

北京工业大学 2022—2023 学年第 2 学期
《数据挖掘》 考试试卷 A 卷

考试说明: 考试时间: 95 分钟

考试形式: 闭卷

适用专业: 软件工程

承诺:

本人已学习了《北京工业大学考场规则》和《北京工业大学学生违纪处分条例》，承诺在考试过程中自觉遵守有关规定，服从监考教师管理，诚信考试，做到不违纪、不作弊、不替考。若有违反，愿接受相应的处分。

承诺人: _____

学号: _____

班号: _____

卷面成绩汇总表(阅卷教师填写)

题号	一	二	三	四	五	六	七	八	九	十	...	总成绩
满分												
得分												

得分

一、单项选择题 (本大题共 20 小题, 每题 2 分, 共 40 分)

1、1886 年, 英国遗传学家 Francis Galton 在研究人类身高的时候, 发现一个有趣的现象: 父母平均身高高于人群平均值的时候, 他们孩子的身高会比父母低一点, 而父母平均身高低于人群平均值的时候, 他们孩子的身高会比父母高一点, 这对应于数据挖掘中的哪个概念? ()

- A. 预测
- B. 关联
- C. 回归
- D. 回归

2、设 $X=\{1, 2, 3\}$ 是频繁项集, 则可由 X 产生 () 个关联规则:

- A. 3
- B. 5

C. 6

D. 4

3、假设属性 income 的最大最小值分别是 12000 元和 98000 元。利用最大最小规范化的方法将属性的值映射到[3, 5]的范围内。对属性 income 的 73600 元将被转化为：（ ）

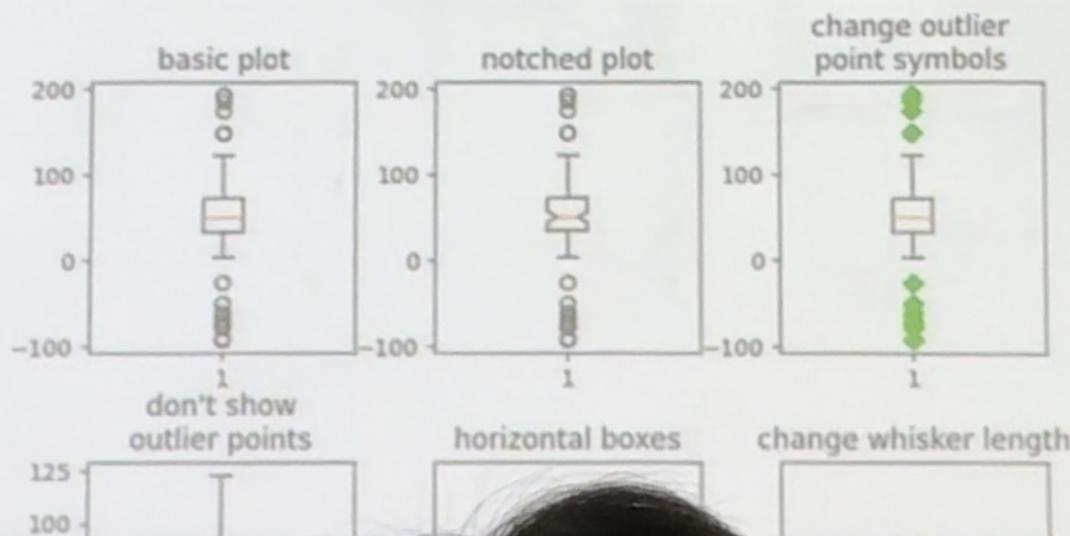
A. 4.577

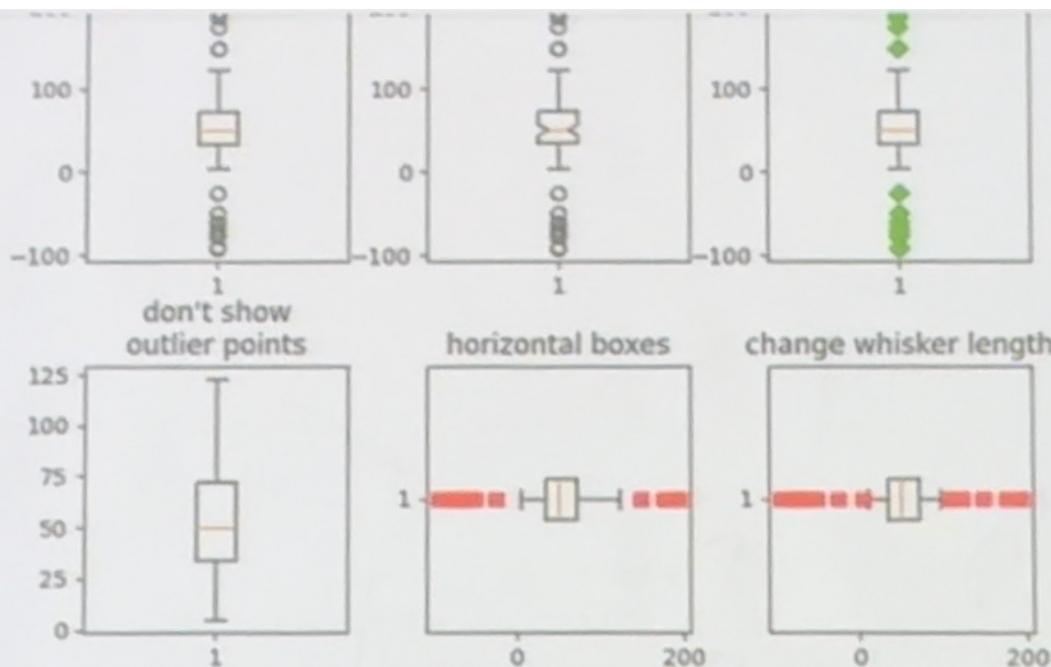
B. 4.432

C. 3.828

D. 2.711

4、下图为箱图的六种可视化表现方式，每一种箱图中间横线代表：（ ）





A. 数据均值

B. 数据中值

C. 数据二分位数

D. 数据四分位数

5、在 2022-2023 第 2 学期《数据挖掘》课程的教授过程中，我们发现以下 Python 著名库中的相关测试模块是有问题的：（ ）

A. sklearn.cluster 中的 DBSCAN 模块

B. scipy.optimize 模块

C. numpy 包中的正态分布模块

D. pandas 包中缺失值处理模块

6、对北京工业大学附近京客隆超市销售记录进行分析，顾客在选购“啤酒”的

6、对北京工业大学附近京客隆超市销售记录进行分析，顾客在选购“啤酒”的同时往往会搭配选择一些卤味熟食，如“红肠”、“猪头肉”等，因此两者搭配一起卖往往能提高各自的销售额，上述这种分析方法在数据挖掘课程中被称为：（ ）

- A. 聚类
- B. 关联
- C. 预测
- D. 分类

7、数据挖掘在具体的实操过程中，一般不考虑以下哪一项：（ ）

第 2页 / 共14页

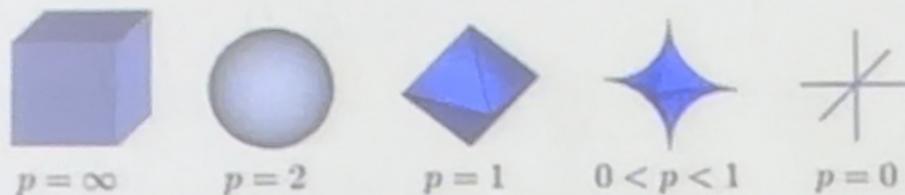
北京工业大学 2022 —2023 学年第 2 学期《数据挖掘》考试试卷

- A. 应用场景
- B. 数据归属

- C. 算法设计
- D. 算法实现

- A. 应用场景
- B. 数据归属
- C. 算法设计
- D. 算力资源

8、针对范数计算，如下图所示，在 $p = \infty$ 条件下的范数被称为：（ ）

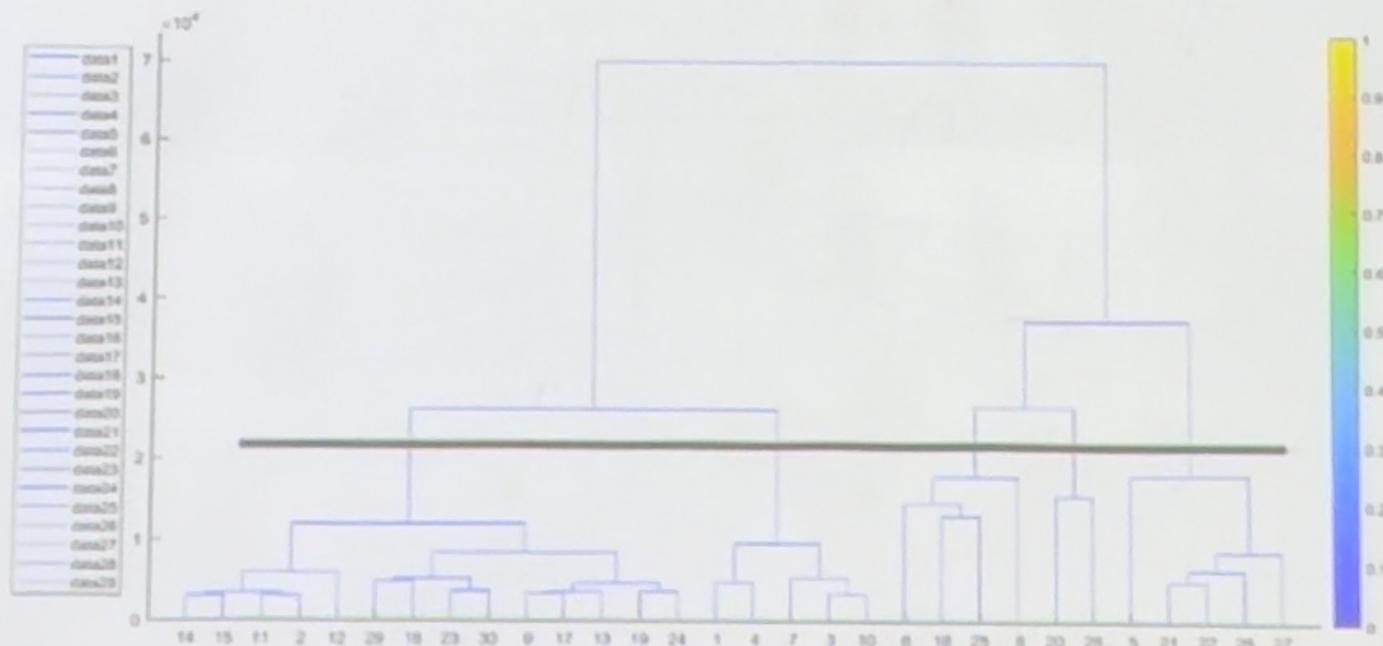


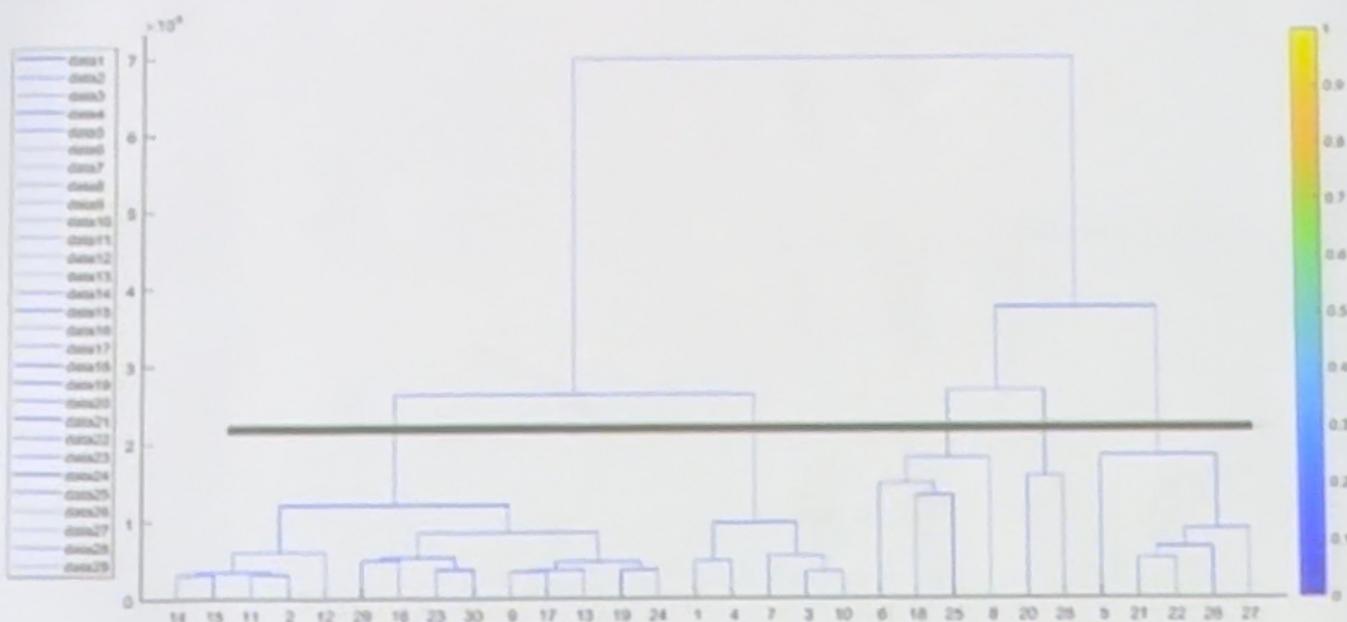
- A. 欧氏距离
- B. 切比雪夫距离
- C. 闵科夫斯基距离
- D. 马氏距离

9、考虑安斯库姆四重奏问题（Anscombe's Quartet），四组不同的数据一元线性回归方程竟然完全一致，通过以下散点图可以明显展示其分布差异，这说明（ ）在数据分析过程中的重要性。

- A. 数据预处理 B. 数据回归建模
C. 数据质量 D. 数据可视化
- 10、以下几种大数据数据量的表示单位中，哪个表示的数据量最大？（ ）
- A. 1 ZB B. 1 HB
C. 1 EB D. 1 PB
- 11、饮食 POI 从味觉上揭示了一个城市发展的态势与偏好。采用高德爬虫采集的 2019 年餐饮类 POI 数据进行城市集群分析，首先根据一个城市整体餐饮类 POI 总数（总量）、一个城市的中餐厅数据（中餐厅）、外国餐厅数量（外国餐厅）、

快餐厅数量（快餐店）、以及一个城市的甜品饮品店数量（甜品饮品店），采用层次聚类法得到如下结果，对于给定截线，当前共有（ ）个聚类。





- A. 3
C. 7

- B. 5
D. 2

12、下图每个圆圈代表一个数据记录，这种数据最适合采用哪种聚类算法（ ）





A. K -means

B. K -Medoids

C. DIANA

D. DBSCAN

13、以下哪种方法不适于进行非平衡数据的再平衡处理（ ）：

A. ENN

B. SMOTE

C. ADASYN

D. GAM

14、在过去 3 年新冠疫情防控过程中，为了尽可能地检测出新冠患者\评判疫苗检测效果，需要尽可能提升（ ）指标

- A. PRECISION
- B. ACCURACY
- C. RECALL
- D. F1-SCORE

15、一般来讲，以下哪种平均数在计算过程中最小？（ ）

- A. 调和平均数
- B. 平方平均数
- C. 几何平均数
- D. 算数平均数

16、给定一个数据序列，现在想度量这个序列的中心趋势，以下哪一项一般不作为中心趋势度量标准：（ ）

- A. 众数
- B. 中值
- C. 均值
- D. 偏度

17、对原始数据进行集成、变换、降维等操作，一般是数据挖掘分析过程中哪个阶段的任务？（ ）

17、对原始数据进行集成、变换、降维等操作，一般是数据挖掘分析过程中哪个阶段的任务？（ ）

- A. 频繁模式挖掘
- B. 数据预处理
- C. 数据聚类
- D. 数据分类

18、本学期《数据挖掘》课程更注重培养学生的哪种能力？（ ）

- A. 发现核心问题的能力
- B. 解决问题的能力
- C. 公式推导的能力
- D. 理解算法理论基础的能力

19、与斯坦福、清华、中科大等校《数据挖掘》课程相比，以下哪一项不是本学期《数据挖掘》课程的特色？（ ）

- A. 实用性
- B. 实操性
- C. 实践性
- D. 理论性

20、美国加州房价数据是机器学习领域常用的一个数据集，采用街区概念，街区是美国调查局发布样本数据的最小地理单位(600 到 3000 人)。数据集中有 20,640 个实例，采用前 8 个属性，预测第 9 个属性，即当地街区房价中位数(单位： $\$/m^2$)，

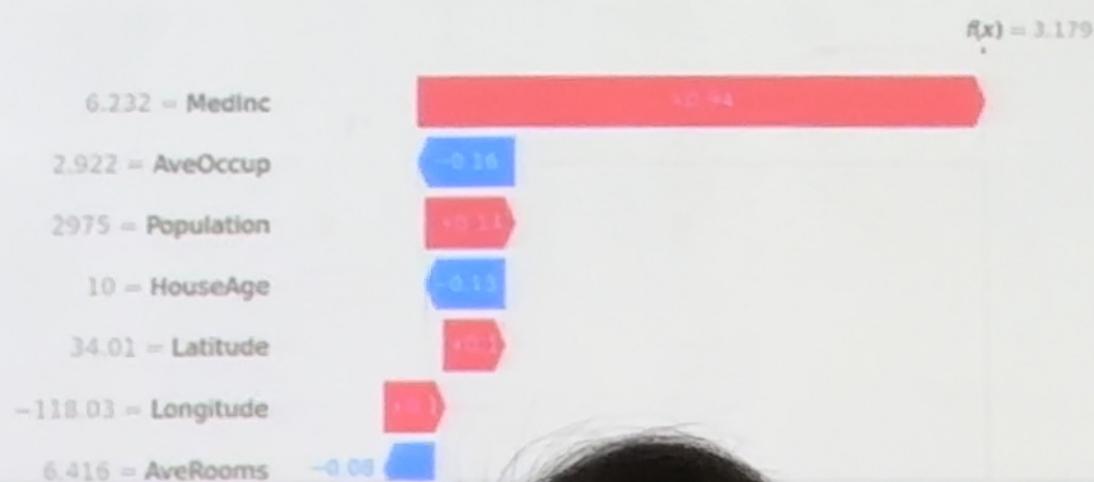
代表房屋大致单价。数据集属性如下表：

个实例，采用前 8 个属性，预测第 9 个属性，即当地街区房价中位数(单位: \$/ m^2)，
代表房屋大致单价，数据集属性如下表所示:

序号	属性	属性缩写	解释
1	longitude	Longitude	房屋经度
2	latitude	Latitude	房屋纬度
3	housingMedianAge	HouseAge	街区房屋年龄中位数
4	totalRooms	AveRooms	房屋总面积 (均值)

5	totalBedrooms	AveBedrms	房屋卧室总面积（均值）
6	population	Population	当地街区人口总数
7	households	AveOccup	当地街区家庭总数
8	medianIncome	MedInc	当地街区收入中位数
9	medianHouseValue	MedHV	当地街区房价中位数

下图是针对波士顿房价数据的 SHAP 可解释性分析结果,根据此 SHAP 结果,以下哪种说法不正确? ()



- A. 街区收入中位数对地区房价的影响最大，说明区位对于房价具有重要影响
- B. 同等条件下，老房子比对应的新房子卖不上价
- C. 同等条件下，房屋面积越大，房屋单价越低
- D. 一个街区家庭总数越多，说明街区人口越多，街区越繁华，房价越贵

得分

二、判断题 (本大题共 20 小题，每题 1 分，共 20 分)

- 1、聚类分析可以看作是一种非监督的分类。 ()
- 2、DBSCAN 能够处理任意形状和大小的簇。 ()
- 3、一组数据一定存在众数。 ()
- 4、Pearson 相关系数不要求原始数据为正态分布。 ()
- 5、KNN 算法不需要全局建模，适用在低维空间。 ()



- 2、DBSCAN 能够处理任意形状和大小的簇。 ()
- 3、一组数据一定存在众数。 ()
- 4、Pearson 相关系数不要求原始数据为正态分布。 ()
- 5、KNN 算法不需要全局建模，适用在低维空间。 ()
- 6、ROC 曲线相比 Precision 和 Recall 等指标更合理，适合处理不平衡数据。 ()

第 6页 / 共14页

6/14

北京工业大学 2022 —2023 学年第 2 学期《数据挖掘》考试试卷

- 7、归一化、中心标准化均属于线性变化。 ()
- 8、PCA 分析假设数据各主特征在正交方向分布。 ()
- 9、深度学习模型如 LSTM、GNN、GRU 等不能用于构建集成学习模型。 ()
- 10、频繁集的所有非空子集必须也是频繁的 ()

- 10、频繁集的所有非空子集必须也是频繁的。 ()
- 11、LOGISTIC 回归可用于多分类任务。 ()
- 12、数据增强之后，数据分析效果一定要比数据增强之前好。 ()
- 13、Apriori 算法是从事务数据库中挖掘单维布尔关联规则的常用算法，该算法利用频繁项集性质的先验知识，从候选项集中找到频繁项集。 ()
- 14、K-Means 是一种产生划分聚类的基于密度的聚类算法，簇的个数由算法自动地确定。 ()
- 15、分类和回归都可用于预测，分类的输出是离散的类别值，而回归的输出是连续数值。 ()
- 16、具有较高支持度的项集具有较高的置信度。 ()
- 17、关联规则挖掘过程是发现满足最小支持度的所有项集代表的规则。 ()
- 18、特征提取技术并不依赖于特定的领域。 ()
- 19、离群点可以是合法的数据对象或者值。 ()
- 20、下表给出了在 5 个时间点观测到的 AllElectronics 和 HighTech (某高技术公司) 的股票价格的简化例子，通过数据挖掘分析，认为如果股市受相同的产业趋

- 19、离群点可以是合法的数据对象或者值。 ()
- 20、下表给出了在 5 个时间点观测到的 AllElectronics 和 HighTech (某高技术公司) 的股票价格的简化例子，通过数据挖掘分析，认为如果股市受相同的产业趋势影响，它们的股价会一起涨跌。 ()

时间点	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

得分

三、简答题 (本大题共 6 小题, 共 50 分)

1、有人认为，优秀的数据分析师一定是名好厨子，请从数据挖掘过程的角度，结合具体数据分析任务，分析上述论断，并给出每个步骤需要注意的事项。

(5 分)

2、PCA 是数据预处理过程中一种常见的数据归约方法，请根据自己的操作经验，用文字写出 PCA 降维分析的分析步骤（6 分）。

3、为了调研了北京工业大学在校大学生周末消费能力，随机抽取了在校 100 名学生周末消费数据，如下表所示，根据此数据，估计北京工业大学在校大学生周末消费能力的平均金额（5 分）。

消费金额：(元)	人数(个)
0-100	35
100-300	25
300-500	20
500-800	15
800 以上	5

4、某机械厂生产机器，设有毛坯、粗加工、精加工和装配一共 12 个连续作业的车间。其中有 5 个连续作业工序的不合格率为 3%，有 4 个连续作业工序不合格率为 5%，有 3 个连续作业工序不合格率为 7%，求整个作业车间的平均不合格率是多少？（列出式子即可，不用计算）（4 分）

5、Cleveland 心脏病数据集是机器学习数据分类中经常采用的一个数据集，共有 14 字段，其中前 13 行字段经常作为模型的输入端 X ， $X = \{x_1, \dots, x_{13}\}$ 表示影响心脏病的各个要素，最后第 14 行字段作为模型的输出端 y ，表征具体的心脏病诊断结果。各个字段定义如下：

Feature	Detail
Age	Age in years
Sex	1 for male; 0 for female
Chest pain type	Value1: typical angina. Value2: atypical angina. Value3: non-anginal pain. Value4: asymptomatic
Resting blood pressure	In mm Hg on admission to the hospital
Serum cholesterol	In mg/dl
Fasting blood sugar > 120 mg/dl	1 for true; 0 for false
Resting electrocardiographic results	Value0: normal. Value1: having ST-T wave abnormality (T-wave inversions and/or ST elevation or depression of > 0.05 mV). Value2: showing probable or definite left ventricular hypertrophy by Estes's criteria
Maximum heart rate achieved	centered
Exercise-induced angina	1 for yes; 0 for no
ST depression induced by exercise relative to rest	In mm Hg on admission to the hospital
Number of major vessels	(0-3) colored by fluoroscopy
The slope of the peak exercise ST segment	Value1: upsloping. Value2: flat. Value3: downsloping
Thallium heart scan	3 for normal; 6 for fixed defect; 7 for reversible defect
Diagnosis heart disease (angiographic disease)	Value0: no disease. Value1: heart disease

拟采用多元回归分析的方法，初步建立心脏病预测模型。针对模型的输入端 $X = \{x_1, \dots, x_{13}\}$ ，是否可能存在某个影响因素 x_i ， x_i 与 y 的相关系数为正（或负），

- (1) 采用 2017-2023 年度考研报名人数记录, 建立一元线性回归方程, 并预测 2024 年全国考研报名人数: (5 分)
- (2) 采用一元线性回归方程进行未来 2024 年考研报名人数的预测, 体现了分析区间 2017-2023 内的惯性趋势, 实际上, 实际考研报名人数还受到其他各种因素的影响。根据 2011-2023 年各年度考研人数记录, 判断未来 2024 年考研报名人数走势将会趋向于以下哪条曲线 (A, B 还是 C) ? 请给出相应的判断理由。 (5 分)



6、下表统计了 2011-2023 年我国各年度考研报名人数与录取人数。根据此表，完成以下分析：

年份	报名人数（万人）	报名增长率	录取人数（万人）
2024	?		?
2023	474	3.72%	125
2022	457	21.22%	110.7
2021	377	10.56%	110.07
2020	341	17.59%	100.03
2019	290	21.80%	81.13
2018	238	18.40%	76.25
2017	201	13.56%	72.22
2016	177	7.30%	58.98
2015	164.9	-4.12%	57.06
2014	172	-2.27%	54.87
2013	176	6.30%	54.9
2012	165.6	9.60%	52.13
2011	151.1	7.50%	49.46

- (1) 采用 2017-2023 年度考研报名人数记录，建立一元线性回归方程，并预测

7、休谟 (David Hume) 批判了机器学习领域内的因果判定与推断，认为人类通过经验，只能感受到两件事物发生的“相继关系”，却永远无法发现两者之间的“因果关系”，这对机器学习领域中可解释性机器学习 (XAI)、相关性分析等构成一个很大的挑战；同时，他进一步批判了机器学习领域中的经验归纳过程，认为如果经验能不断重复的话，归纳才是有效的，但人类永远无法证明经验是永恒不变的，这同样对当前知识图谱、历史经验知识库的构建造成了很大的挑战；结合自身的数据挖掘实践，你认为如何应对以上两大挑战？（5 分）