试卷密封线　学院：_____　班级：_____　姓名：_____　学号：_____

XXXXXXX 学院

2020 至 2021 学年第 一 学期

《机器学习》期末考试试题（B 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.关于 Python 的列表，描述错误的选项是（　　）。

A. Python 列表是包含 0 个或者多个对象引用的有序序列

B. Python 列表用中括号[]表示

C. Python 列表是一个可以修改数据项的序列类

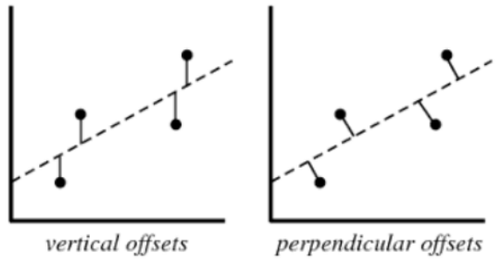
D. Python 列表的长度不可变的

2.目标变量在训练集上的 8 个实际值[1,0,1,1,0,1,0,1]，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. $(\frac{5}{8}log(\frac{3}{8})-\frac{3}{8}log(\frac{5}{8}))$

3.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

4.
```
x = np.arange(10)
print(x[2:])
print(x[-2:])
```
输出结果：（　　）。

A. [1 2 3 4 5 6 7 8 9]
[8 9]　　B. [2 3 4 5 6 7 8 9]
[8 9]

C. [2 3 4 5 6 7 8 9 10]
[9 10]　　D. [1 2 3 4 5 6 7 8 9 10]
[9 10]

5.首次提出“人工智能”是在（　　）年。

A. 1916　　B. 1956　　C. 1960　　D. 1946

6.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

7.下面代码的执行结果是（　　）。
```
ls=[[1,2,3],[[4,5],6],[7,8]]
print(len(ls))
```
A. 4　　B. 3　　C. 8　　D. 1

8.以下哪个是两个数据点 A（0,7）和 B（3,3）之间的曼哈顿距离？（　　）

A. 5　　B. 6　　C. 7　　D. 8

9.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)

10.以下关于字典操作的描述，错误的是（　　）。

A. clear 用于清空字典中的数据　　B. len 方法可以计算字典中键值对的个数

C. keys 方法可以获取字典的值　　D. del 用于删除字典或者元素

11.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

12.Nave Bayes 是一种特殊的 Bayes 分类器,特征变量是 X,类别标签是 C,它的一个假定是（　　）。

A. 各类别的先验概率 P(C)是相等的

B. 以 0 为均值，sqr(2)/2 为标准差的正态分布

C. 特征变量 X 的各个维度是类别条件独立随机变量

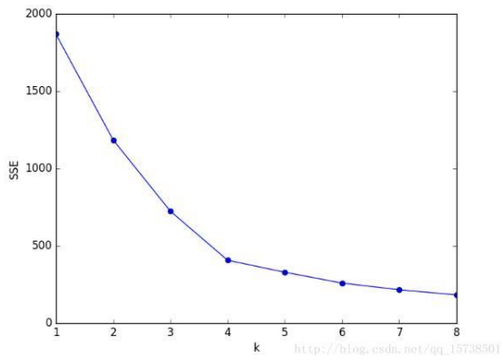
D. P(X|C)是高斯分布

13.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。
14.一般来说，下列哪种方法常用来预测连续独立变量？（　　）
A. 决策树　　B. 线性回归　　C. 朴素贝叶斯　　D. 以上都对
15.以下是两个陈述。以下两个陈述中哪一项是正确的？（　　）
①k-NN 是一种基于记忆的方法，即分类器会在我们收集新的训练数据时立即进行调整。
②在最坏的情况下，新样本分类的计算复杂度随着训练数据集中样本数量的增加而线性增加。
A. ①　　B. ②　　C. ①和②　　D. 这些都不是
16.下面哪句话是正确的？（　　）
A. 增加模型的复杂度，总能减小测试样本误差。
B. 增加模型的复杂度，总能减小训练样本误差。
C. 机器学习模型的精准度越高，则模型的性能越好。
D. 以上说法都不对。
17.以下哪种方法能最佳地适应逻辑回归中的数据？（4 分）
A. 最小二乘法 Least Square Error
B. 极大似然估计方法 Maximum Likelihood
C. 杰卡德距离 Jaccard distance
D. 以上都不对
18.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　　B. 更新权重 w 时，使用的是梯度下降
C. 使用常数项　　D. 多项式阶数
19.关于 kmeans 算法，不正确的是（　　）。
A. 原理简单，容易实现。　　B. K 值很难确定。
C. 聚类效果依赖于聚类中心的初始化。　　D. 对噪音和异常点不敏感。
20.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
21.K-Means 算法无法聚类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
22.在以下不同的场景中,使用的分析方法不正确的有（　　）。
A. 根据商家近几年的成交数据,用线性回归算法拟合出用户未来一个月可能的消费金额公式。
B. 根据用户最近购买的商品信息，用逻辑回归算法识别出淘宝买家可能是男还是女。
C. 用关联规则算法分析出购买了汽车坐垫的买家，是否适合推荐汽车脚垫。
D. 根据商家最近一年的经营及服务数据,用 KNN 算法判断出天猫商家在各自主营类目下所属的商家层级。
23.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
24.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。
25.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　　C. 初始类中心的选取　　D. K 值的选取

得分	评卷人

二、计算题。（本题共 5 小题，共 50 分），

1.（本小题 4 分）已知决策树用字典可以表示为：mytree={'can breath under water?': {0: 'not fish', 1: {'have flippers?': {0: 'not fish', 1: 'is fish'}}}}，请画出相应的决策树。

2.（本小题 8 分）已有训练数据集 DataSet 如下表，使用 KMeans 算法（k=2）对数据进行聚类分析，将数据分成 2 个簇，这里选择点 P1 为第 0 个簇的质心，P2 为第 1 个簇的质心。

DataSet

数据	X 值	Y 值
P1	1	2
P2	6	6
P3	2	1
P4	5	5
P5	7	4
P6	3	2

（1）计算第一次迭代时，每个样本点的聚类评估指标（clusterAssement），将样本点 P4 的结果填入表格。（每空 1 分，共 2 分）

clusterAssement

数据	对应样本所在的簇的序号 P1(第 0 簇质心)P2(第 1 簇质心)	样本距离所在簇质心的距离平方 dist（x,C_i）^{2}
P4		

（2）第一次迭代结束后，哪些点属于第 0 簇？（2 分）哪些点属于第 1 簇？（2 分）

（3）第一次迭代结束后，更新簇的质心，新的质心是什么？（2 分）

3. （本小题 13 分）岭回归算法实现。

（1）岭回归算法采用哪种正则化来简化模型？写出岭回归算法的损失函数？（2 分）

（2）使用正规方程求解，写出岭回归系数的公式。（1 分）

（3）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 RidgeRegres 实现功能：通过正规方程求解回归系数。（10 分）

```
def ridgeRegres(xArr, yArr, lam):
```

4. （本小题 15 分）假设你给出了以下数据，其中 x 和 y 是 2 个输入变量，而 Class 是因变量。

x	y	Class
−1	1	−
0	1	+
0	2	−
1	−1	−
1	0	+
1	2	+
2	2	−
2	3	+

（1）写出 KNN 算法思想的基本步骤。（5 分）

（2）使用 2D 空间显示上述数据，画出训练数据的散点图，用‘〇’表示负样本，‘+’表示正样本。（2 分）

（3）假设你要使用 KNN（k=3）中的欧氏距离来预测新数据点 x = 1 和 y = 1 的类别。该数据点属于哪个类别？（需按照问题（1）写出计算过程）（5 分）

（4）在问题（3）中，使用 KNN 算法时令 k=7，那么 x = 1 和 y = 1 属于哪个类别？（1 分）为什么？（2 分）

5. （本小题 10 分）使用决策树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。使用 ID3 算法构建一个决策树模型时，哪个属性适合做根节点？

（注：所有对数计算均选择 $\log_2$）

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
阴	凉	高	强	是

附件：$\log_2$ 对数计算表：

x	0	1	1/2	1/3	2/3	1/4	3/4	1/5	2/5
$\log_2$（x）	-∞	0	-1	-1.585	-0.585	-2	-0.415	-2.322	-1.322
x	3/5	4/5	1/6	5/6	1/7	2/7	3/7	4/7	5/7
$\log_2$（x）	-0.737	-0.322	-2.585	-0.263	-2.807	-1.807	-1.222	-0.807	-0.485
x	6/7	1/8	3/8	5/8	7/8	1/9	2/9	4/9	5/9
$\log_2$（x）	-0.222	-3	-1.415	-0.678	-0.193	-3.170	-2.170	-1.170	-0.848
x	7/9	8/9	1/10	3/10	7/10	9/10			
$\log_2$（x）	-0.363	-0.170	-3.322	-1.737	-0.515	-0.152			