《机器学习》期末考试复习题库（含答案）

一、单选题

1.混淆矩阵的真负率公式是为

A、TP/(TP+FN)

B、FP/(FP+TN)

C、FN/(TP+FN)

D、TN/(TN+FP)

答案：D

2.一个包含 n 类的多分类问题，若采用一对剩余的方法，需要拆分成多少次？

A、n

B、1

C、n-1

D、n+1

答案：C

解析：答案解析：在一对剩余（One-versus-Rest，简称 OVR）的方法中，对于一个包含 n 类的多分类问题，会将其中一类作为正例，其余 n-1 类作为反例，依次构建 n 个二分类模型。所以需要拆分成 n-1 次。因此，选项 C 是正确的答案。

3.哪一个是机器学习的合理定义？

A、机器学习是计算机编程的科学

B、机器学习从标记的数据中学习

C、机器学习是允许机器人智能行动的领域

D、机器学习能使计算机能够在没有明确编程的情况下学习

答案：D

4.对 Boosting 模型的描述错误的是

A、采用串行训练模式

B、增加被错误分类样本的权值

C、通过改变训练集进行有针对性的学习

D、基础分类器采用少数服从多数原则进行集成

答案：D

解析：Boosting 是一种集成学习方法，它通过串行训练多个基础分类器来提高模型的性能。在每次迭代中，Boosting 算法会根据上一次迭代的结果调整训练集的权重，使得被错误分类的样本在后续迭代中得到更多的关注。基础分类器的集成方式通常是通过加权投票或其他方式来综合多个分类器的预测结果，而不是少数服从多数原则。因此，选项 D 是错误的。

5.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

6.下列哪种方法可以用来缓解过拟合的产生：()。

A、正则化

B、增加更多的特征

C、以上都是

D、增加模型的复杂度

答案：A

7.7.以下哪个不是原型聚类算法（）

A、K 均值算法

B、学习向量量化 LVQ

C、高斯混合聚类

D、PCA 算法

答案：D

解析：答案解析：K 均值算法、学习向量量化 LVQ、高斯混合聚类都属于原型聚类算法。而 PCA 算法（主成分分析）主要用于数据降维，通过线性变换将原始数据变换到一组各维度线性无关的表示上，从而提取数据的主要特征，并非原型聚类算法。因此，选项 D 是正确答案。

8.在回归模型中，下列哪一项在权衡欠拟合(under-fitting)和过拟合(over-fitting）中影响最大？

A、多项式阶数

B、更新权重 w 时，使用的是矩阵求逆还是梯度下降

C、使用常数项

答案：A

9.一对一法分类器，k 个类别需要多少个 SVM：

A、k(k-1)/2

B、k(k-1)

C、k

D、k!

答案：A

解析：一对一法是一种解决多类别分类问题的方法。在一对一法中，对于k个类别，需要构建$k(k-1)/2$个二分类器，每个二分类器用于区分两个类别。具体来说，对于每个类别i，需要构建$k-1$个二分类器，其中第j个二分类器用于区分类别i和类别j。这样,总共需要构建的二分类器数量为: $$\begin{align*}&\frac{k(k-1)}{2}\\=&\frac{k\times(k-1)}{2}\\=&\frac{k^2-k}{2}\\=&\frac{k(k-1)}{2}\end{align*}$$因此，选项 A 是正确答案。

10. 下列贝叶斯网结构中不属于三种典型的依赖关系

A、同父结构

B、选择结构

C、顺序结构

D、V 型结构

答案：B

解析：答案解析：在贝叶斯网中，存在三种典型的依赖关系，分别是同父结构、顺序结构和 V 型结构。同父结构中，多个子节点共享同一个父节点；顺序结构体现节点之间的先后顺序依赖；V 型结构反映了两个子节点通过共同父节点产生的依赖。而选择结构并非贝叶斯网中的典型依赖关系。所以，正确答案是选项 B。

11. 下列两个变量之间的关系中，那一个是线性关系

A、学生的性别与他（她）的数学成绩

B、人的工作环境与他的身体健康状况

C、儿子的身高与父亲的身高

D、正方形的边长与周长

答案：D

12.下面符合特征选择标准的是（）

A、越少越好

B、越多越好

C、选择能够反映不同事物差异的特征

D、以上均不对

答案：C

解析: 特征选择的目的是选取最能有效区分不同类别或事物的特征。如果特征过少，可能无法充分体现事物的特点和差异; 特征过多可能会引入噪声和冗余信息，增加计算负担且不一定能提高准确性。而选择能够反映不同事物差异的特征，才是最关键和有意义的，这样能更好地进行分析和判断。所以选项 C 正确，选项 A、B 过于绝对和片面，选项 D 错误。因此答案是 C。

13.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

B、垂向偏移

C、两种偏移都可以

D、以上说法都不对

答案：A

14.以下哪个是 PCA 算法的主要应用？

A、聚类

B、分类

C、距离度量

D、数据压缩

答案：D

解析：PCA（PrincipalComponentAnalysis）算法即主成分分析算法，是一种常用的数据分析方法。PCA 算法的主要目的是将高维数据投影到低维空间中，同时尽可能保留数据的方差信息。通过这种方式，可以实现数据的降维，减少数据的维度，同时保留数据的主要特征。在数据压缩方面，PCA 算法可以将高维数据投影到低维空间中，从而实现数据的压缩。通过这种方式，可以减少数据的存储空间，同时提高数据的传输和处理效率。在聚类和分类方面，虽然 PCA 算法可以用于数据的预处理和特征提取，但它并不是一种直接的聚类或分类算法。在聚类和分类中，通常需要使用其他算法，如 K-Means 算法、SVM 算法等。在距离度量方面，PCA 算法可以用于计算数据之间的距离，但它并不是一种专门的距离度量算法。在距离度量中，通常需要使用其他算法，如欧几里得距离、余弦相似度等。因此，选项 D 是正确的答案。

15.1.将数据集 D 进行适当处理，产生出训练集 S 和测试集 T，有哪些常见的做法：

A、留出法

B、交叉验证法

C、自助法

D、以上都是

答案：D

16.在机器学习中，学得的模型适用于新样本的能力称为（）

A、分析能力

B、泛化能力

C、训练能力

D、验证能力

答案：B

解析: 在机器学习中，模型的重要作用是对未曾见过的新样本进行准确预测或分类。泛化能力就是指模型从已有的训练数据中学习到的知识和规律，应用到新的、未见过的数据上并取得良好效果的能力。一个具有良好泛化能力的模型，能够有效地处理实际场景中的各种新情况。而分析能力、训练能力、验证能力都不能准确描述学得的模型适用于新样本的能力。所以，答案是 B 选项。

17.线性回归能完成的任务是

A、预测离散值

B、预测连续值

C、分类

D、聚类

答案：B

18.假设现在只有两个类，这种情况下 SVM 需要训练几次？

A、1

B、2

C、3

D、4

答案：A

19.对决策树进行剪枝处理的主要目的是什么

A、避免欠拟合

B、提高对训练集的学习能力

C、避免过拟合，降低泛化能力

D、避免过拟合，提升泛化能力

答案：D

20.若某学习器预测的是离散值，则此类学习任务称为（）

A、分类

B、聚类

C、回归

D、强化学习

答案：A

解析:在机器学习中，根据预测值的类型，可以将学习任务分为分类和回归两类。分类任务的目标是预测离散的类别标签，例如将邮件分为垃圾邮件和正常邮件，将图像分为猫、狗、汽车等类别。而回归任务的目标是预测连续的数值，例如预测房价、股票价格等。聚类任务则是将数据集中的样本分成若干个簇，使得同一个簇内的样本相似度较高，而不同簇之间的样本相似度较低。强化学习则是通过与环境的交互来学习最优的决策策略，以获得最大的累积奖励。因此，若某学习器预测的是离散值，则此类学习任务称为分类，选项 A 正确。

21.在 SVM 中,margin 的含义是()

A、差额

B、损失误差

C、幅度

D、间隔

答案：D

22.KNN 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、弱监督学习

答案：A

解析：KNN 算法是一种基于实例的学习算法，它通过计算新数据与训练数据之间的距离，来确定新数据的类别。在 KNN 算法中，每个训练数据都被标记了一个类别，因此 KNN 算法属于一种有监督学习算法。有监督学习是指从有标记的训练数据中学习模型，以便对新的数据进行预测或分类。在 KNN 算法中，训练数据的标记信息被用于确定新数据的类别，因此 KNN 算法是一种有监督学习算法。因此，正确答案是选项 A。

23.在构造决策树时，以下哪种不是选择属性的度量的方法

A、信息值

B、信息增益

C、信息增益率

D、基尼指数

答案：A

解析: 在决策树算法中，选择属性的度量方法主要有信息增益、信息增益率和基尼指数。这些方法的目的是评估每个属性对于分类的贡献程度，以便选择最优的属性作为决策节点。信息增益衡量了一个属性在划分数据集时能够减少的不确定性程度。信息增益率则对信息增益进行了归一化处理，以避免偏向于具有较多取值的属性。基尼指数则是一种衡量数据集不纯度的指标，通过选择使基尼指数最小的属性来进行划分。而信息值并不是一种常见的选择属性的度量方法。因此，正确答案是 A。

24.以下关于 Sigmoid 的特点说法错误的是()。

A、Sigmoid 函数计算量小

B、趋向无穷的地方，函数变化很小，容易出现梯度消失的现象

C、可以将函数值的范围压缩到[0,1]

D、函数处处连续

答案：A

25.BP 算法总结错误的是（）。

A、当前层的连接权值梯度，取决于当前层神经元阈值梯度和上一层神经元输出

B、算法只要知道上一层神经元的阈值梯度，就能计算当前层神经元的阈值梯度和连接权值梯度

C、隐层的阈值梯度只跟本层的神经元输出值有关

D、隐层阈值梯度取决于隐层神经元输出、输出层阈值梯度和隐层与输出层的连接权值

答案：C

26.以下哪个不是常见的决策树算法

A、ID3

B、C4.5

C、ART

D、BSCAN

答案：D

27.假设我们使用原始的非线性可分版本的 Soft-SVM 优化目标函数。我们需要做什么来保证得到的模型是线性可分离的？

A、C=0

B、C=1

C、正无穷大

D、C 负无穷大

答案：C

28.不属于 KNN 算法要素的是:

A、k 值的选择

B、距离度量

C、分类决策的规则

D、训练样本的个数

答案：D

解析：KNN 算法是一种基本的分类与回归方法，其主要要素包括 k 值的选择、距离度量和分类决策的规则。k 值的选择会影响算法的性能和结果，不同的距离度量方式会影响样本之间的相似度计算，而分类决策的规则则决定了如何根据邻居

的类别来确定待分类样本的类别。训练样本的个数并不是 KNN 算法的要素之一，而是影响算法性能的一个因素。因此，正确答案是 D。

29. 下列中为判别模型的是（）

A、高斯混合模型

B、隐含马尔科夫模型

C、GAN 模型

D、逻辑回归模型

答案：D

解析：答案解析：判别模型是直接对条件概率 P(y|x)进行建模，旨在寻找不同类别之间的决策边界。逻辑回归模型就是通过输入特征 x 来预测输出类别 y 的概率，直接对 P(y|x)进行建模。而高斯混合模型、隐含马尔科夫模型和 GAN 模型更多地是对数据的分布或生成过程进行建模，属于生成模型。所以，选项 D 是判别模型，是正确答案。

30. 关于 logistic 回归和 SVM 不正确的是（）

A、Logistic 回归目标函数是最小化后验概率

B、Logistic 回归可以用于预测事件发生概率的大小

C、SVM 目标是结构风险最小化

D、SVM 可以有效避免模型过拟合

答案：A

31. 机器学习这个术语是由()定义的?

A、rthurSamuel

B、GuidovanRossum

C、JamesGosling

D、以上都不是

答案：A

32.下列方法中，属于无监督学习的为（）

A、线性回归

B、K 均值

C、神经网络

D、决策树

答案：B

监督学习是指在没有标记的数据集上进行学习的方法，目的是发现数据中的潜在模式或结构。在选项中，K 均值算法是一种典型的无监督学习方法，它通过将数据分组为不同的簇，自动发现数据中的内在分组模式，而不需要事先给定数据的类别标签。线性回归、神经网络和决策树通常在有监督学习中应用，需要有已知的输出标签来进行模型的训练和预测。所以，答案选 B。

33.以下关于机器学习描述错误的是?

A、是一门涉及统计学、系统辨识、逼近理论、神经网络、优化理论、计算机科学、脑科学等诸多领域的交叉学科

B、研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能

C、器学习强调三个关键词:算法、模型、训练

D、基于数据的机器学习是现代智能技术中的重要方法之一

答案：C

解析: 机器学习是一门涉及多领域的交叉学科，它研究计算机模拟或实现人类学习行为以获取新知识或技能，A、B 选项正确。基于数据的机器学习也是现代智能技术中的重要方法之一，D 选项正确。而 C 选项中“器学习”表述错误，应为“机器学习”。因此，正确答案是 C。

34. 下面关于 SVM 算法叙述不正确的是（）

A、SVM 在解决小样本、非线性及高维模式识别问题中具有优势

B、SVM 是一种基于经验风险最小化准则的算法

C、SVM 求得的解为全局唯一最优解

D、SVM 最终分类结果只与少数支持向量有关

答案：B

解析：SVM 是一种有监督的学习模型，它的基本思想是找到一个最优的超平面，将不同类别的样本分开。以下是对每个选项的分析: -**A 选项**：SVM 在解决小样本、非线性及高维模式识别问题中具有优势，这是因为它可以通过核函数将数据映射到高维空间，从而更好地处理非线性问题。-**B 选项**：SVM 是一种基于结构风险最小化准则的算法，而不是经验风险最小化准则。结构风险最小化准则考虑了模型的复杂度和泛化能力，而经验风险最小化准则只考虑了模型在训练集上的表现。-**C 选项**：SVM 求得的解为全局唯一最优解，这是因为它通过求解一个凸二次规划问题来找到最优的超平面。-**D 选项**：SVM 最终分类结果只与少数支持向量有关，这是因为支持向量是离超平面最近的样本点，它们对分类结果的影响最大。因此，选项 B 是不正确的。

35. 假如我们使用 Lasso 回归来拟合数据集，该数据集输入特征有 100 个（X1，X2，…，X100）。现在，我们把其中一个特征值扩大 10 倍（例如是特征 X1），然后用相同的正则化参数对 Lasso 回归进行修正。那么，下列说法正确的是？

A、特征 X1 很可能被排除在模型之外

B、特征 X1 很可能还包含在模型之中

C、无法确定特征 X1 是否被舍弃

D、以上说法都不对

答案：B

36. 混淆矩阵中的 TP=16，FP=12，FN=8，TN=4，准确率是

A、四分之一

B、二分之一

C、七分之四

D、三分之二

答案：B

解析：准确率的计算公式为：准确率=（TP+TN）/（TP+TN+FP+FN），其中 TP 表示真正例，TN 表示真负例，FP 表示假正例，FN 表示假负例。将题目中给出的值代入公式，得到准确率=（16+4）/（16+4+12+8）=20/40=1/2。因此，选项 B 是正确答案。

37. 神经网络算法有时会出现过拟合的情况，那么采取以下哪些方法解决过拟合更为可行（）。

A、为参数选取多组初始值，分别训练，再选取一组作为最优值

B、增大学习的步长

C、减少训练数据集中数据的数量

D、设置一个正则项减小模型的复杂度

答案：D

38.1.下面不属于过拟合原因的是

A、特征维度过多

B、模型假设过于复杂

C、训练数据过多

D、噪声过多

答案：C

解析：过拟合是指模型在训练数据上表现良好，但在新数据上表现不佳的现象。造成过拟合的原因主要有以下几点：-特征维度过多：过多的特征可能导致模型过于复杂，从而容易过拟合。-模型假设过于复杂：过于复杂的模型可能会拟合训练数据中的噪声，而不是真正的模式。-噪声过多：训练数据中存在过多的噪声，模型可能会过度拟合这些噪声，而忽略了真正的模式。而训练数据过多通常不会导致过拟合，相反，更多的训练数据可以帮助模型更好地学习数据中的模式，从而提高模型的泛化能力。因此，选项 C 不属于过拟合的原因。

39.SVM 算法的性能取决于：

A、以上所有

B、软间隔参数

C、核函数的参数

D、核函数的选择

答案：A

解析：答案解析：SVM（支持向量机）算法的性能受到多个因素的综合影响。软间隔参数决定了对异常点和噪声的容忍程度。核函数的选择决定了数据在高维空间的映射方式，不同的核函数适用于不同特征的数据集。核函数的参数则会进一步影响核函数的效果和模型的复杂度。综上所述，SVM 算法的性能取决于以上所有选项，即选项 A 是正确的。

40.假设你有以下数据：（0,2）（2,2）（3,1）输入和输出都只有一个变量。使用线性回归模型（y=wx+b）来拟合数据。那么使用留一法（Leave-OneOut）交叉验证得到的均方误差是多少？

A、10/32

B、39/27

C、49/27

D、55/27

答案：C

解析：首先，通过给定的数据进行线性回归拟合得到模型。然后，使用留一法交叉验证，依次将每个数据点作为测试集，其余数据点用于训练模型，并计算测试集的均方误差。经过详细的计算和分析，最终得出的结果是 49/27。因此，选项 C 是正确答案。

41.对于在原空间中线性不可分问题，支持向量机（）。

A、无法处理

B、将数据映射到核空间中

C、在原空间中寻找非线性函数的划分数据

D、在原空间中寻找线性函数划分数据

答案：B

42.下列关于过拟合的说法错误的是

A、过拟合是指模型在训练集上表现很好，但是在交叉验证集和测试集上表现一般

B、解决过拟合可以采用 Dropout 方法

C、解决过拟合可以采用参数正则化方法

D、数据集扩增不能用来解决过拟合问题

答案：D

解析: 过拟合是指模型在训练集上表现很好，但在交叉验证集和测试集上表现一般或较差的现象，A 选项正确。Dropout 方法通过在训练过程中随机忽略一些神经元，减少神经元之间的共适应性，从而缓解过拟合，B 选项正确。参数正则化方法通过对模型的参数进行约束或惩罚，减少模型的复杂度，从而避免过拟合，C 选项正确。数据集扩增可以通过增加训练数据的数量和多样性，来减少模型对训练数据的过度拟合，D 选项错误。因此，答案选 D。

43.关于维数灾难说法错误的是？

A、高维度数据可使得算法泛华能力变得越来越弱

B、降低高维度数据会对数据有所损伤

C、高维度数据增加了运算难度

D、高维度数据难以可视化

答案：A

解析: A 选项错误，高维度数据会导致模型复杂度增加、过拟合风险增大等问题，从而使得算法泛化能力变弱，而不是越来越强。B 选项正确，降低维度可能会损

失一些信息。C 选项正确，高维数据运算量会大幅增加，带来运算难度。D 选项正确，高维度数据很难直观地进行可视化展示。所以说法错误的是 A。

44. 下列有关 SVM 和 LR 说法不正确的是（）

A、SVM 是分类模型，LR 是回归模型

B、SVM 和 LR 都是分类模型

C、SVM 是判别式模型

D、LR 判别式模型

答案：A

解析：SVM（SupportVectorMachine）和 LR（LogisticRegression）都可以用于分类问题，因此选项 A 不正确，选项 B 正确。SVM 是判别式模型，它直接学习决策边界，而不考虑数据的生成过程，因此选项 C 正确。LR 也是判别式模型，它通过学习特征与类别之间的线性关系来进行分类，因此选项 D 正确。综上所述，不正确的说法是选项 A。

45. 若 svm 出现欠拟合，以下合适的做法是

A、使用更 powful 的 kernel

B、增加训练样本

C、使用 L2 正规化

D、做数据增强

答案：A

解析：当 SVM 出现欠拟合时，使用更强大（powerful）的核函数（kernel）可以增加模型的复杂度和表达能力，有助于改善欠拟合情况。增加训练样本不一定能解决欠拟合问题，有可能仍然无法很好地拟合。使用 L2 正规化通常是防止过拟

合的手段。数据增强主要用于增加数据的多样性，对欠拟合的改善作用不直接。所以 A 选项正确。

46. 谷歌新闻每天收集非常多的新闻，并运用()方法再将这些新闻分组，组成若干类有关联的新闻。于是，搜索时同一组新闻事件往往隶属同一主题的，所以显示到一起。

A、关联规则

B、聚类

C、回归

D、分类

答案：B

47. 关于 BP 算法信号前向传播的说法正确的是（）。

A、BP 算法在计算正向传播输出值时需要考虑激活函数

B、P 算法信号前向传播的计算量跟输入层神经元数目无关

C、BP 算法只有在隐层才有激活函数

D、BP 算法信号传播的顺序是输出层、隐层、输入层。

答案：A

48. 以下有关随机森林算法的说法错误的是：

A、随机森林算法的分类精度不会随着决策树数量的增加而提高

B、随机森林算法对异常值和缺失值不敏感

C、随机森林算法不需要考虑过拟合问题

D、决策树之间相关系数越低、每棵决策树分类精度越高的随机森林模型分类效果越好

答案：C

解析: 随机森林是一种常用的机器学习算法，它由多个决策树组成。以下是对每个选项的分析: A.通常情况下，随着决策树数量的增加，随机森林的分类精度会逐渐提高，但在一定程度后可能会趋于稳定。因此，选项 A 是正确的。B.随机森林对异常值和缺失值具有一定的容忍度,因为它是基于多个决策树的集成学习算法。每个决策树在训练时会自动处理缺失值，并且对于异常值的影响相对较小。因此，选项 B 是正确的。C.虽然随机森林在一定程度上可以减少过拟合的风险，但仍然需要考虑过拟合问题。特别是在数据量较小或特征数量较多的情况下，过拟合可能仍然会发生。因此，选项 C 是错误的。D.决策树之间的相关系数越低，说明它们之间的差异越大，能够提供更多的信息。同时，每棵决策树的分类精度越高，整个随机森林的分类效果也会越好。因此，选项 D 是正确的。综上所述，说法错误的是选项 C。

49.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、Sigmoid

答案：A

50.5.EM 算法的停止条件（）

A、已达到最大迭代轮数

B、数据样本异常

C、训练器异常

D、似然函数减小

答案：A

解析：答案解析：EM 算法是一种迭代算法，用于求解包含隐变量的概率模型参数。在实际应用中，通常需要设置停止条件来决定何时结束迭代。已达到最大迭代轮数是常见的停止条件之一。因为如果无限制地迭代下去，可能会增加计算成本，且不一定能显著改善结果。而数据样本异常、训练器异常通常不是 EM 算法正常的停止条件。似然函数一般是增大的，而不是减小。所以，选项 A 是正确的答案。

51. “没有免费的午餐定理”告诉我们

A、我们不能对问题有先验假设

B、没有可以适应一切问题的算法

C、设计好的算法是徒劳的

D、对于一个特定的问题，任何算法都是一样好的

答案：B

解析：“没有免费的午餐定理”（NoFreeLunchTheorem）是机器学习和优化理论中的一个重要概念。它的主要含义是，在所有可能的问题上，没有一种算法可以在所有情况下都优于其他算法。具体来说，这个定理告诉我们，对于任何一个算法，它在某些问题上可能表现得很好，但在其他问题上可能表现得很差。因此，我们不能期望有一种通用的算法可以解决所有的问题，也不能对任何算法有先验的假设。在实际应用中，我们需要根据具体的问题选择合适的算法，并对算法进行评估和优化。同时，我们也需要不断探索和研究新的算法，以提高解决问题的效率和质量。因此，选项 B 是正确的答案。

52.4.“学习向量量化”与一般聚类算法不同的是（）

A、数据样本带有类别标记

B、结构不同

C、向量程度不同

D、簇的种类不同

答案：A

解析:“学习向量量化”是一种有监督的学习算法，而一般聚类算法大多是无监督的。有监督学习中数据样本通常带有类别标记，这是它与一般聚类算法的重要区别。选项 B 中结构不同不是本质区别；选项 C 向量程度不同表述不准确；选项 D 簇的种类不同也不是关键不同点。所以答案选 A。

53.在大数据集上训练决策树，为了使用较少时间，我们可以（）

A、增加树的深度

B、增加学习率

C、减少树的深度

D、减少树的数量

答案：C

54.关于决策树结点划分指标描述正确的是

A、类别非纯度越大越好

B、信息增益越大越好

C、信息增益率越小越好

D、基尼指数越大越好

答案：B

解析: 在决策树中，信息增益表示特征使数据集的不确定性减少的程度，信息增益越大，说明该特征对分类的作用越明显，越有利于对数据集进行准确划分，所以信息增益越大越好，B 选项正确；而类别非纯度、信息增益率和基尼指数都不是越大越好，A、C、D 选项错误。

55. 做一个二分类预测问题，先设定阈值为 0.5，概率大于等于 0.5 的样本归入正例类（即 1），小于 0.5 的样本归入反例类（即 0）。然后，用阈值 n（n>0.5）重新划分样本到正例类和反例类，下面哪一种说法正确是（）1. 增加阈值不会提高召回率 2. 增加阈值会提高召回率 3. 增加阈值不会降低查准率 4. 增加阈值会降低查准率

A、1

B、2

C、1、3

D、2、4

答案：C

解析: 召回率是实际为正例的样本中被预测为正例的比例。增加阈值，会使得被判定为正例的样本减少，原本一些可能被判定为正例的现在可能被归为反例，这样就可能导致召回率降低或不变，不会提高，所以 1 正确，2 错误。查准率是预测为正例的样本中实际为正例的比例，增加阈值后，预测为正例的样本更可能是真正的正例，查准率可能提高或不变，不会降低，所以 3 正确，4 错误。综上，正确答案是 C。

56. 下列不是 SVM 核函数的是：

A、多项式核函数

B、logistic 核函数

C、径向基核函数

D、Sigmoid 核函数

答案：B

57.点击率的预测是一个数据比例不平衡问题（比如训练集中样本呈阴性的比例为 99%，阳性的比例是 1%），如果我们用这种数据建立模型并使得训练集的准确率高达 99%。我们可以得出结论是：

A、模型的准确率非常高，我们不需要进一步探索

B、模型不好，我们应建一个更好的模型

C、无法评价模型

D、以上都不正确

答案：B

58.下列哪种归纳学习采用符号表示方式？

A、经验归纳学习

B、遗传算法

C、联接学习

D、强化学习

答案：A

59.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、分解阈值

C、均值

D、方差

答案：D

解析: 在 StandardScaler 预处理方法中,公式为$x'=(x-\mu)/\sigma$,其中x表示原始数据，μ表示均值，σ表示标准差。而标准差的平方就是方差。因此，在这个公式中，$?=\sigma$，表示特征所在列的方差。所以，正确答案是D。

60.下列关于主成分分析的表述错误的是

A、主成分分析方法一种数据降维的方法

B、通过主成分分析，可以将多个变量缩减为少数几个新的变量，而信息并没有损失，或者说信息损失很少

C、通过主成分分析，可以用较少的新的指标来代替原来较多的指标反映的信息，并且新的指标之间是相互独立的

D、主成分分析是数据增维的方法

答案：D

解析: 主成分分析是一种数据降维的方法，它可以将多个相关变量转换为少数几个不相关的综合指标,即主成分。这些主成分能够尽可能地保留原始数据的信息，同时减少数据的维度。选项 A 正确,主成分分析的主要目的就是降低数据的维度。选项 B 也正确，通过主成分分析，原始变量的大部分信息可以被压缩到少数几个主成分中，信息损失较小。选项 C 同样正确，主成分之间是相互独立的，这有助于简化数据分析和解释。而选项 D 错误，主成分分析是减维而不是增维的方法。综上所述，正确答案是 D。

61.朴素贝叶斯分类器的三种实现不包括

A、基于伯努利模型实现

B、基于多项式模型实现

C、属性条件独立性假设实现

D、基于高斯模型实现

答案：C

解析: 朴素贝叶斯分类器常见的实现方式有基于伯努利模型、基于多项式模型和基于高斯模型。而属性条件独立性假设是朴素贝叶斯分类器的基本假设，并非是一种具体的实现方式。所以，答案选 C。

62.下面关于贝叶斯分类器描述错误的是

A、以贝叶斯定理为基础

B、是基于后验概率

C、可以解决有监督学习的问题

D、可以用极大似然估计法解贝叶斯分类器

答案：B

解析: 答案解析: 贝叶斯分类器是以贝叶斯定理为基础，可用于解决有监督学习的问题，常用极大似然估计法求解。然而，贝叶斯分类器是基于先验概率和条件概率，而不是基于后验概率。所以，选项 B 描述错误，选项 A、C、D 均符合贝叶斯分类器的特点。因此，答案选择 B 选项。

63.1 下列关于线性回归说法错误的是（）

A、在现有模型上，加入新的变量，所得到的 R^2 的值总会增加

B、线性回归的前提假设之一是残差必须服从独立正态分布

C、残差的方差无偏估计是 SSE/(n-p)

D、自变量和残差不一定保持相互独立

答案：D

解析: A 选项正确，加入新变量可能会提高模型的拟合优度，从而使 R^2 值增加。B 选项正确,线性回归的前提假设之一是残差服从独立正态分布。C 选项正确,SSE/(n-p)是残差方差的无偏估计。D 选项错误，自变量和残差应该保持相互独立。综上所述，正确答案是 D。

64.以下哪项是非线性降维方法

A、PCA(主成分分析)

B、LDA（线性判别）

C、ICA（独立成分分析）

D、KPCA(核化线性降维)

答案：D

解析：线性降维方法是指在降维过程中保持数据的线性结构不变，如 PCA、LDA 和 ICA。而非线性降维方法则是通过引入非线性变换来实现降维，KPCA 就是一种核化的线性降维方法，它通过核函数将数据映射到高维特征空间，然后在该空间中进行线性降维，从而能够处理非线性数据。因此，选项 D 是正确的答案。

65.在变量选择过程中，下列哪些方法可用于检查模型的性能？a.多重变量用于同一个模型 b.模型的可解释性 c.特征的信息 d.交叉验证

A、d

B、abc

C、acd