姓名：_____ 学号：_____ 班级：_____ 学院：_____

试卷密封线 试卷密封线

XXXXXXXX 学院

2020 至 2021 学年第 一 学期

《机器学习》期末考试试题（A 卷）

题　目	一	二	三	总分	核分人
得　分					

注：答案请填写在答题卡内，最终答案以答题卡为准

得分	评卷人

一、选择题。（本题共 25 小题，每小题 2 分，共 50 分）

1.贝叶斯公式正确的说法是（ ）

A. P(B|A)=P (A| B)*P(A)/P(B)　　B. P(B|A)=P (A| B)*P(A)/P(AB)

C. P(B|A)=P (A| B)*P(B)/P(AB)　　D. P(B|A)=P (A| B)*P(B)/P(A)

2.a=np.arange(2,15)print(a[1:7:3])输出结果：（ ）

A. [2　5]　　B. [2　5　8]　　C. [1 7 3]　　D. [3 6]

3.Python 语言语句块的标记是(　)

A. 分号　　B. 逗号　　C. 缩进　　D. /

4.Python 内置函数（ ）函数可以返回列表、元组、字典、集合、字符串以及 range 对象中所有元素的个数。

A. len　　B. count　　C. size　　D. shape

5.以下程序的输出结果是（ ）

```
lcat =[“狮子”,“猎豹”,“虎猫”,“花豹”,“孟加拉虎”,“美洲豹”,“雪豹”]
for s in lcat:
    if “豹” in s:
        print(s,end=" ")
        continue
```

A. 雪豹　　B. 猎豹 花豹 美洲豹 雪豹

C. 猎豹　　D. 猎豹
花豹
美洲豹
雪豹

6.在图灵测试中,如果有超过(　)的测试者不能分清屏幕后的对话者是人还是机器,就可以说这台计算机通过了测试并具备人工智能。

A. 30%　　B. 40%　　C. 50%　　D. 60%

7.关于 k-NN 算法，以下哪个选项是正确的？（ ）

A. 可用于分类　　B. 可用于回归　　C. 可用于分类和回归

8.以下两个距离（欧几里得距离和曼哈顿距离）已经给出，我们通常在 K-NN 算法中使用这两个距离。这些距离在点 A（x1，y1）和点 B（x2，Y2）之间。你的任务是通过查看以下两个图形来标记两个距离。关于下图，以下哪个选项是正确的？（ ）

A. 左为曼哈顿距离，右为欧几里得距离　　B. 左为欧几里得距离，右为曼哈顿距离

C. 左或右都不是曼哈顿距离　　D. 左或右都不是欧几里得距离

9.一家公司建立了一个 KNN 分类器，该分类器在训练数据上获得 100%的准确性。当他们在客户端上部署此模型时，发现该模型根本不准确。以下哪项可能出错了？（ ）

注意：模型已成功部署，除了模型性能外，在客户端没有发现任何技术问题。

A. 可能是模型过拟合　　B. 可能是模型未拟合

C. 不能判断　　D. 这些都不是

10.首次提出“人工智能”是在（ ）年。

A. 1916　　B. 1956　　C. 1960　　D. 1946

11.朴素贝叶斯算法，描述正确的是：（ ）

A. 计算先验概率

B. 它假设特征分量之间相互独立

C. 对于给定的待分类项 X={a1,a2,...,an}，求解在此项出现的条件下各个类别 yi 出现的概率，哪个 P(X|yi)最大，就把此待分类项归属于哪个类别。

12.下面关于 ID3 算法中说法错误的是（ ）。

A. ID3 算法要求特征必须离散化

B. ID3 算法是一个二叉树模型

C. 选取信息增益最大的特征，作为树的根节点

D. 信息增益可以用熵，而不是 GINI 系数来计算

13.以下哪些方法不可以直接来对文本分类？（ ）

A. 决策树　　B. K-Means　　C. KNN　　D. 朴素贝叶斯

14.一监狱人脸识别准入系统用来识别待进入人员的身份，此系统一共包括识别 4 种不同的人员：狱警，小偷，送餐员，其他。下面哪种学习方法最适合此种应用需求。（ ）

A. 二分类问题　　B. 多分类问题　　C. 回归问题　　D. 聚类问题

15.关于 L1、L2 正则化下列说法正确的是？（ ）

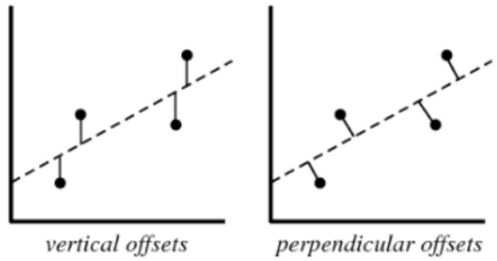
A. L2 正则化得到的解更加稀疏

B. L2 正则化技术又称为 Lasso Regularization

C. L1 正则化得到的解更加稀疏

D. L2 正则化能防止过拟合，提升模型的泛化能力，但 L1 做不到这点

16.下列哪一种偏移，是我们在线性回归模型计算损失函数，例如均方差损失函数时使用的？（ ）



图中横坐标是输入 X，纵坐标是输出 Y。

A. 垂直偏移（vertical offsets）　　B. 垂向偏移（perpendicular offsets）

C. 两种偏移都可以　　D. 以上说法都不对

17.影响基本 K-均值算法的主要因素，不包括( )

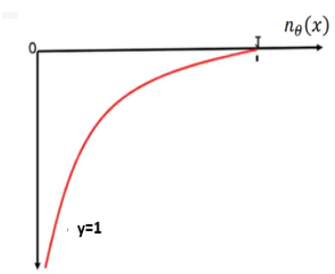
A. 样本输入顺序　　B. 聚类准则（划分簇的原则）　　C. 初始类中心的选取

18.逻辑回归将输出概率限定在 [0,1] 之间。下列哪个函数起到这样的作用？（ ）

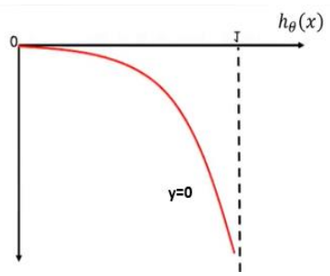
A. Leaky ReLU 函数　　B. Sigmoid 函数　　C. tanh 函数　D. ReLU 函数

19.在逻辑回归中，以下哪个图像显示 y = 1（样本真实值为正样本）的代价函数？（ ）

A.



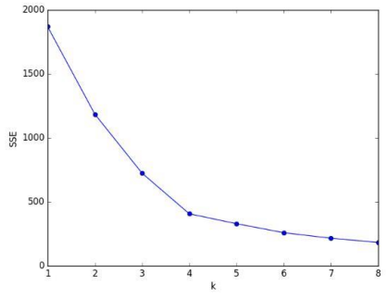
B.



20.K-Means 算法无法聚以下哪种形状的样本？（ ）

A. 圆形分布　　B. 凸多边形分布　　C. 带状分布　　D. 螺旋分布

21.图中是选取不同 k 值时，对应的 SSE(sum of the squared errors，误差平方和)，k 值为多少最好。（ ）



A. 2　　B. 4　　C. 6　　D. 8

22.KNN 算法思想的基本步骤：（ ）

（1）计算已知类别中数据集的点与当前点的距离。

（2）选取与当前点距离最小的 k 个点。

（3）按照距离递增次序排序。

（4）返回前 k 个点出现频率最高的类别作为当前点的预测分类。

（5）确定前 k 个点所在类别的出现频率。

A. 1　3　2　5　4　　B. 1　2　3　4　5

C. 3　1　2　4　5　　D. 2　3　1　5　4

23.如果在大型数据集上训练决策树。为了花费更少的时间来训练这个模型，下列哪种做法是正确的？（ ）

A. 减少树的数量　　B. 增加树的深度　　C. 增加学习率　　D. 减小树的深度

24.目标变量在训练集上的 8 个实际值 [0,0,0,1,1,1,1,1]，目标变量的熵是多少？（ ）

A. $-(\frac{5}{8}log(\frac{5}{8})+\frac{3}{8}log(\frac{3}{8}))$

B. $(\frac{5}{8}log(\frac{5}{8})+\frac{3}{8}log(\frac{3}{8}))$

C. $(\frac{3}{8}log(\frac{5}{8})+\frac{5}{8}log(\frac{3}{8}))$

D. $(\frac{5}{8}log(\frac{3}{8})-\frac{3}{8}log(\frac{5}{8}))$

25.在回归模型中，下列哪一项对于欠拟合（under-fitting）和过拟合（over-fitting）影响最大？（ ）

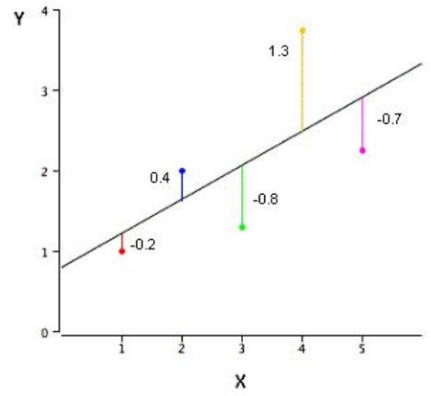
A. 更新权重 w 时，使用的是矩阵求逆还是梯度下降

B. 多项式阶数　　C. 使用常数项

得分	评卷人

二、计算题。（本题共 5 小题，共 50 分），

1.（本小题 4 分）下面这张图是一个简单的线性回归模型,图中标注了每个样本点预测值与真实值的差。计算 SSE（Sum of Squared Error，平方误差之和）。



2.（本小题 8 分）已知逻辑回归模型得到一组逻辑回归结果，要求：

（1）假设阈值为 0.6，写出预测结果。（2 分）

（2）计算出损失函数的值（即真实值与预测值之间的损失值）。（6 分）

逻辑回归结果	逻辑回归预测结果	真实结果
0.40		1
0.65		0
0.20		0
0.80		1
0.70		1

3.（本小题 10 分）预测 20 个西瓜中哪些是好瓜，这 20 个西瓜中实际有 15 个好瓜，5 个坏瓜。某个模型预测的结果是：16 个好瓜，4 个坏瓜。其中，预测的 16 个好瓜中有 14 个确实是好瓜，预测的 4 个坏瓜中有 3 个确实是坏瓜。

注：好瓜为正例，坏瓜为反例。

（1）画出混淆矩阵（4 分）

（2）什么是精确率（Presion），计算准确率 P（2 分）

（3）什么是召回率（Recall），计算召回率 R（2 分）

（4）写出 F1 值的计算公式，求出本例中 F1 值。（2 分）

4.（本小题 12 分）线性回归算法实现。

（1）写出正规方程求解回归系数的公式。（2 分）

（2）已知样本的特征矩阵为 X，目标值向量为 Y=[y0,y1,y2,...,ym]。

编写函数 standRegres 实现功能：通过正规方程求解回归系数。（10 分）

```
def standRegres(xArr,yArr):
```

5.（本小题 16 分）使用朴素贝叶斯分类预测一个未知样本的分类。数据样本用属性"天气","温度","湿度"和"风力"描述。目标分类属性"是否适合打网球"具有两个不同值(即"是","否")。设 C1 对应于分类"是否适合打网球"="是"，而 C2 对应于分类"是否适合打网球"="否"。我们的预测样本为 X = ("天气"="雨","温度"="凉","湿度"="高","风力"="弱")

根据朴素贝叶斯算法计算出 X 的属于不同分类目标值的概率，判断最终的预测结果。

天气	气温	湿度	风力	适合打网球吗？
晴	热	高	弱	否
晴	热	高	强	否
阴	热	高	弱	是
雨	适宜	高	弱	是
雨	凉	正常	弱	是
雨	凉	正常	强	否
阴	凉	正常	强	是
晴	适宜	高	弱	否
晴	凉	正常	弱	是
雨	适宜	正常	弱	是
晴	适宜	正常	强	是
阴	适宜	高	强	是
阴	热	正常	弱	是
雨	适宜	高	强	否