データサイエンティスト検定™
リテラシーレベル

書籍（第2版）掲載の模擬試験（45問）の解説

※『最短突破　データサイエンティスト検定（リテラシーレベル）公式リファレンスブック　第2版』に掲載した問題の解説です。「初版」に掲載した模擬試験の解説ではありませんので、ご注意ください。

問題番号	解答	解説	該当ページ
Q1	c	問題のベン図では、集合Aと集合Bの両方に含まれる要素の集合が表されています。これは積集合と呼ばれ、「AかつB」と表現されます。	59
Q2	a	標準偏差は分散の平方根を取った値であるため、まずは分散を求めます。データの平均は(5＋1＋0－1＋6)/5＝2.2と求められます。ここから、分散は{(5－2.2)^2＋(1－2.2)^2＋(0－2.2)^2＋(－1－2.2)^2＋(6－2.2)^2}/5＝7.76となります。この平方根を取ると、標準偏差は2.79とわかります。	29
Q3	a	今回の問題では、100m走のタイムと給与の額には、一方が変化すれば他方が変化するという関係性(相関関係)があることを主張しており、原因と結果の関係(因果関係)があるとは主張できないことに注意が必要です。b～dは、いずれも因果関係を主張しており、相関関係を主張しているaが適切だとわかります。	34
Q4	b	変数xによる偏微分のため、xが含まれていない項の偏微分の値はすべて0となります。xが含まれているのは$8x^2$とxの2つのみであり、それぞれの偏微分の結果は16xと1になるため、bが適切だとわかります。	55
Q5	b	指数関数とは、$y=a^x$の関係を持つ関数を指します。このとき、xが1増えるとaの指数が1増える、つまりxが1増えるとaを底としたyの対数が1増えると言い換えることができます。よって、指数関数はy軸を対数にした片対数グラフが直線となります。また、先ほどの指数関数を対数関数で表現すると、$\log_a y=x$と表せます。このとき、x軸をaを底とした対数、つまり$\log_a x$で取ると、x＝yの直線関係が描けることがわかります。よって、対数関数はx軸を対数にした片対数グラフとなります。なお、両対数グラフとは、xもyも対数を取って表現したグラフのことをいいます。もし、わからなくなった場合は、具体的な値をaに代入し、実際にグラフを書いてみるとよいでしょう。	41

問題番号	解答	解説	該当ページ
Q6	d	Precision（適合率）は、正例と予想したレコードのうち、実際に正例であるものの割合を指します。問題の混同行列では、故障すると予測されたものは左列の値を足した110＋20＝130で、そのうち実際に故障しているものは110レコードであるため、Precisionは110/130＝0.85と求められます。	80
Q7	c	交差検証では、全データをランダムにブロックに分け、そのうちの1ブロックを検証データ、残りを学習データとしてモデルを作成します。このケースでは、用意されているデータが1万件で、そのうち学習データを8千件としているため、検証データは2千件です。よって、データセットを10000/2000＝5ブロックに分割しており、5ブロックそれぞれが検証データとして使用されるため、学習と評価の実行を5回行うことになります。	136
Q8	a	仮説検定では、否定したい仮説を帰無仮説、主張したい仮説を対立仮説に設定します。この問題では、「夏期講習を受講した生徒の学力が受講前の学力よりも高くなっている」ことを主張したいため、これを対立仮説に設定します。また、この主張の否定（帰無仮説）は、「夏期講習を受講した生徒の学力が受講前の学力よりも高くなっていない」、つまり「夏期講習前後で生徒の学力に差はない」ということなので、適切な組み合わせはaとなります。	85
Q9	a	デンドログラムを見ていくつかのグループに分ける際には、指定したグループの数だけクラスターが存在している箇所に横線を引くことで判断しやすくなります。クラスターが5つ存在するのは、Aのクラスターと（B, C）のクラスターが結合したときです。その箇所に横線を引いてグループを確認すると、5つのグループは（ABC）（D）（E）（F）（G）です。よって、aは不適切だとわかります。なお、類似度の高さは、階層がどの高さでまとめられるかで判断します。BとCは最初にまとめられる（最下部でまとめられる）ので、類似度が高いと判断します。	93

問題番号	解答	解説	該当ページ
Q10	d	a. 原点を0以外に設定すると、棒の長さと実際のスケールがずれ、誤った解釈を導く可能性があるため不適切です。 b. ヒストグラムは1つの定量属性のみを受け入れるため、各層のデータを違う色で表現する、つまり定性属性を追加するのは不適切です。 c. 3D円グラフで立体的に示すと、実際の割合と視覚的な感覚がずれ、誤った解釈を導く可能性があるため不適切です。 d. 増加していることを強調するために矢印を重ねることは、わかりやすく伝わる適切な表現の1つです。	118
Q11	c	時系列データでは、値の変化に様々な要素が関係するため、それらを分解して解釈する必要があります。時系列データを分析する場合、ばらつきや外れ値をいきなり見るのではなく、大局的に長期トレンドや季節変動などの取り出したい傾向を見極めることを優先することが大事です。また、ばらつきや外れ値は、短期的変動である場合とノイズである場合の両方のケースがあるので、いきなり削除対象とするのはよくない判断ともいえます。	125
Q12	a	データの中に外れ値が存在する場合、すべてのデータを用いて算出する統計量は影響を受けます。選択肢のうち、データのすべてを用いる統計量は平均と、平均を用いて算出する標準偏差です。よって、平均が外れ値に引っ張られるとしているaが適切であるとわかります。	123
Q13	a	問題文から、学習用データに対する予測精度が非常に良いにもかかわらず、テストデータに対する予測精度が低いことがわかり、過学習の状態だといえます。過学習を対策するときは、bのようにパラメーターを増やすとモデルの自由度を高くなり、より過学習が起きやすくなることが多いため、aのように正則化を加えるなどしてモデルの自由度を下げることが多く、aが適切であるとわかります。	131

問題番号	解答	解説	該当ページ
Q14	b	深層学習では、学習用の大量のデータを渡すことで、特徴量を自動的に抽出することができます。そのため、人間が特徴量を定義する必要がなく、bは不適切であるとわかります。	144
Q15	b	パディングとは、不足する部分を適当な色のピクセルで埋め合わせる処理を指し、これによって画像サイズを揃えることができます。例えば、画像データの縁に0を挿入することで画像サイズを変えないようにする処理をゼロパディングといいます。なお、a、c、dはそれぞれ、リサイズ、トリミング、正規化を表しています。	151
Q16	d	公開鍵暗号化方式では、公開鍵と秘密鍵の2つを利用し、情報の伝達を行います。送信者は、受信者が公開した公開鍵を用いて受信者に暗号化したデータを送信し、受信者は自身のみが保有している秘密鍵を使って復号します。よって、公開鍵を不特定多数の相手に公開しても秘密鍵を非公開にしているため情報は守られており、dが適切であるとわかります。	213
Q17	a	ER図では、データのまとまりであるエンティティ間の関係性を結合によって表現することができますが、データの値自体の関係性を説明することはできません。よって、データ間の相関関係を説明することはできないため、aが不適切であるとわかります。	170
Q18	c	問題文のデータ定義では、繰り返される項目がなく、レコード単位の情報になっているため、第一正規化の条件を満たしています。さらに、第二正規化の条件を満たしています。一方で、会員レベルIDで会員レベルを一意に特定できるため、会員レベルは会員レベルIDに推移関数従属しています。よって、第三正規化の条件は満たしておらず、cが適切であるとわかります。	171

問題番号	解答	解説	該当ページ
Q19	b	Hadoopでは、ネットワークで接続した複数のコンピューターで分担して処理を行います。Hadoopのデータ処理の仕組みに用いられているMapReduceアプリケーションでは、巨大なデータを処理できる反面、特定のレコードを指定したデータの更新はストレージに対する読み書きが何度も発生するため、あまり推奨されません。よって、bが適切であるとわかります。	175
Q20	b	SQLで未入力を表す場合は、「""」や「''」で表現をします。また、未成年は20歳未満と言い換えられるため、これらをandでつないだbが適切であるとわかります。	178
Q21	d	文字列を日付型に変換する関数としては、TO_DATE関数があります。これは、第一引数に変換したい文字列、第二引数に変換後のフォーマットを指定することで日付型に変換することができる関数です。今回は年(Year)・月(Month)・日(Day)をYYYYMMDDで指定しています。	186
Q22	c	BIツールによるデータ可視化の際には、対象となるデータを選定し、目的に合わせたグラフで可視化することが重要です。また、必要な情報だけを可視化するためにデータを絞り込み、理解しやすいようにグラフの体裁を整えることも求められます。よって、a、b、dは比較的重要度が高い作業です。対してcの「データ作成年月日」は、必ずしも「データそのものが示す日付」と一致するわけではなく、必要性が低い作業だとわかります。	194
Q23	c	クラウド提供者が提供する機械学習のマネージドサービスは、各クラウドベンダーがオリジナルの分析支援ツール等を提供するケースもあるため、オープンソースで提供されているとはいえません。また、高負荷な処理が可能なマネージドサービス等では、使用した分だけ料金が発生する場合があります。よって、cが不適切だとわかります。	163

問題番号	解答	解説	該当ページ
Q24	d	0の入力というエラー系のテストであり、プログラムの内部構造を意識しない検証のためブラックボックスとなり、dが正解となります。	201
Q25	c	SELECT以降の記述から、都道府県と平均年収が表示されることがわかるため、bかcに絞られます。また、ORDER BY以降の記述で、DESCと記載があるため、結果を平均年数の降順で表示することを命令しています。よって、cが適切であるとわかります。	207
Q26	a	分析プロジェクトを成功させるためには、目的の明確化と、現状把握が重要です。プロジェクトがどのような状況に置かれており、何が課題なのかを明確にすることで、その後のアプローチが外れにくくなります。また、実際の分析に移る際も、受領したデータの確認やデータ取得過程の把握を怠ってしまうと誤った結論を導いてしまう可能性があります。よって、受け取ったデータを確認せず、やみくもに機械学習を実行しているaが不適切であるとわかります。	226
Q27	d	aは、近隣地域の学校行事やクリスマス等があるとお客さんが増えるという主張から導かれる仮説です。また、金曜夜や土日の昼にお客さんが多く、月曜日に少ないという主張からbとcの仮説を導くことができます。一方で、DMについては、これまでのヒアリングからは情報が得られていないため、dが誤りであるとわかります。	227
Q28	a	本題のケースのようなデータが一定期間欠損している場合は、依頼主への確認や分析対象からの除外をすることが必要です。aは、データの欠損を自身の解釈で埋めてしまうことになってしまい、その結果正しいデータを用いずに分析しているため、不適切であるとわかります。	230

問題番号	解答	解説	該当ページ
Q29	c	aは、近所で販売している/していない化粧品ブランド名は、知っている化粧品ブランド名に含まれるため、不適切です。bは、年代は生年月日に含まれるため、重複が起きています。dも、居住している地域が、居住している都道府県に含まれるため、重複が起きています。よって、cが適切であるとわかります。このようなMECEな設問設計は、回答データの質の向上につながるため、入念なチェックが必要です。	236
Q30	c	データやAI技術の発展による悪用や倫理的な問題としては、aやbに該当するAIによるフェイクの生成、dに該当する差別的なAIの構築などが挙げられます。cは、AIの悪用や倫理的な問題ではなく、精度が低いAIの利用に関する問題であるため、cが正解となります。	231
Q31	b	特化型AIとは、文字通り特定の分野・作業に特化したAIです。よって、bとcのいずれかになります。なお、強い/弱いAIという区分はAIが自意識を持つかどうかによって区分されるため、特化型/汎用型AIとは分類方法が根本的に異なることに注意が必要です。よって、bが適切であるとわかります。	247
Q32	b	むやみにデータを収集するのではなく、分析対象となるデータのあたりを付けて、効率的かつ効果的に行う必要があります。aやdは分析の仮説を元に検討しており、cは費用面での仮説を検討しています。bは依頼元に任せているだけで、分析担当として考えていないため、この中で適切でないものと考えられます。	254
Q33	a	データを分析した結果、仮説と異なる結果が得られた場合は、「データ処理やロジックにミスがないかを確認すること(d)」と「仮説と異なる結果を受け入れ、再度思考すること(b、c)」の2つが重要です。一方で、最初に立てた仮説を正しくするために、データを勝手に操作すること(a)は、分析者として不適切な対応であるといえます。	257

問題番号	解答	解説	該当ページ
Q34	a	データの可視化や分析の結果が得られたら、仮説通りの結果になっているか、またそうでない場合はどのような原因が考えられるかを解釈する必要があります。aの選択肢にある異常値や外れ値が見つかった場合、そのデータが集計ミスによるものか、適切な集計の結果得られている異常値や外れ値なのかを、データ取得者や関係者に確認する必要があります。分析をする都合だけで、異常値や外れ値を削除してしまうと、本来得られるはずであった貴重な情報を見逃してしまうこともあります。よって、aが不適切であるとわかります。	258
Q35	d	データ分析によって構築したサービスは、適切なタイミングで想定している改善が得られているかを確認する必要があります。サービス内容や環境の変化によってこれまでと異なるデータが発生し、モデルの精度が落ちることがあるため、モデルの再構築やメンテナンスが必要となります。よって、aとdが残りますが、定期的なメンテナンスを納品先に提案しているdのほうがより適切であるとわかります。	259
Q36	b	AIは人間のように高い精度で分析結果を示してくれますが、人間の行動を学習するようなAIは、時に人間のようなヒューマンエラーを犯すこともあります。よって、bが適切でないということがわかります。	267
Q37	a	ロボットの動作技術は、AI（人工知能）やIoTと密接に結びついており、これらの活用例の一つといえます。そのため、ロボットがこれらの技術に取って代わるという主張は誤りであり、aが不適切であるとわかります。	267

問題番号	解答	解説	該当ページ
Q38	a	CPSによるデータ駆動型社会では、センサーなどのデータ取得装置が収集したデータをコンピューターで解析し、定量的な視点を社会の変化に活かしていくことが求められます。そのため、IoTによるモノのデジタル化・ネットワーク化によって、様々なものの情報を取得・伝達することが可能になることが非常に重要です。bはAIを搭載したロボット、cはDX（デジタルトランスフォーメーション）に近い説明、dはIoT（Internet of Things）に特化した説明であり、いずれも不十分です。よって、aが適切であるとわかります。	267
Q39	b	オープンデータの普及によって、誰もが使うことのできる情報の量が増えることで、それらを利用した新しいサービスにつながる期待ができます。一方で、企業や個人の機密情報に当たるデータはオープンデータに不適切です。選択肢の中で、各支店の売上データには顧客名などの機密情報が含まれている可能性があるため、公開する前に該当情報の削除を行う必要があります。よって、bが不適切であるとわかります。	267

問題番号	解答	解説	該当ページ
Q40	b	今回のデータは、最小値4、最大値58、データの個数が30となっており、度数を集計するための区間の大きさである階級幅を1にすると、4～58まで55個の階級が存在してしまい、データの個数に比べて多すぎます。また、逆に階級幅を60にすると、階級は1個しか存在せず、データの分布を示せません。よってaとdは不適切です。次に階級幅を30とすると、15個ずつの階級が2個存在することを示せますが、このデータの分布を適切に示せているわけではありません。よって、bの10で区切ると、データの分布を把握することができ、適切であるとわかります。 また、与えられたデータからヒストグラムの階級数の目安を求める公式として、スタージェスの公式が挙げられます。階級数をk、サンプルサイズをNとすると、スタージェスの公式は$k=\log_2 N+1$と表せます。今回のサンプルサイズN＝30を代入すると、k＝5.9となり、およそ6個の階級が適切であるとわかります。データの最小値は4、最大値は58のため、階級幅の目安はおよそ(58－4)/6＝9と求められ、これに最も近いbが適切であると判断することもできます。ただし公式に入れたものはあくまで目安であり、必ず元データを見て自身で検討することが必要です。	272
Q41	a	相関関係とは、2つの物事の間で一方が変化すれば他方も変化するような関係をいいます。また、因果関係とは2つ以上の物事が原因と結果の関係にあることをいいます。ここでは、犯罪の発生と貧困の発生にどのような関係があるかを考えます。一般的には、「貧困率が高い地域ほど生活のために盗難などの犯罪が起きやすいのではないか？」と考えることができ、貧困率が高いほど犯罪発生率が高く、これらには原因と結果の関係性があると仮説が立てられます。よって、相関関係も因果関係もあると考えることができ、aが適切であるとわかります。	272

問題番号	解答	解説	該当ページ
Q42	c	観察単位をグループにまとめ、そのグループの全世帯を対象とした調査を行う方法は、調査対象を地区（集落）ごとに分類することから、集落抽出法と呼ばれます。よって、cが適切であるとわかります。aの単純無作為抽出法は、母集団からランダムにサンプルを抽出する方法です。bの層化無作為抽出法は、母集団をいくつかのグループに分け、各層の中からランダムにサンプルを抽出する方法で、全世帯を対象としたものではありません。dの多段抽出法は、母集団からのサンプル抽出段階が複数あり、それぞれの層の中からランダムにサンプルを抽出する方法です。a、b、dは、無作為による偶然性を利用することで、母集団の特性に似た標本を選ぶことを期待できます。一方で今回のように、出現頻度が低い事象についても把握したい場合は、集落抽出法が用いられます。	272
Q43	c	多変量のデータの可視化の際には、変数の数によって適切な可視化方法を検討する必要があります。特に変数が4つ以上の場合には、平行線を活用した平行座標プロットや、複数の散布図を整列して表示する散布図行列などが用いられます。cのヒートマップは、2次元データの大きさなどの情報を色や濃淡で表現する手法であり、4変数のデータを適切に可視化することができません。よって、cが不適切であるとわかります。	272
Q44	c	ECサイトに訪れるお客様のアクセスログを活用する際には、データ活用に対する同意の取得、個人情報の加工、漏洩した場合の対応の設定など、あらゆる準備が求められます。cは、削除要請があった場合は該当顧客のデータを全期間削除する必要があるため、削除要請前のデータであっても利用することは誤りであり、cが不適切であるとわかります。	276

問題番号	解答	解説	該当ページ
Q45	d	予測的データ分析を用いることで、過去のデータを利用して定期的に起きる事象の予測や検知を行うことが可能です。なお、予測対象となる事象は、その要因がある程度明らかになっているものである必要があります。aのように1か月先の未来の事象を正確に予測したり、bのように突発的な事象を予測したりすることは難しいとされています。また、予測的データ分析は予測した事象を未然に防ぐことが目的のため、cのように予測を観察するのにとどめ利用しないのも不適切です。よって、dが適切であるとわかります。	267