aws training and certification

AWS 공인 기계 학습 – 전문 분야
AWS Certified Machine Learning – Specialty
(MLS-C01) 시험 샘플 문항

1) 어떤 Machine Learning 팀이 Amazon S3에 대형 CSV 데이터 세트 여러 개를 보유하고 있습니다. 지금까지 Amazon SageMaker Linear Learner 알고리즘으로 빌드한 모델은 비슷한 크기의 데이터 세트로 훈련시키는 데 몇 시간이 걸렸습니다. 이 팀의 리더는 훈련 프로세스의 속도를 높여야 합니다.

Machine Learning 전문가는 무엇으로 이 문제를 해결할 수 있습니까?

A. Amazon SageMaker 파이프 모드를 사용한다.
B. Amazon Machine Learning을 사용하여 모델을 훈련시킨다.
C. Amazon Kinesis를 사용하여 데이터를 Amazon SageMaker로 스트리밍한다.
D. AWS Glue를 사용하여 CSV 데이터 세트를 JSON 형식으로 변환한다.

2) 다음 두 문장으로 구성된 말뭉치 코퍼스에서 유니그램과 바이그램을 모두 사용하는 단어 빈도-역문서 빈도(tf-idf) 행렬을 빌드했습니다.

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

tf-idf 행렬에는 어떤 차원이 있습니까?

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

3) 한 회사가 Amazon S3에 저장하는 모든 데이터 세트의 관리 시스템을 설정하려고 합니다. 이 회사는 데이터 변환 작업의 실행 및 데이터 세트와 관련된 메타데이터의 카탈로그 유지 관리를 자동화하고 싶어합니다. 이 솔루션에 필요한 설정 및 유지 관리 작업은 최소 수준이어야 합니다.

이 회사는 어떤 솔루션으로 원하는 목표를 달성할 수 있습니까?

A. Apache Hive가 설치되어 있는 Amazon EMR 클러스터를 생성한다. 그런 다음 Hive 메타스토어와 스크립트를 만들어 변환 작업을 일정에 따라 실행한다.
B. AWS Glue 크롤러를 만들어 AWS Glue 데이터 카탈로그를 채운다. 그런 다음 AWS Glue ETL 작업을 작성하고, 데이터 변환 작업의 일정을 정한다.
C. Apache Spark가 설치되어 있는 Amazon EMR 클러스터를 생성한다. 그런 다음 Apache Hive 메타스토어와 스크립트를 만들어 변환 작업을 일정에 따라 실행한다.
D. 데이터를 변환하는 AWS Data Pipeline을 만든다. 그런 다음 Apache Hive 메타스토어와 스크립트를 만들어 변환 작업을 일정에 따라 실행한다.

4) 한 데이터 과학자가 여러 파라미터를 변경하여 훈련 프로세스 중에 모델을 최적화하는 작업을 하고 있습니다. 이 데이터 과학자는 동일한 파라미터로 여러 번 실행하는 동안 손실 함수가 서로 다르지만 안정적인 값으로 수렴되는 것을 관찰했습니다.

이 데이터 과학자가 훈련 프로세스를 개선하려면 어떻게 해야 합니까?

A. 학습률을 높인다. 배치 크기를 동일하게 유지한다.
B. 배치 크기를 줄인다. 학습률을 낮춘다.
C. 배치 크기를 동일하게 유지한다. 학습률을 낮춘다.
D. 학습률을 변경하지 않는다. 배치 크기를 늘린다.

5) 한 데이터 과학자가 서로 다른 이진 분류 모델을 평가하려고 합니다. 비즈니스 관점에서, 거짓 긍정 결과는 거짓 부정 결과보다 5배 더 많은 비용이 듭니다.

이 모델은 다음 기준에 따라 평가해야 합니다.
1) 재현율이 80% 이상이어야 함
2) 거짓 긍정 비율이 10% 미만이어야 함
3) 비즈니스 비용을 최소화해야 함

이 데이터 과학자는 이진 분류 모델을 만든 후 해당하는 혼동 행렬을 생성합니다.

이 요구 사항에 맞는 모델을 나타내는 혼동 행렬은 무엇입니까?

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) 한 데이터 과학자가 로지스틱 회귀를 사용하여 사기 탐지 모델을 빌드합니다. 모델 정확성은 99%이지만, 사기 사례의 90%는 이 모델로 감지되지 않습니다.

어떤 작업을 통해 이 모델이 사기 사례를 10% 이상 명확하게 감지하도록 할 수 있습니까?

A. 언더샘플링을 사용하여 데이터 세트의 균형 조정
B. 클래스 확률 임계값 낮추기
C. 정규화를 사용하여 과적합 줄이기
D. 오버샘플링을 사용하여 데이터 세트의 균형 조정

7) 한 회사가 사기 탐지 모델을 빌드하는 데 관심이 있습니다. 현재 데이터 과학자는 사기 사례 수가 적은 탓에 충분한 양의 정보가 없습니다.

유효한 사기 사례를 가장 많이 탐지할 수 있는 방법은 무엇입니까?

A. 부트스트래핑을 사용하는 오버샘플링
B. 언더샘플링
C. SMOTE를 사용하는 오버샘플링
D. 클래스 가중치 조정

8) 한 Machine Learning 엔지니어가 Amazon SageMaker Linear Learner 알고리즘을 사용하여 지도 학습 작업을 위한 데이터 프레임을 준비하려고 합니다. 이 ML 엔지니어는 대상 레이블 클래스가 매우 불균형하고 여러 특성 열에 누락된 값이 있다는 것을 알았습니다. 전체 데이터 프레임에서 누락된 값의 비율은 5% 미만입니다.

누락된 값으로 인한 편향을 최소화하려면 이 ML 엔지니어는 무엇을 수행해야 합니까?

A. 각 누락된 값을 동일한 행에 있는 누락되지 않은 값의 평균 값 또는 중간 값으로 바꾼다.
B. 누락된 값이 포함된 관측치는 데이터의 5% 미만을 나타내므로 삭제한다.
C. 각 누락된 값을 동일한 열에 있는 누락되지 않은 값의 평균 값 또는 중간 값으로 바꾼다.
D. 각 특성에 대해 다른 특성을 기반으로 지도 학습을 사용하여 누락된 값의 근사치를 계산한다.

9) 한 회사가 의사 결정 트리를 사용하여 제품을 안전하거나 안전하지 않다고 평가하는 제품에 대한 고객 의견을 수집했습니다. 훈련 데이터 세트에는 ID, 날짜, 전체 후기, 전체 후기 요약, 이진 태그(안전함/안전하지 않음) 등의 특성이 있습니다. 훈련 중에 누락된 특성이 있는 데이터 샘플이 삭제되었습니다. 몇 개의 인스턴스에서 테스트 세트에 전체 후기 텍스트 필드가 빠진 것이 발견되었습니다.

이 사용 사례에서 누락된 특성이 있는 테스트 데이터 샘플을 가장 효과적으로 처리하는 작업 순서는 무엇입니까?

A. 전체 후기 텍스트 필드가 누락된 테스트 샘플을 삭제한 후, 테스트 세트를 실행한다.
B. 요약 텍스트 필드를 복사하고 이것으로 누락된 전체 후기 텍스트 필드를 채운 후 테스트 세트를 실행한다.
C. 의사 결정 트리보다 누락된 데이터를 더 잘 처리하는 알고리즘을 사용한다.
D. 합성 데이터를 생성하여 데이터가 누락된 필드를 채운 후, 테스트 세트를 실행한다.

10) 한 보험 회사가 청구 건의 규정 준수 검토를 자동화하려고 합니다. 사람이 검토하면 비용이 많이 들고 오류가 발생할 수 있기 때문입니다. 이 회사에는 대량의 청구 모음과 각 청구에 대한 규정 준수 레이블이 있습니다. 각 청구는 영어 문장 몇 개로 구성되어 있고, 대부분은 복잡한 관련 정보가 포함되어 있습니다. 관리 팀에서는 Amazon SageMaker 기본 제공 알고리즘을 사용하여 각 청구를 읽고 청구가 규정을 준수하는지 여부를 예측하도록 훈련시킬 수 있는 Machine Learning 지도 모델을 설계하고 싶어합니다.

다운스트림 지도 작업을 위한 입력값으로 사용할 특성을 청구에서 추출하려면 어떤 접근 방식을 택해야 합니까?

A. 전체 데이터 세트의 청구에서 토큰 딕셔너리를 추출한다. 훈련 세트의 각 청구에서 발견된 토큰에 원-핫 인코딩을 적용한다. 추출된 특성 공간을 Amazon SageMaker 기본 제공 지도 학습 알고리즘에 입력값으로 보낸다.
B. 훈련 세트의 청구에 Word2Vec 모드의 Amazon SageMaker BlazingText를 적용한다. 다운스트림 지도 작업에 대한 입력으로 추출된 특성 공간을 보낸다.
C. 훈련 세트의 레이블이 지정된 청구에 분류 모드의 Amazon SageMaker BlazingText를 적용하여 각각 규정 준수 및 규정 미준수 레이블에 해당하는 청구 특성을 추출한다.
D. 훈련 세트의 청구에 Amazon SageMaker Object2Vec를 적용한다. 다운스트림 지도 작업을 위한 입력값으로 추출된 특성 공간을 보낸다.

답

1) A – Amazon SageMaker 파이프 모드는 데이터를 컨테이너로 직접 스트리밍하여 훈련 작업의 성능을 높입니다. (지원 정보는 이 링크를 참조하십시오.) 파이프 모드에서는 훈련 작업이 Amazon S3에서 직접 데이터를 스트리밍합니다. 스트리밍을 통해 훈련 작업의 시작 시간을 더 앞당기고 처리량을 늘릴 수 있습니다. 또한 파이프 모드에서는 훈련 인스턴스에 필요한 Amazon EBS 볼륨의 크기를 줄일 수 있습니다. B는 이 시나리오에 해당되지 않습니다. C는 스트리밍 수집 솔루션이지만, 이 시나리오에는 해당되지 않습니다. D는 데이터 구조를 변형합니다.

2) A – 문장 2개, 고유한 유니그램 8개, 고유한 바이그램 8개가 있으므로 결과는 (2,16)입니다. 구문은 "Please call the number below"와 "Please do not call us"입니다. 각 단어는 개별적으로(유니그램) "Please," "call," "the," "number," "below," "do," "not," "us"입니다. 고유한 바이그램은 "Please call," "call the," "the number," "number below," "Please do," "do not," "not call," "call us"입니다. tf–idf 벡터화는 이 링크에 설명되어 있습니다.

3) B – AWS Glue가 정답입니다. 이 옵션은 서버리스이므로 최소한의 설정 및 유지 관리만 하면 되며, 인프라 관리가 필요 없습니다. 지원 정보는 이 링크를 참조하십시오. A, C, D 모두 문제를 해결할 수 있는 솔루션이지만, 구성을 위해 더 많은 단계를 거쳐야 하고 실행 및 유지 관리하는 데 더 많은 운영 오버헤드가 필요합니다.

4) B – 손실 함수에 매우 굴곡이 많고 훈련이 멈춘 지점에 극솟값이 여러 개 있을 가능성이 가장 높습니다. 배치 크기를 줄이면 데이터 과학자가 확률적으로 극솟값의 안장점에서 벗어나는 데 도움이 됩니다. 학습률을 낮추면 글로벌 손실 함수의 최솟값 초과를 방지할 수 있습니다. 설명은 이 링크의 논문을 참조하십시오.

5) D – 다음 계산이 필요합니다.

TP = 참 긍정
FP = 거짓 긍정
FN = 거짓 부정
TN = 참 부정
FN = 거짓 부정

재현율 = TP/(TP + FN)

거짓 긍정 비율(FPR) = FP/(FP + TN)

비용 = 5 * FP + FN

	A	B	C	D
재현율	78 / (78 + 22) = 0.78	79 / (79 + 21) = 0.79	90 / (90 + 10) = 0.9	82 / (82 + 18) = 0.82
거짓 긍정 비율	9 / (9 + 91) = 0.09	1 / (1 + 99) = 0.01	4 / (4 + 96) = 0.04	2 / (2 + 98) = 0.02
비용	5 * 9 + 22 = 67	5 * 1 + 21 = 26	5 * 4 + 10 = 30	5 * 2 + 18 = 28

옵션 C와 D는 재현율이 80% 이상이고 FPR이 10% 미만이지만, D가 가장 경제적입니다. 지원 정보는 이 링크를 참조하십시오.

6) B – 클래스 확률 임계값을 줄이면 모델이 더 민감해지므로, 더 많은 사례(이 경우 사기 사례)를 긍정 클래스로 표시하게 됩니다. 그러면 사기 탐지의 가능성이 높아집니다. 그러나 그 대신 정밀도는 낮아지게 됩니다. 이 링크에 있는 논문의 토론 단원에 관련 내용이 나와 있습니다.

7) C – 완전히 채워지지 않은 데이터 세트가 있는 경우, SMOTE(Synthetic Minority Over-sampling Technique)가 소수 클래스에 합성 데이터 요소를 추가하여 새로운 정보를 추가합니다. 이 시나리오에서는 이 기술이 가장 효과적입니다. 지원 정보는 이 링h크의 4.2 단원을 참조하십시오.

8) D – 지도 학습을 사용하여 다른 특성의 값을 기반으로 누락된 값을 예측합니다. 지도 학습 방식마다 성능이 서로 다를 수 있지만, 올바르게 구현된 모든 지도 학습 방식은 응답 A 및 C에 제시된 바와 같이 평균 또는 중간 근사치와 동일하거나 더 나은 근사치를 제공해야 합니다. 누락된 값의 대치법에 적용되는 지도 학습은 활발한 연구 분야입니다. 예는 이 링크를 참조하십시오.

9) B – 이 경우에 전체 후기 요약에 일반적으로 전체 후기에 대한 가장 자세한 문구가 포함되어 있으며, 누락된 전체 후기 텍스트 필드를 대신합니다. 지원 정보는 이 링크의 1,627페이지, 이 링크, 이 링크를 참조하십시오.

10) D – Amazon SageMaker Object2Vec는 단어용 Word2Vec 임베딩 기술을 문장 및 단락 등 더 복잡한 객체에 사용할 수 있도록 일반화합니다. 지도 학습 작업은 레이블이 있는 전체 청구 수준에서 이루어지고 단어 수준에는 사용 가능한 레이블이 없으므로 Word2Vec 대신에 Object2Vec를 사용해야 합니다. 지원 정보는 이 링크와 이 링크를 참조하십시오.