로지스틱 회귀

고려대학교 석 준 희

목차

- 로지스틱 회귀 (Logistic Regression)
- 인공신경망으로의 확장

로지스틱 회귀

분류 문제의 표현

- 회귀문제 (Regression): Y가 연속형 변수일 때
- 분류문제 (Classification): Y가 범주형 변수일 때

- 클래스 혹은 범주를 숫자로 표현이 가능할까?
  - Y = A or B → Y = 1 or 2 : 가능함
  - Y = A, B or C → Y = 1, 2 or 3 : 불가능함
- 기본적으로 회귀 문제로 치환하여 해결하는 것이 불가능

- 많은 분류의 문제들이 클래스 Y를 직접 예측하기 보다는 Y가 특정 클래스일 확률 $\Pr[Y = k|X]$를 예측하고자 함

- 분류 문제 모델링

$$\Pr[Y = k|X] \sim f_k(X)$$

이진 분류 (Binary Classification)

- 이진 분류: 범주/클래스가 2개인 경우

- 예제: 시험합격 ~ 공부시간
  - 종속변수(Y): 시험합격, 합격(1) or 불합격(0)
  - 독립변수(X): 공부시간, 숫자

- 데이터 행렬과 예측 결과

모델로부터 예측된 확률

	공부시간	합격여부	$\Pr[Y=1\|X]$	$\Pr[Y=0\|X]$	$f_1(X) = \widehat{\Pr}[Y=1\|X]$	$f_0(X) = \widehat{\Pr}[Y=0\|X]$
1	0.50	불합격	0	1	0.05	0.95
2	3.30	합격	1	0	0.71	0.29
3	1.75	합격	1	0	0.23	0.77
4	3.00	불합격	0	1	0.62	0.38

로지스틱 회귀 (Logistic Regression)

- 클래스에 대한 확률을 시그모이드(sigmoid) 함수를 이용하여 모델링
- 이진분류(Y = 1 or 0)에 대하여

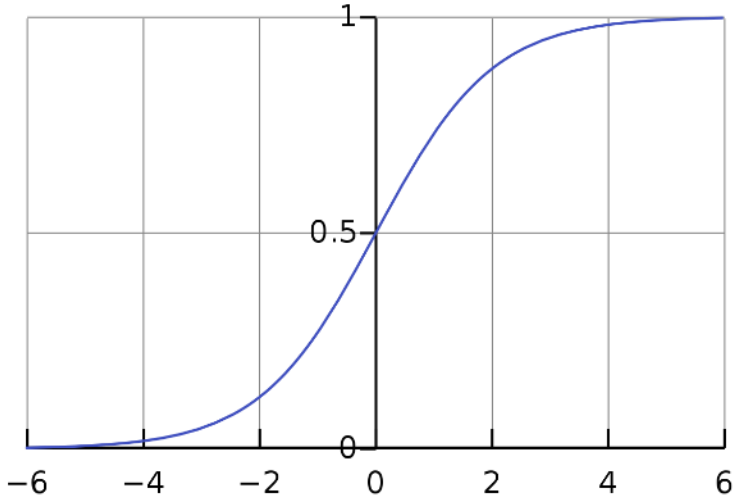
$$\Pr[Y=1|X] = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

$$\Pr[Y=0|X] = \frac{1}{1+e^{\beta_0+\beta_1 X}} = \frac{e^{-(\beta_0+\beta_1 X)}}{1+e^{-(\beta_0+\beta_1 X)}}$$

$$\log\left(\underbrace{\frac{\Pr[Y=1|X]}{\Pr[Y=0|X]}}_{\text{odds}}\right) = \beta_0+\beta_1 X$$

$\underbrace{\qquad\qquad\qquad\qquad}_{\text{logit = log(odds)}}$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

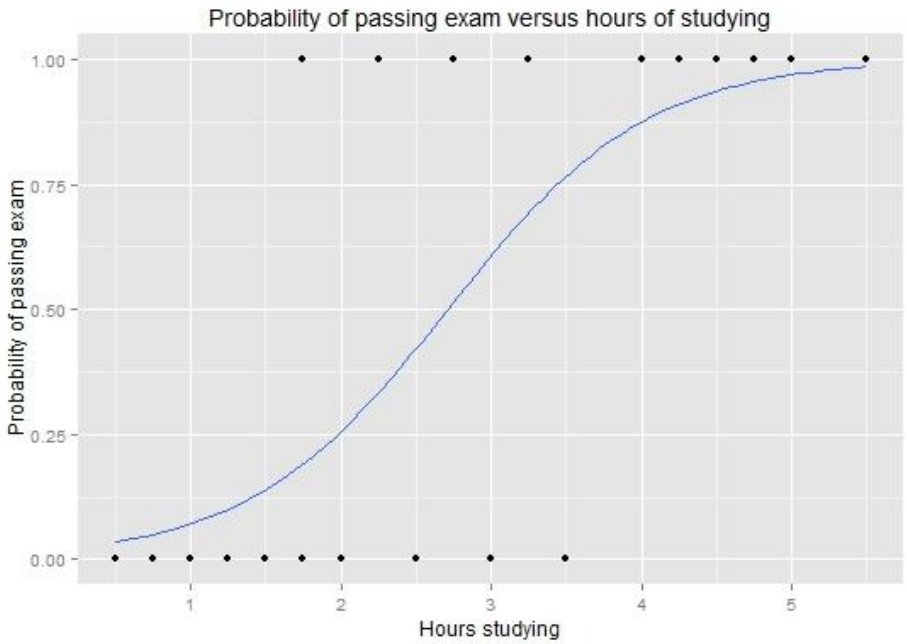


https://en.wikipedia.org/wiki/Sigmoid_function

로지스틱 회귀 (Logistic Regression)

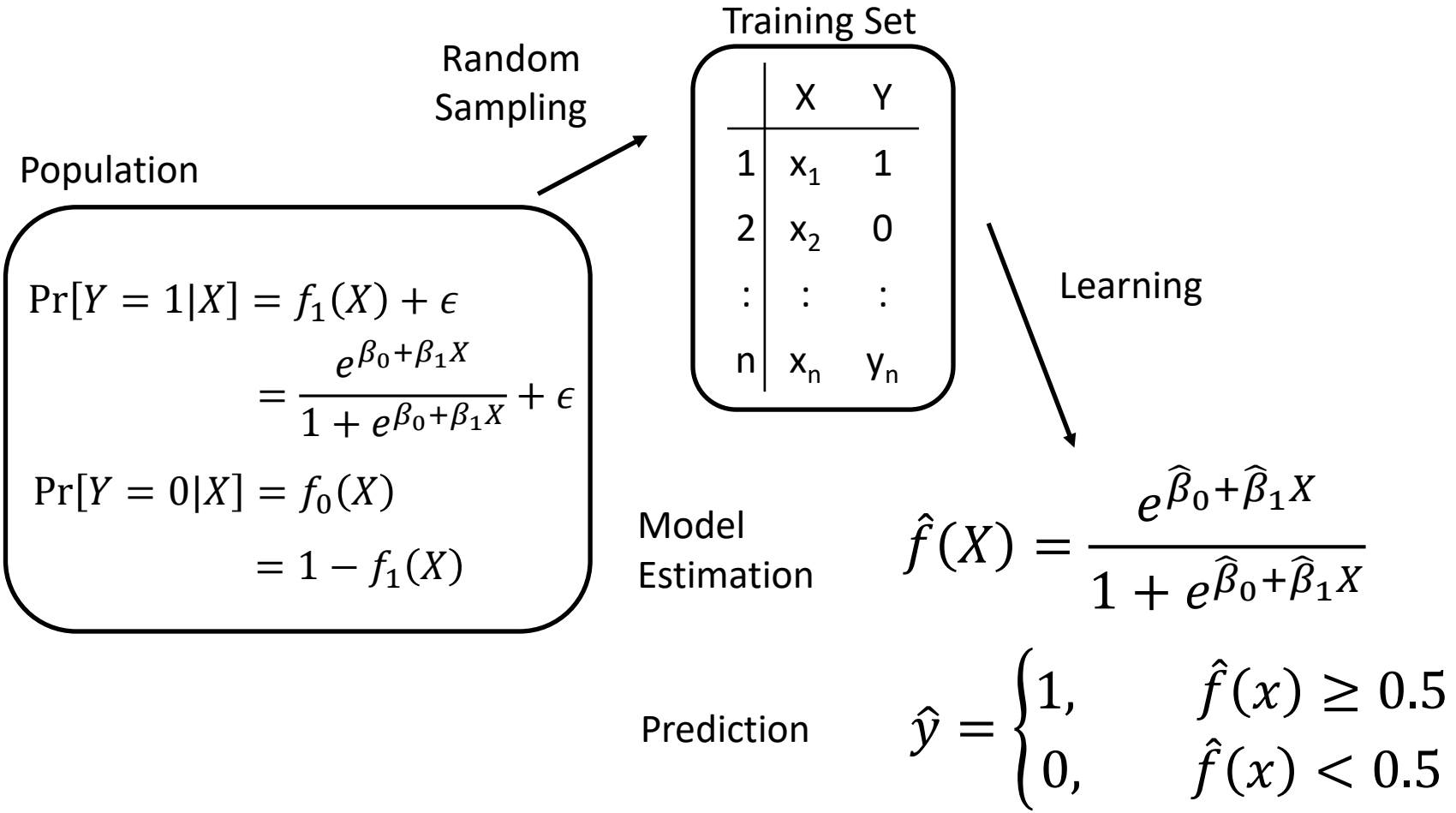
- 예시

	공부시간	합격여부	$\Pr[Y=1\|X]$	$\Pr[Y=0\|X]$	$f_1(X)=\widehat{\Pr}[Y=1\|X]$	$f_0(X)=\widehat{\Pr}[Y=0\|X]$
1	0.50	불합격	0	1	0.05	0.95
2	3.30	합격	1	0	0.71	0.29
3	1.75	합격	1	0	0.23	0.77
4	3.00	불합격	0	1	0.62	0.38



로지스틱 회귀 (Logistic Regression)

- 이진 분류 + 한 개의 독립변수: 확률 예측 후 확률에 따라서 분류를 예측



파라메터 추정

- 우도(Likelihood): 어떤 모델을 가정했을 때 현재 데이터를 관측할 확률
- 로지스틱 회귀에서의 우도

모델: $\Pr[Y=1|X] = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, $\Pr[Y=0|X] = \frac{1}{1+e^{\beta_0+\beta_1 X}}$

우도: $l = \prod_{i:y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_i}} \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$

- 예시

	공부시간	합격여부
1	0.50	불합격
2	3.30	합격
3	1.75	합격
4	3.00	불합격

$$l = \frac{1}{1+e^{\beta_0+\beta_1\times 0.5}} \times \frac{e^{\beta_0+\beta_1\times 3.3}}{1+e^{\beta_0+\beta_1\times 3.3}} \times \frac{e^{\beta_0+\beta_1\times 1.75}}{1+e^{\beta_0+\beta_1\times 1.75}} \times \frac{1}{1+e^{\beta_0+\beta_1\times 3.0}}$$

파라메터 추정

- 최대우도법 (Maximum Likelihood)
  - 왜 하필 우리는 이 데이터를 관측하고 있을까? 그것은 이 데이터를 관측할 확률이 제일 높기 때문이다!
  - 우도를 제일 크게 만드는 파라메터가 진짜 파라메터

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmax}(l) = \text{argmax}\left(\prod_{i:y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_i}} \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right)$$

- 우도의 미분을 통해 최적의 파라메터를 찾을 수 있지만 계산이 어렵기 때문에, 상대적으로 계산이 쉬운 우도의 로그값을 최대화하는 파라메터를 찾음
- 로그 함수의 특성상 $f(x)$를 최대화 하는 것은 $\log(f(x))$를 최대화 하는 것과 같음

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmax}(l) = \text{argmax}(\log l)$$

$$\frac{\partial \log l}{\partial \beta_0} = 0 \qquad \frac{\partial \log l}{\partial \beta_1} = 0 \quad \Rightarrow \quad \hat{\beta}_0, \hat{\beta}_1$$

파라메터 추정

- 우도의 표현은 수식적으로 다음과 같이 변형 가능

$$p_i = \Pr[Y = 1 | X = x_i] = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$l = \prod_{i:y_i=0} (1 - p_i) \prod_{i:y_i=1} p_i = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}$$

- 크로스 엔트로피 (Cross-Entropy) 손실
  - 최대우도법은 $\log l$ 을 최대화하는 것이 목표
  - 이는 $-\log l$ 을 최소화하는 것과 동일하여 손실함수로 정의하는 것이 가능
  - 크로스 엔트로피 손실로 정의하고 이를 최소화하여 모델을 훈련

$$L(\beta_0, \beta_1) = -\log l = -\sum_i [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

파라메터 추정

- 시험합격 ~ 공부시간 예시

$$L = -\log l = \log\left(1 + e^{\beta_0+\beta_1\times 0.5}\right) - (\beta_0 + \beta_1 \times 3.3) + \log\left(1 + e^{\beta_0+\beta_1\times 3.3}\right)$$
$$-(\beta_0 + \beta_1 \times 1.75) + \log\left(1 + e^{\beta_0+\beta_1\times 1.75}\right) + \log\left(1 + e^{\beta_0+\beta_1\times 3.0}\right)$$

$$\frac{\partial L}{\partial \beta_0} = \frac{e^{\beta_0+\beta_1\times 0.5}}{1 + e^{\beta_0+\beta_1\times 0.5}} - 1 + \frac{e^{\beta_0+\beta_1\times 3.3}}{1 + e^{\beta_0+\beta_1\times 3.3}}$$
$$-1 + \frac{e^{\beta_0+\beta_1\times 1.75}}{1 + e^{\beta_0+\beta_1\times 1.75}} + \frac{e^{\beta_0+\beta_1\times 3.0}}{1 + e^{\beta_0+\beta_1\times 3.0}} = 0$$

$$\frac{\partial L}{\partial \beta_1} = 0.5 \times \frac{e^{\beta_0+\beta_1\times 0.5}}{1 + e^{\beta_0+\beta_1\times 0.5}} - 3.3 + 3.3 \times \frac{e^{\beta_0+\beta_1\times 3.3}}{1 + e^{\beta_0+\beta_1\times 3.3}}$$
$$-1.75 + 1.75 \times \frac{e^{\beta_0+\beta_1\times 1.75}}{1 + e^{\beta_0+\beta_1\times 1.75}} + 3.0 \times \frac{e^{\beta_0+\beta_1\times 3.0}}{1 + e^{\beta_0+\beta_1\times 3.0}} = 0$$

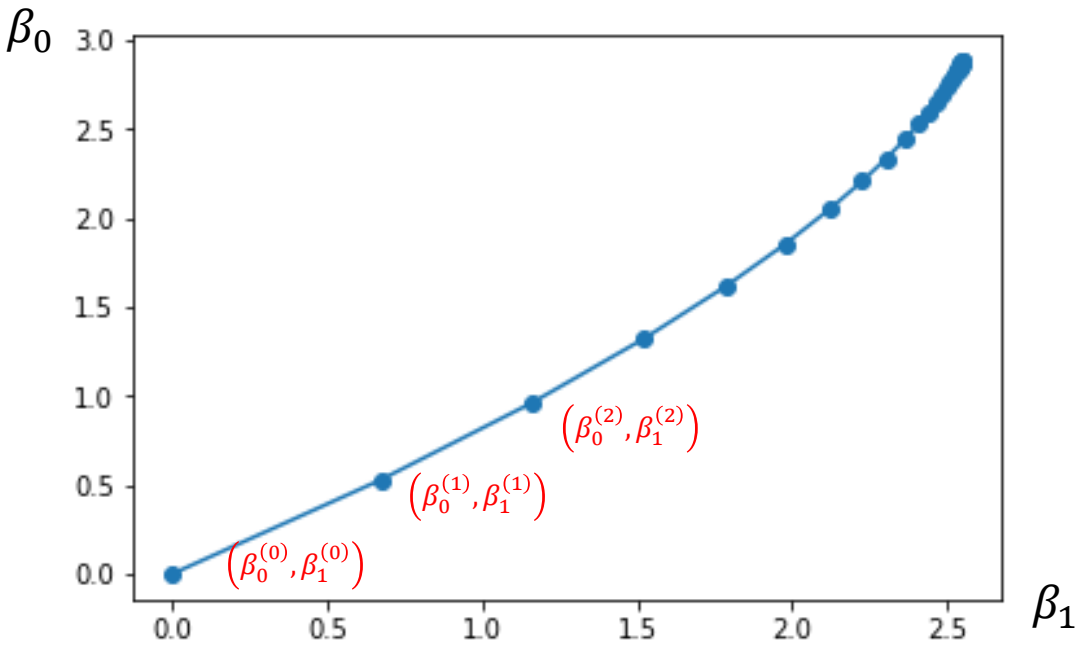
파라메터 추정

- 선형 회귀와 달리 로지스틱 회귀의 파라메터는 closed form으로 계산되지 않음
  - 수치적으로 풀어야 함

- 경사하강법 (Gradient Descent)
  - 임의의 파라메터에서 시작하여 경사(미분값)을 따라 파라메터를 업데이트
  - 인공신경망에서 더 자세히 다룸

$\beta^{(0)}$: random position

$$\beta^{(i+1)} = \beta^{(i)} - \lambda \frac{\partial L(\beta^{(i)})}{\partial \beta}$$



일반적인 로지스틱 회귀

- 하나 이상의 독립 변수에 대해서 선형 회귀와 같이 확장

단순 회귀
Simple Regression

$$\log\left(\frac{\Pr[Y=1|X]}{\Pr[Y=0|X]}\right) = \beta_0 + \beta_1 X \qquad \Pr[Y=1|X] = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

일반 회귀
Multiple Regression

$$\log\left(\frac{\Pr[Y=1|X]}{\Pr[Y=0|X]}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\Pr[Y=1|X] = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}$$

- 선형회귀에서 사용되었던 다양한 기법을 바로 적용 가능
  - 가변수를 이용한 범주형 변수의 표현
  - 상호작용의 고려
  - 고차원 변수의 도입
  - 모델의 복잡성에 대한 고려도 동일

분류 모델의 평가

- 분류 모델을 평가하는 일반적인 지표: 훈련 및 평가 데이터에 공통적으로 사용 가능
- 정확도 (Accuracy): 각 분류를 정확히 맞춘 비율, 1-(오류율, Error Rate)

$$Acc = 1 - ER = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- 혼동행렬 (Confusion Matrix): 진짜 분류와 예측된 분류를 행렬의 형태로 표현

predicted \ truth	A	B	C	D
A	**70**	10	15	5
B	8	**67**	20	5
C	0	11	**88**	1
D	4	10	14	**72**

- 정확도의 경우 간단하지만 정확한 정보를 주지 못하는 경우가 많음
  - 예) 1%의 확률로 발생하는 질병의 경우, 모두 음성으로 예측하면 정확도는 99%

이진 분류 모델의 평가

- 정확도의 경우 간단하지만 정확한 정보를 주지 못하는 경우가 많음
  - 예) 1%의 확률로 발생하는 질병의 경우, 모두 음성으로 예측하면 정확도는 99%
- 이진 분류의 경우 혼동 행렬을 바탕으로 좀 더 다양한 평가 지표를 사용

- 이진 분류에 대한 2x2 혼동 행렬
  - 양성 (Positive): 우리가 찾고 싶은 클래스 (e.g. 질병, 부도 등)
  - 음성 (Negative): 다른 하나의 클래스 (e.g. 정상, 신용 등)

	Truly Positive	Truly Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

- 정확도(Accuracy): $Acc = \frac{TP+TN}{TP+FP+FN+TN}$
- 재현율(Recall, True Positive Rate): $R = TRP = \frac{TP}{TP+FN}$
- 정밀도(Precision, Positive Predictive Value): $P = PPV = 1 - FPR = \frac{TP}{TP+FP}$
- F1 Score: $F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} = \left(\frac{P^{-1}+R^{-1}}{2}\right)^{-1} = \frac{TP}{TP+(FP+FN)/2}$

False Positive Rate

이진 분류 평가 예시

- 분류 예측 결과: 모델은 확률을 예측, 예측 확률을 바탕으로 분류를 결정

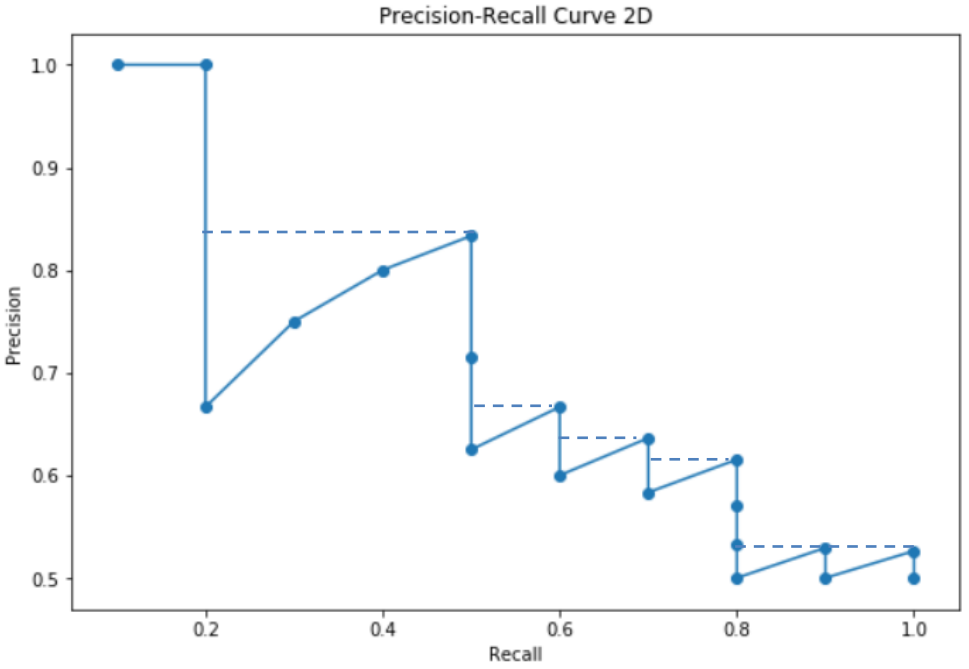
	Y	$\widehat{\Pr}[Y=1\|X]$	$\hat{Y}$			
			≥0.5	≥0.4	≥0.2	≥0.8
1	1	0.85	1	1	1	1
2	0	0.36	0	0	1	0
3	1	0.48	0	1	1	0
4	1	0.66	1	1	1	0
5	0	0.12	0	0	0	0
		정확도	0.80	1.00	0.80	0.60
		재현율	0.67	1.00	1.00	0.33
		정밀도	1.00	1.00	0.75	1.00
		F1	0.80	1.00	0.86	0.50

- 일반적으로 0.5를 기준으로 분류를 결정하지만, 반드시 그럴 필요는 없음
- 재현율과 정밀도를 모두 높이고 싶지만, 둘은 일반적으로 반대 방향으로 변함

이진 분류 모델의 평가

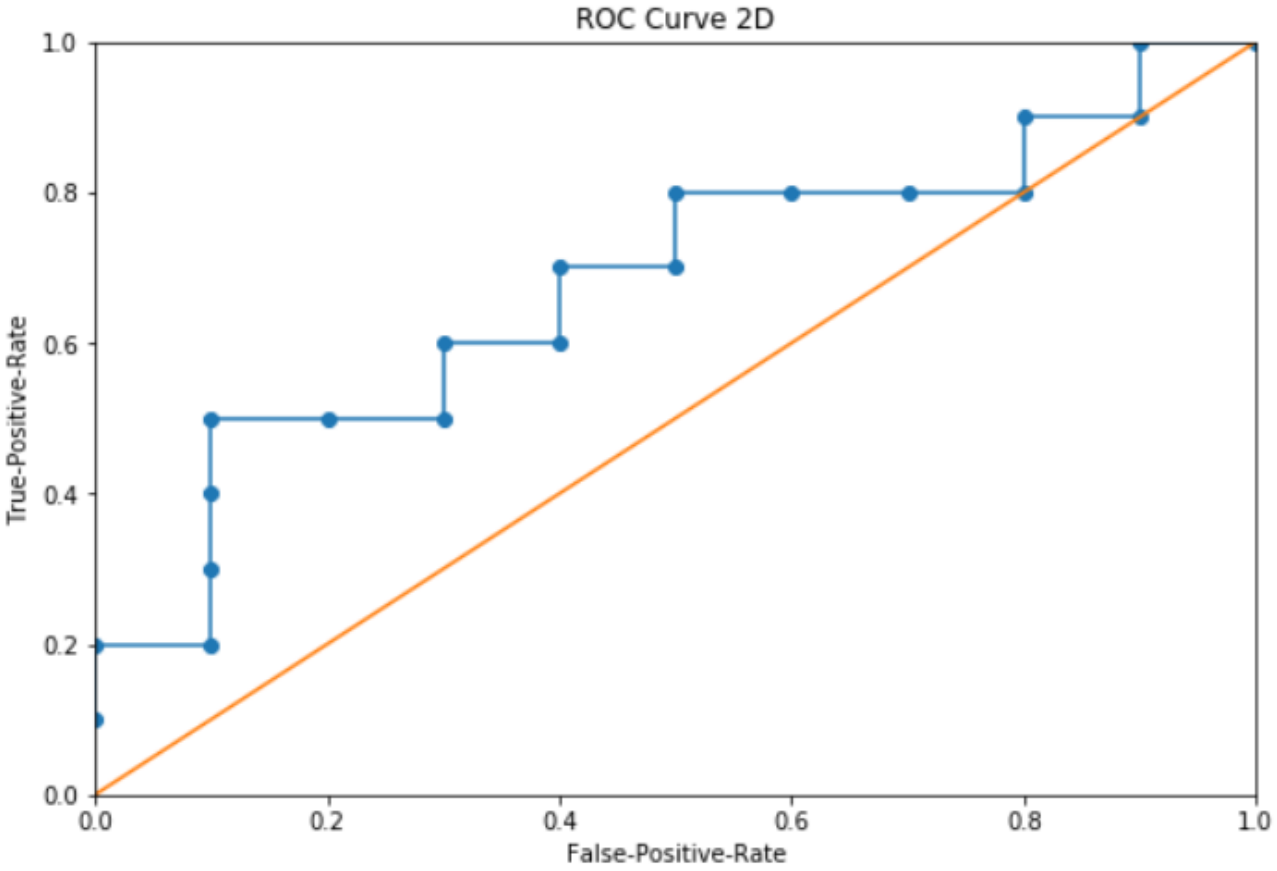
- 같은 확률 예측에 대하여 분류 기준을 변화시킴으로써 서로 다른 성능을 얻을 수 있음
- Precision-Recall Curve: 분류 기준의 변화에 따라 (재현율, 정밀도) 쌍을 표시한 그래프
  - 재현율이 증가함에 따라 정밀도는 대체로 감소하지만 꼭 그렇지는 않음
  - AP (Average Precision): 단조감소한 PR Curve의 면적

	True Label	Pr	Pr≥0.9	Pr≥0.8	Pr≥0.7	Pr≥0.1
0	positive	0.900	positive	positive	positive	positive
1	positive	0.800	negative	positive	positive	positive
2	negative	0.700	negative	negative	positive	positive
3	positive	0.600	negative	negative	negative	positive
4	positive	0.550	negative	negative	negative	positive
5	positive	0.540	negative	negative	negative	positive
6	negative	0.530	negative	negative	negative	positive
7	negative	0.520	negative	negative	negative	positive
8	positive	0.510	negative	negative	negative	positive
9	negative	0.505	negative	negative	negative	positive
10	positive	0.400	negative	negative	negative	positive
11	negative	0.390	negative	negative	negative	positive
12	positive	0.380	negative	negative	negative	positive
13	negative	0.370	negative	negative	negative	positive
14	negative	0.360	negative	negative	negative	positive
15	negative	0.350	negative	negative	negative	positive
16	positive	0.340	negative	negative	negative	positive
17	negative	0.330	negative	negative	negative	positive
18	positive	0.300	negative	negative	negative	positive
19	negative	0.100	negative	negative	negative	positive



이진 분류 모델의 평가

- Precision-Recall curve: 분류 기준의 변화에 따라 (재현율, 정밀도) 쌍을 표시한 그래프
- Receiver-Operating Characteristic (ROC) curve: (TPR,FPR) 쌍을 표시
  - FPR이 증가함에 따라 TPR은 항상 같거나 증가
  - AUC (Area Under Curve): ROC curve 아래쪽의 면적



예제: 부도 예측

- Output: default
- Input: balance, income
- Data description

	Default: Yes	Default: No	
Train	174	4,826	5,000
Test	159	4,841	5,000

- Model: default ~ balance + income

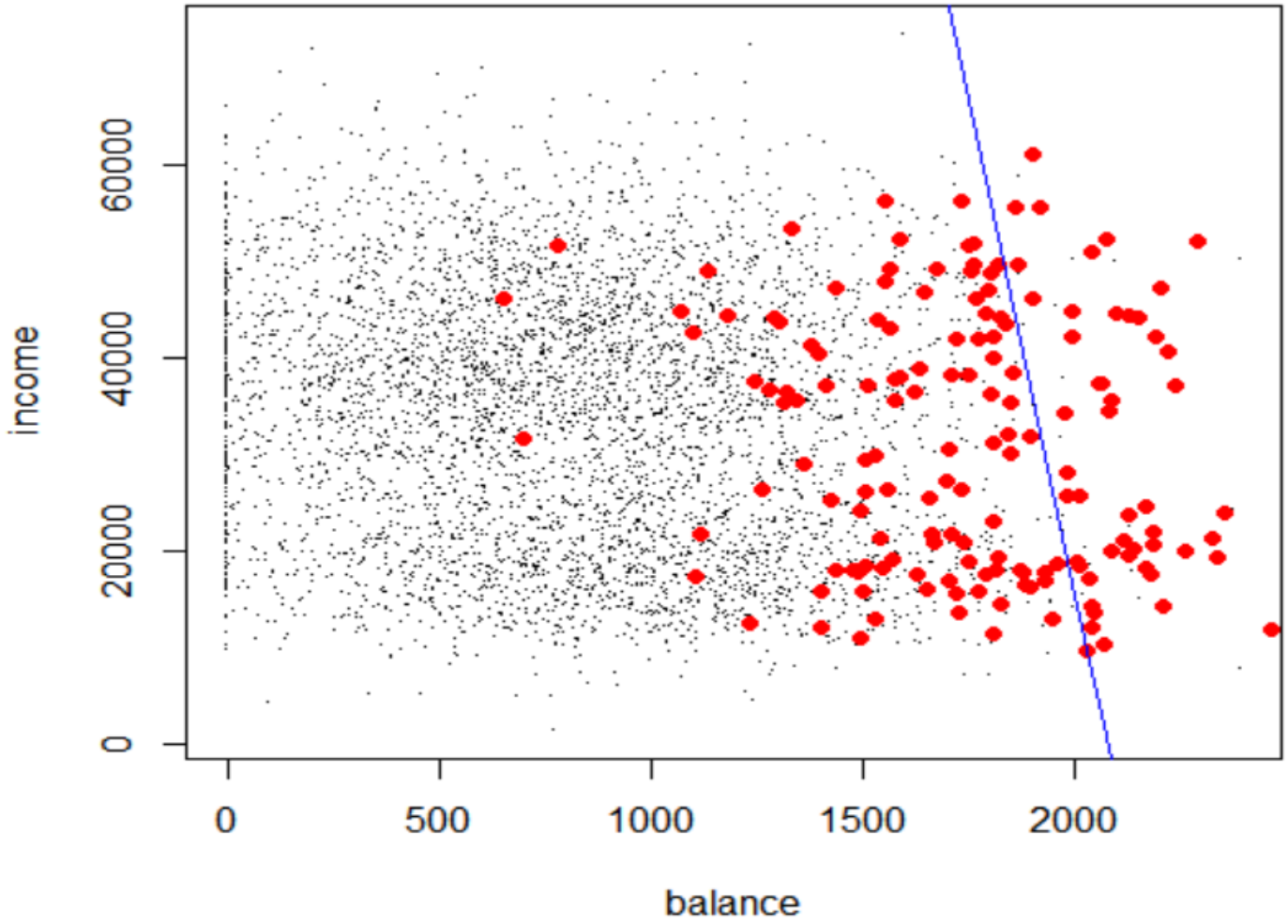
Train Set	Cond. Yes	Cond. No
Pred. Yes	59	19
Pred. No	115	4,807

Test Set	Cond. Yes	Cond. No
Pred. Yes	48	22
Pred. No	111	4,819

	Acc	Recall	Precision	FPR	F1	
Train	0.97	0.34	0.76	0.00	0.47	
Test	0.97	0.30	0.69	0.00	0.42	

예제: 부도 예측

- 모델의 판단 경계 (decision boundary): default ~ balance + income
  - 판단의 경계는 항상 선형 (linear): 선형 모델!!



예제: 부도 예측

- 어느 모델이 더 좋은 모델일까?

	Acc	Recall	Precision	FPR	F1	
모델 1	0.97	0.34	0.51	0.00	0.41	
모델 2	0.97	0.30	0.69	0.00	0.42	

- 질병 예측이라면?
- 서비스 가입 여부라면?

다중 분류 (Multiclass Classification)

- 세 개 분류에 대한 문제는 2개의 이진 분류 문제로 풀 수 있음
- Y = A, B or C 에 대하여, A vs. C와 B vs. C로 나누고 확률을 계산

$$\log\left(\frac{\Pr[Y=A|X]}{\Pr[Y=C|X]}\right) = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2$$

$$\log\left(\frac{\Pr[Y=B|X]}{\Pr[Y=C|X]}\right) = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2$$

$$\Pr[Y=A|X] + \Pr[Y=B|X] + \Pr[Y=C|X] = 1$$

- 아래의 확률 모델과 동일

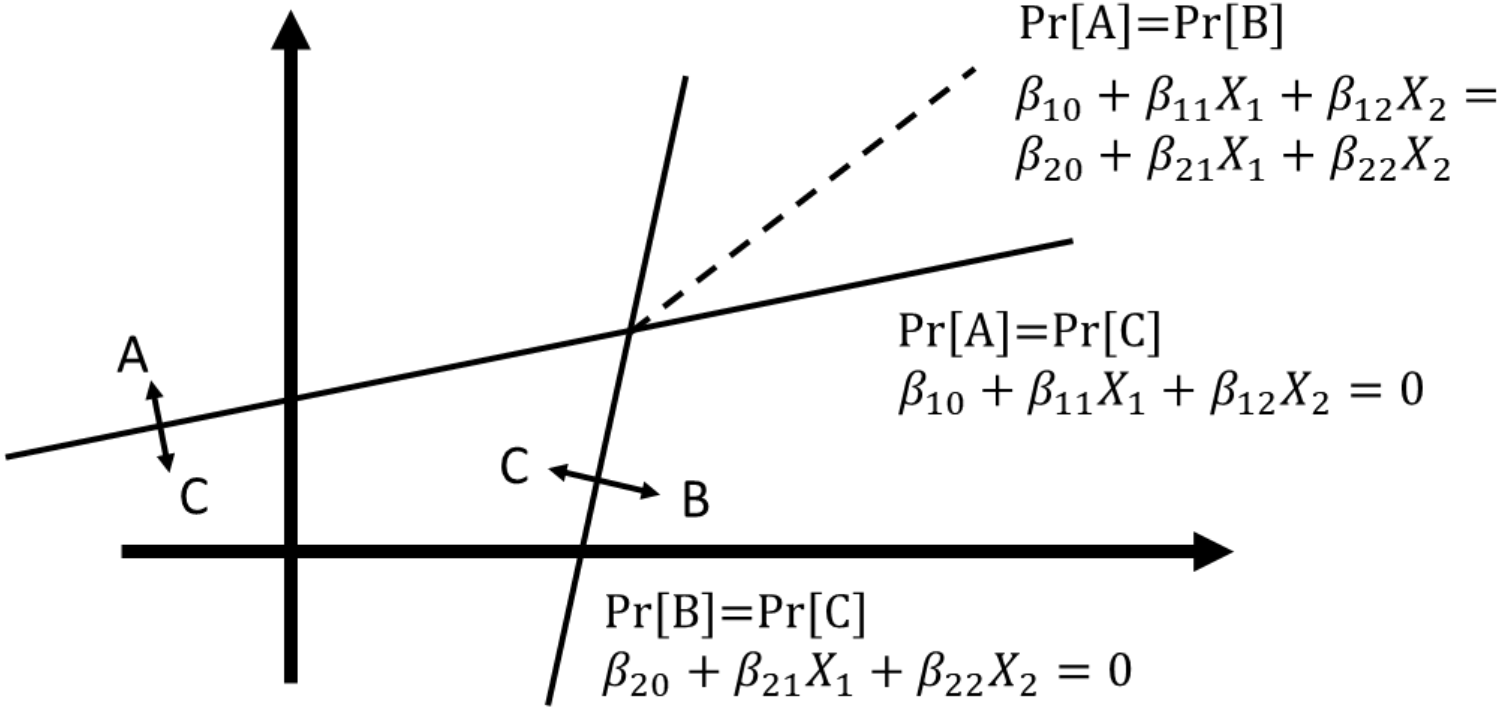
$$\Pr[Y=A|X] = \frac{e^{\beta_{10}+\beta_{11}X_1+\beta_{12}X_2}}{1 + e^{\beta_{10}+\beta_{11}X_1+\beta_{12}X_2} + e^{\beta_{20}+\beta_{21}X_1+\beta_{22}X_2}}$$

$$\Pr[Y=B|X] = \frac{e^{\beta_{20}+\beta_{21}X_1+\beta_{22}X_2}}{1 + e^{\beta_{10}+\beta_{11}X_1+\beta_{12}X_2} + e^{\beta_{20}+\beta_{21}X_1+\beta_{22}X_2}}$$

$$\Pr[Y=C|X] = \frac{1}{1 + e^{\beta_{10}+\beta_{11}X_1+\beta_{12}X_2} + e^{\beta_{20}+\beta_{21}X_1+\beta_{22}X_2}}$$

다중 분류 (Multiclass Classification)

- 전체 표본 공간을 직선(선형)으로 세 개의 구역으로 나누어 분류



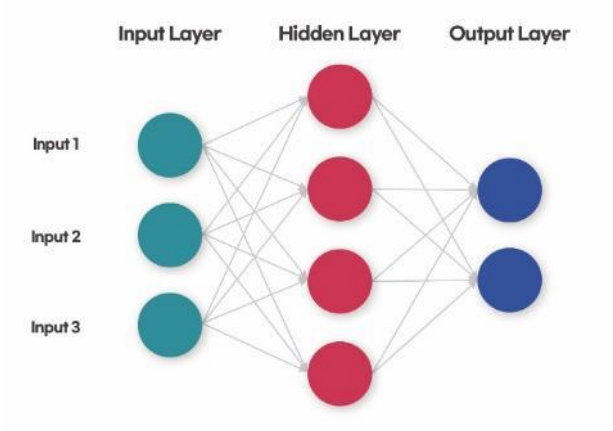
- 다중 분류 모델의 평가
  - 정확도와 혼동 행렬은 그대로 사용 가능
  - 재현율, 정밀도 등은 각 분류별로 계산 가능
    - 예) A, B, C의 경우 A vs. B/C 로 A에 대한 지표 계산

로지스틱 회귀

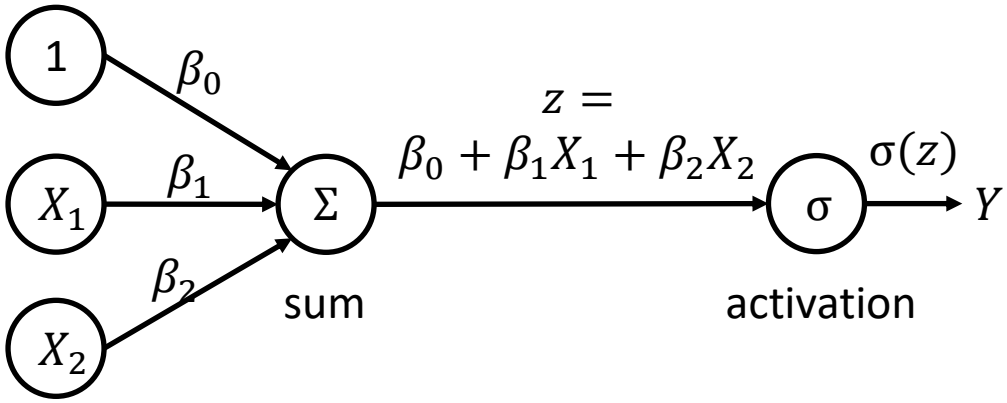
인공신경망으로의 확장

로지스틱 회귀와 퍼셉트론

- 퍼셉트론(perceptron): 인공신경망을 이루는 기본 단위



인공신경망



퍼셉트론

  - 다양한 활성화 함수가 사용될 수 있으나 시그모이드가 일반적으로 사용

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

  - 결국, $Y = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}} = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$

- 로지스틱 회귀와 퍼셉트론은 기본적으로 동일한 모델
  - 로지스틱 회귀는 인공신경망을 이루는 기본적인 단위
  - 인공신경망의 훈련을 위해 일반적으로 크로스 엔트로피 손실을 사용

KOREA UNIVERSITY

감사합니다