上海市高等学校信息技术水平考试（二三级）

《数据科学技术及应用》考试大纲

（2022 年版）

一、考试性质

上海市高等学校信息技术水平考试是上海市全市高校统一的教学考试，是检测和评价高校信息技术基础教学水平和教学质量的重要依据之一。该项考试旨在规范和加强上海高校的信息技术基础教学工作，提高学生的信息技术应用能力。考试对象主要是上海市高等学校在校学生。考试每年举行一次，通常安排在当年的十月下旬、十一月上旬的星期六或星期日。凡考试成绩达到合格者或优秀者，由上海市教育委员会颁发相应的证书。

本考试由上海市教育委员会统一领导，聘请有关专家组成考试委员会，委托上海市教育考试院组织实施。

二、考试目标

“数据科学技术及应用”考核学生对数据科学工作流程的理解，应用统计分析、可视化分析、建模分析等方法对数据进行处理，发现有价值信息的综合能力。考试内容涵盖相关理论知识掌和基本方法的应用实践，要求学生具有应用统计分析和机器学习方法解决数据科学实际问题、提出解决方案和决策建议的能力。

三、考试内容和要求

知识领域	知识单元	知识点	要求
数据科学概述	数据处理的工作流程	问题描述，数据准备、数据探索、建模预测、结果可视化	理解
	大数据技术	大数据 4V 特性、大数据技术	了解
	数据分析工具	Python 编程、Anaconda 集成环境	掌握
多维数据与运算	多维数据	创建一维、二维数组对象	掌握
		数组元素索引、切片、条件筛选	掌握
	多维数据运算	基本算术运算、函数与矩阵运算	理解
		随机生成函数	掌握
数据汇总	数据文件读写	CSV、txt、Excel 文件的读写	掌握

知识领域	知识单元	知识点	要求
与统计	数据清洗和规整化	表结构数据组织、索引和筛选	掌握
		缺失、重复数据处理	掌握
		数据合并	掌握
		数据排序	掌握
	统计分析	统计的基本概念、常用统计量的意义	理解
		常用统计函数，算术运算、聚合	掌握
可视化数据探索	常用可视化分析图形	散点图、柱形图、饼图、直方图、概率密度图、箱形图、折线图、半对数图	掌握
	绘图	创建图形对象、绘制图形	掌握
		绘制子图	掌握
		图元设置	理解
		图形文件保存	理解
机器学习建模分析	机器学习基础	有监督学习、无监督学习	理解
		训练集、测试集、划分方法	掌握
	回归分析	回归分析任务、线性回归的基本原理	理解
		线性回归建模、性能分析	掌握
	分类分析	分类分析任务	掌握
		逻辑回归建模，性能分析	了解
		决策树建模、性能分析	掌握
		集成学习原理、集成学习建模	理解
		支持向量机建模、性能分析	了解
	聚类分析	聚类分析目标	掌握
		K-Means 聚类基本原理	理解
		K-means 建模、性能指标、肘部原理	掌握
神经网络和深度学习	神经网络	感知器、前馈神经网络结构	理解
		神经网络分类与回归建模	掌握
	深度学习	深度学习基础知识、建模	理解
图文语音与时序数据处理	文本数据处理	自然语言处理的基本知识	掌握
		中文文本处理步骤：分词、词性标注、特征提取、语言表示模型	理解
		文本分类的基本方法	了解
	数字图像处理	数字图像表示方法	掌握
		图像数据存储	理解
		卷积神经网络、图像分类	了解
	语音识别	语音数据表示	理解
		语音识别的基本方法	了解
	时序数据处理	时序数据的特征提取、表示	掌握
		基于神经网络的时序预测	了解
案例分析	应用实例数据组织与处理	数据科学应用场景、行业领域数据组织方式、应用实例适应性的探索性分析技术以及机器学习建模分析方法	综合应用
创新应用	创新、创意应用	创新、创意应用	理解

备注：

知识与技能的学习考核要求分为**了解**、**理解**、**掌握和综合应用**四个层次，其含义分别为：

- 了解：知道某原理、现象、方法或技术的存在及特点（比如一些复杂原理、新现象、新技术、新工具等）。
- 理解：懂得某原理、现象、方法或技术的核心知识和使用方法。
- 掌握：熟知并能运用某原理、方法或技术解决问题。
- 综合应用：能选择和运用几种相关原理、方法或技术解决问题

四、试卷结构

题号	题型	题量	分值	考核内容	考核目标
一	单选题	12 题	24 分	数据处理方法 大数据概念 多维数据 表结构数据 汇总统计 可视化分析 图文声音序列数据处理	数据组织分析能力 机器学习建模分析能力 持续学习新技术能力
二	多选题	5 题	10 分	表结构数据 汇总统计 可视化分析 图文声音序列数据处理	数据组织分析能力 机器学习建模分析能力 持续学习新技术能力
三	简答题	2 题	16 分	数据科学应用场景、行业领域数据组织方式、适用的处理方法和技术	数据科学思维能力
四	程序填空题	4-5 题	52 分	根据实际问题，应用合理方法实现探索性分析	数据组织分析能力
五	综合应用题	7-10 小题	48 分	根据实际应用产生的数据集和提出的分析目标，选择合理的探索性分析技术以及机器学习常用方法，编写程序，实现分析功能，并对实现结果进行解释说明	数据组织分析能力 机器学习建模分析能力 领域应用方案及原型设计能力 创新应用能力
合 计		30-34 题	150 分		

五、相关说明

1. 考试时间：150 分钟。
2. 试卷总分：满分 150 分。
3. 等第：不合格、二级合格、三级合格、三级优秀。各等第分数线由考委会划定。

4. 考试方式：考试采用基于网络环境的无纸化上机考试。

5. 考试环境：

- 上海市高等学校信息技术水平考试通用平台。
- 开发语言：Python。开发软件版本环境：
- Anaconda 3.5.1 以上（适用 Python 3 版本）

6. 建议学时数：48 学时。

7. 参考教材：

- 《数据科学技术与应用》，宋晖、刘晓强主编，电子工业出版社，2018。

8. 先修课程：任意一种高级程序设计语言。

六、题型示例

1.单选题

【例】建模分析时，通常用于训练的样本数量_____测试集的样本数量。

A. 小于　　B. 等于　　C. 大于　　D. 小于等于

【参考答案】C

【能力目标】理解使用机器学习算法建模，性能评估的基本原则和方法，考核机器学习建模分析能力

【知识内容】训练集、测试集、划分方法

2. 多选题

【例】关于饼图的描述，错误的是_____。

A. 描述总体的样本值的构成比

B. 饼图每个扇形表示一类样本占总体的百分比

C. 描述总体的各样本区间的样本数量

D. 饼图反映多个总体取值之间的数量关系

【参考答案】CD

【能力目标】理解可视化图形分析的目标，考核数组分析能力

【知识内容】常用可视化分析图形

3. 填空题

【例】统计量“_____”描述样本个体距离均值的离散程度 。

【参考答案】方差

【能力目标】理解统计量“方差”的含义，分析目标

【知识内容】统计的基本概念、常用统计量的意义

4. 论述题

【例】请描述自己专业领域某个具体场景所涉及的数据，给出各项数据名称，含义以及数据的类型（连续数值/可选项/文本/图像/视频/声音/时序）等。

【能力目标】从数据分析的角度了解专业领域的数据，组织专业数据，考核数据科学意识

【知识内容】数据的多样性和数据类型

5. 程序填空题

【例】某网商发售的陕西 250g 红富士苹果，方差为 10g。假设苹果的重量服从正态分布，某用户买了 5 箱，每箱 20 个。

1）生成 5*20 的数组保存每箱苹果的实际重量并显示；

2）统计每箱苹果的重量；

3）统计所有苹果中小于 240g 的个数；

源程序代码如下（fill_1.py）：

```
import numpy as np
#设置显示精度为两位小数
np.set_printoptions ( precision=2, suppress=True )
#按照正态分布随机生成 5*20 的数组模拟苹果重量，并输出
apple = np.    （1）    (250,10, size = (5,20) )
print (""1. \n"", apple)
#输出每箱苹果的重量
print(""2.每箱实际重量 \n"", apple.(    （2）    ) )
#统计所有苹果中重量小于 240g 的个数
print(""3.小于 240g 的苹果\n"",     （3）   .sum() )
```

【参考答案】

(1)【random.normal】

(2)【sum(axis = 1)】/【sum(axis = 0)】

(3)【(apple<240)】

【能力目标】掌握使用多维数组组织数据，进行运行，考核数据组织与分析能力

【知识内容】

（1） 创建多维数组

（2） 多维数据运算

（3） 数组筛选

6. 综合应用题

【例】动物具有多种特征，根据这些特征动物可以分为哺乳类、鸟类、爬行类、鱼类和两栖类等。animals.csv 记录了动物的名字、毛发、蛋、腿、尾巴等多种特征（具体说明见“数据集说明”文件），类型被标记为 Mammal（哺乳动物）、bird（鸟类）、others（其他）三大类。

请根据数据集（animals.csv）文件格式，正确获取数据样本进行预处理，统计分析，建立分类模型；请尝试多种算法，比较分类的性能。具体要求如下：

1) 从文件中读出所需的数据，根据分析需求将所需的数据保存到 DataFrame 中；

2) 数据清洗，判断数据集中是否有缺失数据，并采取合适的方法处理；

3) 统计各类动物的数量，并列出每类动物的名字；

4) 数据预处理，将'type'的值转换为数值类型；

5) 使用散点图矩阵分析动物类型与蛋、尾巴特征间的相关性，并计算他们之间的相关系数；

6) 选择合适的数据列作为特征和标签形成数据集用于训练分类模型，并将分为训练集和测试集；

7) 在训练集上建立分类模型，在测试集上测试模型预测的准确性。 在已学习的分类方法（决策树、支持向量机、神经网络）中试用两种算法；

8) 根据第 7）步的运行结果，说明两种算法在动物分类数据上的性能。请将结果用文字描述在程序文件给出的注释行中。″

【能力目标】理解应用场景需求，选择分析方法和技术，实现分析目标，考核领域应用方案及原型设计能力

【知识内容】

（1） 数据汇总统计、可视化分析

（2） 机器学习，有监督学习，神经网络