互动试题题库

满分：123 分

姓名：　　班级：　　学号：

1. 判断题 （ 1.0 分 ）

在基于 SGD 随机梯度下降算法的神经网络中，每次打乱数据是非常重要和必不可少的。

( ) 对

( ) 错

正确答案：对

2. 判断题 （ 1.0 分 ）

决策树的适用面较广，对于分类应用和回归应用，决策树都可以被用来构建模型。

( ) 对

( ) 错

正确答案：对

3. 判断题 （ 1.0 分 ）

由于贝努利贝叶斯比适合于贝努利（二项分布）分布，因此，贝努利贝叶斯只能用于二分类任务。

( ) 对

( ) 错

正确答案：错

4. 判断题 （ 1.0 分 ）

通常，我们认为对于一个系统来说，误差越小越好，因此无论是泛化误差还是经验误差，都是越小越好。

( ) 对

( ) 错

正确答案：错

5. 判断题 （ 1.0 分 ）

常见的性能度量标准有很多，例如：均方误差、准确率、错误率、精度、查全率、查准率等，其中均方误差只能用于回归模型。

( ) 对

( ) 错

正确答案：错

6. 判断题 （ 1.0 分 ）

Anaconda 是 Python 语言最著名的第三方库之一，它可以实现基于矩阵的数据处理、科学运算、可视化以及机器学习等功能。

( ) 对

( ) 错

正确答案：错

7. 判断题 （ 1.0 分 ）

StratifiedKFold 算法比较适合大规模数据集。

( ) 对

( ) 错

正确答案：错

8. 判断题 （ 1.0 分 ）

GridSearchCV 类可以实现交叉验证和网格搜索的整合。

( ) 对

( ) 错

正确答案：对

9. 判断题 （ 1.0 分 ）

线性模型，不仅可以用来模拟线性关系的数据集，同时也可以用来模拟非线性关系的数据集，甚至是高度非线性关系的数据集。

( ) 对

( ) 错

正确答案：对

10. 判断题 （ 1.0 分 ）

在线性回归模型中， 参数 w 表示的是特征的权重，它可以用来衡量某个特征的重要性。

( ) 对

( ) 错

正确答案：对

11. 判断题 （ 1.0 分 ）

随机森林算法利用训练数据构建了一系列的决策树，它根据损失函数最大化原则建立决策树模型。

( ) 对

( ) 错

正确答案：错

12. 判断题 （ 1.0 分 ）

支持向量机只能用于二分类，无法应用到多分类中。

( ) 对

( ) 错

正确答案：错

13. 判断题 （ 1.0 分 ）

套索回归（Lasso Regression），是目前机器学习中性能最好的模型。

( ) 对

( ) 错

正确答案：错

14. 判断题 （ 1.0 分 ）

batch_size 用于设置单次加入进行训练的样本数，它适用于 SGD 随机梯度下降算法，一般设置为内存或显存所能支持的最大容量。

( ) 对

( ) 错

正确答案：对

15. 判断题 （ 1.0 分 ）

与决策树不同，在构建随机森林的时候，树的深度参数 max_depth 只能等于 1.

( ) 对

( ) 错

正确答案：错

16. 判断题 （ 1.0 分 ）

这学期的《机器学习》课程，只要期末考及格就可以通过。

( ) 对

( ) 错

正确答案：错

17. 判断题 （ 1.0 分 ）

SVM 的核心因素也是它的根本原理是最大化类间间隙。

( ) 对

( ) 错

正确答案：对

18. 判断题 （ 1.0 分 ）

由于随机森林是由 1 个或多个决策树构成，因此随机森林的性能总是优于决策树。

( ) 对

( ) 错

正确答案：错

19. 判断题 （ 1.0 分 ）

网格搜索算法是一种基于穷举搜索的方法，非常耗时，所以它不太适合于大规模数据集的调参。

( ) 对

( ) 错

正确答案：错

20. 判断题 （ 1.0 分 ）

大规模数据集不存在样本不平衡的问题。

( ) 对

( ) 错

正确答案：错

21. 判断题 （ 1.0 分 ）

聚类是有监督的过程，而分类是无监督的过程。

( ) 对

( ) 错

正确答案：错

22. 判断题 （ 1.0 分 ）

激活函数对于神经网络来说是必须的。

( ) 对

( ) 错

正确答案：对

23. 判断题 （ 1.0 分 ）

一般来说，岭回归和套索回归的性能要优于最基本的线性回归模型。

( ) 对

( ) 错

正确答案：错

24. 判断题 （ 1.0 分 ）

在 SVM 数据集中，二维平面的分类线，三维空间的分界面都可以称为超平面。

( ) 对

( ) 错

正确答案：对

25. 判断题 （ 1.0 分 ）

数据的最大方差给出了数据最重要的信息。

( ) 对

( ) 错

正确答案：对

26. 判断题 （ 1.0 分 ）

y^hat 一般用来表示预测结果的正确值。

( ) 对

( ) 错

正确答案：错

27. 判断题 （ 1.0 分 ）

基于 KNN 的回归算法所生成的模型不需要进行训练。

( ) 对

( ) 错

正确答案：对

28. 判断题 （ 1.0 分 ）

与决策树不同，在构建随机森林的时候，树的深度参数 max_depth 只能等于 1.

( ) 对

( ) 错

正确答案：错

29. 判断题 （ 1.0 分 ）

基于高斯模型的径向基核和非线性的多项式核的支持向量机一定比基于线性核的支持向量机更好。

( ) 对

( ) 错

正确答案：错

30. 判断题 （ 1.0 分 ）

由于贝努利贝叶斯比适合于贝努利（二项分布）分布，因此，贝努利贝叶斯只能用于二分类任务。

( ) 对

( ) 错

正确答案：错

31. 判断题 （ 1.0 分 ）

随机森林算法利用训练数据构建了一系列的决策树，它根据损失函数最大化原则建立决策树模型。

( ) 对

( ) 错

正确答案：错

32. 判断题 （ 1.0 分 ）

决策树的适用面较广，对于分类应用和回归应用，决策树都可以被用来构建模型。

( ) 对

( ) 错

正确答案：对

33. 判断题 （ 1.0 分 ）

在主成分分析中，要求每个主成分之间是不相交的，即正交的。

( ) 对

( ) 错

正确答案：对

34. 判断题 （ 1.0 分 ）

分层采样算法是指一种非常特殊的数据划分算法，它可以有效解决原始数据集类别不平衡的问题。

( ) 对

( ) 错

正确答案：错

35. 判断题 （ 1.0 分 ）

PCA 主成分分析不仅能够降低数据的复杂性，实现降维，最重要的是在这个过程中没有任何信息的损失。

( ) 对

( ) 错

正确答案：错

36. 判断题 （ 1.0 分 ）

由于随机森林是由 1 个或多个决策树构成，因此随机森林的性能总是优于决策树。

( ) 对

( ) 错

正确答案：错

37. 单选题 （ 1.0 分 ）

用于调用 sklearn 库中用于完成 KNN 分类的子库是哪一个？

A. sklearn.datasets.make_blobs

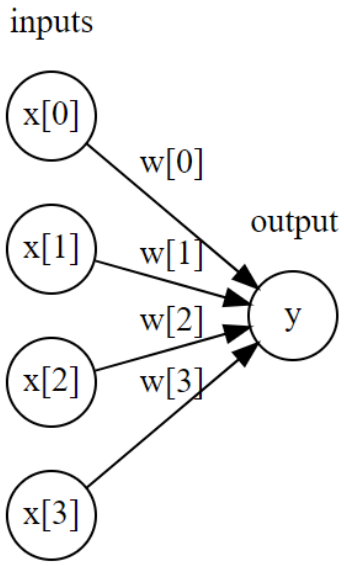
B. sklearn.neighbors.KNeighborsRegressor

C. sklearn.neighbors.KNeighborsClassification

D. sklearn.datasets.make_regression

正确答案：C

38. 单选题 （ 1.0 分 ）



如图所示，假设输入 X[i]分别等于{0,1,2,3}, 权值 w[i]分别等于{0,1,2,3}, 偏移量 b 等于 1。

试求, 输出 y。

A. 0

B. 6

C. 15

D. 36

正确答案：C

39. 单选题 （ 1.0 分 ）

下列误差和错误中，哪一项是由于训练样本的错误而导致？

A. 泛化误差

B. 偏差

C. 方差

D. 噪声

正确答案：D

40. 单选题 （ 1.0 分 ）

下列激活函数中，能够实现将特征限制到区间 [-1, 1] 的是哪一个？

A. Tanh

B. Logistic

C. ReLU

D. pReLU

E. Sigmoid

正确答案：A

41. 单选题 （ 1.0 分 ）

以下子库可以用来载入 train_test_split 训练测试集依赖的是（ ）。

A. sklearn.datasets

B. sklearn.model_selection

C. sklearn.ensemble

D. matplotlib.pyplot

E. numpy

正确答案：B

42. 单选题 （ 1.0 分 ）

以下公示，是哪一种模型的数学表达？

$$f(x) = \begin{cases} P(x=1) = & p \\ p(x=0) = & 1-p \end{cases}$$

$$s.t. 0 < p < 1$$

A. 线性回归模型

B. 岭回归模型

C. 套索回归模型

D. 贝努利贝叶斯模型

E. 高斯被贝叶斯模型

F. 多项式贝叶斯模型

正确答案：D

43. 单选题 （ 1.0 分 ）

StandardScaler 预处理方法可以表示为? = (?−?)/?，其中?表示特征所在列的(　　).

A. 最大值

B. 分界阈值

C. 均值

D. 方差

正确答案：D

44. 单选题 （ 1.0 分 ）

朴素贝叶斯是一种典型的基于概率的机器学习方法，它利用了：（　　）

A. 先验概率

B. 后验概率

C. 以上都是

D. 以上都不是

正确答案：C

45. 单选题 （ 1.0 分 ）

下列误差和错误中，哪一项是由于真实样本和训练样本之间的差异所造成？

A. 泛化误差

B. 偏差

C. 方差

D. 噪声

正确答案：A

46. 单选题 （ 1.0 分 ）

下列哪一个库是 Python 机器学习库？

A. scikit-learn

B. scikit

C. scipy

D. pandas

E. matplotlib

正确答案：A

47. 单选题 （ 1.0 分 ）

以下代码的输出结果是（ ）。

```
import numpy as np
n = 10
num = np.arange(0, n)
print(num)
```

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

B. [0 1 2 3 4 5 6 7 8 9]

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

D. [0 1 2 3 4 5 6 7 8 9]

正确答案：B

48. 单选题 （ 1.0 分 ）

matplotlib 是 python 中最重要的绘图库，其中 plt.show()函数一般放在绘图函数 plt.plot()的（　　）。

A. 前面

B. 后面

C. 前后都行

D. 可以省略

正确答案：B

49. 单选题 （ 1.0 分 ）

对于线性模型来说，如果数据集中的样本具有 m 个特征，那么标准线性模型的参数的个数为（　　）。

A. m-1

B. m

C. m+1

D. 不确定

正确答案：C

50. 单选题 （ 1.0 分 ）

在机器学习库 sklearn 中，用于存储样本拆分函数 train_test_split()的子库是哪一个?

A. sklearn.datasets

B. sklearn.neighbors

C. sklearn.model_selection

D. matplotlib.pyplot

正确答案：C

51. 单选题 （ 1.0 分 ）

以下公示，是哪一种模型的数学表达？

$$f(x) = \begin{cases} P(x=1) = \\ p(x=0) = \end{cases}$$

$$s.t. 0 < p < 1$$

A. 线性回归模型

B. 岭回归模型

C. 套索回归模型

D. 贝努利贝叶斯模型

E. 高斯被贝叶斯模型

F. 多项式贝叶斯模型

正确答案：D

52. 单选题 （ 1.0 分 ）

KNN(K 近邻算法）属于一种典型的（　　）算法。

A. 监督学习

B. 无监督学习

C. 半监督学习

D. 弱监督

正确答案：A

53. 单选题 （ 1.0 分 ）

当我们使用 make_blobs 生产数据集的时候，哪一个属性用于设置样本的数量。

A. n_samples

B. n_features

C. centers

D. random_state

正确答案：A

54. 单选题 （ 1.0 分 ）

针对回归样本生成函数 sklearn.datasets.make_regression()，下列用于设置生成样本个数的超参数是（ ）。

A. n_samples=100

B. n_features=100

C. centers=100

D. noise=100

E. random_state=100

正确答案：A

55. 单选题 （ 1.0 分 ）

二值图像是指将图像上的每一个像素只有两种可能的取值或灰度等级状态，人们经常用黑白、B&W、单色图像表示二值图像。能够将图像转化为二值黑白图的预处理方法是哪一个？

A. Normalizer

B. MaxAbsScaler

C. Binarizer

D. RobustScaler

E. StandardScaler

F. MinMaxScaler

正确答案：C

56. 单选题 （ 1.0 分 ）

在进行线性模型训练的时候，假设 X 具有 10 个特征，请问，下列哪一个符号可以获得最终模型的第 5 个权重参数。

A. coef_(5)

B. coef_[4]

C. coef_[5]

D. coef_(5)

正确答案：B

57. 单选题 （ 1.0 分 ）

SVM 一个重要的特征是，当样本在原始特征空间中线性不可分时，我们可以（　　）。

A. 使用多个分类器一起做联合决策

B. 将特征映射到高维空间再做决策

C. 采用层次化的方法进行决策

D. 自动放弃部分样本

正确答案：B

58. 单选题 （ 1.0 分 ）

如果某模型经过一定程度的简化后，最终只有一个特征变量。则，这个模型将被简化为：（ ）。

A. 一个点的坐标

B. 一条直线的方程

C. 一条曲线的方程

D. 一个平面的方程

正确答案：B

59. 单选题 （ 1.0 分 ）

数据集的划分对于获得优秀的模型至关重要，下列数据集不能出现在模型超参数选择和特征选择的是（ ）。

A. 训练集

B. 验证集

C. 测试集

D. 以上都是

正确答案：C

60. 单选题 （ 1.0 分 ）

根据以下输出结果，判断原始的 python 代码。

类别[0]的概率值为：0.73598

A. print("类别[0]的概率值为：{:.5f}".format(model.predict_proba(X_new)[0]0]))

B. print("类别[0]的概率值为：{:.5f}".format(model.predict(X_new)))

C. print("类别[0]的概率值为：{:.5f}".format(model.score(X_new)))

D. print("类别[0]的概率值为：{:.5f}".format(model.scores(X_new)))

正确答案：A

61. 单选题 （ 1.0 分 ）

下列数据预处理方法可以实现将输入数据的均值置为 0，方差置为 1。

A. StandardScaler

B. MinMaxScaler

C. MaxAbsScaler

D. RobustScaler

E. Normalizer

F. Binarizer

正确答案：A

62. 单选题 （ 1.0 分 ）

下列哪一个方法用于实现算法的训练。

A. model.score()

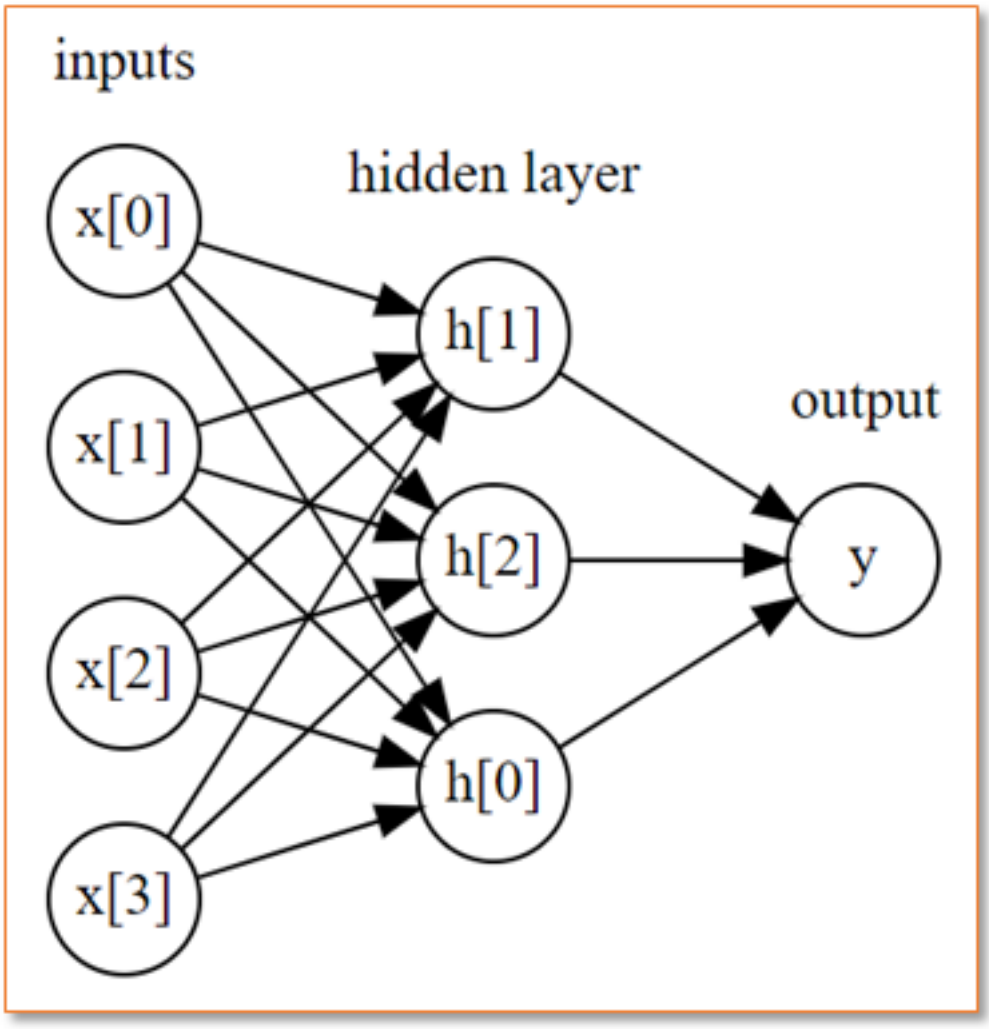
B. model.fit()

C. model.predict()

D. model.scatter

正确答案：B

63. 单选题 （ 1.0 分 ）



如图所示，假设：1. 输入 X[i]分别等于{0,1,2,3},2. 输入层与隐藏层之间的权值 w[i,j]分别等于{0,1,2,3}, 即 x[0]后面的权值 w 都等于 0，x[1]后面的权值都等于 1，依此类推。3. 隐藏层与输出层之间权值 w[i]分别等于{1,2,0},即隐层单元 h[1]后面的权值 w 等于 1，隐层单元 h[2]后面的权值 w 等于 2，隐层单元 h[0]后面的权值 w 等于 0.4. 偏移量 b 等于 1。

试求, 输出 y。

A. 18

B. 28

C. 30

D. 46

正确答案：D

64. 单选题 （ 1.0 分 ）

在下列数据集中，用于调整模型的超参数的是：（ ）。

A. 训练集

B. 验证集

C. 测试集

D. 以上都是

正确答案：B

65. 单选题 （ 1.0 分 ）

对于图像数据来说，通常像素的数值范围是 0~255，但是为了获得更好的性能，我们常常需要将像素的数值转换到 0~1 之间。以下可以实现该功能的方法是哪一个？

A. StandardScaler

B. RobustScaler

C. Binarizer

D. MinMaxScaler

E. MaxAbsScaler

F. Normalizer

正确答案：D

66. 单选题 （ 1.0 分 ）

在 scikit-learn 工具包中，下列哪一个符号用来表示线性回归的权重参数（ ）。

A. coef_

B. intercept_

C. coef

D. intercept

正确答案：A

67. 单选题 （ 1.0 分 ）

下列用于衡量模型在“未来”样本上的性能时所产生的误差，称为：

A. 训练误差

B. 未来误差

C. 泛化误差

D. 经验误差

正确答案：C

68. 单选题 （ 1.0 分 ）

下列数据拆算法中，比较适合用于中小型数据集的是（　　）。

A. StratifiedShuffleSplit

B. ShuffleSplit

C. StratifiedKFold

D. Leave One Out

E. Holdout

正确答案：C

69. 单选题 （ 1.0 分 ）

在利用机器学习进行建模并对自然界的各种现象进行预测时，以下哪一个是步骤是必不可少。

A. 数据载入

B. 数据预处理

C. 模型评分

D. 可视化分析

正确答案：A

70. 单选题 （ 1.0 分 ）

在主成分分析中，第一个新坐标轴的选择是由（　　）决定。

A. 特征值最大的方向

B. 方差最大的方向

C. 响应最强的隐变量

D. 信息量最大的数据源

正确答案：B

71. 单选题 （ 1.0 分 ）

在支持向量机中最大间隙 trick 被称为（　　）。

A. 数据 data

B. 分类器 classifier

C. 优化 optimization

D. 超平面 hyperplane

E. 核化 kernelling

正确答案：C

72. 单选题 （ 1.0 分 ）

在 KNN 算法中，确定最优 K 值时，关键看（　　）。

A. 数据样本

B. 编写代码的程序，例如：Python，Javascript，C++，Matlab

C. 由程序员凭习惯确定

D. 掷骰子确定

正确答案：A

73. 单选题 （ 1.0 分 ）

下列激活函数中，能够实现将负数部分全部归零的是哪一个？

A. Tanh

B. Logistic

C. ReLU

D. pReLU

E. Sigmoid

正确答案：C

74. 单选题 （ 1.0 分 ）

已知：

1. 买彩票中奖的概率为 1%

2. 被天上掉下来的苹果砸中脑袋的概率为 0.1%

3. 如果被天上掉下来的苹果砸中脑袋后，再买彩票中奖的概率为 99%

那么，请问，买彩票中奖时，被天上掉下来的苹果砸中的概率是多少？

A. 9.9%

B. 100%

C. 0.01%

D. 0.99%

正确答案：A

75. 单选题 （ 1.0 分 ）

在下列的交叉验证算法中，哪一个算法，除了测试集以外，平均将数据集划分为 k 份，并每次选择 1 份作为验证集。

A. K 折交叉验证

B. 留一法

C. 留出法

D. 随机采样法

E. 分层采样法

正确答案：A

76. 单选题 （ 1.0 分 ）

以下用于表示 ReLU 激活函数的数学表达式是哪一个？

A. $$f(x)=\frac{1}{1+e^{-x}}$$

B. $$f(x)=\frac{1-e^{-2x}}{1+e^{-2x}}$$

C. $$f(x)=max(0,x)$$

D. $$f(x)=log(e^{x}+1)$$

正确答案：C

77. 单选题 （ 1.0 分 ）

在 scikit-learn 中，K-Means 算法是基于以下哪一种距离计算公式计算数据对象间的距离？

A. 马氏距离

B. 汉明距离

C. 欧氏距离

D. 信息熵

正确答案：C

78. 单选题 （ 1.0 分 ）

下列数据拆分算法，比较适合用于大规模数据集的是（ ）。

A. 留出法

B. K 折交叉法

C. 留一法

D. 分层采样法

正确答案：A

79. 单选题 （ 1.0 分 ）

神经网络进入第二次寒冬，一方面是因为神经网络自身发展遇到了瓶颈，另一方面也是因为俄罗斯学者 Vladmir Vapnik 在 1963 年提出了性能远超当时神经网络的（ ）。

A. 支持向量机模型

B. 朴素贝叶斯模型

C. 随机森林模型

D. K 近邻模型

正确答案：A

80. 单选题 （ 1.0 分 ）

下列可以用来载入文本格式数据集的方法是：（　　）

A. pandas.read_csv

B. pandas.read_txt

C. np.read_csv

D. np.read_txt

正确答案：A

81. 单选题 （ 1.0 分 ）

以下哪一个选项所描述的内容是支持向量。

A. 训练集中所有的样本点

B. 测试集中所有的样本点

C. 距离超平面最近的样本点

D. 超平面

正确答案：C

82. 单选题 （ 1.0 分 ）

以下（　　）语句用于实现从 scipy 库中引入 sparse 稀疏矩阵子库。

A. import scipy as sparse

B. import sparse from scipy

C. import scipy.sparse from *

D. from scipy import sparse

正确答案：D

83. 单选题 （ 1.0 分 ）

高斯朴素贝叶斯在很多时候都具有不错的性能，它是一种基于概率的方法，这种方法使用了（　　）。

A. 先验概率

B. 后验概率

C. AB 都错

D. AB 都对

正确答案：D

84. 单选题 （ 1.0 分 ）

1986 年，Geoffrey Hinton 和 David Rumelhart 联合在 Nature 上发表论文，将(　　)用于神经网络模型，实现了对权重参数的快速计算。

A. 多层感知机模型

B. BP 算法

C. 支持向量机

D. 卷积神经网络

正确答案：B

85. 单选题 （ 1.0 分 ）

```
[11]: # 设定随机森林中树的数量，此处 = 6
      forest = RandomForestClassifier(n_estimators = 6, random_state = 3, n_jobs = -1)
      forest.fit(X_train, y_train)
```

```
[11]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                             max_depth=None, max_features='auto', max_leaf_nodes=None,
                             min_impurity_decrease=0.0, min_impurity_split=None,
                             min_samples_leaf=1, min_samples_split=2,
                             min_weight_fraction_leaf=0.0, n_estimators=6, n_jobs=-1,
                             oob_score=False, random_state=3, verbose=0,
                             warm_start=False)
```

通常，随机森林需要较为复杂的运算过程，因此我们可以通过并行算法来进行提速，以下用于设置并行计算的 CPU 核心个数的参数是（　　）。

A. max_leaf_nodes

B. min_samples_leaf

C. random_state

D. n_jobs

E. verbose

正确答案：D

86. 单选题 （ 1.0 分 ）

“长尾效应”是一种非常严重的数据问题，它主要是指（　　）。

A. 样本的特征数量远大于样本数，形成长矩形矩阵

B. 样本数远大于样本的特征数量，形成长矩形矩阵

C. 样本每个类别的数量严重不平衡，导致部分数据很多，部分数据很少

D. 图像样本的分辨率长度远大于宽度，或宽度远大于长度

正确答案：C

87. 单选题 （ 1.0 分 ）

以下代码可以用来在训练数据集上完成决策树模型训练的是：（ ）。

A. model = tree.DecisionTreeClassifier(X_train,y_train)

B. model.fit(X_train,y_train)

C. model.score(X_train, y_train)

D. model.predict_proba(X_train,y_train)

正确答案：B

88. 单选题 （ 1.0 分 ）

决策树算法也可以被看作是一种基于概率的方法，被认为是定义在特征空间与类别空间上的（ ）分布。

A. 先验概率

B. 后验概率

C. 条件概率

D. 边缘概率

E. 联合概率

正确答案：C

89. 单选题 （ 1.0 分 ）

决策树算法也可以被看作是一种基于概率的方法，被认为是定义在特征空间与类别空间上的（ ）分布。

A. 先验概率

B. 后验概率

C. 条件概率

D. 边缘概率

E. 联合概率

正确答案：C

90. 单选题 （ 1.0 分 ）

给定如下 pandas 数据表，可以正确实现表 1 向表 2 转换的语句是（ ）。

假设，data_frame 是一个基于 pandas 建立的 DataFrame 数据表。表 1

	姓名	归属国	年龄	武力值
0	张飞	蜀国	33	98
1	赵云	蜀国	28	97
2	夏侯惇	魏国	32	94
3	太史慈	吴国	30	92

表 2

	姓名	归属国	年龄	武力值
0	张飞	蜀国	33	98
1	赵云	蜀国	28	97
3	太史慈	吴国	30	92

A. data_frame - (data_frame['归属国'] == '魏国')

B. data_frame[data_frame.归属国 != '魏国']

C. data_frame(['归属国' != '魏国'])

D. data_frame('归属国' - '魏国')

正确答案：B

91. 单选题 （ 1.0 分 ）

下列超参数中，哪一个决定了一个算法中权重更新的快慢?

A. 学习率 learning_rate

B. 最大迭代次数 max_iter

C. 动量 momentum

D. 隐层神经元的数量 hidden_layer_sizes

正确答案：A

92. 单选题 （ 1.0 分 ）

在利用机器学习进行建模并对自然界的各种现象进行预测时，以下哪一个是步骤是必不可少。

A. 数据载入

B. 数据预处理

C. 模型评分

D. 可视化分析

正确答案：A

93. 单选题 （ 1.0 分 ）

KNN 的根本任务是（　　）。

A. 求出距离待测样本最近的 K 个样本的索引

B. 计算待测样本和数据集中每个样本的距离

C. 预测出待测样本所属的类别

D. 利用数据集获得 KNN 模型

正确答案：C

94. 单选题 （ 1.0 分 ）

以下（ ）是 Python 的基础科学计算库，它以 Array 数组为基础，可以用来存储和处理各种大型矩阵，同时提供各种数学运算，包括线性代数、傅里叶变换、产生伪随机序列等。

A. Pandas

B. Matrix

C. Numpy

D. Scipy

E. Array

正确答案：C

95. 单选题 （ 1.0 分 ）

支持向量机的核心任务是：（ ）

A. 最大化类间距

B. 最小化类内距

C. 以上都是

D. 以上都不是

正确答案：C

96. 单选题 （ 1.0 分 ）

下列交叉验证算法中， 哪一个算法可以实现每个类别的样本在划分后都具有相同的分布。

A. 分层采样法

B. K 折交叉验证

C. 留一法

D. 随机采样法

正确答案：A

97. 单选题 （ 1.0 分 ）

```
[11]: # 设定随机森林中树的数量，此处 = 6
      forest = RandomForestClassifier(n_estimators = 6, random_state = 3, n_jobs = -1)
      forest.fit(X_train, y_train)
```

```
[11]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                             max_depth=None, max_features='auto', max_leaf_nodes=None,
                             min_impurity_decrease=0.0, min_impurity_split=None,
                             min_samples_leaf=1, min_samples_split=2,
                             min_weight_fraction_leaf=0.0, n_estimators=6, n_jobs=-1,
                             oob_score=False, random_state=3, verbose=0,
                             warm_start=False)
```

通常，随机森林需要较为复杂的运算过程，因此我们可以通过并行算法来进行提速，以下用于设置并行计算的 CPU 核心个数的参数是（　　）。

A. max_leaf_nodes

B. min_samples_leaf

C. random_state

D. n_jobs

E. verbose

正确答案：D

98. 单选题 （ 1.0 分 ）

随着训练程度的加深，学习器的拟合能力将会逐渐增强，（ ）将会逐渐降低。

A. 泛化误差

B. 偏差

C. 方差

D. 噪声

正确答案：B

99. 单选题 （ 1.0 分 ）

以下标志着“人工智能”这一学科产生的会议是?

A. 国际机器学习大会

B. 博鳌亚洲论坛

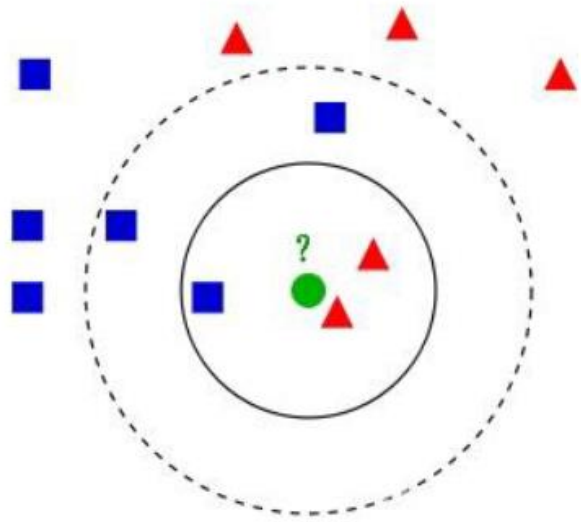
C. 达特茅斯会议

D. 世界人工智能大会（AI World）

正确答案：C

100. 单选题 （ 1.0 分 ）

下图中，当 K=5 时，绿色样本的类别是（ ）。



A. 红色三角形

B. 蓝色正方形

C. 两个类型都有可能

D. 无法判定

正确答案：B

101. 单选题 （ 1.0 分 ）

对于以下代码，我们可以得到以下结论（ ）。

>> print(data.shape)

(200, 500)

A. 数据集 data 包含 500 个样本，每个样本 200 种特征

B. 数据集 data 包含 200 个样本，每个样本 500 种特征

C. 数据集 data 包含 200 个样本，每个样本 200 种特征

D. 数据集 data 包含 500 个样本，每个样本 500 种特征

正确答案：B

102. 单选题 （ 1.0 分 ）

考虑一个雷达预警系统，它时刻在扫描空域，将敌机看作是正样本，将天空的背景看作是负样本，发现敌机（正样本）就会拉响警报。

假设出现情况：一群海鸟在天空飞翔，雷达判断为敌机，立即报警，结果是虚惊一场。这种情况，应该属于以下哪个类型？

A. 真正，True Positive

B. 真负，True Negative

C. 假正，False Positive

D. 假负，False Negative

正确答案：C

103. 单选题 （ 1.0 分 ）

给定样本 X，以及模型 model，下列语句可以用来绘制模型关于样本 X 的预测关系的代码是（ ）。

A. plt.show()

B. plt.plot(X, y)

C. plt.plot(X, model.predict(X))

D. plt.figure()

正确答案：C

104. 单选题 （ 1.0 分 ）

下列哪个符号表示神经元对整个模型贡献的重要程度?

A. y_hat

B. w

C. b

D. y

正确答案：B

105. 多选题 （ 1.0 分 ）

自然界的大多数事务都满足于（　　）分布。

A. 贝努利分布

B. 高斯分布

C. 多项式分布

D. 二项分布

E. 正态分布

正确答案：B，E

106. 多选题 （ 1.0 分 ）

下列数据分割算法中，可以与 GridSearchCV 混合使用的是（　　）。

A. StratifiedShuffleSplit

B. Leave One Out

C. KFold

D. Holdout

正确答案：A，B，C，D

107. 多选题 （ 1.0 分 ）

对于神经网络模型来说，为了获得更好的判别能力，通常需要考虑调节的超参数包括哪些？

A. 神经网络的宽度，即某一层神经网络的神经元数量

B. 神经网络的深度

C. 选择更好的激活函数

D. 正则化参数（alpha）, 批大小（batch_size）, 学习率（learning_rate）, 动量（momentum）等超参数

正确答案：A，B，C，D

108. 多选题 （ 1.0 分 ）

在机器学习的各种算法中，按照数据标注的完整性及数量来进行分类，机器学习主要可以分为以下几种？

A. 监督学习

B. 无监督学习

C. 半监督学习

D. 弱监督学习

正确答案：A，B，C，D

109. 多选题 （ 1.0 分 ）

自然界的大多数事务都满足于（　　）分布。

A. 贝努利分布

B. 高斯分布

C. 多项式分布

D. 二项分布

E. 正态分布

正确答案：B，E

110. 多选题 （ 1.0 分 ）

下列聚类算法中，需要事先指定聚类中心数的是（　　）。

A. K-Means 聚类

B. Mean-Shift 聚类

C. DBSCAN 聚类

D. 凝聚聚类

正确答案：A，D

111. 多选题 （ 1.0 分 ）

请问，被誉为人工智能领域的“三驾马车”的人分别是：

A. Yann LeCun

B. Geoffrey Hinton

C. Joshua Bengio

D. Andrew Ng

正确答案：A，B，C

112. 多选题 （ 1.0 分 ）

以下对于参数和设置方法正确的是：

A. 参数一般是在模型的学习过程中确定

B. 参数一般是由人工设定

C. 超参数一般是在模型的学习过程中确定

D. 超参数一般是由人工设定

正确答案：A，D

113. 多选题 （ 1.0 分 ）

对于结果混淆矩阵来说，以下描述错误的是：

A. TP 是真正，它表示真实情况为真，预测结果也为真

B. FN 是假负，它表示真实情况为假，预测结果也为假

C. FP 是假正，它表示真实情况为假，预测结果为真

D. TN 是真负，它表示真实情况为真，预测结果也为假

正确答案：B，D

114. 多选题 （ 1.0 分 ）

以下在神经网络发展过程中，具有重要地位的模型包括：（ ）。

A. 感知器模型

B. 决策树模型

C. 前馈神经网络

D. BP 神经网络

E. 支持向量机

F. 卷积神经网络

G. 朴素贝叶斯模型

正确答案：A，C，D，F

115. 多选题 （ 1.0 分 ）

Normalizer 预处理需要对每个样本计算其 p-范数，它具有三种最简形，包括：

A. L0-范数

B. L1-范数

C. L2-范数

D. 无穷范数

正确答案：B，C，D

116. 多选题 （ 1.0 分 ）

常用的模型性能评估方法，包括以下哪几种？

A. 留出法

B. 人工验证法

C. 交叉验证法

D. 自助法

正确答案：A，C，D

117. 多选题 （ 1.0 分 ）

下列方法中，可以实现降维的方法包括（　　）。

A. 主成分分析

B. 因子分析

C. 独立成分分析

D. 凝聚聚类

E. K-Means

正确答案：A，B，C

118. 多选题 （ 1.0 分 ）

以下工作，可以利用机器学习技术来提高效率和性能的有哪些?

A. 自动驾驶

B. 古文献修复

C. 画作鉴别

D. 自动翻译

E. 搜索引擎

F. 文献筛选

正确答案：A，B，C，D，E，F

119. 多选题 （ 1.0 分 ）

以下算法属于监督学习算法的是（ ）。

A. 决策树

B. 朴素贝叶斯

C. 广义线性模型

D. 随机森林

正确答案：A，B，C，D

120. 多选题 （ 1.0 分 ）

超参数的选择是一个反复迭代训练的过程，下列数据集在整个模型评估过程中，不只会被用到一次的是（ ）。

A. 训练集

B. 测试集

C. 验证集

D. 以上都是

正确答案：A，B

121. 多选题 （ 1.0 分 ）

支持向量机可以实现以下哪些任务？

A. 二分类

B. 回归分析

C. 聚类

D. 多分类

正确答案：A，B，C，D

122. 多选题 （ 1.0 分 ）

为了使模型训练正常，应该将超参数 random_state 设置为（ ）。

A. 0

B. 1

C. 8

D. 16

正确答案：A，B，C，D

123. 多选题 （ 1.0 分 ）

下列评价指标中，可以用于分类模型的是（ ）。

A. Accuracy 准确度

B. Precision 精确度

C. Recall 召回率

D. PR 曲线

正确答案：A，B，C，D