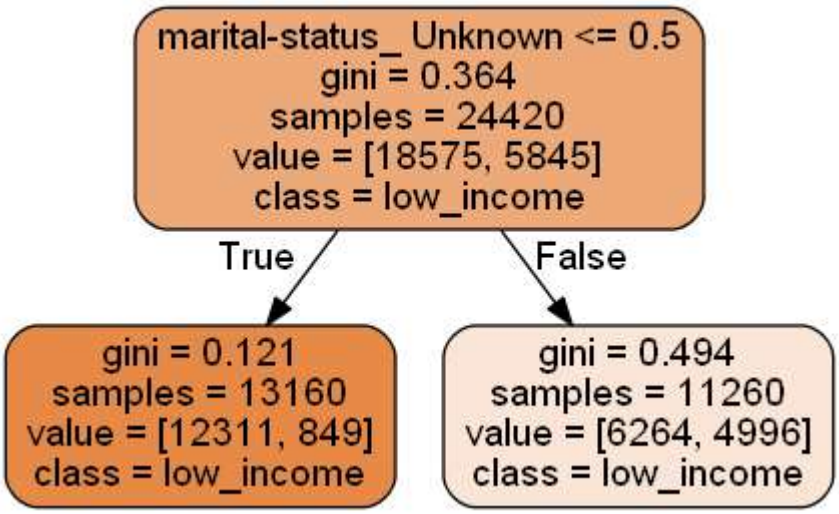
TCT-기술인증테스트
데이터분석
[2021년 #차]

사번 : _____

성명 : _____

1. 다음은 어느 국가의 인구 데이터를 토대로 저소득층과 고소득층을 Decision Tree로 구분한 결과의 일부이다. 다음 문항별로 True / False를 판단하시오.

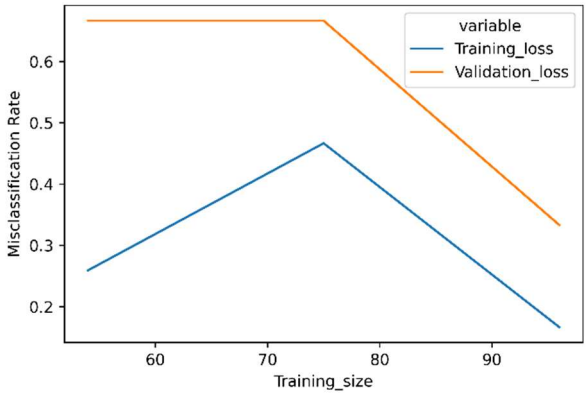


1) Decision Tree에서는 Gini 값이 커질수록 잘 분류된 것이라 할 수 있다. (True / False)
2) Marital-status_Unknown이라는 변수 값이 0.5보다 큰 사람의 49.4%가 고소득층(high_income)이다. (True / False)

(정답) F / F

(해설) Gini 값은 불순도에 대한 값으로 target class 가 아닌 class 에 속하는 규모의 비율이다.

2. 분석가 A가 모델링 과정에서 다음과 같은 Learning curve를 확인하였다. 이 때, 분석가 A가 취해야할 행동으로 적합한 것을 True로 그렇지 않은 것을 False로 표기하시오.



1) 기존 변수의 제곱합 등을 추가하여 Feature를 늘려본다. (True / False)
2) Training_size를 더 늘린다. (True / False)
3) Regularization항의 계수(λ)의 크기를 늘려본다. (True / False)

(정답) T / T / F

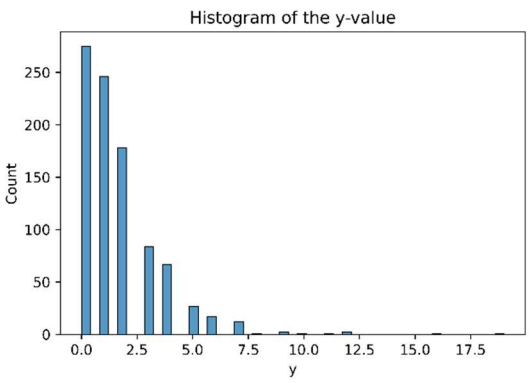
(해설) 현재 under-fitting 상황이므로, Underfitting 을 해결할 수 있는 행동을 하여야 한다.

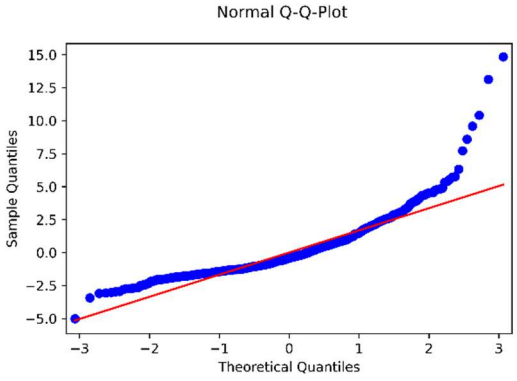
3. (기출활용) 모델의 정확도를 측정하는 MSE(Mean Square Error)는 모델예측 값의 분산(variance), 편차(bias), 그리고 노이즈라고 불리는 natural error로 구성된다. 이들에 대한 다음의 문장들의 참(True), 거짓(False)을 판단하시오.

1) Decision Tree에서 나무의 깊이를 깊게 설정할수록 Variance는 줄어든다. [True / False]
2) Tree 계열 알고리즘 중 RandomForest는 Bias를 줄이기보다는 Variance를 줄임으로써 예측 오차를 줄이고자 한다. [True / False]
3) Boosting 방식이 가진 장점은 Bias를 줄일 수 있다는 점이다. [True / False]

(정답) F / T / T

4. 분석가 A는 특정 매장에 방문하는 고객 수(y-value)를 예측하고자 한다. 선형 회귀(linear regression)를 통해 예측하고자 하였고, 이를 위해 방문 고객 수 분포를 확인하고, 다음과 같은 결과를 얻었다.





OLS Regression Results

Dep. Variable:	y	R-squared:	0.110
Model:	OLS	Adj. R-squared:	0.105
Method:	Least Squares	F-statistic:	22.55
Log-Likelihood:	-1844.1	Prob (F-statistic):	2.28e-21
No. Observations:	915	AIC:	3700.
Df Residuals:	909	BIC:	3729.
Df Model:	5		

다음의 각 문장에 대해 판단하거나, 빈칸을 채우시오.

1) 선형회귀 결과 F-test의 P-value가 매우 작기 때문에, 선형회귀를 통해 얻은 선형 수식을 통해 매장 방문 고객 수를 산출하는데 일단 문제는 없다. (True/False)

2) 방문 고객 수 예측이기 때문에 y 변수의 [_____]과/와 [_____]가 유사한지를 확인한 후, [__________] 알고리즘을 적용하는 것이 좀더 타당하다.

(정답) F, 평균, 분산, 포아송 회귀

(해설) y 값 분포가 skewed 되어 있고, Q-Q Plot 진단으로도 문제가 있으므로, 선형 회귀를 적용하지 않는 것이 타탕하다.

5. Tree 계열 알고리즘을 이용하여 모델링을 하는 다양한 기법들이 있다. 다음은 그 기법들 중 어느 특정 기법을 pseudo 코드로 표현한 것이다. 어느 머신러닝 기법인가?

```
Define X_train, y_train, X_test and y_test

Loop n Do
    Initialize a simple decision tree as T(n)
    Fit T(n) on X_train and y_train
    Make a prediction with X_train
    Refine y_train as differences between the predictions and the actual target value
End Loop

Loop n Do
    Make a prediction for T(n) with X_test
End Loop
Define the final prediction as summing the predict results
```

① Bagging
② Bootstrapping
③ Adaboost
④ Gradient Boosting
⑤ Adam

(정답) 4

6. 2019 년부터 2021 년 5 월까지의 재직자 및 퇴사자의 데이터를 이용하여 매월 직원의 익월 퇴사확률을 예측하는 모델을 DNN(Deep Neural Network) 알고리즘으로 개발하고자 한다. 이를 위해 분석가 A 는 2019 년과 2020 년 데이터를 랜덤하게 나누어 train 과 validation set 으로 사용하였다. 그리고 가장 최근인 2021 년 데이터는 test set 으로 활용하였다. 그리고 각 데이터 셋의 성능이 다음과 같이 얻어졌다. 다음 중 옳은 것을 고르시오.

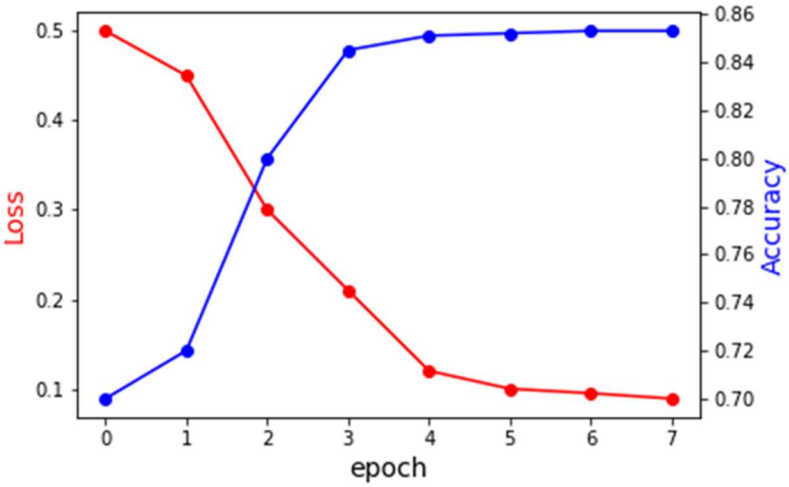
	train	valid	test
Accuracy	95%	93%	70%
Precision	70%	67%	40%
Recall	80%	80%	60%

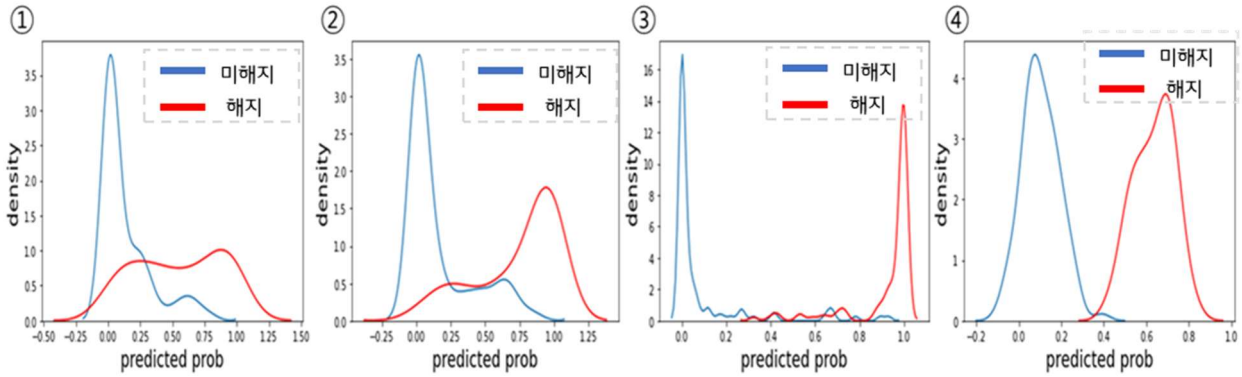
① 현재 모델이 overfitting 되었으므로 Drop out, regularization 등을 적용하여 DNN 알고리즘을 다시 학습해야 한다.
② 각 데이터 셋의 성능만으로는 현재 모델의 문제를 파악할 수 없다. 모델의 문제를 파악하기 위해서는 각 epoch별 train, valid의 성능 변화를 살펴볼 필요가 있다.
③ 현재 모델이 underfitting 되었으므로 테스트 셋의 일부를 학습 셋으로 포함시켜서 재학습을 해야한다.
④ 시간에 따라 데이터의 분포가 변화한 것으로 파악된다. 정확한 모델 성능 확인을 위해서는 시간에 따라 데이터를 학습/검증/테스트 셋으로 나누고 모델의 성능을 확인할 필요가 있다.
⑤ 현재 모델이 overfitting 되었으므로 DNN 모델의 layer 수를 증가한 구조의 DNN 알고리즘을 다시 학습해야 한다.

(정답) 4

(해설) data mismatch 현상이 발견됨.

7. A 통신사는 고객의 최근 6 개월간의 핸드폰 사용패턴을 이용하여 익월 고객의 해지여부를 예측하는 모델을 만들고자 한다. 이를 위하여 분석가 B 는 핸드폰 사용패턴을 설명변수, 익월 고객의 해지여부를 종속변수로 하여 DNN 모델을 만들었다. 모델 학습 과정에서 epoch 별 binary cross entropy 및 accuracy 는 주어진 그래프와 같다. 다음에 주어진 문장의 빈 칸을 채우시오. 단, undersampling 을 적용하여 학습 데이터 셋을 구성하여 학습 데이터에는 해지고객과 비해지고객의 비율이 1:1 이라고 가정한다. 이때, 예측값의 분포는 주어진 그래프 중 어떤 형태에 가까운지 선택하시오.





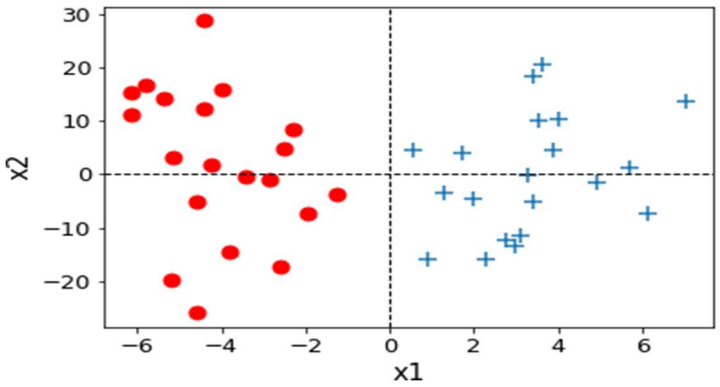
(정답)3

(해설) loss 는 줄어드나 accuracy 에는 변화가 없는 것으로 보아 예측 확률이 0 혹은 1 에 가까워지며 loss 만 줄어들고 있는 상황임을 알 수 있음.

8. 아래 그림과 같은 데이터를 regularized logistic regression 을 이용하여 분류하는 모델을 만들려고 한다. 즉, 다음의 penalized log-likelihood 를 최대로 만들어주는 w1 과 w2 를 구하고자 한다.

$$\sum_{i=1}^{n} \log P(y_i|X_i, w_1, w_2) - \frac{C}{2}(|w_1| + |w_2|)$$

$$where\ \hat{P}(y = 1|X, w_1, w_2) = \frac{1}{1 + \exp(-w_1 x_1 - w_2 x_2)}$$



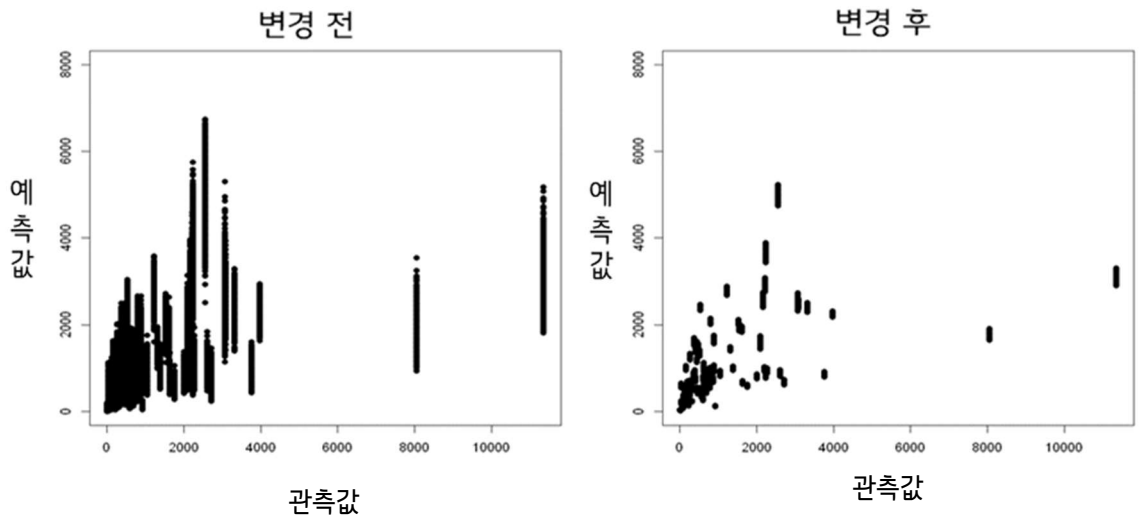
Regularization parameter인 C를 증가시킴에 따라 나타날 것으로 기대되는 현상으로 옳은 것을 고르시오.

① w1이 w2보다 더 빨리 0이 된다.
② w2가 w1보다 더 빨리 0이 된다.
③ w1과 w2가 동시에 0이 된다.
④ w1과 w2 모두 0이 되지 않고, C가 증가함에 따라 감소하기만 할 것이다.
⑤ 현재 주어진 그래프에 주어진 정보만으로 w1과 w2에 대해서 결론을 내리는 것은 불가능하다.

(정답) 2

(해설) X1 만을 이용하여 모델을 만들어도 training loss 를 0 으로 만들 수가 있다. 즉, w2 의 값이 0 이어도 완벽하게 분류가 가능한 상황이므로 w2 가 w1 보다 더 빨리 0 이 될 것이다.

9. 랜덤포레스트 모형으로 A 전자 제품의 판매수량을 예측하고자 한다. 분석가 A 는 일관성 있는 결과를 얻기 위해 모델링을 100 회 반복했고, 그 결과 한 개 실측값에 대해 100 개의 예측값을 얻었다. 실측값(X 축)과 예측값(Y 축)을 이용하여 아래 두 산점도를 출력하였다. 왼쪽은 default setting 으로 예측한 결과이고, 오른쪽은 모형의 hyper-parameter 를 조정하여 다시 예측한 결과이다. 어떤 parameter 를 어떻게 수정한 결과인가?

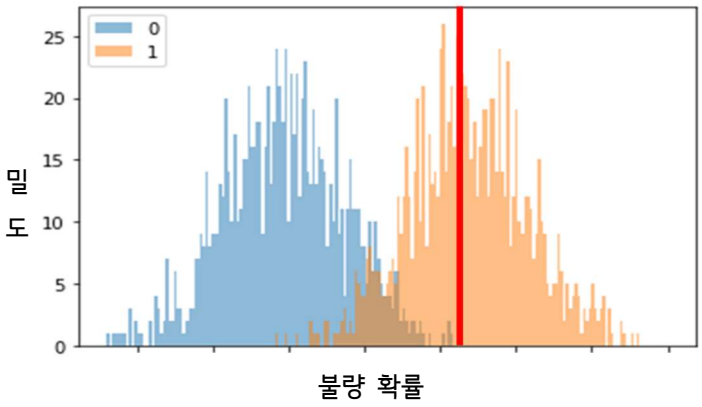


① 트리의 개수를 증가시킴
② 리프 노드의 최소 sample 사이즈를 감소시킴
③ Sample의 크기를 감소시킴
④ Terminal Node의 개수를 증가시킴
⑤ 선택할 X의 개수를 감소시킴

(정답) 1

(해설) tree 의 개수를 증가시키는 경우는 모델 결과값의 편차가 작아짐.

10. A 사는 제품의 불량을 검출하는 모델을 만들고자 한다. A 사는 공정 중간에 제품의 양불판정을 하고, 이 단계에서 불량 판정된 제품은 보정 공정(Reprocess)을 진행하여 다시 양품화 할 수 있다. 다음 그림은 불량여부 판정을 위한 분류모델 생성 후 훈련데이터에 대해 추정된 불량확률의 히스토그램을 출력한 결과이다. 0 은 정상제품 1 은 불량제품을 의미할 때, 분석가가 그림상의 빨간 선을 기준으로 큰 값은 불량, 작은 값은 양품으로 판정하려고 한다. (즉, 빨간선=threshold)
괄호에 주어진 단어 중 하나씩 선정하여 해당 결정에 대해 옳은 문장을 완성하시오.

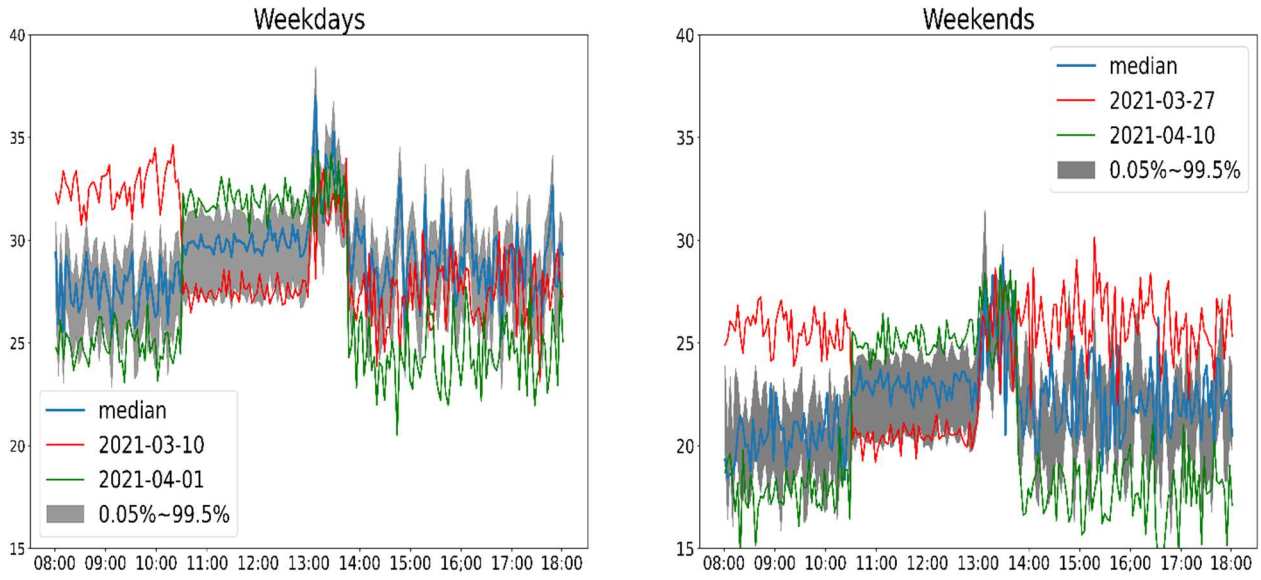


Precision이 (**높/낮**)고 recall이 (**높/낮**)고, (**중간판정/최종판정**) 공정에 적합한 결정임

(정답) 높, 낮, 중간판정

(해설) 확실한 불량에 대해서만 불량으로 판정하는 경우임. 중간검사 공정에 적용하여 보정작업을 진행할 수 있는 경우에 적합한 전략임.

11. 시스템 A 의 메모리 사용량을 모니터링하는 모델을 만들고자 한다. 이를 위하여 분석가 B 는 시스템이 정상적으로 운영된 6 개월간의 데이터를 수집하고 비지도학습 기반의 모니터링 시스템을 만들고자 한다. 이를 위하여 시간대별 패턴을 파악하기 위해 6 개월간 동시간대에 수집된 메모리 사용량의 중앙값, 99% confidence interval 를 계산 후, 다음과 같은 그래프를 얻었다. 다음 중 옳은 것을 고르시오. 단, 빨강선, 녹색선처럼 대표패턴에서 벗어나는 일별 패턴이 다수 존재하는 상황임.



① Outlier를 제외하기 위해 시점을 moving 하며 moving variance를 계산 후, variance가 증가하는 시점의 데이터를 제외한다.
② 주중과 주말의 각 시점별 99% confidence interval을 벗어나는 포인트가 다수인 날(day)을 제외한다.
③ 주중과 주말에 따른 차이가 명확하므로 주말여부만 모델에 반영하면 outlier를 제외할 필요가 없다.
④ 주중과 주말의 각 시점별 99% confidence interval을 벗어나는 포인트를 삭제한다.
⑤ Outlier를 제외하기 위해 시점을 moving 하며 moving average 및 moving variance를 계산 후, variance가 증가하면서 동시에 average의 변화폭이 큰 시점의 데이터를 제외한다.

(정답) 2

(해설) moving variance 를 계산하여 outlier 를 제외한다면 값의 변화가 큰 시점을 outlier 로 판단하게 됨. 이 경우에는 특정 시간에 값이 증가하는 경향이 존재하므로 해당 방법으로 outlier 를 제외할 수 없음. 또한, 주중과 주말의 차이가 명확하여 모델에 반영할 필요는 있으나 정상 데이터만을 모델에 반영해야하므로 여전히 outlier 는 제외할 필요가 있음. 특정 시점만을 제외하는 경우 모델에 데이터의 전체 패턴을 학습시키는 것이 불가능함.

12. 다음은 텍스트 분석 시에 이용할 수 있는 기법에 대한 설명이다. 옳은 것을 고르시오.
    ① Word2Vec과 GloVe는 학습에 사용하는 목적함수는 동일하다.
    ② Wor2Vec는 학습용 데이터에서 단어들이 등장하는 횟수를 반영한 embedding 기법이다.
    ③ LDA(Latent Dirichlet Allocation)에서 문서별 토픽의 분포는 dirichelet 분포를 따른다고 가정하며, 이때 dirichelet 분포 parameter α를 증가시킬수록 대부분의 문서가 적은 수의 토픽을 갖게 될 것이다.
    ④ LSA와 LDA는 토픽 모델링 및 문서 유사도 계산에 활용할 수 있는 알고리즘이다.
    ⑤ Doc2Vec 중 PV-DM과 PV-DBOW는 모두 단어가 input이며, 이후 나올 단어를 예측하여 문서에 대한 embedding vector를 얻는 기법이다.

(정답) 4

(해설) LDA 에서 parameter 를 감소시킬수록 대부분의 문서는 적은 수의 토픽을 갖게 될 것임.

13. 다음은 XAI 에 대한 설명이다. 각 문장의 True/False 를 판별하시오.
    ① Kernel SHAP에서 Base value는 전체 학습 데이터의 y의 예측값의 평균이므로 데이터의 로컬 특성을 Kernal SHAP으로 파악하는 것은 불가능하다. (True / False)
    ② 한 개체의 SHAP value를 모두 더한 값은 모델의 예측값과 동일하다. (True / False)
    ③ 각 변수의 모든 학습 셋에 있는 관측치에 대한 SHAP value의 평균값을 계산하면, 그 값이 변수의 중요도를 나타낸다. (True / False)

(정답) False/False/False

(해설) Base value 는 background dataset 의 예측값의 평균이므로 background dataset 을 조정하여 local 특성을 파악하는 것이 가능함. SHAP value 와 base value 의 합이 모델의 예측값이다. SHAP VALUE 의 절댓값의 평균값이 각 변수의 중요도임.

14. A 회사에서는 자신의 홈페이지 배너 광고를 최적화하기 위해 위해 가장 클릭률이 높은 배너를 찾기 위한 A/B Test 를 하고 있다. A/B Test 를 위해 2 개의 배너 광고를 준비하고 각 배너가 노출된 횟수와 클릭된 횟수를 측정하였다. A 회사 분석팀에서는 위 과정을 통해 얻은 노출횟수와 클릭횟수를 통해 어느 배너가 더 효과적인지를 확률분포를 가정하여 기댓값을 산출하고자 한다. 다음 확률분포 중 어느 분포로 가정하는 것이 타당한가?

    ① 베르누이 분포
    ② 정규 분포
    ③ 감마 분포
    ④ 베타 분포
    ⑤ 지수 분포

(정답) 4

(해설) 2 개의 모수를 가지고, 0~1 사이의 확률변수를 가지는 분포는 베타 분포이므로, 베타 분포를 가정하는 것이 타당하다.

15. [교차출제-AI] 다음 [표]에는 기계 학습(Machine Learning) 모델 학습 시 발생한 문제 상황과 문제를 해결하기 위한 해결책이 나열되어 있다. 과대적합(Overfitting)과 과소적합(Underfitting) 각각에 해당하는 문제 상황과 알맞은 해결책을 [표]에서 찾아 작성하시오.

[표]

문제 상황	(가) 빅데이터를 이용하여 개인의 신용을 평가하는 기계학습 모델을 구상하고 주어진 데이터를 학습한 결과 인종, 나이, 학력, 소득수준, 근무경력 등의 정보를 모두 활용하지 않고 근무경력이 많기만 하면 신용도를 높게 평가하는 모델이 생성되었다. (나) 차종이 경형, 소형, 중형, 대형, 스포츠카의 5종으로 나누어져 있는 차량 데이터셋을 이용해 차종을 분류하려고 한다. 심층 신경망(Deep Neural Network) 모델을 활용하여 학습한 결과 학습 데이터셋 100,000건에 대해 분류정확도(accuracy)가 20% 정도로 나왔다. (다) 일주일 간의 기상 이미지 데이터를 활용하여 다음날의 기상 상태를 예측하기 위해 깊고 복잡한 구조의 심층 신경망(Deep Neural Network) 모델을 만들었더니 노이즈 데이터(noise data)도 학습한 모델이 생성되었다.
해결책	(A) 모델의 복잡도를 증가시킨다. (B) 드롭아웃(dropout)을 적용한다. (C) 평가셋(validation set)에서의 성능 개선이 보이지 않으면 학습을 중단한다.

	과대적합	**과소적합**
문제 상황	1)	2)
해결책	3)	4)

1) _____
2) _____
3) _____
4) _____

(정답) 1) 다
2) 가, 나
3) B, C
4) A

(해설) 과대적합(Overfitting) 및 과소적합(Underfitting)과 관련된 문제이다. '다'는 주어진 데이터만을 잘 학습하고 현장의 데이터에는 추론결과가 좋지 않은 과대적합이 발생한 사례이다. '가, 나'는 주어진 학습 데이터의 패턴을 파악하지 못하여 학습 데이터셋에서의 성능도 적절히 수렴하지 못한 과소적합 사례이다. 과대적합은 regularization 기법을 적용해야 하고 과소적합은 모델의 capacity 를 늘려야 한다.

16. [교차출제-AI] 과대적합(Overfitting)을 방지하기 위해 모델의 복잡도(complexity)를 낮추고 일반화(generalization) 특성을 높이는 기법을 정규화(regularization)라고 한다. 다음 중 정규화를 위한 방법으로 올바르지 않은 것을 고르시오.

① 기계 학습(Machine Learning) 모델에 직접 파라미터(parameter) 값의 범위를 제한하는 추가 조건(constraints)를 부여한다.
② 기계 학습 모델의 목적 함수(objective function)에 결과적으로 파라미터(parameter) 값의 범위가 제한되도록 동작하는 항(term)을 추가한다.
③ 학습 데이터(training data)를 설명할 수 있는 여러 가정(multiple hypotheses)을 조합(combine)한다.
④ 학습 데이터에 포함되어 있는 잡음(noise data)를 제거한다.
⑤ 학습 데이터 분포에 대한 사전 지식(prior knowledge)을 활용하여 추가 학습 데이터를 생성한다.

(정답) 4

(해설) 학습 데이터에 포함된 잡음(noise data) 또한 regularization 효과를 낼 수 있다.

17. [교차출제-DS] A사에서 고객 데이터를 분석하여 업무에 활용하고자 한다. 시스템에 저장된 테이블 및 데이터는 다음과 같이 구성되어 있고, EDA를 진행하기 위한 데이터 추출 조건이 아래와 같이 주어져 있다. 각 추출 조건에 따른 SQL 매핑시, 잘못된 데이터를 발생시키는 SQL을 고르시오(추출 조건에 없는 기준은 고려하지 않으며, 빈칸은 NULL을 의미함)

TABLE_A

ID	성별	나이	이름
A	남	30	홍길동
B	여	53	안동근
C	남	34	김아랑
D	남	29	이민혁
E		78	최근무

TABLE_B

ID	지역	연봉
D	서울	
E	부산	5400
A		3800
F	대전	2900
G	구미	6900

TABLE_C

ID	방문일	구매수
E	20210701	3
B	20210713	4
G	20210402	6
F	20210704	1
B	20210503	3

[추출 조건]

① 성별(남/여) 평균 연봉. 성별이 없는 경우 없는 경우 "제외"로 처리
② "ID, 나이"별 구매 수 합
③ 나이별 평균 연봉. 연봉이 없는 경우 0
④ TABLE_A 기준, "ID, 성별, 나이" 별 연봉 합. 연봉이 없는 경우 0
⑤ "이름, ID"별 최초 방문일, 이름이 없는 경우 '-'로 표시.

① SELECT A.성별, AVERAGE(B.연봉) FROM TABLE_A A, TABLE_B B
WHERE A.ID = B.ID AND A.성별 IS NOT NULL GROUP BY A.성별

② SELECT A.ID, A.나이, SUM(C.구매수) FROM TABLE_A A, TABLE_C C
WHERE A.ID = C.ID GROUP BY A.ID, A.나이

③ SELECT A.나이, AVERAGE(NVL(B.연봉,0)) FROM TABLE_A A, TABLE_B B
WHERE A.ID = B.ID GROUP BY A.나이

④ SELECT A.ID, A.성별, A.나이, SUM(NVL(B.연봉,0)) FROM TABLE_A A, TABLE_B B
WHERE A.ID = B.ID GROUP BY A.ID, A.성별, A.나이

⑤ SELECT NVL(A.이름,'-'), NVL(A.ID, C.ID), MIN(C.방문일) FROM TABLE_A A FULL OUTER JOIN
TABLE_C C
ON A.ID = C.ID GROUP BY NVL(A.이름,'-'), NVL(A.ID, C.ID)

(정답) 4

(해설) 4 번의 경우 JOIN 으로 TABLE_B 에 없는 ID 는 제외되어, TABLE_A 의 모든 ID 를 추출하지 못함

수고하셨습니다.