1. 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

2. 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

3. 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

4．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

5. 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

6. 以下公示，是哪一种模型的数学表达？



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

7. 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

8. 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

9. 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

10. 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

11. 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

12. 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

13. 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

14. 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

15. 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

16. 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

17. 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

18. 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

19. 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

20. 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

21. 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

22. 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

23. 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

24. 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

25. 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

26. 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

27. 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

28. 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

29. 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

30. 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

31. 以下用于表示 ReLU 激活函数的数学表达式是哪一个？

A.

B.

C.

D.

正确答案：C

32. 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

33. 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

34 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

35. 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

36. 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

37. 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

38. 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

39. 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

41. 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

42. 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

43. 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

44. 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

45. 下图中，KNN算法当 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

46. 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

47. 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

48. 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

49. 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

50. 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

正确答案：B

51. 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

52. 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

53. 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

54. 正则化（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

55. 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

56. 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

E. K-Means

正确答案：A，B，C

57. 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

58. 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

59. 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，C

60. 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

61. 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

62. 使用决策树分类时，如果输入的某个特征的值是连续的，通常使用二分法对连续属性离散化，即根据是否大于/小于某个阈值进行划分。如果采用多路划分，每个出现的值都划分为一个分支，这种方式的最大问题是

A. 计算量太大

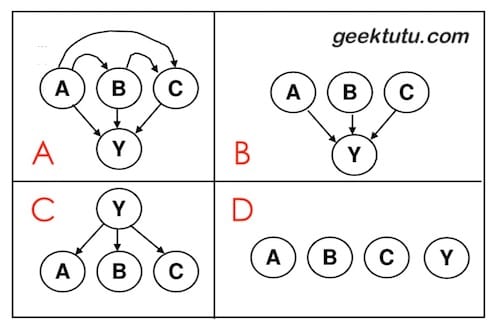
B. 训练集和测试集表现都很差

C. 训练集表现良好，测试集表现差

D. 训练集表现差，测试集表现良好

答案：C

连续值通常采用二分法，离散特征通常采用多路划分的方法，但分支数不宜过多。 连续特征每个值都划分为一个分支，容易过拟合，泛化能力差，导致训练集表现好，测试集表现差。

62. 输入属性是 A, B, C，输出属性是 Y。若使用朴素贝叶斯分类(Naive Bayes Classifier)，下面哪张图能表示朴素贝叶斯分类的假设。

答案：C

朴素贝叶斯分类器是机器学习一个特别质朴而深刻的模型：当要根据多个特征而非一个特征对数据进行分类时，假设这些特征相互独立，然后利用条件概率乘法法则得到每一个分类的概率，选择概率最大的那个作为输出。

回顾一下贝叶斯公式 P(Y|A,B,C) = P(A,B,C|Y) \* P(Y) / P(A,B,C)，如果需要求在A,B,C条件下Y的概率，则需要知道先验概率P(A,B,C)和P(Y)，和在Y条件下，A,B,C的概率。

63. 对神经网络(neural network)而言，哪一项对过拟合(overfitting)和欠拟合(underfitting)影响最大。

A 隐藏层节点(hidden nodes)数量

B 学习速率(learning rate)

C 初始权重

D 每一次训练的输入个数固定

答案：A

过拟合和欠拟合与神经网络的复杂程度有关，模型越大越容易过拟合。隐藏层节点数量直接决定了模型的大小与复杂程度。

64. (超范围问题) 对多项式回归(polynomial regression)而言，哪一项对过拟合(overfitting)和欠拟合(underfitting)影响最大。

A 多项式的度(polynomial degree)

B 是否通过矩阵求逆/梯度下降学习权重

C 高斯噪声方差(variance of the Gaussian noise)

D 每一次训练的输入个数固定

答案A

多拟合/欠拟合与模型复杂度有关，和模型复杂度最相关的是多项式的度。举一个极端的例子，度为1，则是线性回归，y=kx+b，一条直线分类，很容易欠拟合。那度比较大时，则能表示更为复杂的曲线，容易过拟合。

65.（超范围知识点）

问题：特征缩放(feature scaling)的作用？

A. 降低每次迭代计算成本加速梯度下降

B. 防止陷入局部最优

C. 加快了标准方程(Normal Equation)的求解

D. 相同的训练效果迭代次数更少，从而加速了梯度下降。

答案:D

如果某个特征比其他特征值大得多，则需要付出额外的迭代成本。因此训练前，进行特征缩放可以避免额外的迭代，快速达到预期效果。最常用的方法，比如归一化。将所有的特征的值缩放到0-1之间。

66. 在下面的二元标签数据集上训练线性支持向量机(Support Vector Machine, SVM)模型

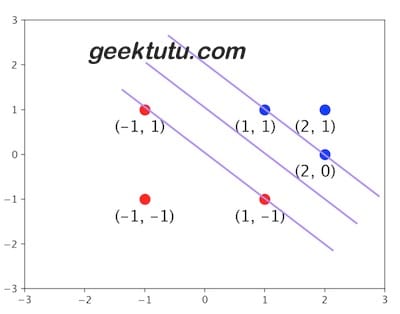
+：(-1, 1), (1, -1), (-1, -1)  
-：(1, 1), (2, 0), (2, 1)

请问，这个模型中的支持向量是哪些？

A (-1, 1), (1, 1), (2, 1)

B (-1, 1), (-1, -1), (2, 1)

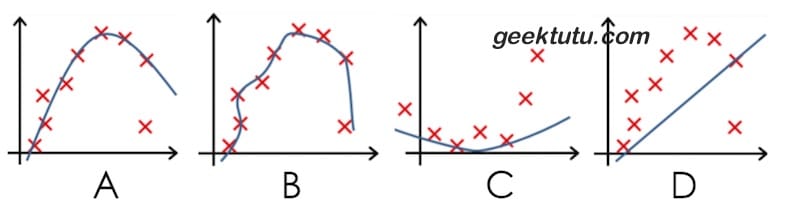
C (-1, 1), (1, -1), (1, 1), (2,0)

答案：

C 先将坐标点画出来，问自己一个问题，是不是所有的点对于分割线的位置起决定性作用？

其实在特别远的区域，无论增加多少个样本点，也不会影响分割线的位置，因为分割线是由几个关键点决定的（图上四个），这几个关键点支撑起了一个分割超平面，这四个关键点，就是支持向量。

67. 下列哪个模型过拟合(Overfit)。

答案：B

拟合曲线紧跟数据集且复杂很高，表明过度拟合了训练集。

68. 下列不属于降维算法的是？

A. 主成分分析(Principal Component Analysis，PCA)

B. 线性判别分析(Linear Discriminant Analysis, LDA)

C. 逻辑回归(Logistic Regression, LR)

答案C

PCA 和 LDA 都是常用的降维算法。LDA作用于带标签数据，PCA作用于无标签数据。

LDA的原理是，将带上标签的数据（点），通过投影的方法，投影到维度更低的空间中，使得投影后的点，会形成按类别区分，一簇一簇的情况，相同类别的点，将会在投影后的空间中更接近。

PCA的原理是，通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。

69. 下列关于极大似然估计（Maximum Likelihood Estimate，MLE），说法正确的是（多选）？

 A. MLE 可能并不存在

B. MLE 总是存在

C. 如果 MLE 存在，那么它的解可能不是唯一的

D. 如果 MLE 存在，那么它的解一定是唯一的

答案：AC

解析：如果极大似然函数 L(θ) 在极大值处不连续，一阶导数不存在，则 MLE 不存在，另一种情况是 MLE 并不唯一，极大值对应两个 θ。

70. 如果我们说“线性回归”模型完美地拟合了训练样本（训练样本误差为零），则下面哪个说法是正确的？

A. 测试样本误差始终为零

B. 测试样本误差不可能为零

C. 以上答案都不对

答案：C

解析：根据训练样本误差为零，无法推断测试样本误差是否为零。值得一提是，如果测试样本样本很大，则很可能发生过拟合，模型不具备很好的泛化能力！

71. 下列关于线性回归分析中的残差（Residuals）说法正确的是？

A. 残差均值总是为零

B. 残差均值总是小于零

C. 残差均值总是大于零

D. 以上说法都不对

答案：A

解析：线性回归分析中，目标是残差最小化。残差平方和是关于参数的函数，为了求残差极小值，令残差关于参数的偏导数为零，会得到残差和为零，即残差均值为零。

72. 下列哪一项能反映出 X 和 Y 之间的强相关性？

A. 相关系数为 0.9

B. 对于无效假设 β=0 的 p 值为 0.0001

C. 对于无效假设 β=0 的 t 值为 30

D. 以上说法都不对

答案：A

解析：相关系数的概念我们很熟悉，它反映了不同变量之间线性相关程度，一般用 r 表示。

73. 下列哪些假设是我们推导线性回归参数时遵循的（多选）？

A. X 与 Y 有线性关系（多项式关系）

B. 模型误差在统计学上是独立的

C. 误差一般服从 0 均值和固定标准差的正态分布

D. X 是非随机且测量没有误差的

答案：ABCD

解析：在进行线性回归推导和分析时，我们已经默认上述四个条件是成立的。

74. 为了观察测试 Y 与 X 之间的线性关系，X 是连续变量，使用下列哪种图形比较适合？

A. 散点图

B. 柱形图

C. 直方图

D. 以上都不对

 答案：A

解析：散点图反映了两个变量之间的相互关系，在测试 Y 与 X 之间的线性关系时，使用散点图最为直观。

75. 一般来说，下列哪种方法常用来预测连续独立变量？

A. 线性回归

B. 逻辑回顾

C. 线性回归和逻辑回归都行

D. 以上说法都不对

答案：A

解析：线性回归一般用于实数预测，逻辑回归一般用于分类问题。

76. 个人健康和年龄的相关系数是 -1.09。根据这个你可以告诉医生哪个结论？

A. 年龄是健康程度很好的预测器

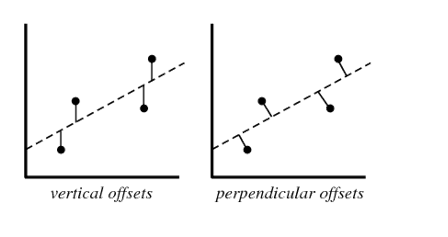
B. 年龄是健康程度很糟的预测器

C. 以上说法都不对

答案：C

解析：因为相关系数的范围是 [-1,1] 之间，所以，-1.09 不可能存在。

77. 下列哪一种偏移，是我们在最小二乘直线拟合的情况下使用的？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

答案：A

解析：线性回归模型计算损失函数，例如均方差损失函数时，使用的都是 vertical offsets。perpendicular offsets 一般用于主成分分析（PCA）中。

78. 假如我们利用 Y 是 X 的 3 阶多项式产生一些数据（3 阶多项式能很好地拟合数据）。那么下列说法正确的是（多选）?

A. 简单的线性回归容易造成高偏差（bias）、低方差（variance）

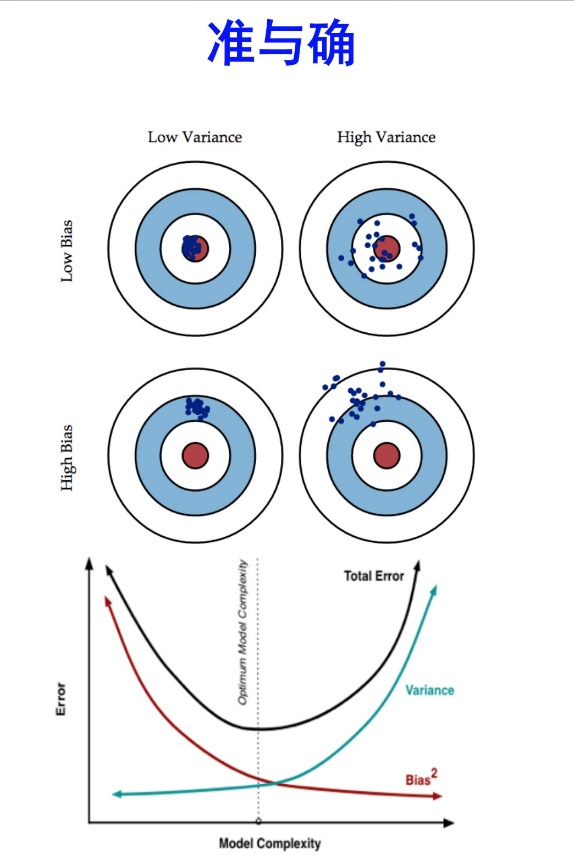
B. 简单的线性回归容易造成低偏差（bias）、高方差（variance）

C. 3 阶多项式拟合会造成低偏差（bias）、高方差（variance）

D. 3 阶多项式拟合具备低偏差（bias）、低方差（variance）

答案：AD

解析：偏差和方差是两个相对的概念，就像欠拟合和过拟合一样。如果模型过于简单，通常会造成欠拟合，伴随着高偏差、低方差；如果模型过于复杂，通常会造成过拟合，伴随着低偏差、高方差。用一张图来形象地表示偏差与方差的关系：



偏差（bias）可以看成模型预测与真实样本的差距，想要得到 low bias，就得复杂化模型，但是容易造成过拟合。方差（variance）可以看成模型在测试集上的表现，想要得到 low variance，就得简化模型，但是容易造成欠拟合。实际应用中，偏差和方差是需要权衡的。若模型在训练样本和测试集上都表现的不错，偏差和方差都会比较小，这也是模型比较理想的情况。

79. 假如你在训练一个线性回归模型，有下面两句话：

     1. 如果数据量较少，容易发生过拟合。

    2. 如果假设空间较小，容易发生过拟合。

关于这两句话，下列说法正确的是？

A. 1 和 2 都错误

B. 1 正确，2 错误

C. 1 错误，2 正确

D. 1 和 2 都正确

答案：B

解析：先来看第 1 句话，如果数据量较少，容易在假设空间找到一个模型对训练样本的拟合度很好，容易造成过拟合，该模型不具备良好的泛化能力。

再来看第 2 句话，如果假设空间较小，包含的可能的模型就比较少，也就不太可能找到一个模型能够对样本拟合得很好，容易造成高偏差、低方差，即欠拟合。

80. 关于特征选择，下列对 Ridge 回归和 Lasso 回归说法正确的是？

A. Ridge 回归适用于特征选择

B. Lasso 回归适用于特征选择

C. 两个都适用于特征选择

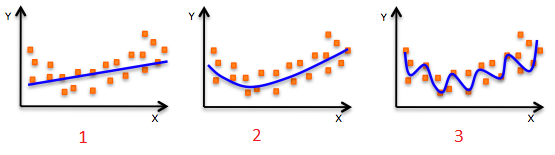
D. 以上说法都不对

答案：B

解析：上一题我们已经介绍过，Lasso 回归会让一部分回归系数刚好可以被约束为 0，起到特征选择的效果。

Ridge 回归又称岭回归，它是普通线性回归加上 L2 正则项，用来防止训练过程中出现的过拟合。L2 正则化效果类似上一题左图，限定区域是圆，这样，得到的回归系数为 0 的概率很小，很大概率是非零的。因此，比较来说，Lasso 回归更容易得到稀疏的回归系数，有利于舍弃冗余或无用特征，适用于特征选择。

89. 下面三张图展示了对同一训练样本，使用不同的模型拟合的效果（蓝色曲线）。那么，我们可以得出哪些结论（多选）？



A. 第 1 个模型的训练误差大于第 2 个、第 3 个模型

B. 最好的模型是第 3 个，因为它的训练误差最小

C. 第 2 个模型最为“健壮”，因为它对未知样本的拟合效果最好

D. 第 3 个模型发生了过拟合

E. 所有模型的表现都一样，因为我们并没有看到测试数据

答案：ACD

解析：1、2、3 模型分别对应的多项式阶数由小到大，即模型由简单到复杂。模型越简单，容易发生欠拟合；模型越复杂，容易发生过拟合。第 1 个模型过于简单，出现欠拟合；第 3 个模型过于复杂，对训练样本拟合得很好，但在测试样本上效果会很差，即过拟合；第 2 个模型最为“健壮”，在训练样本和测试样本上拟合效果都不错！

90. 如果 Y 是 X（X1，X2，...，Xn）的线性函数：Y = β0 + β1X1 + β2X2 + ··· + βnXn，则下列说法正确的是（多选）？

A. 如果变量 Xi 改变一个微小变量 ΔXi，其它变量不变。那么 Y 会相应改变 βiΔXi。

B. βi 是固定的，不管 Xi 如何变化

C. Xi 对 Y 的影响是相互独立的，且 X 对 Y 的总的影响为各自分量 Xi 之和

答案：ABC

解析：这题非常简单，Y 与 X（X1，X2，...，Xn）是线性关系，故能得出 ABC 结论。

91. 构建一个最简单的线性回归模型需要几个系数（只有一个特征）？

A. 1 个

B. 2 个

C. 3 个

D. 4 个

答案：B

解析：最简单的线性回归模型，只有一个特征，即 Y = aX + b，包含 a 和 b 两个系数。

92. 加入使用逻辑回归对样本进行分类，得到训练样本的准确率和测试样本的准确率。现在，在数据中增加一个新的特征，其它特征保持不变。然后重新训练测试。则下列说法正确的是？

A. 训练样本准确率一定会降低

B. 训练样本准确率一定增加或保持不变

C. 测试样本准确率一定会降低

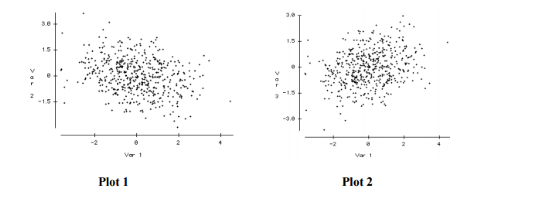
D. 测试样本准确率一定增加或保持不变

答案：B

解析：在模型中增加更多特征一般会增加训练样本的准确率，减小 bias。但是测试样本准确率不一定增加，除非增加的特征是有效特征。

这题对应的知识点也包括了增加模型复杂度，虽然会减小训练样本误差，但是容易发生过拟合。

93. 假设变量 Var1 和 Var2 是正相关的，那么下面那张图是正确的？图中，横坐标是 Var1，纵坐标是 Var2，且对 Var1 和 Var2 都做了标准化处理。



A. Plot 1

B. Plot 2

答案：B

解析：显然，Plot 2 显示出 Var2 与 Var1 是正相关的，例如 Var2 = Var1。Plot 1 显示出 Var2 与 Var1 是负相关的，例如 Var2 = -Var1。

94. 假设一个公司的薪资水平中位数是 $35,000，排名第 25% 和 75% 的薪资分别是 $21,000 和 $ 53,000。如果某人的薪水是 $1，那么它可以被看成是异常值（Outlier）吗？

A. 可以  
B. 不可以  
C. 需要更多的信息才能判断  
D. 以上说法都不对

答案：C  
解析：异常值（Outlier）指样本中的个别值，其数值明显偏离它（或他们）所属样本的其余观测值，也称异常数据，离群值。目前人们对异常值的判别与剔除主要采用物理判别法和统计判别法两种方法。  
所谓物理判别法就是根据人们对客观事物已有的认识，判别由于外界干扰、人为误差等原因造成实测数据值偏离正常结果，在实验过程中随时判断，随时剔除。  
统计判别法是给定一个置信概率，并确定一个置信限，凡超过此限的误差，就认为它不属于随机误差范围，将其视为异常值剔除。当物理识别不易判断时，一般采用统计识别法。  
该题中，所给的信息量过少，无法肯定一定是异常值。

95. 自变量x是独立变量，因变量y是非独立的，以下关于两个变量的回归和相关的说法，哪个是正确的？

A. 变量x和y之间的关系是对称的

B. 变量x和y之间的关系是不对称的

C. 两变量的相关性不是对称的，但回归是对称的

D. 两变量的相关性是对称的，但回归不是对称的.

答案: D

Correlation is a statistic metric that measures the linear association between two variables. It treats y and x symmetrically.

Regression is setup to predict y from x. The relationship is not symmetric.

96. 在其他条件都一样的情况下，样本的数量是如何影响过拟合的？

1. 样本数量少，容易出现过拟合

2. 样本数量少，不容易出现过拟合

3. 样本数量多，容易出现过拟合

4. 样本数量多，不容易出现过拟合

A. 1 and 4

B. 2 and 3

C. 1 and 3

D. None of theses

答案: A

In particular, if we have very few observations and it’s small, then our models can rapidly overfits data. Because we have only a few points and as we’re increasing in our model complexity like the order of the polynomial, it becomes very easy to hit all of our observations.

On the other hand, if we have lots and lots of observations, even with really, really complex models, it is difficult to overfit because we have dense observations across our input.

97. 假设有一个**Ridge**模型，调大系数以降低模型复杂性时，以下有关偏差和方差的表述哪些是正确的？

A. 当系数很小时，偏差降低，方差降低

B. 当系数很小时，偏差降低，方差提高

C. 当系数很小时，偏差提高，方差降低

D. 当系数很小时，偏差提高，方差提高

答案: C

If lambda is very large it means model is less complex. So in this case bias is high and variance in low.

98. 假设有一个使用**Ridge正则项的**偏复杂的回归模型拟合数据集，调小系数以降低模型复杂性时，以下有关偏差和方差的表述哪些是正确的？

A. 当系数很小时，偏差降低，方差降低

B. 当系数很小时，偏差降低，方差提高

C. 当系数很小时，偏差提高，方差降低

D. 当系数很小时，偏差提高，方差提高

答案: B

If  lambda is very small it means model is complex. So in this case bias is low and variance is high because model will overfit the data.

99. 以下关于ridge回归的表述，哪些正确?

如果则模型就是一般的线性回归模型

如果则模型不同于一般的线性回归模型

如果则模型的参数趋向于零

如果则模型的参数趋向于无穷大

A. 1 and 3

B. 1 and 4

C. 2 and 3

D. 2 and 4

答案: A

Specifically, we can see that when lambda is 0, we get our least square solution. When lambda goes to infinity, we get very, very small coefficients approaching 0.

100. 以下算法中哪些算法是没有闭解的?

A. Ridge 回归

B. Lasso 回归

C. Ridge 和 Lasso 回归

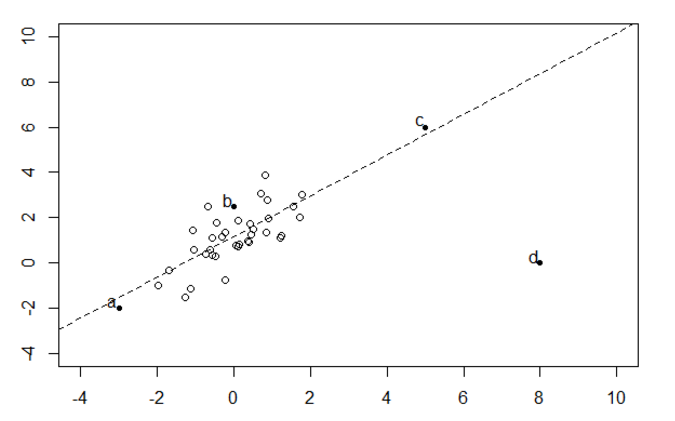
D. 以上两者

答案: B

The Lasso does not admit a closed-form solution. The L1-penalty makes the solution non-linear. So we need to approximate the solution.

If you want to read more about closed form solutions, refer this [link](http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf).

101.  考虑下图数据集：



在进行线性拟合时，哪一个黑点对拟合结果的影响最大？

A) a

B) b

C) c

D) d

答案:D

Linear regression is sensitive to outliers in the data. Although c is also an outlier in given data space but it is closed to the regression line(residual is less) so it will not affect much.

102. 在一个只有一个变量的简单的线性逻辑回归模型中，如果输入变量有一个单位的变化，输出变量的将如何变化？

A: 变化一个单位

B. 无变化

C. 发生变化的大小等于模型的截距值

D. 发生变化的大小等于模型的斜率值

答案: D

Equation for simple linear regression: Y=a+bx. Now if we increase the value of x by 1 then the value of y would be a+b(x+1) i.e. value of y will get incremented by b.

103. 逻辑回归的输出值是一个概率值，取值返回是[0,1]. 逻辑回归使用了一下的哪个函数转换为概率值的？

A. Sigmoid

B. Mode

C. Square

D. Probit

答案: A

Sigmoid function is used to convert output probability between [0,1] in logistic regression.

104. 假设我们用逻辑回归解决n-分类问题，我们用一对其余方法（OVR），以下说法哪个是正确的？

A. 需要n个二分类器

B. 需要n-1个二分类器

C. 需要1个二分类器

D. 以上都不对.

答案: A

If there are n classes, then n separate logistic regression has to fit, where the probability of each category is predicted over the rest of the categories combined.

Take a example of 3-class(-1,0,1) classification. Then need to train 3 logistic regression classifiers.

-1 vs 0 and 1

0 vs -1 and 1

1 vs 0 and -1

105. 数据预处理阶段，对数值特征归一化或标准化，理论上不会对哪种模型产生很大影响。

A. 决策树

B. k-means

C. kNN

​答案

A k-means和kNN(k-NearestNeighbor)都需要使用距离。而决策树对于数值特征，只在乎其大小排序，而非绝对大小。不管是标准化或者归一化，都不会影响数值之间的相对大小。