December 10, 2025

▣ **강의명:** CS109A: 데이터 과학 입문

▣ **주차:** Lecture 14

▣ **교수명:** Pavlos Protopapas, Kevin Rader, Chris Gumb

▣ **목적:** Lecture 14의 핵심 개념 학습

□ 핵심 요약

본 문서는 로지스틱 회귀의 심화 주제를 다룹니다. 단순 모델을 넘어 다중 로지스틱 회귀, 상호작용항의 해석, 정규화(Ridge)를 통한 과적합 방지 방법을 배웁니다.
또한, 로지스틱 회귀를 분류(Classification) 문제에 활용하는 방법, 즉 결정 경계(Decision Boundary)의 개념과 다중 클래스(K>2) 분류를 위한 OvR, 다항 로지스틱 회귀(Softmax)를 학습합니다.
마지막으로, 분류 모델의 성능을 평가하는 핵심 지표인 혼동 행렬(Confusion Matrix), 민감도, 특이도, ROC 커브, AUC의 개념을 상세히 설명합니다.

Contents

1 로지스틱 회귀 복습 (Review)

1.1 왜 로지스틱 회귀인가?

선형 회귀(Linear Regression)는 예측 값이 연속적인 숫자(예: 집값, 온도)일 때 사용합니다. 하지만 우리가 예측하려는 대상(Y)이 '성공/실패', '합격/불합격', '생존/사망'처럼 두 가지 범주 중 하나라면(Binary) 선형 회귀는 적합하지 않습니다.

로지스틱 회귀(Logistic Regression)는 Y가 범주형일 때, 특히 이진(binary) 분류 문제에서 사용됩니다.

주의사항

오해 피하기: 선형 회귀를 이진 분류에 쓰면 안 되는 이유

선형 회귀 모델($Y = \beta_0 + \beta_1 X$)을 그대로 사용하면 두 가지 큰 문제가 발생합니다.

- **범위 초과:** Y는 0 또는 1이어야 하지만, 선형 회귀의 예측 값은 1을 넘거나 0보다 작아질 수 있습니다. 이는 '확률'로 해석할 수 없게 만듭니다.
- **관계 왜곡:** 0과 1 사이의 관계가 직선적(linear)이라고 가정하지만, 실제로는 특정 지점에서 급격히 변하는 S자 형태(비선형)일 가능성이 높습니다.

1.2 핵심 아이디어: 확률을 직접 모델링하지 않는다

로지스틱 회귀는 $P(Y = 1|X)$ (성공 확률)을 직접 모델링하는 대신, '확률'을 변형한 **'로그-오즈(Log-Odds)'**를 선형 회귀 형태로 모델링합니다.

1. **확률 (Probability, P):** 0과 1 사이의 값. (예: $P = 0.8$)
2. **오즈 (Odds):** 성공 확률 / 실패 확률. (예: $Odds = 0.8/(1 - 0.8) = 4$). 0부터 무한대(∞)까지의 값을 가집니다. "실패보다 성공할 확률이 4배 높다"는 의미입니다.

$$\text{Odds} = \frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{P}{1 - P}$$

3. **로그-오즈 (Log-Odds) 또는 로짓(Logit):** 오즈에 자연로그(ln)를 취한 값. (예: $\ln(4) \approx 1.386$). 음의 무한대($-\infty$)부터 양의 무한대($+\infty$)까지 모든 값을 가질 수 있습니다.

$$\text{Logit}(P) = \ln(\text{Odds}) = \ln\left(\frac{P}{1 - P}\right)$$

로그-오즈는 선형 회귀의 예측 값처럼 범위에 제한이 없으므로, 이를 X에 대한 선형 결합으로 모델링할 수 있습니다.

$$\underbrace{\ln\left(\frac{P}{1 - P}\right)}_{\text{Log-Odds}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

1.3 추정: 최대가능도 추정법 (MLE)

선형 회귀는 β를 찾기 위해 오차제곱합(SSE)을 최소화했습니다(최소제곱법, OLS).

로지스틱 회귀는 **최대가능도 추정법 (Maximum Likelihood Estimation, MLE)**을 사용합니다. 직관

적으로, "우리가 가진 데이터(Y 값들)가 관찰될 확률을 가장 높게 만드는 β 값을 찾자"는 의미입니다.

이는 수학적으로 **'음의 로그-가능도(Negative Log-Likelihood)'**를 최소화하는 것과 같으며, 이 손실 함수(Loss Function)를 **'이진 교차 엔트로피 (Binary Cross-Entropy)'**라고 부릅니다.

$$\text{Loss (Binary Cross-Entropy)} = -\sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

* y_i: 실제 값 (0 또는 1) * p_i: 모델이 예측한 $P(Y_i = 1)$ 확률

선형 회귀와 달리 한 번에 풀리는 해(Closed-form solution)가 없으며, **경사 하강법(Gradient Descent)**과 같은 수치 최적화(Numerical Optimization) 기법을 사용해 β 값을 반복적으로 찾아나갑니다.

2 로지스틱 회귀의 추론 (Inference)

추론(Inference)의 목적은 모델의 계수(β)가 통계적으로 유의미한지, 그리고 그 값의 신뢰구간(Confidence Interval)이 어느 정도인지 파악하는 것입니다.

2.1 선형 회귀(t-분포) vs 로지스틱 회귀(Z-분포)

- **선형 회귀:** 오차항의 분산(σ^2)을 별도로 '추정'해야 합니다. 이 불확실성 때문에 β의 분포는 t-분포를 따릅니다.
- **로지스틱 회귀:** Y가 베르누이 분포($Y \sim \text{Bernoulli}(P)$)를 따릅니다. 베르누이 분포의 분산은 $P(1-P)$로, 평균(P)이 정해지면 분산이 '공짜로' 결정됩니다. 별도로 추정할 분산이 없으므로, (샘플이 충분히 크다면) β는 정규분포(Z-분포)를 따른다고 가정합니다.

□ 예제:

Z-분포 vs t-분포의 실제적 차이
95% 신뢰구간을 계산할 때,
- **Z-분포 (로지스틱):** $\hat{\beta} \pm \mathbf{1.96} \times$ (표준오차)
- **t-분포 (선형):** $\hat{\beta} \pm \mathbf{t_{\alpha/2,df}} \times$ (표준오차) (보통 2에 가까운 값)

실제 계산에서는 큰 차이가 없으나, 통계적 근거가 다릅니다. 'statsmodels' 라이브러리는 이러한 Z-통계량, p-value, 신뢰구간을 제공해줍니다.

2.2 계수(Coefficient) 해석하기

β 값 자체보다 e^{β} (지수 변환) 값이 훨씬 직관적입니다.

- $\boldsymbol{\beta_j}$: X_j가 1단위 증가할 때, **로그-오즈(Log-Odds)**가 β_j 만큼 **증가(덧셈)**합니다.
- $\boldsymbol{e^{\beta_j}}$: X_j가 1단위 증가할 때, **오즈(Odds)**가 e^{β_j} 만큼 **곱해집니다(배수)**.

이를 **오즈비 (Odds Ratio, OR)**라고 부릅니다. * $e^{\beta_j} > 1$: X_j가 증가하면 성공 오즈가 증가합니다. (긍정적 관계) * $e^{\beta_j} = 1$: X_j는 성공 오즈와 관계 없습니다. ($\beta_j = 0$) * $e^{\beta_j} < 1$: X_j가 증가하면 성공 오즈가 감소합니다. (부정적 관계)

□ 예제:

예: 이진 예측변수 (Binary Predictor) 해석 (성별과 심장병)
심장병 발병($Y = 1$) 여부를 성별로 예측하는 모델을 가정합니다. (기준 그룹: 남성)

$$\ln(\text{Odds}) = \beta_0 + \beta_1 \cdot \text{Female} \quad (\text{Female} = 1, \text{Male} = 0)$$

여기서 $\beta_0 = 0.214$이고 $\beta_1 = -1.272$라고 가정합시다.

1. β_0**의 해석 (기준 그룹):**
   - $\beta_0 = 0.214$는 **남성(기준 그룹)**의 **로그-오즈**입니다.
   - $e^{\beta_0} = e^{0.214} \approx 1.24$는 **남성**의 **오즈**입니다.
   - (심장병에 걸릴 오즈가 안 걸릴 오즈보다 1.24배 높다.)
2. β_1**의 해석 (차이):**

- $\beta_1 = -1.272$는 **여성**의 로그-오즈가 **남성**보다 1.272만큼 **낮다**는 의미입니다.
- $e^{\beta_1} = e^{-1.272} \approx 0.28$은 **오즈비(Odds Ratio)**입니다.
- **해석:** "다른 조건이 같다면, 여성의 심장병 발병 **오즈**는 남성의 **0.28배 (즉, 72% 낮다)**."

주의사항

오즈(Odds)와 확률(Probability)을 혼동하지 마세요!
계수(β) 해석은 항상 **오즈(Odds)** 관점에서 이루어집니다. "여성의 심장병 발병 확률이 72% 낮다"라고 해석하면 **틀립니다**. 확률로 변환하려면 $P = \text{Odds}/(1 + \text{Odds})$ 공식을 사용해야 하며, 이는 기준이 되는 확률(baseline probability)에 따라 그 변화량이 달라지는 비선형(non-linear) 관계입니다.

3 다중 로지스틱 회귀와 상호작용

3.1 다중 로지스틱 회귀 (Multiple Logistic Regression)

선형 회귀와 마찬가지로, 여러 개의 예측 변수($X_1, \ldots, X_p$)를 사용하여 모델을 구성할 수 있습니다.

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

해석: β_j (또는 e^{β_j})의 의미는, **"다른 모든 예측 변수(X_k)를 통제(일정하게 유지)했을 때"** X_j가 1단위 변할 때의 로그-오즈 (또는 오즈비) 변화량입니다.

주의점: 선형 회귀와 동일한 문제들, 즉 **다중공선성(Multicollinearity)**과 **과적합(Overfitting)**이 여기서도 동일하게 발생합니다.

3.2 상호작용 (Interactions)

상호작용 항은 "한 변수(X_1)가 결과(Y)에 미치는 영향이 다른 변수(X_2)의 값에 따라 달라질 때" 사용됩니다.

□ **예제:**

예: 상호작용 해석 (Age × Female)
심장병 모델에 나이(Age)와 성별(Female), 그리고 둘의 상호작용 항을 추가합니다.

$$\ln(\text{Odds}) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Female} + \beta_3 (\text{Age} \times \text{Female})$$

이 모델은 성별에 따라 두 개의 다른 모델로 분리됩니다.

1. **남성 (Male, Female=0)의 모델:** Female=0을 대입하면 β_2, β_3 항이 사라집니다.

$$\ln(\text{Odds})_{\text{Male}} = \beta_0 + \beta_1 \text{Age}$$

(절편: β_0, 나이의 기울기: β_1)

2. **여성 (Female, Female=1)의 모델:** Female=1을 대입합니다.

$$\ln(\text{Odds})_{\text{Female}} = \beta_0 + \beta_1 \text{Age} + \beta_2 (1) + \beta_3 (\text{Age} \times 1)$$

$$\ln(\text{Odds})_{\text{Female}} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{Age}$$

(절편: $\beta_0 + \beta_2$, 나이의 기울기: $\beta_1 + \beta_3$)

계수 해석:

- β_1: **남성(기준 그룹)**의 나이 1살 증가에 따른 로그-오즈 변화량.
- β_3: **여성**의 나이 1살 증가에 따른 로그-오즈 변화량이 **남성 대비** 얼마나 **다른지 (그 차이)**

만약 β_3가 0이라면 (상호작용이 없다면), 두 그룹의 나이 기울기는 β_1로 동일할 것입니다.

4 분류와 결정 경계 (Classification & Decision Boundary)

4.1 확률에서 분류로: 임계값 (Threshold)

로지스틱 회귀는 $P(Y=1|X)$ 확률을 예측합니다. 이를 0 또는 1의 분류로 바꾸려면 **'임계값(Threshold)'** (보통 0.5)을 정해야 합니다.

- $P(Y=1) \geq 0.5$ 이면, $\hat{Y}=1$ (성공)으로 분류한다.
- $P(Y=1) < 0.5$ 이면, $\hat{Y}=0$ (실패)로 분류한다.

4.2 결정 경계 (Decision Boundary)

결정 경계는 모델의 예측이 $\hat{Y}=0$에서 $\hat{Y}=1$로 바뀌는 지점, 즉 $P(Y=1)=0.5$가 되는 지점의 선 또는 면을 의미합니다.

$P=0.5$라는 것은 어떤 의미일까요?

- $P=0.5$
- $\text{Odds} = P/(1-P) = 0.5/0.5 = 1$
- $\ln(\text{Odds}) = \ln(1) = 0$

즉, 결정 경계는 **로그-오즈가 0이 되는 지점**입니다.

$$\ln(\text{Odds}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = 0$$

주의사항

오해 피하기: 결정 경계는 항상 선형인가?
결정 경계가 선형(직선, 평면)일 수도 있고, 비선형(곡선, 곡면)일 수도 있습니다. 이는 **모델에 어떤 항을 포함했는지**에 달려있습니다.

- **선형 경계:** 모델이 $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$ 처럼 X의 1차항만 포함하면, 결정 경계는 X_1과 X_2 공간에서 **직선**이 됩니다.
- **비선형 경계:** 모델이 $X_1^2, X_2^2, X_1 X_2$ 같은 **다항식(Polynomial) 항이나 상호작용 항**을 포함하면, (예: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 = 0$) 결정 경계는 X_1과 X_2 공간에서 **곡선 (원, 타원 등)**이 됩니다.

비선형 항을 추가함으로써 로지스틱 회귀는 복잡한 데이터 패턴도 분류할 수 있게 됩니다.

[이미지 삽입: 왼쪽은 직선으로 두 클래스를 나누는 결정 경계, 오른쪽은 곡선(원형)으로 두 클래스를 나누는 결정 경계를 보여줌]

5 정규화 (Regularization)

모델에 다항식 항이나 상호작용 항을 많이 추가하면 결정 경계가 매우 복잡해지면서 훈련 데이터에만 꼭 맞는 **과적합(Overfitting)**이 발생할 수 있습니다.

정규화(Regularization)는 모델의 복잡도에 페널티를 부과하여 과적합을 방지하는 기법입니다. 계수(β)의 크기가 너무 커지지 않도록 손실 함수(Loss Function)에 **페널티 항(Penalty Term)**을 추가합니다.

5.1 손실 함수 + L2 (Ridge) 페널티

로지스틱 회귀의 손실 함수(Binary Cross-Entropy)에 L2 페널티(계수 제곱의 합, Ridge)를 더합니다.

$$\text{Loss}_{\text{Regularized}} = \underbrace{\text{Loss (Binary Cross-Entropy)}}_{\text{모델이 데이터에 얼마나 잘 맞는지}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{모델이 얼마나 복잡한지 (페널티)}}$$

- λ (람다): 정규화의 강도를 조절하는 하이퍼파라미터입니다.
  - $\lambda = 0$: 페널티 없음 (표준 로지스틱 회귀).
  - $\lambda \to \infty$: 페널티가 매우 강해져 모든 β가 0에 가까워집니다 (모델이 매우 단순해짐).
- 페널티는 보통 절편(β_0)을 제외하고 적용됩니다.

주의사항

sklearn의 'C' 파라미터 이해하기

'sklearn.linear$_{m}$odel.*LogisticRegression*' λ 대신 C라는 파라미터를 사용합니다. C는 λ의 역수 ($C = 1/\lambda$) 개념입니다.

- **높은 'C' (예: $C = 100$) $\implies$ 낮은 λ:** 정규화(페널티)가 **약합니다**. 모델이 복잡해지고 훈련 데이터에 더 강하게 맞춰집니다 (과적합 위험).
- **낮은 'C' (예: $C = 0.01$) $\implies$ 높은 λ:** 정규화(페널티)가 **강합니다**. 모델이 단순해지고 β 계수들이 0에 가까워집니다 (과소적합 위험).

최적의 C 값은 **교차 검증(Cross-Validation)**을 통해 찾아야 합니다.

```python
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV

# C 값이낮을수록정규화가강해짐
logreg = LogisticRegression(penalty='l2', C=0.1)

# 교차검증으로최적의  C 값을찾을수도있음
params = {'C': [0.01, 0.1, 1, 10, 100]}
grid_search = GridSearchCV(LogisticRegression(penalty='l2'), params, cv=5)
# grid_search.fit(X_train, y_train)
# print(grid_search.best_params_)
```

Listing 1: sklearn에서 Ridge 정규화를 적용한 로지스틱 회귀

6 다중 클래스 로지스틱 회귀 (Multiclass)

Y의 범주가 3개 이상일 때 (예: 'CS', 'Stat', 'Other') 사용하는 방법입니다. 여기서는 순서가 없는 **명목형(Nominal)** 범주를 가정합니다.

6.1 접근법 1: One-vs-Rest (OvR)

가장 간단하고 직관적인 방법입니다. 범주가 K개일 때, K개의 독립적인 이진 분류기를 만듭니다.

- **분류기 1:** 'CS' vs 'Not CS' (즉, 'Stat' + 'Other')
- **분류기 2:** 'Stat' vs 'Not Stat' (즉, 'CS' + 'Other')
- **분류기 3:** 'Other' vs 'Not Other' (즉, 'CS' + 'Stat')

새로운 데이터가 들어오면, 3개의 분류기를 모두 돌려서 각각의 확률(또는 점수)을 계산한 뒤, 가장 높은 확률(점수)을 보인 클래스로 예측합니다. ('sklearn'의 'multi$_c lass =' ovr'$')

6.2 접근법 2: 다항 로지스틱 회귀 (Multinomial)

OvR과 달리, K개의 클래스를 한 번에 처리하는 단일 모델을 만듭니다. 하나의 클래스(예: 'Other')를 **기준(Reference) 클래스**로 정합니다. 그리고 $K-1$개의 로그-오즈 모델을 만듭니다.

- **모델 1:** $\ln\left(\frac{P(\text{CS})}{P(\text{Other})}\right) = \beta_0^{(1)} + \beta_1^{(1)}X + \dots$
- **모델 2:** $\ln\left(\frac{P(\text{Stat})}{P(\text{Other})}\right) = \beta_0^{(2)} + \beta_1^{(2)}X + \dots$

('sklearn'의 'multi$_c lass =' multinomial'$')

6.3 Softmax: 점수를 확률로 변환하기

OvR이든 다항 로지스틱이든, 각 클래스 k에 대한 '점수(Score)' 또는 '로짓(Logit)' s_k가 나옵니다. 이 점수들은 합쳐도 1이 되지 않기 때문에 확률로 사용하기 어렵습니다.

소프트맥스(Softmax) 함수는 이 점수(s_k)들을 0과 1 사이의 값으로 변환하고, 모든 클래스의 확률 총합이 1이 되도록 정규화해줍니다.

$$P(Y = k|X) = \frac{e^{s_k}}{\sum_{j=1}^{K} e^{s_j}}$$

6.4 다중 클래스 분류 (K>2)

임계값 0.5는 더 이상 의미가 없습니다. 대신 **'다수결 원칙(Plurality Wins)'**을 사용합니다. Softmax를 통해 계산된 K개의 확률 중, **가장 높은 확률을 가진 클래스**로 예측합니다.

예: $P(\text{CS}) = 0.2$, $P(\text{Stat}) = 0.4$, $P(\text{Other}) = 0.4$ $\implies$ 가장 높은 확률이 0.4로 두 개이므로, 둘 중 하나를 선택 (혹은 추가 규칙 적용). 만약 $P(\text{Stat}) = 0.41$, $P(\text{Other}) = 0.39$ 였다면 $\hat{Y} = \text{Stat}$로 예측.

7 분류 모델 평가 (Evaluation)

모델을 만들었다면, 이 모델이 얼마나 '좋은' 분류기인지 평가해야 합니다. 단순 정확도(Accuracy)는 특히 데이터가 불균형할 때(예: 99%가 'No', 1%가 'Yes') 성능을 오해하게 만듭니다.

7.1 혼동 행렬 (Confusion Matrix)

분류 결과(예측)와 실제 값을 2×2 표로 정리한 것입니다.

		모델의 예측 (Predicted)	
		Negative (0)	**Positive (1)**
실제 값 (Actual)	**Negative (0)**	**True Negative (TN)**	**False Positive (FP)**
	Positive (1)	**False Negative (FN)**	**True Positive (TP)**

- **True Positive (TP): 정답.** 실제 Positive, 예측 Positive (예: 암환자를 암으로 진단)
- **True Negative (TN): 정답.** 실제 Negative, 예측 Negative (예: 건강한 사람을 건강하다고 진단)
- **False Positive (FP): 1종 오류.** 실제 Negative, 예측 Positive (예: 건강한 사람을 암으로 오진)
- **False Negative (FN): 2종 오류.** 실제 Positive, 예측 Negative (예: 암환자를 건강하다고 오진) ← 치명적 오류!

7.2 핵심 평가지표

1. **민감도 (Sensitivity) = 재현율 (Recall) = True Positive Rate (TPR)**

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad \text{(실제 Positive 중 맞춘 비율)}$$

의미: 실제 암환자 중 몇 %를 '암'이라고 잡아냈는가? (FN을 줄이는 데 초점) 의료 진단에서 매우 중요합니다. (놓치면 안 됨)

2. **특이도 (Specificity) = True Negative Rate (TNR)**

$$\text{Specificity} = \frac{TN}{TN + FP} \quad \text{(실제 Negative 중 맞춘 비율)}$$

의미: 실제 건강한 사람 중 몇 %를 '건강'이라고 판단했는가? (FP를 줄이는 데 초점) FP의 비용이 클 때 (예: 스팸 필터가 중요한 메일을 스팸 처리) 중요합니다.

3. **정밀도 (Precision) = Positive Predictive Value (PPV)**

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{(예측 Positive 중 맞춘 비율)}$$

의미: 모델이 '암'이라고 예측한 사람들 중, 실제 암환자는 몇 %인가?

4. **False Positive Rate (FPR)**

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{TN + FP} \quad \text{(실제 Negative 중 틀린 비율)}$$

의미: 건강한 사람 중 몇 %를 '암'이라고 잘못 예측했는가?

주의사항

베이즈 정리와 낮은 유병률(Prevalence) 문제

베이즈 정리에 따르면, 아무리 테스트기(모델)의 민감도(99%)와 특이도(99%)가 높아도, 질병 자체가 매우 희귀하다면(예: 유병률 0.1%), 테스트 결과가 '양성(Positive)'이 나왔더라도 실제 환자일 확률(PPV/정밀도)은 매우 낮을 수 있습니다. 대부분이 False Positive이기 때문입니다.

7.3 임계값(Threshold)의 트레이드오프

분류 임계값 0.5는 절대적인 기준이 아닙니다. 임계값을 조절하면 민감도(TPR)와 특이도(1-FPR)가 반비례 관계(Trade-off)로 움직입니다.

- **임계값을 낮추면 (예: 0.5 → 0.3):** 모델이 'Positive'라고 더 쉽게 예측합니다. TP ↑ (좋음), FN ↓ (좋음) $\implies$ **민감도(TPR) 상승** FP ↑ (나쁨), TN ↓ (나쁨) $\implies$ **FPR 상승 (특이도 하락)** (예: "일단 암일 가능성이 조금만 있어도 양성으로 판정" → FN은 줄지만 FP가 늘어남)
- **임계값을 높이면 (예: 0.5 → 0.7):** 모델이 'Positive'라고 더 보수적으로 예측합니다. TP ↓ (나쁨), FN ↑ (나쁨) $\implies$ **민감도(TPR) 하락** FP ↓ (좋음), TN ↑ (좋음) $\implies$ **FPR 하락 (특이도 상승)** (예: "확실히 암일 때만 양성으로 판정" → FP는 줄지만 FN이 늘어남)

7.4 ROC 커브와 AUC

ROC 커브 (Receiver Operating Characteristic Curve)는 이 트레이드오프를 시각화한 그래프입니다.

- **X축:** False Positive Rate (FPR) (1 - 특이도)
- **Y축:** True Positive Rate (TPR) (민감도)

모든 가능한 임계값(0에서 1까지)에 대해 (FPR, TPR) 좌표를 찍어서 연결한 선입니다.

[이미지 삽입: *ROC* 커브 그래프. *(0,0)*에서 *(1,1)*을 잇는 점선(*Random Classifier*), *(0,0)*에서 *(0,1)*을 거쳐 *(1,1)*로 가는 실선(*Perfect Classifier*), 그리고 그 사이를 지나는 실제 모델의 *ROC* 커브를 보여줌]

- **완벽한 모델 (Perfect Classifier):** (0, 1) 지점을 통과합니다 (FPR=0, TPR=1).
- **랜덤 모델 (Random Classifier):** $y = x$ 대각선. (FPR과 TPR이 같음)
- **좋은 모델:** 커브가 왼쪽 위 (0, 1)에 최대한 가까이 붙습니다.

AUC (Area Under the Curve)는 이 ROC 커브 아래의 면적입니다. 0부터 1 사이의 값을 가지며, 모델의 전체적인 성능을 하나의 숫자로 요약해줍니다.

- **AUC = 1.0:** 완벽한 분류기
- **AUC = 0.5:** 쓸모없는 분류기 (랜덤 추측)
- **AUC ≈ 0.8~0.9:** 매우 좋은 분류기

AUC는 임계값에 상관없이 모델이 'Positive' 샘플을 'Negative' 샘플보다 얼마나 더 높은 확률로 예측하는지(순서)를 나타내는 지표입니다.

8 핵심 용어 정리

Table 1: 로지스틱 회귀 및 분류 평가 핵심 용어

용어	원어	쉬운 설명
오즈	Odds	성공 확률 / 실패 확률. $(P/(1-P))$
로그-오즈	Log-Odds	오즈에 자연로그를 취한 값. $\ln(P/(1-P))$. 로지스틱 회귀의 Y 값.
오즈비	Odds Ratio (OR)	X가 1단위 증가할 때, 오즈가 몇 '배' 변하는지. (e^{β})
MLE	Max Likelihood Estimation	데이터가 관찰될 확률을 최대화하는 β를 찾는 추정 방식.
이진 교차 엔트로피	Binary Cross-Entropy	로지스틱 회귀의 손실 함수(Loss Function). 음의 로그-가능도.
결정 경계	Decision Boundary	예측 클래스가 0에서 1로 바뀌는 경계선. $P = 0.5$ (즉, Log-Odds=0)인 지점.
정규화	Regularization	모델 복잡도에 페널티를 주어 과적합을 막는 기법. (예: L2/Ridge)
C 파라미터	C (in sklearn)	$1/\lambda$. 정규화 강도의 역수. (C가 낮을수록 정규화가 강함)
OvR	One-vs-Rest	K개 클래스 분류 시, K개의 이진 분류기('A' vs 'Not A')를 만듦.
다항 회귀	Multinomial Regression	K개 클래스 분류 시, 1개의 기준 클래스 대비 K-1개 모델을 만듦.
소프트맥스	Softmax	K개의 클래스 점수(Logit)를 총합 1인 확률로 변환하는 함수.
혼동 행렬	Confusion Matrix	모델의 예측(TP, FP, FN, TN)과 실제 값을 비교한 표.
민감도 (재현율)	Sensitivity (Recall)	실제 '성공' 중 모델이 '성공'으로 맞춘 비율. $(TP/(TP+FN))$
특이도	Specificity	실제 '실패' 중 모델이 '실패'로 맞춘 비율. $(TN/(TN+FP))$
정밀도	Precision	모델이 '성공' 예측 중 실제 '성공'인 비율. $(TP/(TP+FP))$
ROC 커브	ROC Curve	모든 임계값에 대해 (FPR, TPR)을 그린 그래프.
AUC	Area Under the Curve	ROC 커브 아래 면적. 1에 가까울수록 좋은 모델.

9 학습 체크리스트

title=최종 점검 체크리스트

- ☐ 왜 이진 분류 문제에 선형 회귀 대신 로지스틱 회귀를 써야 하는지 설명할 수 있는가?
- ☐ 확률(P), 오즈(Odds), 로그-오즈(Log-Odds)의 관계를 설명할 수 있는가?
- ☐ β_1 계수와 e^{β_1} (오즈비)의 해석상 차이를 설명할 수 있는가?
- ☐ 다중 로지스틱 회귀에서 β_j를 해석할 때 "다른 변수를 통제할 때"라는 조건이 왜 붙는지 아는가?
- ☐ 상호작용 항($X_1 \times X_2$)이 모델의 절편과 기울기에 각각 어떤 영향을 미치는지 설명할 수 있는가?
- ☐ 결정 경계(Decision Boundary)가 $P = 0.5$ 지점, 즉 $X\beta = 0$ 지점과 같다는 것을 수학적으로 유도할 수 있는가?
- ☐ 모델에 다항식 항을 추가하면 결정 경계가 어떻게 변하는지 아는가?
- ☐ 정규화(Regularization)가 필요한 이유(과적합 방지)를 설명할 수 있는가?
- ☐ sklearn의 'C' 파라미터가 낮을수록 정규화가 강해진다는 것을 아는가? ($C \approx 1/\lambda$)
- ☐ 다중 클래스 분류의 2가지 접근법 (OvR, Multinomial)을 비교할 수 있는가?
- ☐ Softmax 함수의 역할(점수 $\rightarrow$ 확률 정규화)을 아는가?
- ☐ 혼동 행렬의 TP, FP, FN, TN이 각각 무엇을 의미하는지 아는가?
- ☐ 민감도(재현율), 특이도, 정밀도의 차이를 (공식과 의미) 설명할 수 있는가?
- ☐ 분류 임계값(Threshold)을 조절하면 민감도와 특이도가 어떻게 변하는지(Trade-off) 아는가?
- ☐ ROC 커브의 X축(FPR)과 Y축(TPR)이 무엇이며, AUC가 왜 0.5면 '랜덤'인지 설명할 수 있는가?

10 1페이지 요약 (1-Page Summary)

title=1. 로지스틱 회귀 모델 Y가 0 또는 1일 때, 성공 확률 P를 직접 모델링하지 않고, **로그-오즈**를 X에 대한 선형식으로 모델링합니다.

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

β는 MLE (최대가능도 추정법) 또는 경사하강법으로 추정합니다.

title=2. 계수 해석 β_j는 X_j가 1단위 증가할 때 **로그-오즈**의 **덧셈** 변화량입니다. e^{β_j} (오즈비)는 X_j가 1단위 증가할 때 **오즈**의 **곱셈** 변화량 (배수)입니다.

title=3. 결정 경계 (Decision Boundary) 분류 임계값을 $P = 0.5$로 두면, 결정 경계는 $\ln(\text{Odds}) = 0$이 되는 지점, 즉 $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = 0$이 됩니다. X의 1차항만 있으면 직선, 다항식/상호작용항이 있으면 곡선이 됩니다.

title=4. 정규화 (Regularization) 과적합을 막기 위해 손실 함수(이진 교차 엔트로피)에 페널티 항을 추가합니다. (L2/Ridge) $\text{Loss}_{\text{Reg}} = \text{Loss} + \lambda \sum \beta_j^2$. 'sklearn'에서는 $C \approx 1/\lambda$를 사용하며, C가 낮을수록 정규화가 강합니다. 최적의 C는 교차 검증(Cross-Validation)으로 찾습니다.

title=5. 다중 클래스 (Multiclass) K>2

- **OvR (One-vs-Rest):** K개의 이진 분류기('A' vs 'Not A')를 만듦.
- **Multinomial:** 1개의 기준 클래스 대비 K-1개 모델을 만듦.
- **Softmax:** K개의 점수(Logit)를 총합 1인 확률로 변환.
- **분류:** 가장 높은 확률을 가진 클래스로 예측 (Plurality Wins).

title=6. 분류 평가 (Evaluation)

- **혼동 행렬:** TP, FP, FN, TN
- **민감도(TPR):** $\frac{TP}{TP+FN}$ (실제 P 중 예측 P)
- **특이도(TNR):** $\frac{TN}{TN+FP}$ (실제 N 중 예측 N)
- **정밀도(PPV):** $\frac{TP}{TP+FP}$ (예측 P 중 실제 P)
- **ROC 커브:** X축=FPR (1-특이도), Y축=TPR (민감도). 모든 임계값에서의 성능 시각화.
- **AUC:** ROC 커브 아래 면적. 1에 가까울수록 좋은 분류기.