8 차시	총 10 문제			연습: □	과제 : ☑	평가 : □

1. 이진 분류 문제에서 Precision 과 Recall 간의 상충 관계를 조정하는 데 사용되는 지표는?

① 정확도(Accuracy)

② F1 스코어

③ 손실 함수 값

④ ROC AUC

정답: 2 번

해설

- **Precision(정밀도)**와 **Recall(재현율)** 사이에는 일반적으로 **상충관계(trade-off)**가 존재합니다.
  - 정밀도를 높이려면 모델이 긍정 예측(Positive)하는 기준을 엄격하게(높은 임계값) 하여 잘못된 양성(위양성)을 줄이려 하지만, 이로 인해 실제 양성을 놓칠(Recall 감소) 가능성이 커집니다.
  - 반대로 재현율을 높이려면 모델이 더 쉽게 양성으로 예측(낮은 임계값)하여 놓치는 양성을 줄이지만, 오탐(Precision 감소)이 늘어날 수 있습니다.
- **F1 스코어**는 **정밀도(Precision)**와 **재현율(Recall)**을 **조화평균(Harmonic Mean)** 방식으로 결합한 지표로, 두 값 간의 균형을 평가하는 데 유용합니다.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2. 이진 분류에서 ROC 곡선의 AUC 값이 0.5 인 경우, 모델의 상태를 어떻게 평가할 수 있나요?

① 완벽한 예측 모델

② 랜덤 추측과 동일

③ 데이터 과적합

④ 완전한 오분류

정답: 2 번

해설

- **ROC(Receiver Operating Characteristic) 곡선**에서 **AUC(Area Under the Curve)** 값이 **0.5**라는 것은, 모델이 **임의 추측(random guessing)** 수준의 분류 성능을 보인다는 뜻입니다.
- **AUC가 1**에 가까울수록 분류 성능이 우수하며, **0.5**는 사실상 동전을 던지는 것과 동일한 예측 능력을 의미합니다.
- 따라서 AUC=0.5이면 **"랜덤 추측과 동일"**한 성능으로 평가할 수 있습니다.

3 이진 분류 문제에서 시그모이드 활성화 함수의 출력값이 무엇을 나타내나요?

① 클래스 확률

② 클래스 레이블

③ 손실 함수 값

④ 모델의 학습률

정답: 1번

해설

- **시그모이드 활성화 함수**는 입력값을 0과 1 사이의 값으로 매핑합니다.
- 이 값은 **해당 샘플이 양성(Positive) 클래스에 속할 확률**을 나타냅니다.
- 보통 이진 분류 문제에서, 시그모이드 함수의 출력이 0.5보다 크면 양성 클래스로 분류하고, 0.5보다 작으면 음성 클래스로 분류합니다.
- 따라서 시그모이드의 출력값은 **클래스 확률**로 해석됩니다.

4. 다음 중 이진 분류 문제에서 주로 사용되는 손실 함수는 무엇인가요?

① 교차 엔트로피

② 평균 제곱 오차

③ 로그 손실

④ 평균 절대 오차

정답: 1 번

해설

- 이진 분류 문제에서는 모델이 출력한 확률과 실제 레이블(0 또는 1) 사이의 차이를 평가하기 위해 **Binary Cross-Entropy Loss** (또는 log loss라고도 함)를 주로 사용합니다.
- **교차 엔트로피**는 모델이 예측한 확률 분포와 실제 분포 간의 차이를 측정하며, 로그 함수를 사용하여 오차를 계산합니다.
- 반면, **평균 제곱 오차**와 **평균 절대 오차**는 주로 회귀 문제에서 사용됩니다.
- 옵션 ③ '로그 손실'도 이진 분류에서 사용되는 손실 함수의 다른 이름이지만, 일반적으로 딥러닝 분야에서는 **교차 엔트로피**라는 용어를 더 많이 사용합니다.

5. 다음 중 클래스 불균형 문제를 해결하기 위한 방법으로 옳지 않은 것은?

① 오버샘플링

② 언더샘플링

③ 모델 복잡도 증가

④ 가중치 조정

정답: 3번

해설

- **오버샘플링**:
  소수 클래스의 데이터를 인위적으로 늘려서 클래스 비율을 맞추는 방법입니다.
- **언더샘플링**:
  다수 클래스의 데이터를 일부 제거하여 클래스 비율을 맞추는 방법입니다.
- **가중치 조정**:
  손실 함수에 클래스별 가중치를 부여하여, 소수 클래스에 더 큰 페널티를 주는 방식입니다.
- **모델 복잡도 증가**:
  모델의 복잡도를 증가시키는 것은 일반적으로 클래스 불균형 문제를 해결하는 방법이 아닙니다. 오히려 과적합(overfitting)을 유발할 수 있으며, 클래스 불균형에 따른 문제를 직접적으로 개선하지 않습니다.

따라서, 클래스 불균형 문제를 해결하기 위한 방법으로는 **모델 복잡도 증가**가 적절하지 않습니다.

6. 이진 분류 모델 평가에서 재현율(Recall)이 높은 경우 의미하는 것은?

① 실제 양성 데이터를 잘 예측함

② 실제 음성 데이터를 잘 예측함

③ 오차가 줄어듦

④ 모델의 속도가 빠름

정답: 1 번

해설

- **재현율(Recall)**은 모델이 실제 양성(Positive) 데이터를 얼마나 잘 찾아내는지를 나타내는 지표입니다.
- 수식으로 표현하면,

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

로 정의됩니다.

- **높은 재현율**은 실제 양성 데이터 중 많은 부분을 모델이 양성으로 올바르게 예측했다는 의미입니다.
- 따라서, 재현율이 높다는 것은 **실제 양성 데이터를 잘 예측한다**는 뜻입니다.

나머지 선택지들은 재현율의 정의와 맞지 않습니다.

7. 이진 분류 문제에서 데이터의 클래스 비율이 극단적으로 다른 경우 적합한 평가지표는?

① 정확도(Accuracy)

② F1-score

③ 정밀도

④ 재현율

정답: 2 번

해설

- **정확도(Accuracy)**는 클래스 불균형 문제에서 다수 클래스에 편향되어 높은 값이 나올 수 있어, 전체 성능을 제대로 평가하지 못할 수 있습니다.
- **정밀도(Precision)**와 **재현율(Recall)**는 각각 양성 예측의 정확성과 실제 양성을 얼마나 잘 검출하는지 평가하지만, 단독으로는 모델의 전반적인 성능을 나타내기 어렵습니다.
- **F1-score**는 정밀도와 재현율의 조화평균으로, 클래스 불균형 상황에서 소수 클래스의 예측 성능을 더 균형 있게 평가할 수 있습니다.

따라서, 클래스 비율이 극단적으로 다른 이진 분류 문제에서는 **F1-score**가 가장 적합한 평가지표입니다.

8. 다음은 이진 분류 딥러닝 학습을 위한 초기 하이퍼파라메터를 나타낸다. 빈 칸 [ ] 을 채우시오.

```
# 학습률
lr = 0.01

# 초기화
net = Net(n_input, n_output)

# 손실 함수
criterion = [                ]

# 최적화 함수: 경사 하강법
optimizer = optim.SGD(net.parameters(), lr=lr)

# 반복 횟수
num_epochs = 10000
```

정답:

```
# 학습률
lr = 0.01

# 초기화
net = Net(n_input, n_output)

# 손실 함수
criterion = nn.BCELoss() or nn.BCEWithLogitsLoss()

# 최적화 함수: 경사 하강법
optimizer = optim.SGD(net.parameters(), lr=lr)

# 반복 횟수
num_epochs = 10000
```

9. 다음은 이진 분류를 위한 비용함수이다. 빨간 색으로 표시된 부분이 가질 수 있는 최대값과 최소값을 쓰시오

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i log(h_\theta(x_i)) + (1-y_i)log(1-h_\theta(x_i))$$

$$where, h_\theta(x_i) = \frac{1}{1+e^{-\theta x}}, \quad y \in 0, 1$$

정답:

최댓값: 0

최솟값: -∞

- **최대값**: 0 ($h_\theta(x_i) = 1$ 일 때, 그리고 $y_i = 1$)
- **최소값**: $-\infty$ ($h_\theta(x_i) \to 0$ 일 때, 그리고 $y_i = 1$)

만약 $y_i = 0$이면, 그 항은 $0 \cdot \log(\dots)$이므로 **항상 0**입니다.

왜 그런가?

1. **이진 분류**에서 $y_i \in \{0, 1\}$
2. **로지스틱 회귀**에서 $h_\theta(x_i) = \sigma(\theta^T x_i)$은 $(0, 1)$ 범위의 확률로 해석
3. 빨간색 부분 $y_i \log\big(h_\theta(x_i)\big)$를 살펴보면,
   - $y_i = 1$ 인 경우 → 이 항은 $\log\big(h_\theta(x_i)\big)$
     - $h_\theta(x_i)$가 1에 가까우면 $\log(1) = 0$ → **최대값** 0
     - $h_\theta(x_i)$가 0에 가까우면 $\log(0) \to -\infty$ → **최소값** -∞
   - $y_i = 0$ 인 경우 → 이 항은 $0 \cdot \log\big(h_\theta(x_i)\big) = 0$

따라서, **$y_i \log\big(h_\theta(x_i)\big)$**가 가질 수 있는 값의 범위는 **$[-\infty, 0]$**이고, 실제로 최댓값은 0, 최솟값은 $-\infty$입니다.

10. 다음은 악성 종양 (Target = 1) 과 양성 종양 (Target = 0) 을 분류하는 이진 분류모델에서 얻어진 혼동행렬이다. 정확도(Accuracy), 재현율 (Recall), 정밀도 (Precision) 를 각 각 계산하시오. (소수점 3 자리 반올림)

		예측값	
		악성	양성
실제값	악성	5	4
	양성	6	2

정답:

정확도(Accuracy) ≈ 0.412

재현율(Recall) ≈ 0.556

정밀도(Precision) ≈ 0.455

주어진 혼동행렬(Confusion Matrix)은 다음과 같이 해석할 수 있습니다:

	예측: 악성(1)	예측: 양성(0)
실제 악성(1)	TP = 5	FN = 4
실제 양성(0)	FP = 6	TN = 2

- **TP**(True Positive): 실제 악성인데 악성으로 예측 (5개)
- **FN**(False Negative): 실제 악성인데 양성으로 예측 (4개)
- **FP**(False Positive): 실제 양성인데 악성으로 예측 (6개)
- **TN**(True Negative): 실제 양성인데 양성으로 예측 (2개)

이제 정확도(Accuracy), 재현율(Recall), 정밀도(Precision)를 계산해 봅시다.

---

1. 정확도 (Accuracy)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{5 + 2}{5 + 2 + 6 + 4} = \frac{7}{17} \approx 0.4118 \ (\text{약 } 0.412)$$

---

2. 재현율 (Recall)

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{5}{5 + 4} = \frac{5}{9} \approx 0.5556 \ (\text{약 } 0.556)$$

---

3. 정밀도 (Precision)

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{5}{5 + 6} = \frac{5}{11} \approx 0.4545 \ (\text{약 } 0.455)$$