aws training and certification

AWS 认证机器学习 – 专项
AWS Certified Machine Learning – Specialty
(MLS-C01) 考试样题

1) 机器学习团队在 Amazon S3 中有多个大型 CSV 数据集。过去，使用 Amazon SageMaker 线性学习器算法构建的模型，在类似大小的数据集上进行训练需要花费数小时时间。团队的负责人需要加快这一训练过程。

为了解决这个问题，机器学习专家可以采取什么操作？

A. 使用 Amazon SageMaker 管道模式。
B. 使用 Amazon Machine Learning 训练模型。
C. 使用 Amazon Kinesis 将数据流式传输到 Amazon SageMaker。
D. 使用 AWS Glue 将 CSV 数据集转换为 JSON 格式。

2) 从包含以下两个句子的文字语料库，构建了一个同时使用一元和二元语言模型的词频-逆文档频文本频率（tf – idf）矩阵：

1. Please call the number below。
2. Please do not call us。

tf – idf 矩阵的维度是多少？

A. (2, 16)
B. (2, 8)
C. (2, 10)
D. (8, 10)

3) 一家公司正在设置一个系统来管理存储在 Amazon S3 中的所有数据集。该公司希望自动对数据运行转换任务并维护与数据集有关的元数据目录。该解决方案应尽可能减少设置和维护工作。

哪个解决方案可以让公司实现这些目标？

A. 创建安装了 Apache Hive 的 Amazon EMR 集群。然后，创建 Hive 元存储以及按计划运行转换任务的脚本。
B. 创建 AWS Glue 爬网程序以填充 AWS Glue 数据目录。然后，编写一个 AWS Glue ETL 任务，并为数据转换任务设置计划。
C. 创建安装了 Apache Spark 的 Amazon EMR 集群。然后，创建 Apache Hive 元存储以及按计划运行转换任务的脚本。
D. 创建用于转换数据的 AWS Data Pipeline。然后，创建 Apache Hive 元存储以及按计划运行转换任务的脚本。

4) 一位数据科学家在训练过程中通过改变多个参数来优化模型。这位数据科学家观察到，在使用相同参数的多次运行过程中，损失函数收敛到不同的稳定值。

这位数据科学家应采取什么操作来改进训练过程？

A. 增加学习速率。保持相同的批次大小。
B. 减小批次大小。降低学习速率。
C. 保持相同的批次大小。降低学习速率。
D. 不更改学习速率。增大批次大小。

5) 一位数据科学家在评估不同的二元分类模型。一个假阳性结果的成本是假阴性结果的 5 倍（从业务的角度来看）。

模型应根据以下标准进行评估：
1) 查全率必须至少为 80%
2) 假阳性比率不能超过 10%
3) 必须尽可能减少业务成本

在创建各个二元分类模型之后，这位数据科学家生成了对应的混淆矩阵。

哪个混淆矩阵代表满足要求的模型？

A. TN = 91, FP = 9
   FN = 22, TP = 78
B. TN = 99, FP = 1
   FN = 21, TP = 79
C. TN = 96, FP = 4
   FN = 10, TP = 90
D. TN = 98, FP = 2
   FN = 18, TP = 82

6) 一位数据科学家使用逻辑回归来构建欺诈检测模型。虽然模型准确率为 99%，但模型未检测到 90% 的欺诈案例。

哪些操作肯定可以帮助模型检测到超过 10% 的欺诈案例？

A. 使用欠采样来平衡数据集
B. 降低类别概率阈值
C. 使用正则化来减少过度拟合
D. 使用过采样来平衡数据集

7) 一家公司有兴趣构建欺诈检测模型。目前，由于欺诈案例数较少，数据科学家没有足够的信息。

哪种方法最有可能检测到最多的有效欺诈案例？

A. 使用自举方法进行过采样
B. 欠采样
C. 使用 SMOTE 进行过采样
D. 分类权重调整

8) 一位机器学习工程师在为使用 Amazon SageMaker 线性学习器算法的监督式学习任务准备数据帧。ML 工程师注意到目标标签分类极不平衡，多个特征列缺少值。整个数据帧中缺失值的比例低于 5%。

该 ML 工程师应采取什么方法来尽可能减少由于缺失值造成的偏差？

A. 用相同行中非缺失值的平均值或中值替换各个缺失值。
B. 删除包含缺失值的观察数据，因为这些值代表不到 5% 的数据。
C. 用相同列中非缺失值的平均值或中值替换各个缺失值。
D. 对于各个特征，根据其他特征使用监督式学习来估算缺失值。

9) 一家公司使用决策树收集客户对其产品的评论，其评级为安全或不安全。训练数据集具备以下各项特征：id、日期、完整评论、完整评论摘要以及安全/不安全二元标签。在训练期间，丢弃缺少特征的任意数据样本。在很少的情况下，发现测试集缺少完整评论文本字段。

对于此使用案例，可解决缺少特征的测试数据样本的最有效做法是什么？

A. 丢弃缺少完整评论文本字段的测试样本，然后对整个测试集运行。
B. 复制摘要文本字段并用这些字段填充缺失的完整评论文本字段，然后对整个测试集运行。
C. 使用比决策树能更好地处理缺失数据的算法。
D. 生成合成数据来填充缺少数据的字段，然后对整个测试集运行。

10) 一家保险公司需要自动执行索赔合规性审核，因为人工审核成本高且易于出错。该公司要处理大量的索赔，每个索赔有一个合规性标签。每个索赔由几个英文句子组成，其中许多句子包含复杂的相关信息。管理层希望使用 Amazon SageMaker 内置算法来设计机器学习监督式模型，该模型可进行训练来读取各个索赔并预测索赔是否合规。

应该使用什么方法从索赔中抽取特征，以便用作下游监督式任务的输入？

A. 从整个数据集中的索赔派生令牌字典。对在训练集中的各个索赔中发现的令牌，应用独热编码。将派生的特征空间作为输入发送到 Amazon SageMaker 内置监督式学习算法。
B. 将 Word2Vec 模式的 Amazon SageMaker BlazingText 应用到训练集中的索赔。将派生的特征空间作为输入发送到下游监督式任务。
C. 将分类模式的 Amazon SageMaker BlazingText 应用到训练集中的已标记索赔，以便用于分别派生与合规及不合规标签对应的索赔的特征。
D. 将 Amazon SageMaker Object2Vec 应用到训练集中的索赔。将派生的特征空间作为输入发送到下游监督式任务。

答案

1) A – Amazon SageMaker 管道模式将数据直接流式传输到容器，这可以改进训练任务的性能。（有关支持信息，请参阅此链接。）在管道模式中，您的训练任务直接从 Amazon S3 流式传输数据。流式传输可以为训练任务提供更快的启动时间和更好的吞吐量。利用管道模式，您还可以为训练实例减小 Amazon EBS 卷的大小。B 不适用于此场景。C 是流式提取解决方案，但不适用于此场景。D 会转换数据结构。

2) A – 这里有 2 个句子、8 个不同的一元语言模型和 8 个不同的二元语言模型，因此结果是 (2,16)。短语为“Please call the number below”和“Please do not call us”。单独的每个单词（一元语言模型）包括“Please”、“call”、“the”、“number”、“below”、“do”、“not”和“us”。唯一二元语言模型包括“Please call”、“call the”、“the number”、“number below”、“Please do”、“do not”、“not call”和“call us”。tf-idf 向量器在此链接中进行介绍。

3) B – AWS Glue 是正确的答案，因为此选项所需的设置和维护工作最少，这是由于它是无服务器类型的，并且不需要管理基础设施。有关支持信息，请参阅此链接。A、C 和 D 均为可解决问题的解决方案，但需要更多的步骤进行配置，并且运行和维护所需的运营开销较高。

4) B – 损失函数的曲线很可能非常弯，并且具有多个局部最小值，训练会在这些位置卡住。减少批次大小有助于数据科学家随机地摆脱局部最小值鞍点。降低学习速率可以防止过度处理全局损失函数最小值。有关说明，请参阅此链接中的文章。

5) D – 需要以下计算：

TP = 真阳性
FP = 假阳性
FN = 假阴性
TN = 真阴性
FN = 假阴性

查全率 = TP/(TP + FN)

假阳性比率 (FPR) = FP/(FP + TN)

成本 = 5 * FP + FN

	A	B	C	D
查全率	78 / (78 + 22) = 0.78	79 / (79 + 21) = 0.79	90 / (90 + 10) = 0.9	82 / (82 + 18) = 0.82
假阳性比率	9 / (9 + 91) = 0.09	1 / (1 + 99) = 0.01	4 / (4 + 96) = 0.04	2 / (2 + 98) = 0.02
成本	5 * 9 + 22 = 67	5 * 1 + 21 = 26	5 * 4 + 10 = 30	5 * 2 + 18 = 28

选项 C 和 D 的查全率超过 80%，FPR 低于 10%，但 D 最经济高效。有关支持信息，请参阅此链接。

6) B – 降低类别概率阈值使得模型更加敏感，因此会将更多案例标记为正类，在本例中为欺诈案例。这将增加欺诈检测的可能性。但是，其代价是降低查准率。此链接中的文章的“讨论”部分介绍了这一内容。

7) C – 在数据集未完全填充的情况下，合成少数过采样技术（SMOTE）通过将合成数据点添加到少数分类来添加新信息。此技术在本场景中最有效。有关支持信息，请参阅此链接中的第 4.2 部分。

8) D – 根据其他特征的值使用监督式学习来预测缺失值。不同监督式学习方法可能会有不同的表现，但相比平均值或中值近似值，任何正确实施的监督式学习方法应提供相同或更好的近似值，如答案 A 和 C 中的建议。应用到缺失值估算的监督式学习是一个非常活跃的研究领域。有关示例，请参阅此链接。

9) B – 在此案例中，完整评论摘要通常包含整个评论最具说明性的短语，并且能够有效地代表缺少的完整评论文本字段。有关支持信息，请参阅此链接的第 1627 页、此链接和此链接。

10) D – Amazon SageMaker Object2Vec 针对单词到更复杂的对象（例如句子和段落）广泛运用 Word2Vec 嵌入技术。由于监督式学习任务在整个索赔的级别上进行，而索赔存在标签但单词级别没有标签可用，必须使用 Object2Vec 而不是 Word2Vec。有关支持信息，请参阅此链接和此链接。