《数据科学导论》课程

习题集

北京邮电大学

计算机学院（国家示范性软件学院）

目录

一、章节习题

第一章：数据科学导论概述

本章节是数据科学导论第一个章节，题目主要以查找文献资料和启发性问答为主，题目难度较低。

1. 查阅文献并思考，大数据的价值可以体现在哪些方面？
2. 查阅文献并思考，数据科学与统计学有何不同？
3. 查阅文献并思考，数据科学家和数据分析师有什么不同？
4. 查阅文献并思考，数据科学有哪些基本原则？
5. 查阅文献并思考，数据科学与数据密集型科学有什么不同？
6. 查找资料并思考，数据科学家需要具备哪些技能？
7. 查找资料，举出一个数据科学的实践案例。

第二章：数学基础

本章节是数据科学导论的数学基础：

1—9 题是线性代数、概率论和集合论的基础性题目，难度较低；

10 题是牛顿法求解方程根的基础性问题，难度中等，计算量相对较大；

11—13 题是图论的基础性问题，难度中等。

1. 已知$\boldsymbol{A}=\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}$，$\boldsymbol{B}=\begin{bmatrix}1 & 2\\ 0 & 3\end{bmatrix}$，求$\boldsymbol{BA}$和$\boldsymbol{AB}$。
2. 已知$\boldsymbol{a}=[1,2,-1]^T, \boldsymbol{b}=[1,2,3], \boldsymbol{A}=\boldsymbol{ab}$,求$\boldsymbol{A}^n$。
3. 矩阵$\boldsymbol{A}=\begin{bmatrix}1 & 2 & -1\\ 0 & 1 & 3\\ 0 & -0 & 1\end{bmatrix}$，计算矩阵$\boldsymbol{A}$的逆矩阵。
4. 计算矩阵$\boldsymbol{A}=\begin{bmatrix}3 & -1 & -1\\ -1 & 3 & -1\\ -1 & -1 & 3\end{bmatrix}$的特征值。
5. 设矩阵$\boldsymbol{A}=\begin{bmatrix}4 & 2 & 3\\ 1 & 1 & 0\\ -1 & 2 & 3\end{bmatrix}$，求矩阵$\boldsymbol{B}$满足$\boldsymbol{AB}=\boldsymbol{A}+2\boldsymbol{B}$
6. 设A，B为两个随机事件，$P(A)=0.4, P(A\cup B)=0.7$，若A，B互不相容，求B的概率，若A，B相互独立，求B的概率。
7. 从[0,1]中随机取两个数，求它们的和小于 5/4，积小于 1/4 的概率。
8. 四个人分别给四个不同的朋友写信，他们写的信随机投给一个未收到信的人，求四封信全部正确和全部错误投递出的概率。
9. 证明$S=\{(x_1,x_2)|x_1+2x_2\geq 1, x_1-x_2\geq 1\}$是凸集。
10. **试说明牛顿法的求解步骤。**

试用牛顿法求解

$$\min f(x) = \frac{1}{2}x_1^2 + \frac{9}{2}x_2^2, \quad x^{(0)} = (9,1)^T$$

11. 简单有向图有 21 条边，3 个度为 4 的结点，其余均为度为 3 的结点，求此图有多少个节点。

12. 证明：设G为简单有向图，图中任何两个结点间有且只有一条有向边相连，证明该图所有结点的入度的平方和等于所有结点出度的平方和。

13. 已知图的邻接矩阵$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$，求该图的补图，并分别计算二者的关联矩阵和谱。

第三章：Python 语言初步

本章节是 python 语言初步，考查学生对 python 语言的掌握能力：

1—5 题是对 python 的基础使用，要求熟练掌握特定函数，难度较低；

6—7 题是综合完成一个具有功能性的小系统，要求有熟练的编程技巧和代码逻辑，难度较高

1. 输出 9 行内容，第 1 行输出 1，第 2 行输出 12，第 3 行输出 123，以此类推，第 9 行输出 123456789。
2. 随机生成 20 个学生的成绩，并定义函数来判断这 20 个学生成绩的等级（100-90 为 A，80-90 为 B，其余为 C）。其中生成 1-100 随机数的语句为：score = random.randint(1,100)。
3. 1） 生成一个大文件 data.txt,要求 1200 行，每行随机为 0~20 的整数;

   2） 读取 data.txt 文件统计这个文件中出现频率排前 10 的整数;
4. 输入某年某月某日，判断这一天是这一年的第几天。
5. 利用切片操作，实现一个 trim()函数，去除字符串首尾的空格，注意不要调用 str 的 strip()方法。
6. **斗地主是一款风靡全国的纸牌类游戏，有着广泛的群众基础。一副牌共有 54 张，包括大王、小王以及黑桃、红桃、梅花、方块四种花色各 13 张。请模拟斗地主发牌过程。**

   【功能要求】

   1）三人制斗地主，每人 17 张牌，留有三张底牌；

   2）大小顺序：大小王、2、A 以及 K 到 3（3 为最小）；

   3）同一大小的牌按照黑桃、红桃、梅花、方块的顺序排列；

   4）最后展示出三个玩家的牌，以及底牌；
7. **使用 Python 语言，设计一个小型的学生宿舍管理程序，系统用户为宿舍管理员。**

   【功能要求】

   1）学生信息：学号、姓名、性别(男/女)、宿舍房间号、联系电话

2）系统功能：

(1) 可按学号查找某一位学生的具体信息

(2) 可以录入新的学生信息

(3) 可以显示现有的所有学生信息

【程序要求】

1）使用函数 列表 字典 字符串 条件循环等解决问题

2）程序规模在 80-200 行左右

第四章：数据预处理

本章节是数据预处理，考察对数据预处理方法的掌握情况：

1—6 题是数据预处理方法的基本使用，包含对相关概念的理解，难度中等；

7—8 题除了要理解基本的概念，还要灵活运用，对 python 有较为熟练地掌握，难度较高。

1. 请使用股票的交易信息的数据(数据来源：http://tushare.org/)中的数据集绘制成交价格日线图。
2. 请使用 Kaggle 的房价预测竞赛的数据集(数据集来源：Kaggle—House Prices: Advanced Regression Techniques)中的数据集对其中的缺失值进行处理
3. 假设属性 age 包括如下值：13，15，16，16，19，20，20，21，22，22，25，25，25，30，33，33，35，35，36，40，45，46，52，70。
   （a） 使用个数为 3 的箱，用箱均值平滑以上数据。
   （b） 如何确定数据中的离群点
   （c） 还有什么方法来进行数据平滑
4. 以下规范化方法的值域是什么？
   （a） 最小-最大规范化
   （b） z 分数规范化
   （c） 小数定标规范化
5. 使用 4.3 中的 age 数据，完成以下操作：
   （a） 最小-最大规范化
   （b） z 分数规范化
   （c） 小数定标规范化
   （d） 对于给出的数据，你愿意选择哪种规范化方法。给出你的理由
6. 假设有 12 个价格记录，如下所示：

   5，10，11，13，15，35，50，55，72，92，204，215

   使用下面的方法将其划分为三个箱进行离散化：
   （a） 等频划分
   （b） 等宽划分
   （c） 聚类

7. ChiMerge 算法是一种基于卡方值的自下而上的离散化方法。它依赖于 χ^2 的值：具有最小 χ^2 值的相邻区间合并在一起，直到满足停止标准。请使用鸢尾花数据集（可以在 Sklearn 包中获得）作为待离散化集合，使用 ChiMerge 算法，对四个数值属性分别进行离散化。

8. 对如下问题，使用伪代码或 Python 语言，给出一个算法

（a） 对于标称数据，基于给定模式中属性不同值的个数，自动产生概念分层

（b） 对于数值数据，基于等宽划分的原则，自动产生概念分层

（c） 对于数值数据，基于等频划分的原则，自动产生概念分层

第五章：分析方法初步

本章节是数据分析方法初步，利用所学算法对数据进行分析：

1—3 题是考察基本概念，难度较低；

4—6 题复现完成相关算法，需要对算法有深入的理解，并且能够熟练运用 python 语言，难度较高；

7—13 题是对利用 Sklearn 库中的数据处理函数对实际数据进行处理并实现相应任务，需要对 Sklearn 熟练运用，难度较高。

1. 说出下面任务的 T，P，E。T 代表任务(task)，P 代表任务 T 的性能(performance)，E 代表经验(experience)：

   预测学生是否能够考上研究生；

   预测下节课有多少学生旷课；

   学生根据兴趣聚成几个社团。

2. 列举身边采用机器学习方法解决实际问题的例子。

3. 对糖尿病数据集进行线性回归分析，其中糖尿病数据集取自 Sklearn 中的 datasets，数据集包含 442 个患者的 10 个生理特征（年龄，性别、体重、血压）和一年以后疾病级数指标。通过线性回归，预测糖尿病病情。

4. 实现线性回归算法（不使用已有的机器学习工具包），并使用线性回归实现波士顿房价数据集的分析。

5. 实现朴素贝叶斯算法，并对鸢尾花数据集进行分类检验分类结果。

6. 实现 KNN 算法，并对鸢尾花数据集进行分类检验分类结果。

7. 采用决策树，KNN，朴素贝叶斯，SVM，logistic 回归等分类算法预测病人是否患有乳腺癌，乳腺癌数据集取自 Sklearn 的标准数据集。

8. 使用 KNN 算法对手写数字数据集进行分类，对测试集进行预测并计算其各个分类性能指标。

9. 使用真实的新闻分类数据集采用支持向量机分类算法对其进行分类，最终使用 Sklearn 的自动调参工具对模型进行调优。

10. 手写 K 均值聚类算法，实现对鸢尾花数据集的聚类，然后计算 jaccard 系数作为聚类性能评价指标。

11. **采用乳腺癌数据集进行聚类分析，不使用数据集中的标记数据只使用属性值，K-均值聚类与 AGENS 聚类设置簇的个数为 2，对三种聚类结果都进行可视化，对比乳腺癌数据集中的真实标记值，比较几种聚类方法的性能。**
12. **采用 Sklearn 中自带的手写数字数据集构建神经网络模型，并计算其预测准确率。**
13. **采用 bagging 集成方法对乳腺癌数据集进行分类，其中基学习器分别选取决策树，KNN，SVM 等，对比 5.5 时的准确率，总结集成学习优点。**

第六章：数据科学实践

本章节是数据科学实践内容，对课程案例的深入探索：

1—4 题要求在理解课程案例后，能够用新的方法比较数据信息并且进行更加深入的探索，需要熟练运用前几章所学的内容，难度较高。

参考教材《数据科学导论》第六章内容，对应题目案例。

1. **案例 1 尝试使用数据特征中最有意义的前 k 个特征进行后续的训练，并观察和使用全量特征之间的差别。**
2. **案例 3 中使用数据的统计特征进行分析，请尝试使用循环神经网络这类方法来处理时间序列数据。**
3. **案例 4 中使用岭回归模型预测价格，但是由于模型表达能力的欠缺，会造成效果不佳，尝试使用更加复杂模型进行分析。**
4. **案例 4 提供一个简单的特征拼接来使用特征，请分析如何对特征进行组合能够取得更好的效果？分析不同特征组合之间为什么会造成效果差异。**

第七章：数据科学的重要研究领域

本章节是数据科学的重要研究领域：

1—9 题是概念和知识性问题，除了课本学习到的知识，还需要自己进行深入思考，发散性思考，难度中等。

1. **文本分析的任务都有哪些，举例说明。**
2. **简述图像分析方法与视频分析方法的区别和联系？**
3. **图像分析方法的基本任务是什么？**
4. **查阅资料，了解图像视频分析方法的最新研究进展。**
5. **简述网络结构分析的概念与划分。**
6. **列举几个你生活中的遇到的复杂网络，它们属于同质信息网络还是异质信息网络？**
7. **社交网络分析还在哪些领域发挥了重要作用，请举例说明？**
8. **什么是可视化分析，分类依据是什么，可以分为几类？**
9. **查阅资料，了解社交媒体数据、交通轨迹数据、图数据的最新可视化研究进展。**

第八章：大数据处理技术简介

本章节是大数据处理技术简介：

1—5 题是对课程中基本概念的考察，难度较低；

6—9 题需要对大数据处理技术进行掌握和运用，有一定的工程量，难度较高。

1. 云计算有哪几种服务类型，举例说明。
2. 云计算有哪几种部署方式，各有什么特点？
3. 云计算与其他计算模式有何不同？
4. 虚拟化技术主要有哪几种？
5. 亚马逊云服务提供哪几种服务？
6. **尝试百度智能云等开放平台，根据个人需求建立一个文字识别或图像识别工具。**
7. **假设你现在要建立一个个人网站，比较在本地搭建与使用云服务器搭建的过程区别。**
8. **查阅 Hadoop 文档，实现对 HDFS 中的文件进行单词数量统计。**
9. **查阅 Spark 文档，实现对 HDFS 中的文件进行单词数量统计并从多到少排序。**

二、课程大作业

题目一：商品价格信息预测

- **题目背景**

考虑到网上销售的产品数量，产品定价在规模上变得更加困难。服装有很强的季节性定价趋势，受品牌影响很大，而电子产品价格根据产品规格波动。如何根据以往信息进行合理定价，有效地帮助商家进行商品的销售是一个有意义的问题。

- **分析目标**

通过给出的商品描述、商品类别和品牌信息，并结合训练数据中的商品价格来给新商品定价格。Eg ：

商品名称	品牌名称	商品描述	商品类别
美杜莎羊皮飞行员夹克外套男	Versace	1、时尚衣领设计，经典百搭，休闲舒适；2、简约袖口设计；细节尽显品质；3、精湛的制作工艺，细节彰显品质；4、精心挑选高品质面料，手感好，面料舒适	服饰
新款秋冬季男士韩版潮流连帽帅气夹克衣服	美特斯邦威	一件精心设计、不挑身材的保暖夹克。舒适，时尚，有型。	服饰

显然 Versace 的衣服价格上应该远高于美特斯邦威的衣服，并且在商品描述中，可以发现两者描述有细微差别。（本 project 旨在对文本信息进行分析，提取文本信息中重要信息，推导出和价格之间的潜在关系）

- **数据字段分析**

字段名	含义
train_id / test_id	商品的序号值
name	商品名称
category_name	商品所属类别
item_condition_id	商品当前是否有货
brand_name	商品品牌
shipping	是否包邮
item_description	商品描述
price	商品

- **数据集**

train.csv 训练集 （含 price）

test.csv 测试集 (不含 price) ; label_test.csv 测试集中对应的 price

f_test.csv 最终的评价数据集 （不含 price）

- **评价指标**

评价的使用的是 Mean Squared Logarithmic Error: 计算的方式如下

$$\mathbf{MSLE} = \frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(\alpha_i + 1))^2$$

其中代表测试集的样本数；代表的是预测的商品价格值；代表实际的销售价格。

题目二：健康保险的客户投保预测

- **题目背景**

这是一家为客户提供健康保险的保险公司，现在他们需要你的帮助来建立一个模型来预测过去一年的投保人（客户）是否也会对公司提供的汽车保险感兴趣。

- **分析目标**

通过已投健康保险的客户的个人信息以及车辆信息，来预测该客户是否对公司提供的汽车保险感兴趣。

- **数据字段分析**

id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
Policy*Sales*Channel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested

- **数据集**

Response 为 label，即在测试集中需要预测的部分。

训练集：304888 条

测试集：76221 条

- **评价指标**

AUC(Area Under Curve) 被定义为 ROC 曲线下的面积。

题目三：信用卡意向预测问题

- **题目背景**

GAMMA 银行是一家私人银行，经营各种银行产品，如储蓄账户、活期账户、投资产品、信贷产品等。该行还向现有客户交叉销售产品，为此，客户使用不同的通信方式，如电视广播、电子邮件、网上银行推荐、手机银行等。在这种情况下，GAMMA 客户银行希望将其信用卡交叉销售给现有客户。银行已经确定了一组有资格使用这些信用卡的客户。

现在，银行正在寻求您的帮助，以确定可能对推荐的信用卡表现出更高意向的客户。

- **分析目标**

通过银行收集到的客户属性数据，预测客户是否对当前推出的信用卡感兴趣。

- **数据样例**

ID	Gender	Age	Region_code	Occupation	Channel_Code	Vintage	Credit_Product	Avg_Account_Balance	Is_Active	Is_Lead
NNVBBKZB	Female	73	RG268	Other	X3	43	No	1045696	No	0
IDD62UNG	Female	30	RG277	Salaried	X1	32	No	581988	No	0

- **各字段含义**

Variable	Definition
ID	ID for Customer
Gender	Gender of Customer
Age	Age of Customer
Region_Code	Code of the Region for the Customer
Occupation	Occupation Type for the Customer
Channel_Code	Acquisition Channel Code for the Customer(Encoded)
Vintage	Vintage for the Customer(In Months)
Credit_Product	If the Customer has any active credit product
AvgAccountBalance	Average Account Balance for the Customer in last 12 Months
Is_Active	If the Customer is Active in last 3 Months
Is_Lead(Target)	If the Customer is interested for the Credit Card 0:not interested 1:interested

Is_Lead 即为需要预测的内容

训练集 ： 196580

测试集 ： 49145(without Is_Lead)

- **评价指标**

AUC(Area Under Curve)：ROC 曲线下与坐标轴围成的面积。AUC 的取值范围在 0.5 和 1 之间。AUC 越 接近 1，检测方法的真实性越高。

```
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve()
roc_auc = auc(fpr,tpr)
```

题目四：基于客户行为贷款违约预测

- **题目背景**

GAMMA 银行是一家私人银行，经营各种银行产品，如储蓄账户、活期账户、投资产品、信贷产品等。 银行向客户提供贷款。

现在，银行正在向您寻求帮助，以确定可能对不会违约的客户提供贷款。

- **分析目标**

通过银行收集到的客户属性数据，预测客户贷款后是否在未来会出现违约。

- **数据样例**

id	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	Profession	CITY	STATE	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Risk_Flag
1	1303834	23	3	single	rented	no	Mechanical_engineer	Rewa	Madhya_Pradesh	3	13	0
2	7574516	40	10	single	rented	no	Software_Developer	Parbhani	Maharashtra	9	13	0

- **各字段含义**

Variable	Definition
Income	Income of the user
Age	Age of the user
Experience	Professional experience of the user in years
Married/Single	Whether married or single
House_Ownership	Owned or rented or neither
Car_Ownership	Does the person own a car
Profession	Profession
City	City of residence
STATE	State of residence
CURRENT_JOB_YRS	Years of experience in the current job
CURRENT_HOUSE_YRS	Number of years in the current residence
Risk_flag(Target)	whether there has been a default in the past or not

Risk_flag 即为需要预测的内容

训练集 ： 201600 条

测试集 ： 50400 条

- **评价指标**

AUC(Area Under Curve)：ROC 曲线下与坐标轴围成的面积。AUC 的取值范围在 0.5 和 1 之间。AUC 越 接近 1，检测方法的真实性越高。

```
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve()
roc_auc = auc(fpr,tpr)
```

题目五：黑色星期五销售预测

- **题目背景**

一家零售公司“GAMMA 私人有限公司”希望了解客户针对不同类别的各种产品的购买行为（特别是购买金 额）。他们分享了上个月选定的大批量产品的不同客户的购买摘要。该数据集还包含客户人口统计数据 （年龄、性别、婚姻状况、城市类型、居住城市）、产品详细信息（产品 ID 和产品类别）以及上月的总 购买量。

现在，他们想建立一个模型来预测客户对各种产品的购买量，这将有助于他们针对不同的产品为客户创 建个性化的报价。

- **分析目标**

通过公司收集到的客户属性数据，预测客户在黑色星期五的购买力。

- **数据样例**

User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
1000001	P00069042	F	0-17	10	A	2	0	3			8370
1000001	P00248942	F	0-17	10	A	2	0	1	6	14	15200

- **各字段含义**

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation(Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product*Category*1	Product Category (Masked)
Product*Category*2	Product may belongs to other category also (Masked)
Product*Category*3	Product may belongs to other category also (Masked)
Purchase(target)	Purchase Amount (Target Variable)

Purchase 即为需要预测的内容

训练集 ： 440055 条

测试集 ： 110013 条

- **评价指标**

AUC(Area Under Curve)：ROC 曲线下与坐标轴围成的面积。AUC 的取值范围在 0.5 和 1 之间。AUC 越 接近 1，检测方法的真实性越高。

```
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve()
roc_auc = auc(fpr,tpr)
```