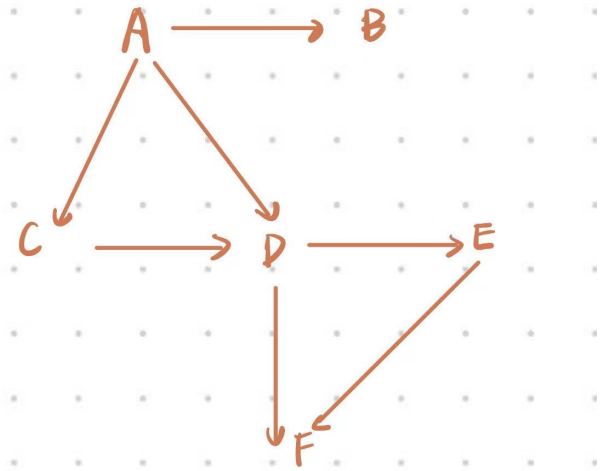
单词编号/谈话编号	01	02	03
1	AAA (G)	CCC (T)	DDD (T)
2	BBB (G)	DDD (R)	CCC (R)
3	DDD (T)	CCC (T)	AAA (G)

单词编号/谈话编号	01	02	03
4	CCC (R)	AAA (T)	BBB (G)
5	CCC (T)	AAA (G)	BBB (T)

（1）画出隐狄利克雷分配算法的示意图，写出该算法生成 M 篇每篇有 N 个特定单词的文档的联合概率
（2）请按照隐狄利克雷分配算法的迭代规则进行一轮迭代，给出一轮迭代后每个词的主题。在统计文档中主题的频率和同一词汇中主题的频率时，为了防止乘 0 发生，请加上小常数 0.01。

题目 6：小阳人（15 分）

让我们短暂地将目光拉回现实世界。最近，第二波新型冠状肺炎疫情来势汹汹，根据钟南山等人的模型预测，这波疫情将在六月底达到峰值。小航同学研究了疫情中几个因素之间的相互关系，这些因素包括：A——社会面阳性病例数量迅速增加；B——发热门诊就诊人数迅速增加；C——任课教师感染新冠；D—同班同学感染新冠；E——舍友感染新冠；F——自己感染新冠。他认为这些因素都可以使用 0-1 变量表征，且变量之间的概率相依关系可以使用一张有向图表示。



（1）写出这张图的拓扑序
（2）如果 D 已知，判断 A 和 F 是否互相独立；如果 C 已知，判断 A 和 E 是否互相独立
（3）已知各个变量间的概率相依关系如下表所示：

-	A=0	A=1
B=1	0.2	0.7

-	A=0	A=1
C=1	0.1	0.6

-	A=0, C=0	A=1, C=0	A=0, C=1	A=1, C=1
D=1	0.03	0.15	0.35	0.87

-	D=0	D=1
E=1	0.16	0.43

-	D=0, E=0	D=1, E=0	D=0, E=1	D=1, E=1
F=1	0.01	0.65	0.49	0.93

求 A=0, B=1, C=1, D=0, E=1, F=1 的概率

（4）现在已知 D=0, E=1, F=1，求 A=0 的概率

题目 7：猫猫统治人类（20 分）

古人云（？），大猫征服荒野，小猫征服人类，谁都无法拒绝不凶不闹不拆家，给撸给抱给亲亲的小猫咪。叶子姐姐希望创建一个神经网络用于区分一张图片是不是猫猫。

（1）最初，叶子姐姐决定使用一个类似于 LeNet 的简单的网络完成分类，这个网络中包含 2 层卷积层、2 层平均池化层、3 层全连接层。请你帮她画出这个网络的大致结构示意图。

（2）叶子姐姐先使用很小的图片测试她的网络。她找到了一张 $5 \times 5 \times 1$ 的灰度图：

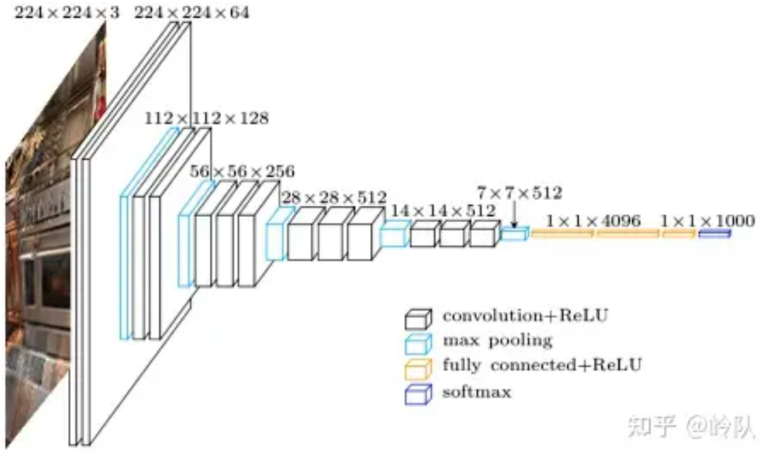
0	3	2	9	8
0	3	4	0	0
0	0	1	0	0
0	3	4	1	0
0	9	3	1	0

现在使用如下卷积核对图像进行卷积：

1	0	-1
1	0	-1
1	0	-1

要求 $stride = 1, padding = 1$。给出卷积后的特征图矩阵，并使用文字表述这个卷积核起到的作用。

（3）叶子姐姐发现她设计的网络效果太差了，于是她准备使用 VGG-19 网络来实现这一任务。她找到了一个用于将输入图像 1000 分类的 VGG-19 网络，网络的示意图如下：



用文字说明这个网络为什么被称为 VGG-19。计算网络中的参数数目（只考虑权重参数，不考虑偏置）

（4）显然，（3）中的网络不能直接用于叶子姐姐的任务。说明叶子姐姐应该如何修改这个网络，使之完成二分类任务？

（5）VGG 19 网络的贡献之一是它是一种使用“块”（可以看作一种集成了几个特定层的单元）的网络。在现代卷积神经网络的发展过程中，还有很多种不同的网络提出了全新的结构。请列举几例，并说明这些网络中包含哪些全新的结构。

（6）叶子姐姐发现她的网络在训练集上的准确率为 97.8%，而测试集上只有 85.9%。请你给她提出至少 3 条建议来解决这一问题。

（7）叶子姐姐现在不仅希望研究一张图片是不是猫猫，她还希望研究“猫猫”一词的百度搜索指数随时间的变化规律。她首先希望将第 t 天的搜索指数 y_t 直接回归到前 5 天的搜索指数 $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}$ 上。简述叶子姐姐用于求解这

个回归问题的算法的流程，写出必要的数学表达式。

（8）叶子姐姐现在希望使用循环神经网络（RNN）解决搜索指数预测问题。画出该网络的结构简图。

（9）叶子姐姐现在不仅想使用 $y_i(i<t)$ 来对 y_t 进行预测，她还使用了季节因素参数 x_1、宠物市场数据 x_2、社交媒体讨论热度 x_3 作为解释变量一同输入神经网络。假设她只使用了一层 RNN 网络，隐藏层的参数数目是 512，计算网络中的参数数目（只考虑权重参数，不考虑偏置）

（10）叶子姐姐发现，自己的网络训练效果一直很差。究其根本原因，是数据中的某些极端点严重干扰了网络的训练。叶子姐姐现在没有时间清洗数据来排除这些极端点了，请你给她提一条建议，帮助她解决这个问题。