# 北京大学信息学院考试试卷

**考试科目：** Python与数据科学导论 **姓名：** _____ **学号：** _____.
**考试时间：** 202* 年 6 月____日 **任课教师：** 胡俊峰

| 题号 | 一 | 二 | 三 | 四 | 五 | 六 | 七 | 八 | **总分** |
|---|---|---|---|---|---|---|---|---|---|
| 分数 | | | | | | | | | |
| 阅卷人 | | | | | | | | | |

以下为试题和答题纸，共　　页。

**一、 多重选择题：（每小题2分，共16分）**

**从候选答案中选择1到多个正确答案，每个选项占0.5分**

1) 考虑numpy的multiply操作，对x和y两个array来说，以下哪些格式可以正常计算不会报错?

A. x: (1,1) y: (2,30)
B. x: (1,30) y: (5,30)
C. x: (1,5) y. (1,30)
D. x: (20,5) y: (5,20)
答案：（ ）

2) 以下哪种训练技巧会使得模型参数变稀疏?

A. L2正则化 B. L1正则化 C. Dropout D. teacher forcing
答案：（ ）

4、常见的将非平稳序列转换成平稳序列的方法有：

A) 对原序列进行差分操作
B) 对原序列进行平方开放操作
C) 对原序列进行取对数运算
D) 对原序列将原序列进行滑动窗口平均操作
答案：（ ）

6、仿关于离散余弦变换（DCT）和图像压缩，下列说法正确的是

A. DCT能够将时序信号投影到频域特征空间
B. 图像处理中常用二维的DCT，等价于在一维DCT的基础上再做一次DCT
C. JPEG是一种有损图像压缩算法，量化是造成损失的最主要原因
D. 在由量化步长构成的矩阵中，左上角对应高频分量，右下角对应低频分量，因为人眼对于低频分量更敏感，所以矩阵左上角的值普遍小于右下角
答案：（ ）

7、下关于协程(routine)，下面哪些说法是正确的?

A. 全局解释锁(GIL)保证任何时刻只有一个协程执行，因此在多协程之间不需要加锁。

B. 协程之间不是并发的关系

C. 每个协程有单独的python解释器实例执行指令

D. 协程常用于I/O通讯，资源管理与操作响应等

答案：（　　）

8、关于Kmeans聚类算法，下列说法正确的是：

A. 可以看作是一种特殊的矩阵分解问题

B. 该算法属于监督学习

C. 最终的聚类结果与初始聚类中心的选择无关

D. 可以使用不同的距离函数和核函数

答案：（　　）

**二、阅读程序并给出运行结果（共30分）**

```
import copy
ls = [1, 2, [3, 4]]
c = copy.copy(ls)
ls[-1].append(5)
ls.append(6)
print(ls)
print(c)
```

请写出上面程序的运行结果：

5、

```
def func_a(func_a_arg_a, func, **kwargs):
    print(func_a_arg_a)
    func(**kwargs)

def func_b(arg_a):
    print(arg_a)

def func_c():
    print('Hello World')

if __name__ == '__main__':
    func_a(func_a_arg_a='temp', arg_a='Hello Python', func=func_b)
    func_a(func_a_arg_a='temp', func=func_c)
```

请写出上面程序的运行结果：

6、

```
def fun(items):
    se = set()
    for it in items:
        if it not in se:
            yield it
            se.add(it)
a = [1, 5, 2, 1, 9, 1, 5, 10]
print(list(fun(a)))
```

请写出上面程序的运行结果：

7、

```
mat_1 = [['a','b','c'],['d','e','f']]
mat_2 = ['1','2','3']
result = ','.join([i+j
    for vec in mat_1
    for i,j in zip(vec, mat_2)])
print(result)
```

请写出上面程序的运行结果：

8、

```
def call_Fun_counter(func):
    def wrapper(*args, **kwargs):
```

```
        wrapper.calls += 1
        return func(*args, **kwargs)
    wrapper.calls = 0
    return wrapper

@call_Fun_counter
def fib(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        return fib(n-1) + fib(n-2)

print(fib(4))
print(fib.calls)
```

请写出上面程序的运行结果：

9、

```
import numpy as np
a = np.array([[1,2,3],[2,3,4]])
b = a
b += 1
print(a)
a = a.T + b[0,1:]
print(a)
```

请写出上面程序的运行结果：

10、

```
def main():
    try:
        func()
        print("function ends")
    except ZeroDivisionError:
        print('Divided By Zero! ')
```

```
    except:
        print('Its an Exception!')

def func():
    print(1/0)

main()
```

请写出上面程序的运行结果：

## 三、Python代码填空（共 16 分）

用代码进行代码填空

(如果一行能完成，尽量写一行，多于一行代码视情况可能会扣0.5-1分)

1、请用列表表达式生成50以内4的倍数（2分）

_____

2、姓名和年龄的list，请实现按照年龄排序，年龄相同再按姓名排序 （2分）

```
lst = [["zs", 19], ["ll", 54],["wa",23], ["df", 23],["xf",23]]

lst2 = _____
```

3、标准化是特征处理的常规方法，对于给定的矩阵X，请按列对其进行标准化（x-mean/std）（2分）

```
X = np.random.random((100, 30))

X_mean = _____(1分)
```

X_std = _____(1分)

4. 垃圾邮件中可能包含恶意的电子邮箱地址，请写出判断一个字符串是否为合法电子邮箱地址的正则表达式。为简化问题，假设电子邮箱必须有且仅有一个@，包含若干大小写字母、数字、短横线-和英文句点.，并且两个英文句点不能相邻。（2分）

————————————————————————————————

5.考虑CNN的卷积操作，输入为 长L*宽L*通道数c，卷积核大小为 长k*宽k，共有m个卷积核，填充为p步长s，问输出层的尺寸（size）和通道数分别为？（2分）

————————————————，————————————————

**四**、深度学习部分（10分）

下面是一段使用NumPy搭建神经网络的代码，损失函数为交叉熵：

```
import numpy as np
def sigmoid(x):
    return

def forward(W_1, W_2, X, Y):
    z_2 = np.dot(X, W_1)
    a_2 = sigmoid(z_2)

    y_pred = sigmoid(z_3)

    J_z_3_grad =
    J_W_2_grad = a_2.T @ J_z_3_grad
    J_a_2_grad = J_z_3_grad @ W_2.T
    a_2_z_2_grad =
    J_z_2_grad =
    J_W_1_grad =
        return y_pred, (J_W_1_grad, J_W_2_grad)
```

（1）（6分）代码填空

（2）（2分）在MNIST数据集的一个较小子集上使用该神经网络进行训练，发现产生了过拟合现象，写出两种合理的解决方式

（3）（2分）训练一个二分类任务时，如果训练数据类别不平衡（正例较多，负例较少），写出两种合理的提高分类准确率的方法

答案：

**五、简答题（共12分）：**

2、简述协程概念以及python中有哪两种实现协程的机制？（4分）

3、用LSTM+attention机制实现的序列生成模型（seq2seq），在实际使用中在输出序列中会容易生成一些重复的单词。请简要分析这种现象的原因（2分），给出你认为合理的解决问题方案（2分）

**函数实现：（共16分）**

1、下面给出了一个二叉树的类型定义

```
class BinaryTree(object):
    def __init__(self,rootObj):
        self.key = rootObj
        self.leftChild = None
        self.rightChild = None

    def insertLeft(self,newNode):
        if self.leftChild == None:
            self.leftChild = BinaryTree(newNode)
        else:
            t = BinaryTree(newNode)
```

```
            t.leftChild = self.leftChild
            self.leftChild = t

    def insertRight(self,newNode):
        if self.rightChild == None:
            self.rightChild = BinaryTree(newNode)
        else:
            t = BinaryTree(newNode)
            t.rightChild = self.rightChild
            self.rightChild = t

    def getRightChild(self):
        return self.rightChild

    def getLeftChild(self):
        return self.leftChild
    def setRootVal(self,obj):
        self.key = obj

    def getRootVal(self):
  return self.key
```

要求：

1）写出语句序列生成一个该类型的实例r，包含3个结点，根节点内容为字符串“+“，左子树节点内容为字符串“15”，右子树内容为字符串“10”（2分）

语句序列：

2）为这个BinaryTree类添加一个成员函数countLeaf方法，实现对实例中节点数的计数，并返回计数值。比如上面那个树的实例，调用该方法返回值为3（2分）

语句序列（包含函数定义和添加成员函数到类中的语句）：

**2、下面是一个可以正常执行的代码环境的部分代码，要求：**

**1）在空白处补充numpy代码，实现用卷积核进行图像边缘提取的操作（8分）**

**2）给出代码中两条print语句的输出结果（2分）**

**在这里给出上面代码中两条print语句的输出结果：**

**3.在机器翻译任务使用的基于LSTM的seq2seq递归网络模型中，经常会使用attention机制进行结果优化。同时在结果生成中会采用beam search算法。**
**请问beam search算法的简单流程：**