2022-2023 学年第一学期信安 201 班《机器学习》期中测验

姓名　　学号

一、单选题(共 30 分，每题 1.5 分)

1. 下列___D__不属于人工智能的研究领域。
A. 模式识别　B. 机器学习　C. 深度学习　D. 编译原理

2. 下面能实现人工智能算法的开发环境有（D）
A. C 语言　B. Java 语言　C. Python 语言　D. 以上都可以

3. 被广泛认为是 ai 诞生的标志的是（C）
A. 1946 年计算机的诞生　B. 1936 年图灵机的出现
C. 1956 年达特茅斯会议　D. 1943 年神经网络的提出

4. 可以避免类型错误的函数是？B
A. str()　B. vars()　C. type()　D. chr()

5. 假设当前示例为：types =['爱情片','动作片','喜剧片']
我们在使用列表时，以下哪个选项，会引起索引错误？D
A. types[-1]　B. types[1]　C. types[2]　D. types[3]

6. 下列对字典操作错误的是（A）
dict = {'Name': 'apple', 'type': 'fruit'};
A. 添加键值对　dict.add['color'] = "red"
B. 更新键值对　dict.update({"color": "red"})
C. 删除键 dict.pop('type')
D. 删除字典元素 dict.clear()

7. 关于 Pandas 的说法错误的是：D
A. Pandas 是一个强大的分析结构化数据的工具集
B. Pandas 的基础是 NumPy
C. read_csv()是 Pandas 中用来读取 CSV 文件的函数
D. Python 官方标准发行版中内置了 Pandas 库，无需另外安装

8. 下列关于 DataFrame 说法正确的是（C）。
A、DataFrame 结构是由行索引和数据组成
B、DataFrame 的行索引类型必须为 int
C、创建一个 DataFrame 对象时需要指定索引
D、DataFrame 每列的数据类型必须是相同的

9. 已知，有如下一个二维数组：
arr2d = np.array([[1, 2, 3],[4, 5, 6],[7, 8, 9]])
如果希望获取元素 5，则可以使用（A）实现。
A. arr2d[1, 1]　B. arr2d[1]　C. arr2d[2]　D. .arr2d[1, 0]

10. 下面的程序运行后，输出的结果为（C）。
```
import numpy as np
arr1 = np.array([[0], [1], [2]])
arr2 = np.array([1, 2, 3])
result = arr1 + arr2
print(result.shape)
```

A. (3, 2)　B. (2, 3)　C. (3, 3)　D. (2, 0)

11. 两个 Series 或 DataFrame 合并时，没有对齐的位置会使用（C）进行补齐。
A. Null　B. 0　C. NaN　D. null_values

12. （B）是指机器学习算法对新鲜样本的适应能力。
A. 模型测试　B. 泛化能力　C. 过拟合　D. 模型训练

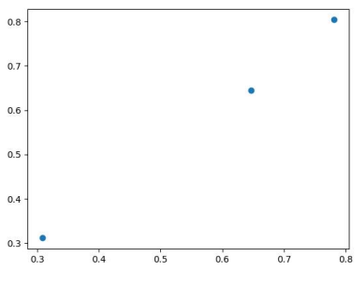
13.
```
import matplotlib.pyplot as plt
from pylab import np
n = 3
Z = np.random.uniform(0,1,n)
plt.pie(Z),plt.show()
```

下列选项为该代码段的运行结果的是（B）

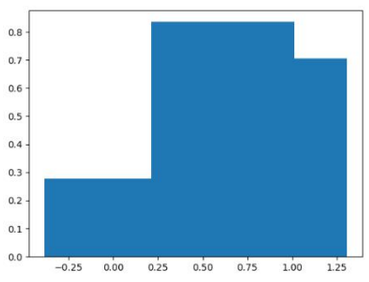
A. [0.16166226 0.75586239 0.22085426]

B.

C.



D.



14. 下列函数中，可以设置当前图形 x 轴范围的是（B）。

A. xlabel() B. xlim() C. title() D. legend()

15. 下列函数中，用于保存当前生成的图表的是（C）。

A. figure() B. hist() C. savefig() D. show()

16. 函数 read_csv()，以（A）作为分隔符

A. ‘,’ B. ‘\t’ C. 空格 D. ‘\d’

17. 请阅读下面一段程序：

```
import pandas as pd
df_obj = pd.DataFrame([[4, -1, -3, 0],
                       [2, 6, -1, -7],
                       [8, 6, -5, 1]])
print(df_obj.sort_values(by=1))
```

执行上述程序后，最终输出的结果为（C）。

D.

	0	1	2	3
1	2	6	-1	-7
0	4	-1	-3	0
2	8	6	-5	1

C.

	0	1	2	3
0	4	-1	-3	0
1	2	6	-1	-7
2	8	6	-5	1

B.

	0	1	2	3
2	8	6	-5	1
0	4	-1	-3	0
1	2	6	-1	7

A.

	3	2	0	1
0	0	-3	4	-1
1	-7	-1	2	6
2	1	-5	8	6

18. 某机器学习模型对训练集的准确率很高，但对测试集则效果不佳，其原因可能是__B_。

A. 欠拟合

B. 过拟合

C. 参数过少

D. 机器性能问题

19. 关于 train_test_split(data)分割后的数据，接收方式正确的是（B）

A. x_train, y_train, x_test, y_test

B. x_train, x_test, y_train, y_test

C. x_test, y_test, x_train, y_train

D. 无所谓，随便怎么接收都行

20. 下面哪个情形不适合作为 K-Means 迭代终止的条件？B

A. 前后两次迭代中，每个聚类中的成员不变

B. 前后两次迭代中，每个聚类中样本的个数不变

C. 前后两次迭代中，每个聚类的中心点不变

D. 前后两次迭代中，每个聚类的均值不变

一、多选题（共 10 分，每题 2 分）

1. Anconda 支持哪些操作系统？（ABC）

A、Windows

B、Mac OS

C、Linux

D、Android

2. 影响聚类算法效果的主要原因有（ABC）
A. 特征选取
B. 模式相似性测度
C. 分类准则
D. 已知类别的样本质量
3. 以下属于 K-Means 的不足的是（ABD）
A. 很难提前选定 K 值。
B. 不适合非凸不规则形状的簇
C. 无法计算较大的数据集
D. 时间复杂度较高
4. 属于 Pyhton 中常用的列表迭代方法的有（ABC）
A. for 循环遍历
B. 按索引序列遍历
C. 按下标遍历
D. 按存储地址遍历

5. Pandas 的数据结构有（ABC）。
A、Series
B、DataFrame
C、Panel
D、Vector

三、 简答题(共 15 分)
1. 对模型进行评价的常用指标有哪些？(3 分)

准确率，错误率、精确率、召回率、均方误差

2. 请简述过拟合出现的原因以及避免策略。(4 分)

过拟合也称过学习，指模型过度学习了训练数据的固有关系。它的直观表现是算法在训练集上表现好，但在测试集上表现不好，泛化能力差。出现过拟合主要是因为训练集再数量级和模型的复杂度不匹配等原因

避免策略：数据扩增、正规化、交叉验证（答出交叉验证即可得分）

3. 请简述强化学习的核心及优缺点。(4 分)
强化学习的核心是评价策略的优劣，从好的动作中学习优的策略，通过更优的策略使得系统输出向更好的方向发展。
优点：1. 无需预备知识 2. 眼光长远考虑持久回报
缺点：奖励函数设计困难

4. 什么是聚类，聚类和分类的区别是什么？(4 分)
聚类是指将不同的对象划分为由多个对象组成的多个类的过程。聚类是一种无监督学习方法
分类是把不同的数据划分开，其过程是通过训练数据集获得一个分类器，再通过分类器去预测未知数据，分类是一种监督学习方法。

四、程序填空题(共 25 分)
1. (9 分)在 python 中用递归算法求 1！+ 2！+ • • • + n！（n 为正整数，从键盘输入）。请阅读如下程序，在空白处选择适当的表达式或语句。

A. n-1	B. n*n-1	C. n*f(n-1)
D. range(1,n+1)	E. s+f(i)	F. f(i)

```
def f(n):
    if n == 0:
        return 1
    else:
        return （1）
    s = 0
    n = int(input("请输入 n:"))
    for i in （2）:
        s = （3）
    print(s)
```

答案： 1. C 2.D 3.E

2. (16 分)从以下候选答案集合中为每空选择一个正确的答案，将其字母编号填入相应空格。候选答案集合如下：

A． filename	B. KMeans(n_clusters = 3)
C. KMeans(n_clusters = 4)	C． Kmeans.labels
D． Kmeans.tags	E． drawing
F． plotting	G． predict

现给定 iris.data 鸢尾花测量数据集，包括花瓣的长度和宽度、花萼的长度和宽度 4 类特征。鸢尾花有三个品种：setosa，versicolor，virginnica。使用 K-Means 算法对数据集样本进行品种聚类。

```
import pandas as pd
from sklearn.cluster import KMeans

filename = 'iris.data'
data = pd.read_csv(__（1）__, header = None)
#生成k-means模型
X = data.iloc[:,0:4].values.astype(float)          #提取数据集中的四个特征数据
kmeans =_____（2）_____                            #模型定义
kmeans.fit(X)                                      #模型学习
print('means.labels_:\n',_____（3）_____)          #输出聚类结果
pd.__（4）___.scatter_matrix(data, c=kmeans.labels_)   #生成散点矩阵图观察聚类效果
plt.show( )
```

1):【A】
2):【B】
3):【D】
4):【G】

五、 编程题(20 分)

1. KNN 算法采用不同特征值之间的距离进行分类，现有如下样本，请使用 KNN 算法对以下数据集（movies.xlsx, sheet_name=0）进行分类，并对给出的测试数据进行预测。

	A	B	C
1	打斗次数	拥抱次数	分类情况
2	36	1	动作片
3	43	2	动作片
4	0	10	爱情片
5	59	1	动作片
6	1	15	爱情片
7	2	19	爱情片

1. 电影类别数据集

	A	B
1	打斗次数	拥抱次数
2	80	3
3	96	8
4	0	60

2. 待预测数据

```
#导包
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier

#读取数据
_____
#将测试数据放入 DataFrame 中
X_test = pd.DataFrame(_____)
#获取属性
_____
#获取类别
_____
#分类器初始化
_____
#对训练集进行训练
_____
#对测试集数据进行预测
_____
```

答案:

```
#导包
import numpy as np
import pandas as pd
```

```
import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier

#读取数据
movie = pd.read_excel('movies.xlsx', sheet_name=0)
#将测试数据放入 DataFrame 中
X_test = pd.DataFrame({'打斗次数':[80,96,0],'拥抱次数':[3,8,60]})
#获取属性
X = movie[['打斗次数','拥抱次数']]
#获取类别
Y = movie[['分类情况']]
#分类器初始化
knn = KNeighborsClassifier()
#对训练集进行训练
knn.fit(X,Y)
#对测试集数据进行预测
knn.predict(X_test)
```