第一章  机器学习概述

1.机器学习研究什么问题，构建一个完整的机器学习算法需要哪些要素？

机器学习主要研究如何选择统计学习模型，从大量已有数据中学习特定经验。构建一个完整的机器学习算法需要三个方面的要素，分别是数据，模型，性能度量准则。

2.可以生成新数据的模型是什么，请举出几个例子

可以生成新数据的模型是生成模型，典型的生成模型有朴素贝叶斯分类器、高斯混合模型、隐马尔可夫模型、生成对抗网络等。

3.监督学习、半监督学习和无监督学习是什么，降维和聚类属于哪一种？

监督学习是指样本集合中包含标签的机器学习，无监督学习是无标签的机器学习，而半监督学习介于二者之间。降维和聚类是无监督学习。

4.过拟合和欠拟合会导致什么后果，应该怎样避免？

过拟合导致模型泛化能力弱，发生明显的预测错误，往往是由于数据量太少或模型太复杂导致，通过增加训练数据量，对模型进行裁剪，正则化的方式来缓解。而欠拟合则会导致模型不能对数据进行很好地拟合，通常是由于模型本身不能对训练集进行拟合或者训练迭代次数太少，解决方法是对模型进行改进，设计新的模型重新训练，增加训练过程的迭代次数。

5.什么是正则化，L1正则化与L2正则化有什么区别？

正则化是一种抑制模型复杂度的方法。L1正则化能够以较大概率获得稀疏解，起到特征选择的作用，并且可能得到不止一个最优解。L2正则化相比前者获得稀疏解的概率小的多，但得到的解更加平滑。

第二章  逻辑回归与最大熵模型

1.逻辑回归模型解决（ B ）

A.回归问题

B.分类问题

C.聚类问题

D.推理问题

2.逻辑回归属于（ B ）回归

A.概率性线性

B.概率性非线性

C.非概率性线性

D.非概率性非线性

3.逻辑回归不能实现（ D ）

A.二分类

B.多分类

C.分类预测

D.非线性回归

4.下列关于最大熵模型的表述错误的是（ B ）

A.最大熵模型是基于熵值越大模型越稳定的假设

B.最大熵模型使用最大熵原理中一般意义上的熵建模以此缩小模型假设空间

C.通过定义最大熵模型的参数可以实现与多分类逻辑回归相同的作用

D.最大熵模型是一种分类算法

5.下列关于模型评价指标的表述错误的是（ C ）

A.准确率、精确率、召回率以及AUC均是建立在混淆矩阵的基础上

B.在样本不平衡的条件下准确率并不能作为很好的指标来衡量结果

C.准确率表示所有被预测为正的样本中实际为正的样本的概率

D.一般来说，置信度阈值越高，召回率越低，而精确率越高

6.简述逻辑回归的原理。

7.最大熵模型的优点和缺点是什么？

1. k-近邻算法

1.k-近邻算法的基本要素不包括（ C ）

A.距离度量

B.k值的选择

C.样本大小

D.分类决策规则

2.关于k-近邻算法说法错误的是（ D ）

A.k-近邻算法是机器学习

B.k-近邻算法是无监督学习

C.k代表分类个数

D.k的选择对分类结果没有影响

3.以下关于k-近邻算法的说法中正确的是（ B ）

A.k-近邻算法不可以用来解决回归问题

B.随着k值的增大，决策边界会越来越光滑

C.k-近邻算法适合解决高维稀疏数据上的问题

D.相对3近邻模型而言，1近邻模型的bias更大，variance更小

4.（ B ）不可以通过无监督学习方式进行训练

A.k-近邻算法

B.决策树

C.RBM

D.GAN

5.以下关于k-近邻算法的说法中，错误的是（ C ）

A.一般使用投票法进行分类任务

B.k-近邻算法属于懒惰学习

C.k-近邻算法训练时间普遍偏长

D.距离计算方法不同，效果也可能有显著差别

6.简述k-近邻算法的步骤。

7.k-近邻算法有哪些优缺点？

第四章 决策树

1.关于机器学习中的决策树学习，说法错误的是（ A ）

A.受生物进化启发

B.属于归纳推理

C.用于分类和预测

D.自顶向下递推

2.在构建决策树时，需要计算每个用来划分数据特征的得分，选择分数最高的特征，以下可以作为得分的是（ D ）

A.熵

B.基尼系数

C.训练误差

D.以上都是

3.在决策树学习过程中，（ D ）可能会导致问题数据（特征相同但是标签不同）

A.数据噪音

B.现有特征不足以区分或决策

C.数据错误

D.以上都是

4.根据信息增益来构造决策树的算法是（ A ）

A.ID3决策树

B.递归

C.归约

D.FIFO

5.决策树构成顺序是（ A ）

A.特征选择、决策树生成、决策树剪枝

B.决策树剪枝、特征选择、决策树生成

C.决策树生成、决策树剪枝、特征选择

D.特征选择、决策树剪枝、决策树生成

6.决策树适用于解决什么样的问题？

7.ID3和CART算法有什么区别？

第五章 朴素贝叶斯分类器

1.朴素贝叶斯分类器的特征不包括（ C ）

A.孤立的噪声点对该分类器影响不大

B.数据的缺失值影响不大

C.要求数据的属性相互独立

D.条件独立的假设可能不成立

2.朴素贝叶斯分类器基于（ B ）假设

A.样本分布独立性

B.属性条件独立性

C.后验概率已知

D.先验概率已知

3.下列关于朴素贝叶斯分类器错误的是（ D ）

A.朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率

B.对小规模的数据表现很好，能个处理多分类任务，适合增量式训练

C.对缺失数据不太敏感，算法也比较简单，常用于文本分类

D.对输入数据的表达形式不敏感

4.朴素贝叶斯分类器为（ A ）

A.生成模型

B.判别模型

C.统计模型

D.预算模型

5.下列关于朴素贝叶斯分类器正确的是（ D ）

A.朴素贝叶斯分类器的变量必须是非连续型变量

B.朴素贝叶斯模型中的特征和类别变量之间也要相互独立

C.朴素贝叶斯分类器对于小样本数据集效果不如决策树好

D.朴素贝叶斯模型分类时需要计算各种类别的概率，取其中概率最大者为分类预测值

6.如何理解朴素贝叶斯分类器中的拉普拉斯平滑？

7.简述朴素贝叶斯算法的原理。

第六章 支持向量机

1.支持向量指的是（ B ）

A.对原始数据进行采样得到的样本点

B.决定分类面可以平移的范围的数据点

C.位于分类面上的点

D.能够被正确分类的数据点

2.下面关于支持向量机(SVM)的描述错误的是（ D ）

A.是一种监督式学习的方法

B.可用于多分类的问题

C.支持非线性的核函数

D.是一种生成式模型

3.下面关于支持向量机(SVM)的描述错误的是（ D ）

A.对于分类问题，支持向量机需要找到与边缘点距离最大的分界线，从而确定支持向量

B.支持向量机的核函数负责输入变量与分类变量之间的映射

C.支持向量机可根据主题对新闻进行分类

D.支持向量机不能处理分界线为曲线的多分类问题

4.支持向量机中margin指（ C ）

A.盈利率

B.损失误差

C.间隔

D.保证金

5.选择margin最大的分类器的原因是（ D ）

A.所需的支持向量个数最少

B.计算复杂度最低

C.训练误差最低

D.有望获得较低的测试误差

6.支持向量机的基本思想是什么？

7.支持向量机如何实现多分类？

第七章 集成学习

1.下列哪个集成学习器的个体学习器存在强依赖关系（ A ）

A.Boosting

B.Bagging

C.Random Forest

D.随机森林

2.下列哪个集成学习器的个体学习器不存在强依赖关系（ C ）

A.Boosting

B.AdaBoost

C.随机森林

D.EM

3.下列（ D ）不是Boosting的特点

A.串行训练的算法

B.基分类器彼此关联

C.串行算法不断减小分类器训练偏差

D.组合算法可以减小分类输出方差

4.下列（ C ）不是Bagging的特点

A.各基础分类器并行生成

B.各基础分类器权重相同

C.只需要较少的基础分类器

D.基于Bootstrap采样生成训练集

5.集成学习的主要思想是（ D ）

A.将多个数据集合集成在一起进行训练

B.将多源数据进行融合学习

C.通过聚类算法使数据集分为多个簇

D.将多个机器学习模型组合起来解决问题

6.集成学习的基本原理是什么？举例说明三种集成学习的应用。

7.集成学习中生成多样性大的个体学习器的方法有哪些？

第八章 EM算法及其应用

1.EM算法是（ B ）学习算法

A.有监督

B.无监督

C.半监督

D.都不是

2.EM算法的E和M指（ A ）

A.Expectation-Maximum

B.Expect-Maximum

C.Extra-Maximum

D.Extra-Max

3.EM算法可以应用于（ D ）

A.学习贝叶斯网络的概率

B.EM-聚类

C.训练HMM

D.以上均可

4.EM算法的核心思想是（ A ）

A.通过不断地求取目标函数的下界的最优值，从而实现最优化的目标。

B.列出优化目标函数，通过方法计算出最优值

C.列出优化目标函数，通过数值优化方法计算出最优值

D.列出优化目标函数，通过坐标下降方法计算出最优值

5.聚类算法包括（ D ）

A.K-means

B.single-linkage

C.Expectation-Maximum

D.以上都有

6.简述EM算法的基本流程。

7.EM算法是如何应用于GMM的？

第九章 降维

1.下列可以用于降维的机器学习方法是（ C ）

A.决策树

B.KNN

C.PCA

D.K-means

2.下列是机器学习中降维任务的准确描述的为（ B ）

A.依据某个准则对项目进行排序

B.将其映射到低维空间来简化输入

C.预测每个项目的实际值

D.对数据对象进行分组

3.下列可以可以通过机器学习解决的任务为（ A ）

A.聚类、降维

B.回归、迭代

C.分类、抽象

D.派生、推荐

4.下列关于主成分分析的表述错误的是（ D ）

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

5.下列关于奇异值分解的表述正确的是（ C ）

A.只有方阵能进行奇异值分解

B.只有非奇异矩阵能进行奇异值分解

C.任意矩阵都能进行奇异值分解

D.对称矩阵的奇异值就是其特征值

6.简述主成分分析的基本原理。

7.简述奇异值分解的基本原理。

第十章 聚类

1.欧氏距离是闵可夫斯基距离阶为（ C ）的特殊情况

A.0.5

B.1

C.2

D.∞

2.在层次聚类中（ C ）

A.需要用户预先设定聚类的个数

B.需要用户预先设定聚类个数的范围

C.对于N个数据点，可形成1到N个簇

D.对于N个数据点，可形成1到N/2个簇

3.关于K-Means算法的表述不正确的是（ B ）

A.算法开始时，K-Means算法需要指定质心

B.K-Means算法的效果不受初始质心选择的影响

C.K-Means算法需要计算样本与质心之间的距离

D.K-means属于无监督学习

4.K-Medoids聚类与K-Means聚类最大的区别在于（ A ）

A.中心点的选取规则

B.距离的计算方法

C.聚类效果

D.应用层面

5.DBSCAN算法属于（ D ）

A.划分聚类

B.层次聚类

C.完全聚类

D.不完全聚类

6.不同的聚类方法分别适用于什么样的场合，请简要分析。

7.试分析K-Means聚类算法收敛的原因。

第十一章 神经网络与深度学习

1.下列关于标准神经元模型的说法错误的是（ A ）

A.具有多个输入端，每个输入端具有相同的权重

B.神经元具有Sigmoid或类似的可导激活函数

C.神经元能够根据误差信号通过梯度下降法调整权重，实现学习

D.具有一个或多个输出端，且输出端信号一致

2.下列关于多层感知机的描述正确的是（ D ）

A.由于激活函数的非线性特点，导致反向传播过程中梯度消失的问题

B.激活函数不必可导

C.没有前馈计算也可以进行反向传播计算

D.ReLU激活函数导致的神经元死亡指的是该节点以后都不可能被激活

3.误差反向传播算法属于（ B ）学习规则

A.无导师

B.有导

C.死记忆

D.混合

4.下列关于卷积神经网络说法错误的是（ A ）

A.是目前网络深度最深、应用最成功的深度学习模型

B.卷积神经网络模拟了人类视觉信息处理的过程

C.图像的卷积，很类似视觉皮层对图像进行某种特定功能的处理

D.模拟大脑的视觉处理过程就是卷积神经网络的思路

5.下列关于生成对抗网络的描述错误的是（ D ）

A.生成对抗网络包括两部分，即生成器和判别器

B.生成对抗网络的判别器进行训练时，其输入为生成器生成的图像和来自训练集中的真实图像，并对其进行判别

C.生成对抗网络的生成器从随机噪声中生成图像（随机噪声通常从均匀分布或高斯分布中获取)

D.既然生成对抗网络是无监督模型，则不需要任何训练数据

6.前馈式神经元网络与反馈式神经元网络有何不同？

7.简述误差反向传播算法。