《计算机应用数学第一次作业》

截止时间：2022.11.10

计算题（40 分，每题 10 分）

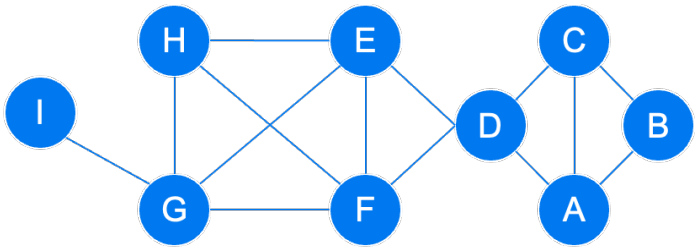
1、

设 X 服从$f(x)=\begin{cases} kx, & 0\le x<3 \\ 2-\frac{x}{2}, & 3\le x\le 4 \\ 0, & others \end{cases}$，求 k, E(X)， Var(X)

2、已知某工厂某批次水泥重量服从正态分布，总体方差为 2.65 公斤， 从该工厂随机抽取 18 袋水泥，其平均重量为 24.9 公斤， 试求该工厂水泥平均重量的 95%和 99%的置信区间。

3、从某学校抽取 20 名高一学生，经测量，这 20 名学生的平均身高为 155cm，标准差为 10cm，假设平均身高服从正态分布，试求该学校高一学生总平均身高的 95%和 99%的置信区间。

4、计算下图中每个节点的 Betweenness centrality，写出计算过程



编程题 （60 分，4 选 2）

说明：建议使用开源工具包，例如 scikit-learn 中有朴素贝叶斯、高斯混合模型等函数实现，sknetwork 中有 PageRank 函数实现……

1、A/B test

数据集：AB_Test_Dataset.csv。

数据描述：见列名。

任务描述：使用 AB-test 计算测试组和对照组直接在 revenue 上的差别，包括 mean、p-value 和 confidence interval 等。注意：需要先对数据在 USER 级别进行聚合，剔除既在 control 也在 variant 的 USER。

要求输出：任务描述中提到的数值，以类似下表的方式展现

	Customers	Credit(USD)	Mean(USD)	%Impact(95% CI)	P-value
C	3,942,457	$1,340,859,230	$340.108		
T	3,940,134	$1,340,199,799	$340.141	0.01%(-0.18%,0.20%)	0.920

2、朴素贝叶斯分类器（Naive Bayes Classifier）
数据集：Bayesian_Dataset_train.csv, Bayesian_Dataset_test.csv。
数据描述：列名分别为 “年纪、工作性质、家庭收入、学位、工作类型、婚姻状况、族裔、性别、工作地点”，最后一列是标签，即收入是否大于 50k 每年。
任务描述：使用朴素贝叶斯（Naïve Bayesian）预测一个人的收入是否高于 50K 每年。
要求输出：1）结果统计，例如 precision、recall、F1 score 等；2）csv 文件，在 test 文件最后增加一列，填入模型预测的收入标签（<=50K 或>50K）
Optional：探索不同参数对结果的影响。

3、高斯混合模型与 EM 算法
数据集：Iris 数据集
数据描述：https://www.kaggle.com/datasets/uciml/iris，可通过 sklearn 直接导入数据集

```
from sklearn import datasets
iris = datasets.load_iris()
```

任务描述：使用高斯混合模型与 EM 算法对数据进行分类计算，mixture components 设置为 3。
要求输出 :不同高斯分布的 mean 和 variance，每个高斯分布对应的权重，plot 出分布的图。EM 算法可以参考
https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html。
Optional：尝试不同的 covariance structures，包括 spherical、diagonal、tied 与 full。

4、PageRank
数据集：PageRank_Dataset.csv。
数据描述：数据集中每一行是一条边<起点 ID，终点 ID>。
任务描述：使用 PageRank 算法计算每个节点的 PageRank 值，参数设置可以参考谷歌 PageRank 算法。
要求输出：1）结果统计，例如 PageRank 分数最高的 20 个 node；2）csv 文件，每一行是一个节点 ID 和它对应的 PageRank 值。
Optional：探索不同参数（β）对结果的影响。