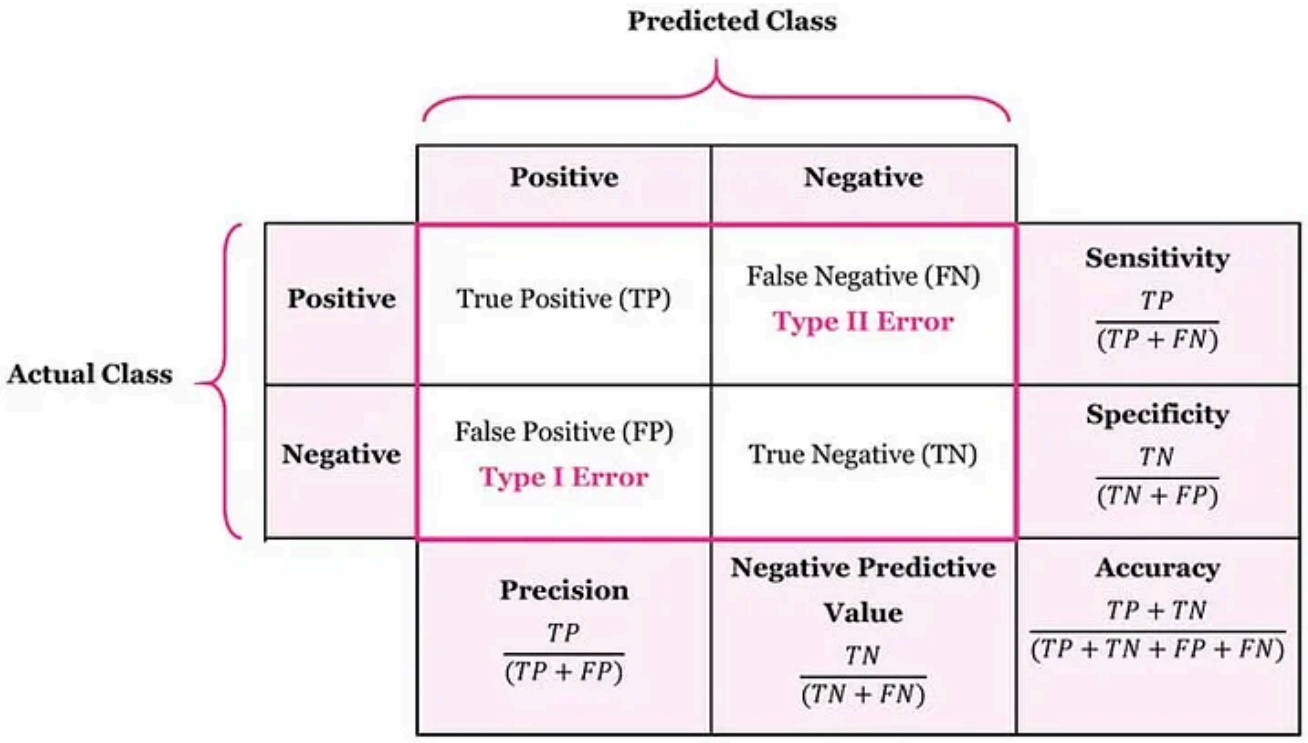
(보안데이터분석) 연습문제_05

1. 다음 중 로지스틱 회귀분석에 대한 설명으로 가장 **옳지 않은** 것은 무엇가?
    1. 종속 변수가 범주형(이분형) 데이터일 때 사용되는 회귀분석 방법이다.
    2. 독립 변수들의 선형 결합을 이용하여 종속 변수의 특정 범주에 속할 확률을 예측한다.
    3. 결과 값은 0과 1 사이의 확률 값으로 해석되며, Odds ratio를 통해 설명력을 파악할 수 있다.
    4. 회귀 계수의 부호는 독립 변수가 종속 변수의 확률에 미치는 영향의 방향을 나타낸다.
    5. 잔차항은 정규분포를 따르며, 등분산성을 만족해야 분석 결과의 신뢰도를 높일 수 있다. ❤ 로지스틱 회귀분석의 잔차항은 정규분포를 따르지 않으며, 등분산성을 가정하지 않다. 종속변수가 sigmoid 함수를 통과해 범주형이기 때문. 최대우도법(Maximum Likelihood Estimation)을 통해 모델을 추정하고, 모델의 적합도는 카이제곱 검정 등을 통해 평가
2. 로지스틱 회귀분석에서 사용되는 Odds, Logit, Log Odds의 관계를 가장 정확하게 설명한 것은 무엇가?
    1. Logit은 Odds의 역수이며, Log Odds는 Logit에 자연로그를 취한 값이다.
    2. Odds는 성공 확률에 대한 실패 확률의 비율이며, Logit은 Odds에 자연로그를 취한 값이고, Log Odds는 Logit의 제곱근이다.
    3. Odds는 성공 확률을 실패 확률로 나눈 값이며, Log Odds는 Logit 값 자체이고, Logit은 Odds에 자연로그를 취한 값이다.
    4. Odds는 실패 확률에 대한 성공 확률의 비율이며, Logit은 Odds에 자연로그를 취한 값이고, Log Odds는 Logit과 동일한 의미로 사용된다. ❤
    5. Odds는 성공 확률과 실패 확률의 합이며, Logit은 Odds의 지수 함수 값이고, Log Odds는 Logit의 역수이다.
3. 서포트 벡터 머신(SVM) 분류 모델의 핵심 개념 중 초평면(Hyperplane), 마진(Margin), 서포트 벡터(Support Vector)에 대한 설명으로 가장 **옳지 않은** 것은 무엇인가?
    1. 초평면은 데이터를 두 개 이상의 클래스로 나누는 결정 경계 역할을 하며, 선형 SVM에서는 직선, 비선형 SVM에서는 곡면이 될 수 있다.
    2. 마진은 초평면과 가장 가까운 각 클래스의 데이터 포인트(서포트 벡터) 사이의 거리로 정의되며, 넓은 마진을 갖는 모델이 일반화 성능이 더 좋다고 여겨진다.

    3. 서포트 벡터는 초평면의 위치와 마진의 크기를 결정하는 데 직접적인 영향을 미치는, 각 클래스에서 초평면에 가장 가까운 데이터 포인트들이다.
    4. SVM은 마진을 최대화하는 최적의 초평면을 찾는 것을 목표로 하며, 이를 통해 훈련 데이터에 과적합되는 것을 방지하고 새로운 데이터에 대한 예측 성능을 향상시킨다.
    5. 서포트 벡터는 결정 경계로부터 멀리 떨어진 데이터 포인트이며, 이들은 초평면의 위치나 마진의 크기에 영향을 미치지 않는다. ❤
4. 서포트 벡터 머신(SVM)에서 사용되는 커널 트릭(Kernel Trick)에 대한 설명으로 가장 **옳은** 것은 무엇인가?
    1. 커널 트릭은 비선형 결정 경계를 만들기 위해 원래의 저차원 데이터를 더 낮은 차원으로 매핑하는 기술이다.
    2. 커널 트릭은 데이터 포인트 간의 실제 내적 값을 명시적으로 계산하지 않고, 고차원 특징 공간에서의 내적 효과를 간접적으로 얻는 방법이다. ❤
    3. 커널 트릭은 선형적으로 분리 가능한 데이터에만 적용할 수 있으며, 비선형 데이터에는 사용할 수 없다.
    4. 커널 트릭은 SVM 모델의 복잡도를 줄이고 훈련 시간을 단축시키는 것을 주요 목표로 한다.
    5. 커널 트릭은 이상치(outlier)에 민감하게 반응하여 모델의 일반화 성능을 저하시키는 경향이 있다.
5. 서포트 벡터 머신(SVM)의 다양한 커널(Linear, Polynomial, Sigmoid, RBF)과 관련된 하이퍼파라미터 및 특징에 대한 설명 중 가장 **옳지 않은** 것은 무엇인가?
    1. **Linear 커널**은 별도의 하이퍼파라미터가 없으며, 데이터가 선형적으로 분리 가능할 때 효율적인 성능을 보인다.
    2. **Polynomial 커널**은 `degree` 라는 하이퍼파라미터를 가지며, 이 값은 결정 경계의 유연성을 조절하고 높은 차원의 비선형성을 모델링할 수 있도록 한다. 다항식의 차수를 조절하여 다양한 형태의 비선형 결정 경계를 만들 수 있다. 높을 수록 과적합 위험이 커지다.
    3. **Sigmoid 커널**은 `gamma` 와 `coef0` 라는 하이퍼파라미터를 가지며, 신경망의 활성화 함수와 유사한 형태를 띠어 특정 조건에서 선형 커널과 유사한 결과를 낼 수도 있다.
    4. **RBF (Radial Basis Function) 커널**은 `gamma` 라는 하이퍼파라미터를 가지며, 이 값이 작을수록 데이터 포인트의 영향 범위가 좁아져 복잡한 결정 경계를 만들 가능성이 높아진다. ❤ `gamma` 값이 클수록 데이터 포인트의 영향 범위가 좁아져 각 데이터 포인트에 민감한 복잡한 결정 경계를 만들 가능성이 높아진다.
    5. **모든 커널**은 공통적으로 규제 파라미터 `C` 를 가지며, 이는 모델의 과적합을 방지하고 일반화 성능을 조절하는 중요한 역할을 한다. `C`는 훈련 데이터에 대한 오류를 얼마나 허용할 것인지를 결정 한다. 규제 파라미터 `C`는 모든 `kernel trick`이 가지고 있다.

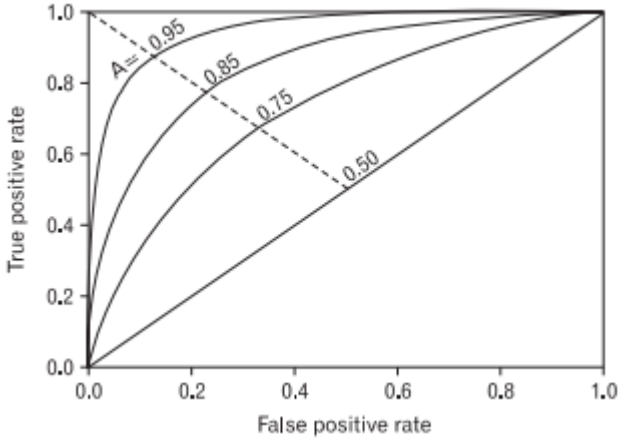
- 다음은 `Confusion Matrix` 이다.



6. Confusion Matrix에서 Precision은 다음 중 어떤 개념을 나타내는가?
    1. 모델이 실제 Negative 인 것을 Negative 로 정확히 분류한 비율
    2. 모델이 Positive로 예측한 것 중에서 실제 Positive인 것의 비율 ❤
    3. 모델이 Positive로 잘못 예측한 것의 비율
    4. 실제 Negative 인 것 중에서 모델이 Positive로 잘못 분류한 비율
    5. 모델이 실제 Positive 인 것을 Positive로 정확히 분류한 비율
7. Confusion Matrix에서 Recall은 다음 중 어떤 개념을 나타내는가?
    1. 모델이 실제 Negative 인 것을 Negative 로 정확히 분류한 비율
    2. 모델이 Positive로 예측한 것 중에서 실제 Positive인 것의 비율
    3. 모델이 Positive로 잘못 예측한 것의 비율
    4. 실제 Negative 인 것 중에서 모델이 Positive로 잘못 분류한 비율
    5. 실제 Positive 인 것 중에서 모델이 Positive로 정확히 분류한 비율 ❤
8. Confusion Matrix에서 F1 score는 다음 중 어떤 개념을 나타내는가?
    1. Precision과 Recall의 조화평균 ❤
    2. Precision과 Recall의 산술 평균
    3. Precision과 Recall의 가중 평균
    4. Precision과 Recall의 곱
    5. Precision과 Recall의 차
9. Confusion Matrix를 기반으로 할 때, 1종 오류(Type I Error)와 2종 오류(Type II Error)에 대한 설명으로 가장 **옳은** 것은 무엇가?
    1. 1종 오류는 실제 Positive인 것을 Negative로 잘못 예측했을 때 발생하며, 이는 False Negative (FN)에 해당합니다. 2종 오류는 실제 Negative인 것을 Positive로 잘못 예측했을 때 발생하며, 이는 False Positive (FP)에 해당한다.

2. 1종 오류는 모델이 Positive라고 예측했지만 실제로는 Negative인 경우로, False Positive (FP)에 해당합니다. 2종 오류는 모델이 Negative라고 예측했지만 실제로는 Positive인 경우로, False Negative (FN)에 해당한다. ❤
3. 1종 오류와 2종 오류는 모델이 정확하게 분류한 경우에 발생하며, 각각 True Positive (TP)와 True Negative (TN)에 해당한다.
4. 1종 오류는 모델의 정확도(Accuracy)를 낮추는 주요 원인이며, 2종 오류는 모델의 재현율(Recall)을 높이는 주요 원인이다.
5. 1종 오류와 2종 오류는 서로 독립적인 개념으로, 하나를 줄이면 다른 하나는 반드시 증가하는 관계는 아니다.

10. 분류 모델의 성능을 평가하는 지표인 ROC (Receiver Operating Characteristic) 곡선과 AUC (Area Under the ROC Curve)에 대한 설명으로 가장 **옳지 않은** 것은 무엇가?
    1. ROC 곡선은 분류 임계값의 변화에 따라 True Positive Rate (TPR)을 Y축에, False Positive Rate (FPR)을 X축에 나타낸 그래프이다.
    2. AUC는 ROC 곡선 아래의 면적을 의미하며, 그 값이 1에 가까울수록 모델이 Positive 클래스를 Negative 클래스보다 훨씬 잘 분류한다고 해석할 수 있다.
    3. 무작위로 예측하는 분류 모델의 ROC 곡선은 좌하단에서 우상단으로 이어지는 대각선에 가까우며, AUC 값은 약 0.5이다.
    4. 좋은 분류 모델은 ROC 곡선이 좌상단에 가깝게 위치하며, AUC 값은 0.5보다 훨씬 큰 값을 가진다.
    5. AUC 값이 0에 가까울수록 모델의 성능이 완벽한 분류와 유사하며, 모든 Positive 샘플은 Positive로, 모든 Negative 샘플은 Negative로 정확하게 예측한다. ❤



11. ROC (Receiver Operating Characteristic) 곡선과 AUC (Area Under the ROC Curve)는 분류 모델의 성능을 평가하는 데 유용하게 사용된다. 이와 관련하여 1종 오류(Type I Error) 및 2종 오류(Type II Error)와 어떤 관계가 있는지 가장 **옳지 않은** 것은 무엇가?

    1. ROC 곡선의 X축인 False Positive Rate (FPR)은 실제 Negative를 Positive로 잘못 예측할 확률로, 이는 1종 오류와 직접적인 관련이 있다.
    2. ROC 곡선의 Y축인 True Positive Rate (TPR 또는 Recall)은 실제 Positive를 Positive로 정확하게 예측할 확률로, 이는 2종 오류를 낮추는 데 기여한다.
    3. 특정 임계값에서 1종 오류를 줄이려고 하면 일반적으로 FPR이 감소하며, ROC 곡선 상에서 좌측으로 이동하는 경향이 있다. 이때 2종 오류를 나타내는 False Negative Rate (FNR = 1 - TPR)은 증가할 수 있다.
    4. AUC는 ROC 곡선 아래의 면적으로, 모델의 전반적인 성능을 나타내며, AUC 값이 높을수록 1종 오류와 2종 오류를 모두 낮추는 방향으로 모델이 작동할 가능성이 높다.
    5. 완벽한 분류 모델은 AUC 값이 1이며, 이때 1종 오류와 2종 오류가 모두 최소화되어 0의 값을 갖는다. AUC 값이 0.5인 모델은 1종 오류와 2종 오류를 최소화하는 최적의 성능을 보인다. ❤
12. ETL (Extract, Transform, Load) 프로세스에서 각 단계의 역할에 대한 설명으로 올바르지 않은 것은 무엇인가?
    1. tract 단계는 데이터를 소스 시스템에서 추출하여 추출된 데이터를 원본 형식으로 가져온다.
    2. ansform 단계는 추출된 데이터를 비즈니스 규칙 및 요구 사항에 따라 변환하고 정제하여 목적지 시스템으로 올바른 형식으로 데이터를 준비한다.
    3. ad 단계는 변환된 데이터를 목적지 시스템으로 적재하여 저장하고, 필요한 경우 데이터를 인덱싱하여 쉽게 검색할 수 있도록 한다.
    4. tract 단계에서는 데이터를 추출하는 데 필요한 쿼리 및 필터링 작업을 수행한다.
    5. ansform 단계에서는 주로 데이터 웨어하우스에 데이터를 적재하는 작업을 수행한다. ❤
13. EDA (Exploratory Data Analysis)에 대한 다음 설명 중 올바르지 않은 것은 무엇지 고르시오?
    1. DA는 데이터셋의 기본적인 통계적 특성을 요약하고 탐색하여 데이터의 패턴 및 구조를 파악하는 과정이다.
    2. DA는 데이터 분석 전에 반드시 수행되어야 하며, 데이터의 품질을 검증하고 잠재적인 문제점을 파악하는 데 도움이 된다.
    3. DA는 시각화 기법 등을 사용하여 데이터의 분포, 상관 관계, 이상치 등을 탐색한다.
    4. EDA는 데이터를 변환하고 모델링하는 데 사용되는 과정으로, 데이터셋의 최종 결과물을 생성한다. ❤
    5. DA를 통해 도출된 인사이트는 데이터를 해석하고 향후 분석 방향을 결정하는 데 사용될 수 있다.

14. MLOps (Machine Learning Operations)에 대한 다음 설명 중 올바르지 않은 것은 무엇지 고르시오?
    1. LOps는 머신러닝 모델의 개발, 배포, 운영 및 유지보수를 자동화하고 효율화하기 위한 개념이다.
    2. MLOps는 소프트웨어 개발의 DevOps 개념을 머신러닝 모델 개발 및 운영에 적용한 것을 말한다.
    3. MLOps는 모델을 개발한 후에만 관련될 뿐, 모델 개발 과정 자체에는 영향을 미치지 않는다. ❤
    4. MLOps는 모델의 생명 주기 전반에 걸쳐 통합된 프로세스를 제공하여 모델의 안정성, 확장성 및 성능을 향상시킨다.
    5. LOps는 CI/CD (Continuous Integration/Continuous Deployment) 및 모델 모니터링과 같은 개념을 포함하여 머신러닝 시스템의 자동화된 배포 및 관리를 지원한다.