(보안데이터분석) 연습문제_06

1. 의사결정 트리 모델에 대한 설명 중 **가장 정확한** 것은 무엇인가?
    1. 데이터를 여러 개의 규칙 기반 하위 집합으로 순차적으로 분할하여 분류 또는 예측을 수행하며, 각 분할은 **특정 기준**에 따라 이루어진다. ❤
    2. 모델 학습 과정에서 정보 이득(Information Gain) 또는 지니 불순도(Gini Impurity) 감소와 같은 불확실성 증가 지표를 사용하여 최적의 분할 기준을 찾는다.
       정보 이득이나 지니 불순도 감소는 불확실성을 감소시키는 방향 규칙
    3. 최종 예측은 리프 노드에 도달한 데이터 샘플들의 다수 클래스 또는 평균값으로 결정되며, 모델의 복잡성을 줄이기 위해 모든 분할을 최대한 깊게 수행한다.
       분항의 깊이가 깊어 질수록 모델의 복잡성이 증가한다
    4. 모델의 로직은 '만약(If) - 그렇다면(Then)' 형태의 규칙들의 집합으로 표현될 수 있으며, 이는 모델의 해석 가능성을 낮추는 주요 단점이다.
       해석가능성의 근거
    5. 과적합(Overfitting)을 방지하기 위해 트리의 깊이나 리프 노드의 최소 샘플 수와 같은 하이퍼파라미터 튜닝은 모델 성능에 큰 영향을 미치지 않는다. 일반화의 큰 영향
2. 의사결정 트리 모델에서 데이터를 분류하거나 예측하는 과정을 시각적으로 나타내는 구조의 주요 구성 요소에 대한 설명 중 **가장 부적절한** 것은 무엇가?
   3) **루트 노드 (Root Node):** 트리의 최상단에 위치하며, 전체 데이터셋을 나타낸다.
   4) **내부 노드 (Internal Node):** 특정 속성에 대한 테스트를 수행하고, 그 결과에 따라 데이터를 자식 노드로 분할한다.
   5) **리프 노드 (Leaf Node):** 더 이상 분할되지 않는 노드로서, 최종 예측 또는 분류 결과를 나타낸ㄷ.
   6) **브랜치 (Branch):** 내부 노드에서 발생하는 테스트의 결과를 나타내며, 자식 노드로 이어지는 경로이다.
   7) **순환 노드 (Cyclic Node):** 특정 조건을 만족하면 이전 노드로 돌아가 데이터 분할 과정을 반복하는 노드이다. ❤ 의사결정트리모델에는 존재하지 않는 노드
3. **분류 모델**로 사용되는 의사결정 트리 모델에서 노드를 분할하는 최적의 기준을 선택하는 데 사용되는 주요 지표가 **아닌** 것은 무엇가?
    1. 정보 이득 (Information Gain)
    2. 지니 불순도 (Gini Impurity)
    3. 엔트로피 (Entropy)

    4. 카이제곱 통계량 (Chi-squared Statistic) 범주형 변수 간의 독립성을 검정하는 데 사용, 범주형 속성의 분할 기준을 평가하는 데 활용할 수 있음
    5. 평균 제곱 오차 (Mean Squared Error) ❤ 의사결정트리 모델을 회귀분석에 사용할 때 사용
4. 정보 이론에서 사용되는 엔트로피에 대한 설명 중 **가장 정확한** 것은 무엇인가?
    1. 엔트로피는 특정 사건이 발생할 가능성의 정도를 나타내는 지표이며, 항상 0과 1 사이의 값을 가진다. 확률이 아니다. 1 이상의 값도 가능하다.
    2. 엔트로피 값이 높을수록 정보의 양이 적고, 예측하기 쉬운 상태를 의미한다. 반대설명
    3. 엔트로피는 데이터 집합 내의 순수도 또는 불확실성을 측정하는 척도로, 모든 클래스의 비율이 균등할수록 높은 값을 가진다.❤ 비균등 할 수록 정보 엔트로피(불순도)가 낮다.
    4. 이산 확률 변수 X의 엔트로피 H(X)는 다음과 같이 계산된다. (단, $p(x_i)$는 x_i가 발생할 확률이다.) $H(X) = \sum_{i=1}^{n} p(xi)log2(p(xi))$ 음의 부호가 없다. 확률의 역수에 대한 로그를 변환하기 때문에 음의 부호가 필요하다.
    5. 정보 이득은 부모 노드의 엔트로피와 자식 노드의 엔트로피의 합으로 계산되며, 이 값이 클수록 정보 획득량이 적음을 의미한다. 부모 엔트로피에서 자식엔트로피의 가중평균을 빼준 차이를 의미하면 차이가 클수록 정보 획득량이 많음을 의미한다.
5. 크로스 엔트로피(Cross-Entropy)에 대한 설명 중 **가장 정확한** 것은 무엇인까?
    1. 크로스 엔트로피는 두 확률 분포가 얼마나 유사한지를 측정하는 지표로, 값이 작을수록 두 분포는 서로 다르다고 해석한다. 차이값, loss 함수로도 사용 따라서 클수로 다름
    2. 크로스 엔트로피는 항상 자기 정보량(Self-Information)보다 작거나 같은 값을 가지며, 이는 정보 이론의 기본적인 속성 중 하나이다. 항상 자기정보량의 평균값 엔트로피보다 크거나 같은 값을 가짐: KL-divergence를 유도할 때 숨어 있던 H(p) $H(p,q) \geq H(p)$
    3. 크로스 엔트로피는 실제 확률 분포 P에 대해, 이를 추정하는 확률 분포 Q를 사용하여 평균 비트 수를 추정하는 척도이다. ❤
    4. 크로스 엔트로피는 분류 문제에서 모델의 성능을 평가하는 데 주로 사용되며, 예측 확률과 실제 레이블 간의 상관관계를 직접적으로 나타낸다. 모델의 예측확률 분포와 실제 레이블의 원-핫 인코딩된 분포 사이의 차이를 측정하는 손실함수로 자주 사용, 직적 관계는 아님
    5. 크로스 엔트로피는 두 확률 분포 사이의 거리 또는 차이를 측정하는 대칭적인 지표로서, $H(p,q) = H(q,p)$의 성질을 가진다. KL-divergence처럼 비대칭적 지표이다. 즉, $H(p,q) \neq H(q,p)$
6. 쿨백-라이블러 발산(Kullback-Leibler Divergence, KL 발산)과 크로스 엔트로피(Cross-Entropy)의 관계에 대한 설명 중 **가장 정확한** 것은 무엇인가? (단, p는 실제 확률 분포, q는 추정된 확률 분포를 나타냄.)
    1. KL 발산은 p와 q 사이의 거리 또는 유사성을 직접적으로 측정하는 대칭적인 지표이며, 크로스 엔트로피와는 독립적인 개념이다. 비유사성 측정, 비대칭 지표

2. KL 발산은 p에 대한 q의 크로스 엔트로피 H(p,q)에서 q의 엔트로피 H(q)를 뺀 값으로 정의됩니다. H(p,q)에서 H(p)를 뺀 것이다.
   $D_{KL}(p \,||\, q) = H(p, q) - H(p)$
3. KL 발산은 항상 0보다 크거나 같은 값을 가지며, p와 q가 동일할 때 최대값을 가진다. 동일할 때 0 : 비유사성 측정
4. 크로스 엔트로피를 최소화하는 것은 KL 발산을 최대화하는 것과 동일한 목표를 가지며, 이는 머신러닝 모델 학습의 핵심 원리 중 하나이다. KL-divergence을 최소화하는 것과 동일
5. KL 발산은 q에 대한 p의 크로스 엔트로피 H(q,p)에서 p의 엔트로피 H(p)를 뺀 값으로 정의된다. ❤

7. 로지스틱 회귀분석 모델의 손실 함수로 널리 사용되는 크로스 엔트로피(Cross-Entropy) 손실에 대한 설명 중 **가장 정확한** 것은 무엇인가? (단, y는 실제 레이블(0 또는 1), $\hat{y}$는 모델의 예측 확률(0과 1 사이의 값)을 나타냄)
   1) 크로스 엔트로피 손실은 실제 레이블과 예측 확률 사이의 유클리드 거리(Euclidean Distance)를 최소화하는 것을 목표로 한다. 유클리드 차이는 회귀분석, 분류는 크로스 엔트로피 차이
   2) 크로스 엔트로피 손실 함수는 볼록(convex) 함수가 아니므로, 경사 하강법(Gradient Descent)과 같은 최적화 알고리즘을 적용했을 때 지역 최적해(local optima)에 빠질 위험이 있다. 볼록(convex)하기 때문에 전역 최적해(Global optima) 가능성이 있다.
   3) 실제 레이블이 1일 경우, 예측 확률 $\hat{y}$가 1에 가까워질수록 크로스 엔트로피 손실 값은 커지고, 0에 가까워질수록 작아진다. 반대이다.
   4) 크로스 엔트로피 손실은 모델의 예측 확률 분포와 실제 레이블의 확률 분포 사이의 차이를 측정하며, 이 차이가 작을수록 손실 값은 작아진다. ❤
   5) 크로스 엔트로피 손실은 다중 클래스 분류 문제에만 적용 가능하며, 이진 분류 문제에서는 다른 형태의 손실 함수를 사용해야 한다. 모두 사용할 수 있다. 다중 클래스 분류에는 범주형 크로스 엔트로피(Categorical Cross-Entropy)의 형태

$$BCE \;=\; -\left[y\, log(\hat{y}) \;+\; (1-y)\, log(1-\hat{y})\right]$$

$$CCE \;=\; -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j\, log(\hat{y_j}) \;+\; (1-y_j)\, log(1-\hat{y_j})$$

8. 다음 중 앙상블(Ensemble) 기법에 대한 설명으로 **가장 적절한 것**은 무엇인가?
   1. 앙상블 기법은 하나의 복잡한 모델을 사용하는 방식으로, 예측 정확도를 높이기 위해 단일 모델의 복잡도를 증가시키는 것이 핵심이다.
   2. 배깅(Bagging)은 약한 학습기들을 순차적으로 학습시키며 오차를 줄여나가는 방식으로, 이전 모델의 오차를 보완하는 데 초점을 맞춘다.
   3. 부스팅(Boosting)은 다수의 강한 학습기를 병렬로 결합하여 예측 성능을 높이는 기법이다.
   4. 대표적 배깅(Bagging)을 활용한 랜덤 포레스트(Random Forest)는 여러 개의 결정 트리를 생성하되, 각 트리는 데이터와 특성의 일부를 무작위로 선택하여 학습한다. ❤

    5. 앙상블 기법은 항상 과적합을 방지하므로, 하이퍼파라미터 튜닝 없이도 안정적인 성능을 보장한다.

9. 다음 중 앙상블 기법의 보팅(Voting), 배깅(Bagging), 부스팅(Boosting), 스태킹(Stacking)과 관련된 설명으로 **가장 부적절한** 것은 무엇인가?
    1. 하드 보팅(Hard Voting)은 최종 예측 결과를 다수결 원칙에 따라 결정하며, 각 모델의 예측값들 중 가장 많이 선택된 값을 최종 예측으로 삼는다.
    2. 소프트 보팅(Soft Voting)은 각 모델이 예측한 클래스별 확률 값을 평균내어, 가장 높은 평균 확률을 갖는 클래스를 최종 예측으로 결정한다.
    3. 부트스트래핑(Bootstrapping)은 배깅(Bagging) 앙상블 기법에서 각 모델 학습에 사용될 데이터셋을 원본 데이터에서 중복을 허용하여 무작위로 추출하는 샘플링 방식이다.
    4. 스태킹(Stacking)은 여러 개의 기반 모델(Base Model)의 예측 결과를 최종 예측하는 또 다른 모델(Meta Model 또는 Aggregator)을 사용하여 앙상블 성능을 향상시키는 기법이다.
    5. 부스팅(Boosting) 기법 중 하나인 에이다부스트(AdaBoost)는 각 약한 학습기가 동일한 가중치를 가지며, 순차적인 학습 과정에서 이전 모델의 오분류된 데이터에 낮은 가중치를 부여하여 다음 모델 학습에 반영한다. ❤ 서로 다른 가중치를 가지며, 이전 모델에서 오분류된 데이터에 더 높은 가중치를 부여

10. 의사결정 트리 모델의 단점과 이를 보완하기 위한 가지치기(Pruning) 및 파라미터 조정(Parameter Tuning)에 대한 설명 중 **가장 부적절한** 것은 무엇인가?
    1. 의사결정 트리는 학습 데이터에 과적합(Overfitting)되기 쉬운 경향이 있으며, 이는 새로운 데이터에 대한 예측 성능을 저하시키는 주요 단점 중 하나이다.
    2. 가지치기(Pruning)는 의사결정 트리의 복잡성을 줄여 과적합을 방지하는 기법으로, 트리의 깊이를 제한하거나 불필요한 노드를 제거하는 방식으로 수행된다.
    3. 의사결정 트리의 파라미터 조정 시, 트리의 최대 깊이(max_depth)를 제한하는 것은 과적합을 방지하는 효과적인 방법 중 하나이지만, 과도하게 제한하면 모델이 충분히 학습되지 않아 성능이 저하될 수 있다.
    4. 의사결정 트리는 데이터의 작은 변화에도 최종 트리 구조가 크게 달라질 수 있어, 모델의 안정성(Stability)이 높고 해석력이 저하되는 단점이 있다. ❤ 최종 트리 구조가 변하는 것은 모델 안정성이 낮다는 특성이다. 트리구조라 직관적이어서 해석이 용이하다.
    5. 의사결정 트리의 분할 기준(예: 정보 이득, 지니 불순도)을 선택하는 것은 파라미터 조정의 한 과정이며, 이는 모델의 성능과 과적합 정도에 영향을 미칠 수 있다.

11. 랜덤 포레스트(Random Forest) 앙상블 모델의 학습 과정에서 이루어지는 특징 선택(Feature Selection), 서브샘플링(Subsampling), 그리고 최종 결과 집계(Result Aggregation)에 대한 설명 중 **가장 부적절한** 것은 무엇인가?

    1. 랜덤 포레스트는 각 개별 결정 트리를 학습할 때, 전체 특성 중 일부를 무작위로 선택하여 해당 서브셋 내에서 최적의 분할 특성을 찾는다. 이는 트리들 간의 다양성을 증진시키는 중요한 메커니즘이다.
    2. 랜덤 포레스트는 과적합을 줄이기 위해, 각 결정 트리를 학습할 때 원본 데이터셋 전체를 사용하는 대신 중복을 허용하는 무작위 추출 방식인 부트스트래핑(Bootstrapping)을 통해 생성된 서브샘플링된 데이터셋을 활용한다.
    3. 회귀(Regression) 문제에서 랜덤 포레스트의 최종 예측은 각 개별 결정 트리의 예측값들을 산술 평균(Arithmetic Mean)하여 얻어지며, 이는 예측의 안정성을 높이는 데 기여한다.
    4. 분류(Classification) 문제에서 랜덤 포레스트의 최종 예측은 일반적으로 각 개별 결정 트리의 예측 레이블 중 가중 평균(Weighted Average) 방식을 통해 결정되며, 각 트리의 정확도를 고려하여 가중치를 부여한다. ❤ 분류문제에 대해서는 하드보팅 방식이다.
    5. 랜덤 포레스트의 특징 무작위 선택 과정은 모델이 특정 소수의 강력한 특성에만 과도하게 의존하는 것을 방지하고, 다양한 특성들이 예측에 기여할 수 있도록 유도하여 일반화 성능을 향상시키는 데 목적이 있다.
12. 스태킹(Stacking) 앙상블 기법에 대한 설명 중 **가장 정확한** 것은 무엇인가?
    1. 스태킹은 여러 개의 약한 학습기들을 순차적으로 학습시켜 이전 모델의 오차를 보완하는 방식으로 최종 예측 성능을 향상시키는 기법이다.
    2. 스태킹은 여러 개의 기반 모델(Base Model)의 예측 결과를 단순 평균하거나 다수결 투표를 통해 최종 예측을 수행하는 앙상블 방법이다.
    3. 스태킹은 각 기반 모델의 예측 결과를 입력으로 사용하는 또 다른 모델(Meta Model 또는 Aggregator)을 학습하여 최종 예측을 수행하는 앙상블 기법이다. ❤
    4. 스태킹에서 각 기반 모델은 서로 독립적으로 학습되며, 메타 모델은 학습 과정에 전혀 관여하지 않고 최종 예측 단계에서만 활용된다.
    5. 스태킹은 과적합 위험을 완전히 제거할 수 있는 강력한 앙상블 기법이므로, 기반 모델의 성능이 낮더라도 최종 예측 성능을 항상 보장한다.
13. XGBoost(Extreme Gradient Boosting) 기법과 그 하이퍼파라미터 조정에 대한 설명 중 **가장 부적절한** 것은 무엇인가?
    1. XGBoost는 경사 부스팅(Gradient Boosting) 알고리즘을 기반으로 하지만, 규제(Regularization)를 강화하고 병렬 처리(Parallel Processing)를 지원하여 성능과 속도를 향상시킨 알고리즘이다.
    2. XGBoost의 주요 하이퍼파라미터 중 `n_estimators`는 생성할 약한 학습기(트리)의 개수를 결정하며, 이 값이 클수록 모델의 복잡도가 증가하여 과적합 위험이 높아질 수 있다.
    3. XGBoost의 하이퍼파라미터 조정 시, `learning_rate`는 각 트리가 이전 트리의 오차를 얼마나 강하게 보정할지를 결정하며, 일반적으로 작은 값을 사용할수록 학습 속도는 느려지지만 안정적인 성능을 얻을 수 있다.
    4. XGBoost는 트리의 분할 시, 손실 함수의 감소뿐만 아니라 트리의 복잡성을 고려하는 정규화 항을 목적 함수에 포함하여 과소적합

(Underfitting)을 방지하고 모델의 일반화 성능을 향상시킨다. ❤ 과적합(Overfitting)을 방지

5. XGBoost의 하이퍼파라미터 중 `max_depth`는 각 트리의 최대 깊이를 제한하는 파라미터이며, 이 값을 너무 크게 설정하면 모델이 학습 데이터에 과적합될 가능성이 높아진다.

14. LightGBM(Light Gradient Boosting Machine) 기법과 그 하이퍼파라미터 조정에 대한 설명 중 **가장 부적절한** 것은 무엇인가?
    1. LightGBM은 XGBoost와 마찬가지로 경사 부스팅 프레임워크 기반이지만, 리프 중심 트리 성장(Leaf-wise Tree Growth) 방식을 사용하여 트리의 균형을 맞추는 데 초점을 맞춘다. ❤
       XGBoost이나 일반트리는 깊이 중심 트리 성장(Depth-wise Tree Growth) 방식을 사용하고 이는 균형을 위함이다. 리프 중심 트리 성장(Leaf-wise Tree Growth) 방식은 손실 감소를 최대화(정확도에 중심)하는 방식이다.
    2. LightGBM은 대규모 데이터셋에서 더 빠른 학습 속도와 더 낮은 메모리 사용량을 달성하기 위해 Gradient-based One-Side Sampling (GOSS) 및 Exclusive Feature Bundling (EFB)과 같은 독자적인 기법을 사용한다.
    3. LightGBM의 주요 하이퍼파라미터 중 `num_leaves`는 각 트리의 최대 리프 노드 수를 결정하며, 이 값이 클수록 모델의 복잡도가 증가하여 과적합 위험이 높아질 수 있다.
    4. LightGBM의 하이퍼파라미터 조정 시, `learning_rate`는 각 부스팅 단계에서 모델을 얼마나 강하게 업데이트할지를 제어하며, 작은 값을 사용하면 학습이 더 안정적이지만 더 많은 부스팅 단계를 필요로 할 수 있다.
    5. LightGBM은 트리의 깊이를 제한하는 `max_depth` 파라미터를 제공하며, 이를 통해 모델의 복잡성을 제어하고 과적합을 방지할 수 있다.
       `num_leaves`와 `max_depth`는 함께 조정하여 최적의 성능을 찾아야 한다.

15. CatBoost 기법과 그 하이퍼파라미터 조정에 대한 설명 중 **가장 부적절한** 것은 무엇인가?
    1. CatBoost는 다른 경사 부스팅 알고리즘과 달리 범주형 특성(Categorical Features)을 별도의 전처리 없이 효과적으로 처리하기 위해 순서 통계(Ordered Target Statistics)와 같은 혁신적인 방법을 사용한다.
    2. CatBoost는 훈련 데이터의 순서에 민감하게 반응하여 예측 성능의 변동성이 크다는 단점을 가지며, 이를 완화하기 위한 별도의 파라미터 조정이 필요하다. ❤ 훈련 데이터 순서에 덜 민감하도록 설계. 다른 의사결정모델 계열보다 순서 통계를 사용 목표 변수 통계의 편향을 줄인다.
    3. CatBoost는 과적합을 방지하기 위해 내장된 강력한 정규화(Regularization) 기법을 제공하며, L1 및 L2 규제 외에도 트리 분할 시 발생하는 기울기 편향(Gradient Bias)을 줄이는 데 초점을 맞춘 규제를 포함한다.

4. CatBoost의 주요 하이퍼파라미터 중 `iterations`는 부스팅 라운드 수, 즉 생성할 트리의 개수를 결정하며, 이 값이 클수록 모델의 복잡성이 증가하고 학습 시간이 길어질 수 있다.
5. CatBoost의 하이퍼파라미터 조정 시, `learning_rate`는 각 트리의 영향력을 조절하며, 작은 값을 사용하면 학습이 더 신중하게 진행되어 과적합 위험을 줄일 수 있지만 수렴에 더 많은 시간이 소요될 수 있다.