aws training and certification

AWS 認定機械学習 – 専門知識
AWS Certified Machine Learning – Specialty
(MLS-C01) 認定試験の質問例

1) ある機械学習チームが、CSV 形式の大規模なデータセットを Amazon S3 内にいくつか格納しています。この程度のサイズのデータセットを使用して、Amazon SageMaker の線形学習アルゴリズムを用いて作成されたモデルをトレーニングする場合、今までは数時間かかっていました。このチームのリーダーは、トレーニングプロセスにかかる時間を短縮する必要があります。

この問題に対処するには、どうすればよいですか。

A. Amazon SageMaker のパイプモードを使用する。
B. Amazon Machine Learning を使用して、モデルをトレーニングする。
C. Amazon Kinesis を使用して、データを Amazon SageMaker にストリーミングする。
D. AWS Glue を使用して、データセットを CSV 形式から JSON 形式に変換する。

2) 次の 2 つの文から成るテキストコーパスから、ユニグラムとバイグラムの両方を使用する Term Frequency – Inverse Document Frequency (TF-IDF、単語出現頻度ー逆文書頻度) マトリックスを作成しました。

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

TF-IDF マトリックスの次元はどれですか。

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

3) ある企業が、Amazon S3 に格納されているすべてのデータセットを管理するためのシステムを構築しようとしています。この企業は、データ変換ジョブを実行する作業と、データセットに関するメタデータのカタログをメンテナンスする作業を自動化したいと考えています。要件は、セットアップ作業とメンテナンス作業の量を最小化することです。

この要件を満たすには、どうすればよいですか。

A. Amazon EMR クラスターを作成し、Apache Hive をインストールする。次に、Apache Hive メタストアとスクリプトを作成し、スケジュールに基づいて変換ジョブを実行する。
B. AWS Glue クローラを作成し、AWS Glue データカタログにデータを書き込む。次に、AWS Glue ETL ジョブを作成し、また、データ変換ジョブのスケジュールを設定する。
C. Amazon EMR クラスターを作成し、Apache Spark をインストールする。次に、Apache Hive メタストアとスクリプトを作成し、スケジュールに基づいて変換ジョブを実行する。
D. データを変換するための AWS Data Pipeline を作成する。次に、Apache Hive メタストアとスクリプトを作成し、スケジュールに基づいて変換ジョブを実行する。

4) データサイエンティストが、モデルトレーニングプロセスで、複数のパラメータ値を調整してモデルを最適化しようとしています。同じパラメータ値を使用してトレーニングを複数回実行したとき、損失関数が、異なってはいるが安定した値に収束することがわかりました。

トレーニングプロセスを改善するには、どうすればよいですか。

A. 学習率を高くする。バッチサイズを変更しない。
B. バッチサイズを小さくする。学習率を低くする。
C. バッチサイズを変更しない。学習率を低くする。
D. 学習率を変更しない。バッチサイズを大きくする。

5) データサイエンティストが、さまざまな二項分類モデルを評価しています。業務の点から見ると、偽陽性結果は偽陰性結果に比べてコストが 5 倍高くなります。

このモデルの評価基準は次のとおりです。

1) リコール率が 80% 以上である。
2) 偽陽性率が 10% 以下である。
3) 業務コストが最小になる。

データサイエンティストは、各二項分類モデルを作成した後、対応する混合行列を生成しました。

評価基準を満たしているモデルを表している、混合行列はどれですか。

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) データサイエンティストがロジスティック回帰を使用して、不正検知モデルを作成しています。モデルの正確性は 99% ですが、不正ケースの 90% は、このモデルで検知されません。

不正ケースの検知率が確実に 10% を上回るようにするには、どうすればよいですか。

A. アンダーサンプリングを使用して、データセットを均衡化する。
B. クラス確率閾値を減らす。
C. 正則化を使用して、オーバーフィッティングを減らす。
D. オーバーサンプリングを使用して、データセットを均衡化する。

7) ある企業が、不正検知モデルの作成に関心を示しています。現在のところ、不正ケースの数が少ないので、データサイエンティストは十分な量のデータを持っていません。

最も多くの不正ケースを検知できる方法はどれですか。

A. ブートストラップを使用したオーバーサンプリング
B. アンダーサンプリング
C. SMOTE を使用したオーバーサンプリング
D. クラス重み付け調整

8) 機械学習エンジニアが、Amazon SageMaker 線形学習アルゴリズムを使用する、教師あり学習タスク用のデータフレームを準備しています。エンジニアは、ターゲットラベルクラスの均衡度が非常に低く、複数の特徴量列の値が欠損していることに気付きました。データフレーム全体における欠損値の割合は、5% 未満です。

欠損値に起因する偏りを最小化するには、どうすればよいですか。

A. 各欠損値を、同じ行内の欠損していない値の平均値または中央値で置き換える。
B. 欠損値はデータの 5% 未満にすぎないため、欠損値が含まれている観測値を削除する。
C. 各欠損値を、同じ列内の欠損していない値の平均値または中央値で置き換える。
D. 各特徴量に対して、他の特徴量に基づく教師あり学習を使用して、欠損値の近似値を求める。

9) ある企業が、自社製品に関する顧客の意見を収集し、ディシジョンツリーを使用して安全かどうかを評価しています。トレーニングデータセット内の特徴量は、ID、日付、レビュー詳細情報、レビュー概要情報、および安全/非安全タグです。トレーニング時、欠損値があるデータサンプルは除去されました。テストデータセットにおいて、少数のレビュー詳細情報テキストフィールドの値が欠損していることがわかりました。

このシナリオにおいて、欠損値があるテストデータサンプルに対処するための、最も効果的な方法はどれですか。

A. レビュー詳細情報テキストフィールドの値が欠損しているテストデータサンプルを除去する。テストデータセットを使用して、モデルを実行する。
B. レビュー概要情報テキストフィールドの値をコピーする。この値を使用して、欠損しているレビュー詳細情報テキストフィールドの値を補完する。テストデータセットを使用して、モデルを実行する。
C. ディシジョンツリーよりも効果的な欠損値処理アルゴリズムを使用する。
D. 合成データを生成する。この値を使用して、欠損しているレビュー詳細情報テキストフィールドの値を補完する。テストデータセットを使用して、モデルを実行する。

10) ある保険会社が、保険金請求の適合性審査を自動化する必要があります。人による審査は高コストであり、またミスが発生しやすいからです。この企業では、保険金請求が大量に発生します。また、各保険金請求に対して適合性ラベルを付加します。各保険金請求は、数個の英文で構成されています。英文の多くは、込み入った関連情報です。経営幹部は、Amazon SageMaker の組み込みアルゴリズムを使用して、教師あり学習モデルを設計し、このモデルをトレーニングしたいと考えています。このモデルの目的は、各保険金請求を読み取り、保険金請求の適合性を予測することです。

保険金請求の中から、ダウンストリームの教師あり学習タスクに対する入力として使用する特徴量を抽出するには、どうすればよいですか。

A. データセット全体内の保険金請求内から、トークンの辞書を抽出する。トレーニングセット内の各保険金請求内にあるトークンに対して、ワンホットエンコーディングを適用する。抽出された特徴量空間を、Amazon SageMaker の組み込み教師あり学習アルゴリズムに入力として送信する。
B. Amazon SageMaker の BlazingText アルゴリズムの Word2Vec モードを、トレーニングセット内の保険金請求に適用する。抽出された特徴量空間を、ダウンストリームの教師あり学習タスクに対する入力として送信する。
C. Amazon SageMaker の BlazingText アルゴリズムの分類モードを、トレーニングセット内のラベル付き保険金請求に適用する。適合ラベルに対する保険金請求、および不適合ラベルに対応する保険金請求に対する、特徴量を抽出する。
D. Amazon SageMaker の Object2Vec アルゴリズムを、トレーニングセット内の保険金請求に適用する。抽出された特徴量空間を、ダウンストリームの教師あり学習タスクに対する入力として送信する。

回答

1) A – Amazon SageMaker のパイプラインモードを使用した場合、データがコンテナに直接ストリーミングされます。これにより、トレーニングジョブのパフォーマンスが向上します。(補足情報については、このリンク先を参照してください。)パイプモードの場合、トレーニングジョブによって Amazon S3 内のデータが直接ストリーミングされます。データがストリーミングされることにより、トレーニングジョブがより早く開始され、また、スループットが向上します。さらに、トレーニングインスタンス用 Amazon EBS ボリュームのサイズが小さくなります。B は、このシナリオには当てはまりません。C は、ストリーミングを使用する方法ですが、このシナリオには当てはまりません。D の場合、データ構造が変換されます。

2) A – 文が 2 個、一意のユニグラムが 8 個、一意のバイグラムが 8 個あります。したがって、結果は (2,16) になります。フレーズは、「Please call the number below」および「Please do not call us」です。一意のユニグラムは、「Please」、「call」、「the」、「number」、「below」、「do」、「not」、および「us」です。一意のバイグラムは、「Please call」、「call the」、「the number」、「number below」、「Please do」、「do not」、「not call」、および「call us」です。TF-IDF によるベクトル化については、このリンク先を参照してください。

3) B – AWS Glue が正解です。サーバーレスなので、セットアップ作業とメンテナンス作業の量が最小になるからです。また、インフラストラクチャを管理する必要もありません。補足情報については、このリンク先を参照してください。A、C、および D でも、この問題を解決することはできますが、構成作業のステップ数が多くなります。また、実行作業とメンテナンス作業の運用コストが高くなります。

4) B – 損失関数の曲がり具合が大きく、かつ、局所的極小値が複数個あるので、トレーニングが行き詰まっています。バッチサイズを小さくした場合、確率的に、局所的極小鞍点から抜け出しやすくなります。また、学習率を低くした場合、損失関数の大域的最小値を通り過ぎてしまう事態を回避できます。説明については、このリンク先の文書を参照してください。

5) D – 次の計算を行う必要があります。

TP = 真陽性
FP = 偽陽性
FN = 偽陰性
TN = 真陰性
FN = 偽陰性

リコール = TP / (TP + FN)
偽陽性率 (FPR) = FP / (FP + TN)
コスト = 5 * FP + FN

	A	B	C	D
リコール	78 / (78 + 22) = 0.78	79 / (79 + 21) = 0.79	90 / (90 + 10) = 0.9	82 / (82 + 18) = 0.82
偽陽性率	9 / (9 + 91) = 0.09	1 / (1 + 99) = 0.01	4 / (4 + 96) = 0.04	2 / (2 + 98) = 0.02
コスト	5 * 9 + 22 = 67	5 * 1 + 21 = 26	5 * 4 + 10 = 30	5 * 2 + 18 = 28

C と D は共に、リコール率が 80% 以上、偽陽性率が 10 % 未満です。ただし、D のほうが費用対効果の面で優れています。補足情報については、このリンク先を参照してください。

6) B – クラス確率閾値を減らすと、モデルの感度が向上するので、陽性クラス（この場合は不正ケース）としてマークされるケースの数が増えます。これにより、不正を検知できる可能性が高まります。その代わり、精度が低下します。詳細については、このリンク先の文書の「DISCUSSION」セクションを参照してください。

7) C – データセット内のデータ量が不十分な場合、Synthetic Minority Over-sampling Technique (SMOTE) を使用すれば、合成データポイントを少数派クラスに追加することによって、データ量を増やすことができます。**このシナリオの場合、この手法が最も効果的です。**補足情報については、このリンク先の文書のセクション 4.2 を参照してください。

8) D – 教師あり学習を使用し、他の特徴量の値に基づいて欠損値を予測します。教師あり学習手法の性能にはばらつきがあります。ただし、教師あり学習手法を適切に実装した場合、A および C で示されている平均値や中央値と同等以上の精度の近似値が得られます。教師あり学習を使用して欠損値を補完する方法は、研究の盛んな分野です。例については、このリンク先を参照してください。

9) B – このシナリオにおいて、レビュー概要情報テキストフィールドには通常、レビュー内容を的確に説明したフレーズが含まれています。つまり、レビュー概要情報テキストフィールド値を、欠損しているレビュー詳細情報テキストフィールド値の代わりとして使用できます。補足情報については、このリンク先の文書の 1627 ページ、このリンク先、およびこのリンク先を参照してください。

10) D – Amazon SageMaker の Object2Vec アルゴリズムは、単語に対する Word2Vec 埋め込み手法を、文や段落などのより複雑なオブジェクトに対しても使用できるよう、汎用化したものです。教師あり学習タスクは、保険金請求全体のレベルで実行されます。各保険金請求にはラベルが付加されていますが、単語レベルでラベルを付加することはできません。したがって、Word2Vec の代わりに Object2Vec を使用する必要があります。補足情報については、このリンク先およびこのリンク先を参照してください。

