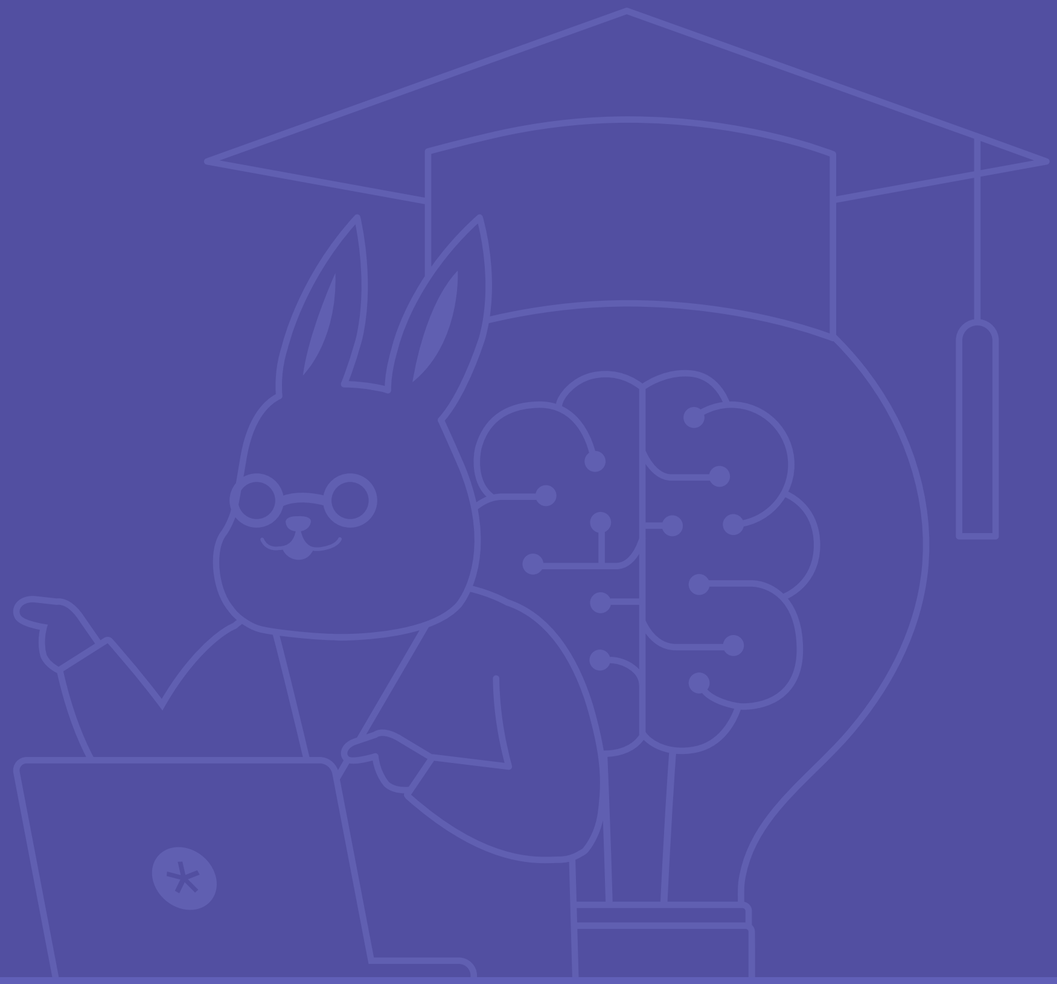
머신러닝 심화

2장 분류



Contents

01. 분류 개념과 로지스틱 회귀

02. Support Vector Machine(SVM)

03. KNN(K-Nearest Neighbor)

04. 나이브 베이즈 분류

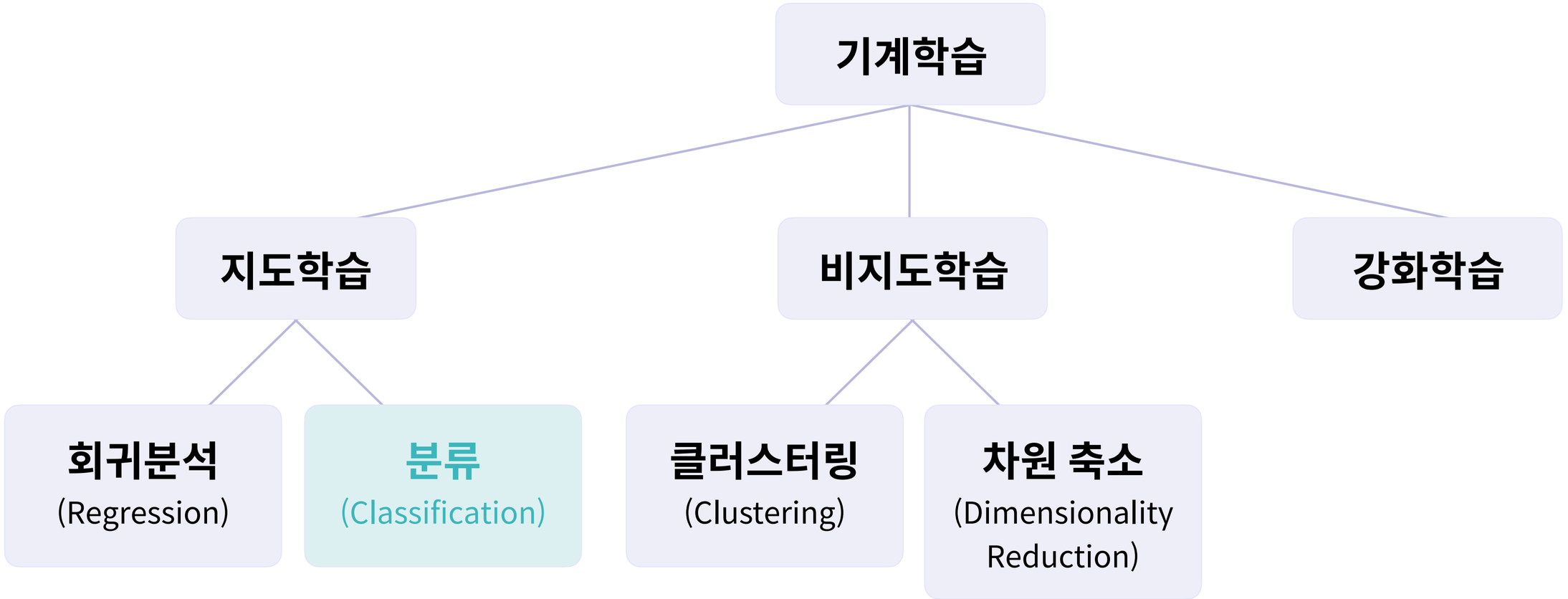
05. 분류 알고리즘 평가 지표(1)

06. 분류 알고리즘 평가 지표(2)

01

분류 개념과 로지스틱 회귀

기계 학습


기계학습
지도학습
비지도학습
강화학습
회귀분석
(Regression)
분류
(Classification)
클러스터링
(Clustering)
차원 축소
(Dimensionality Reduction)

가정해보기

해외 여행을 준비하고 있다고 가정하기

완벽한 여행을 위해 항공 지연을 피하고자 함

기상 정보(구름양, 풍속)를 활용하여
해당 항공의 **지연 여부**를 예측할 수 있다면?

문제 정의와 해결 방안

- 문제 정의

데이터 : **과거 기상 정보(풍속)**와 그에 따른 **항공 지연 여부**
X Y

목표 : 기상 정보에 따른 항공 지연 여부 예측하기

- 해결 방안

분류 알고리즘 활용

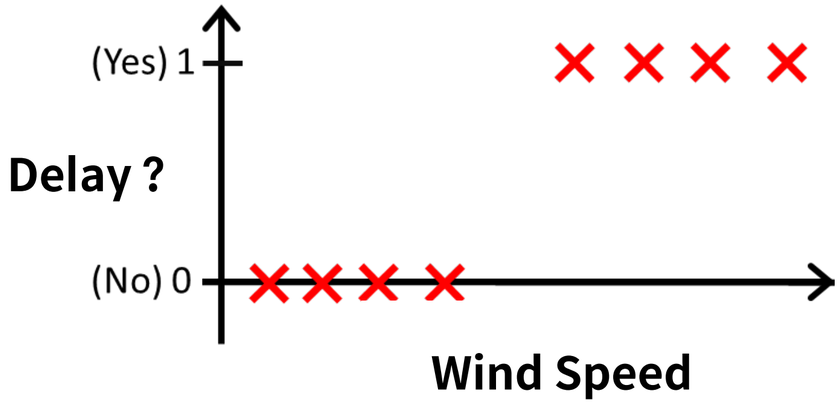
X	Y
풍속(m/s)	지연 여부
2	No
4	Yes
3	No
1	No

분류란?

주어진 입력값이 **어떤 클래스에 속할지**에 대한
결과값을 도출하는 알고리즘

다양한 분류 알고리즘이 존재하며,

예측 목표와 데이터 유형에 따라 적용

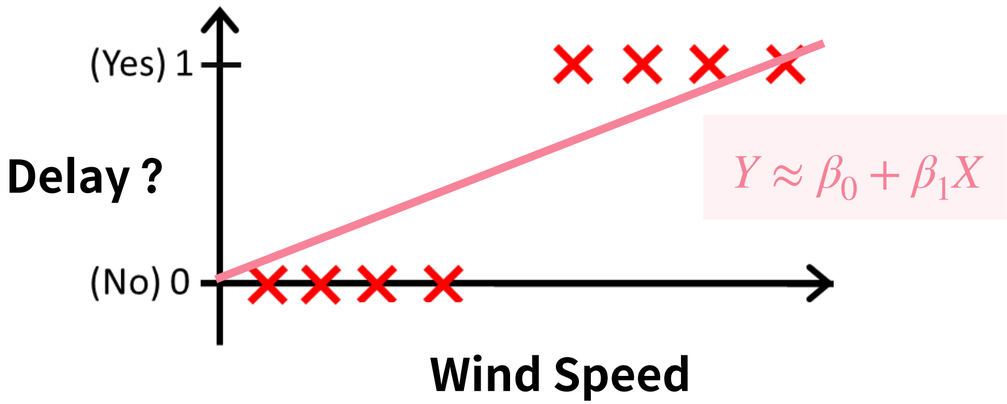


분류 문제에 회귀 알고리즘 적용하기

일반적인 회귀 알고리즘은 분류 문제에
그대로 사용할 수 없다!

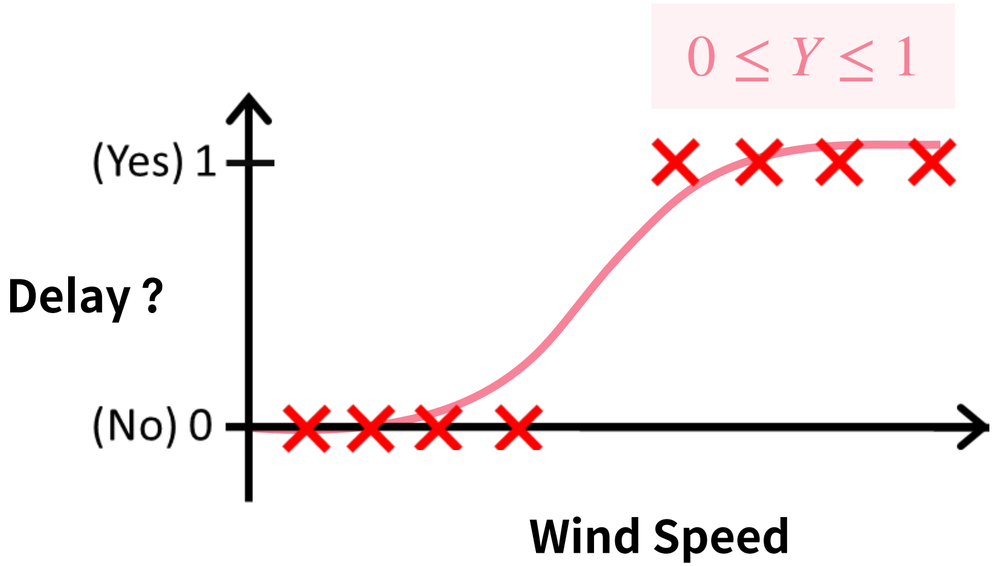
Why?
선형 회귀는 **-∞ ~ +∞**의 값을 가질 수 있음

Q. 우리의 목표는
지연 여부 판별인데 결과값이 1000이라면?



그렇다면 어떻게 해야 할까?

해당 클래스에 속할 확률인
0 또는 1 사이의 값만 내보낼 수 있도록
선형 회귀 알고리즘 수정하기

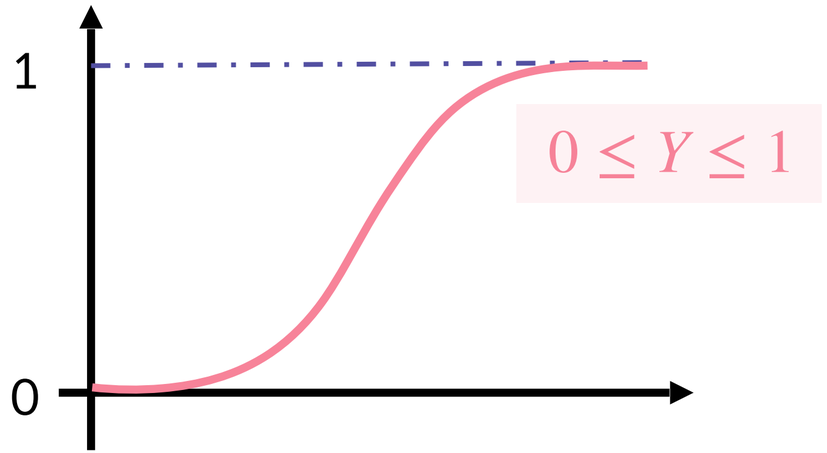


이처럼 분류 문제에 적용하기 위해
출력값의 범위를 수정한 회귀를
로지스틱 회귀(Logistic Regression)라고 함

분류 문제를 위한 회귀, Logistic Regression

이진 분류(Binary Classification) 문제
를 해결하기 위한 모델

최소값 0, 최대값 1로 결과값을 수렴시키기 위해
Sigmoid (logistic) 함수 사용

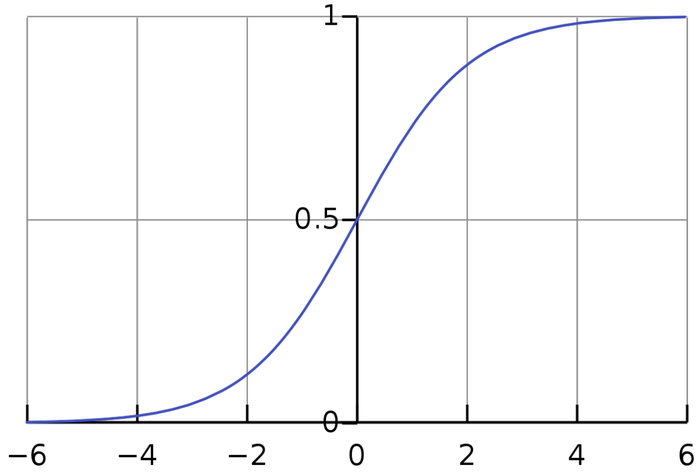


Sigmoid (logistic) 함수

$$g(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{1 + e^{x}}$$

s자형 곡선을 갖는 함수

x 값이 커질 경우 $g(x)$ 값은 점점 1에 수렴하고,
x 값이 작아질 경우 $g(x)$ 값은 점점 0에 수렴함



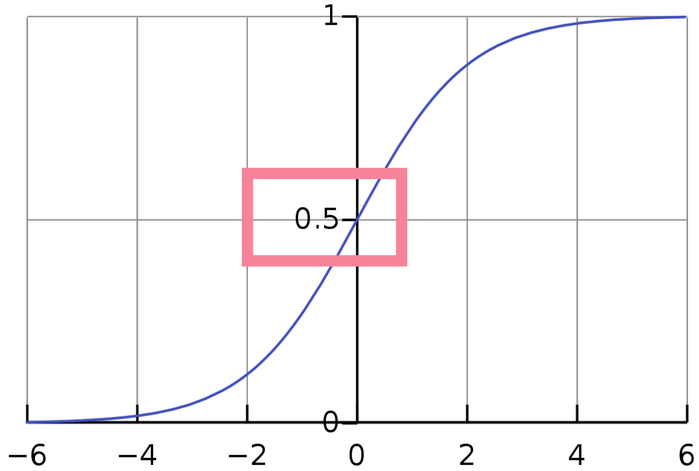
확률 결과값 판별 방법, 결정 경계(Decision Boundary)

결정 경계란, **데이터를 분류하는 기준값**을 의미함

그렇다면, 출력된 확률값을
어떠한 기준으로 클래스에 속한다고 판별해야 할까?

일반적으로 출력값(확률) **0.5**를 기준으로 판별

Sigmoid (logistic) function



로지스틱 회귀 특징

- 주로 2개 값 분류(이진 분류)를 위해 사용
- 선형 회귀를 응용한 분류 알고리즘이기 때문에 선형 회귀의 특징 보유

02

Support Vector Machine(SVM)

문제 2

• 문제 정의

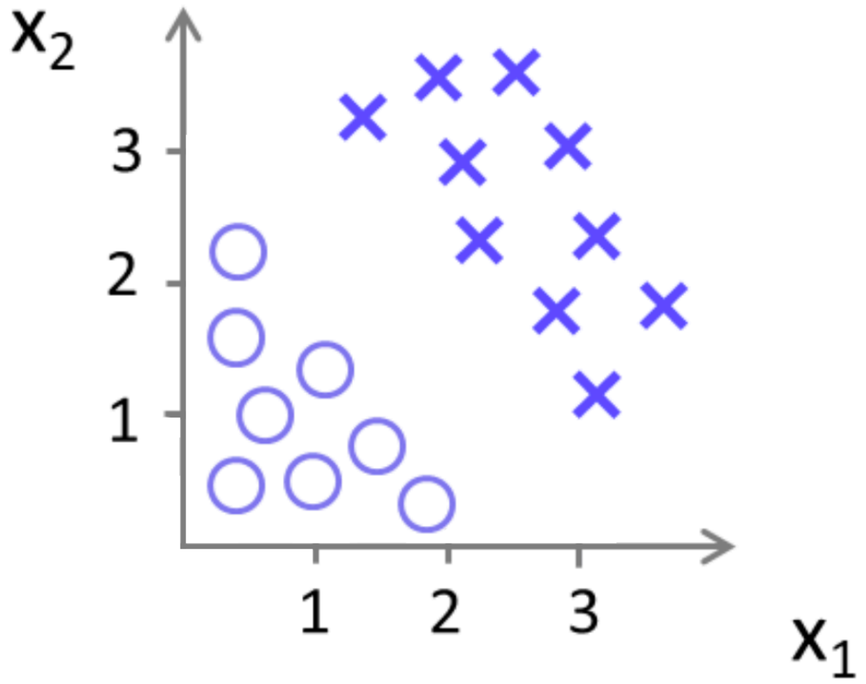
양성(1)과 음성(0)

두 개의 결과 값으로 분류되는 이진분류 문제

ex. 지연 여부 판별, 이상 거래 판별

• 해결 방안

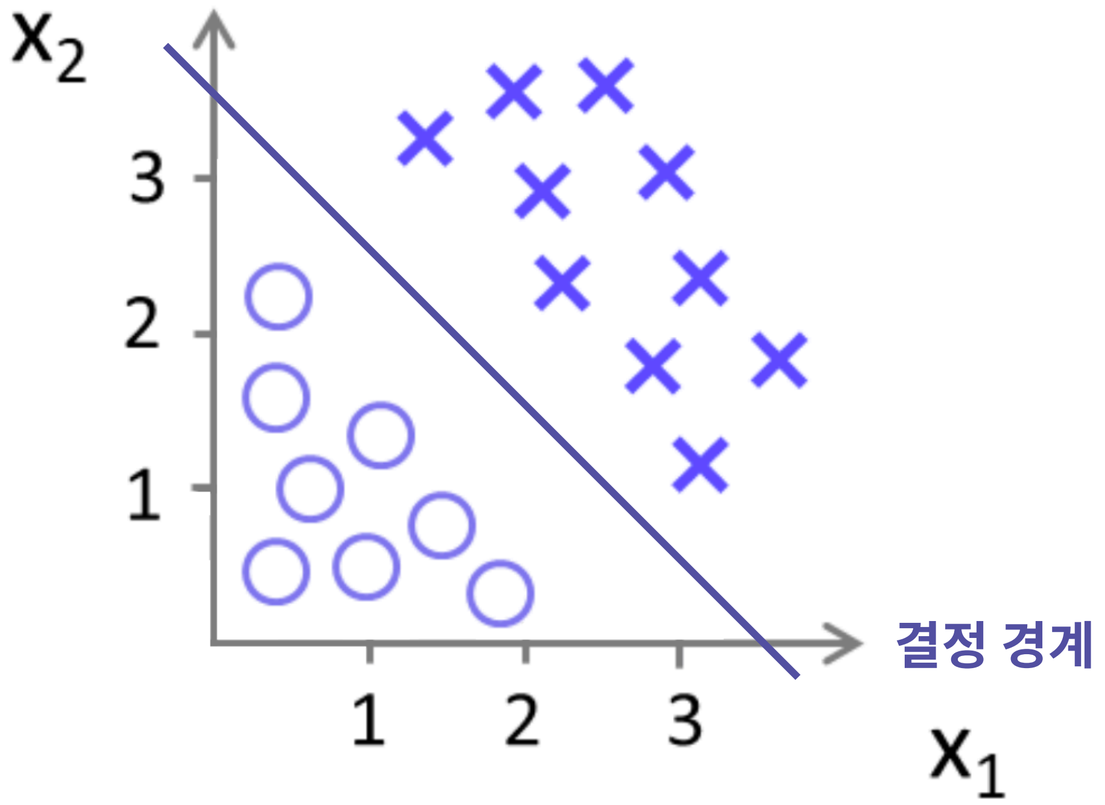
SVM(Support Vector Machine) 분류 알고리즘



SVM(Support Vector Machine)

딥러닝 기술 등장 이전까지
가장 인기 있던 분류 알고리즘

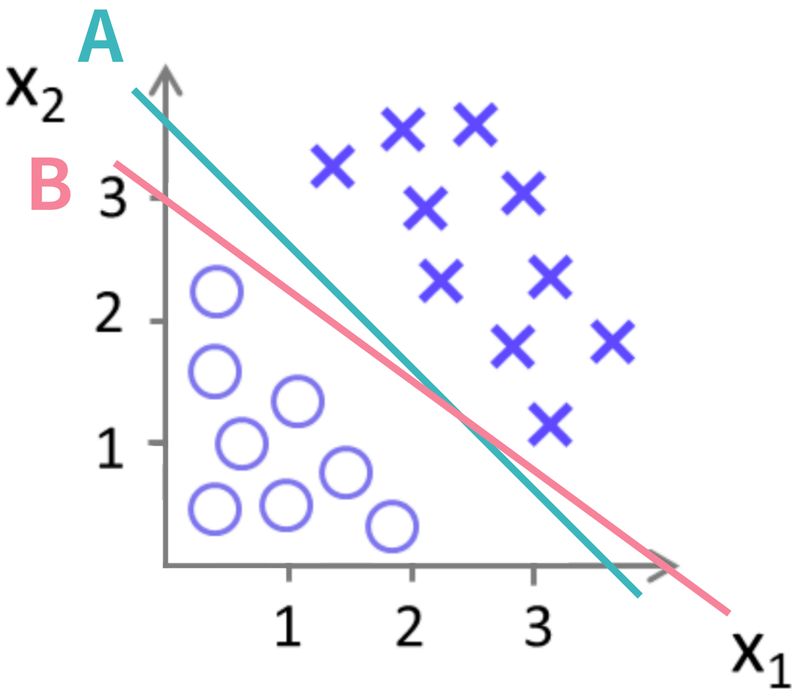
최적의 결정 경계(Decision Boundary)
즉, 데이터를 분류하는 기준 선을 정의하는 모델



최적의 결정 경계(Decision Boundary)

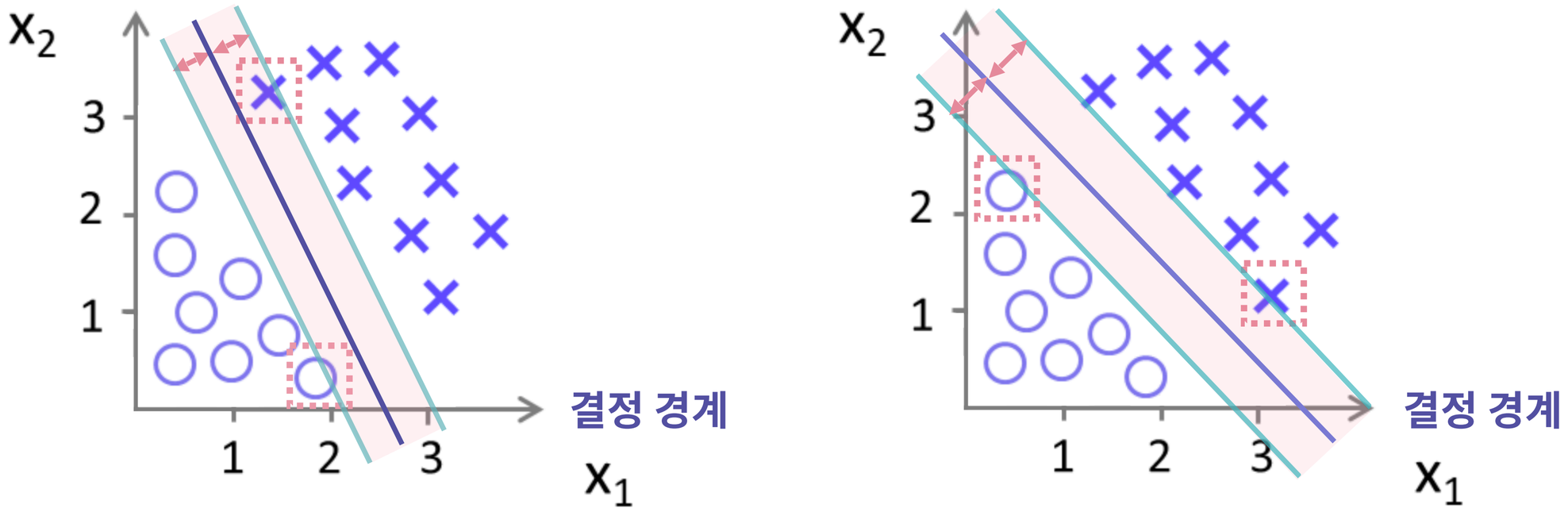
최적의 결정 경계는
데이터 군으로부터 **최대한 멀리 떨어지는 것**

Q. A와 B 중 더 최적의 결정 경계는?
정답은 **A!** 데이터로부터 최대한 멀리 떨어져서 분류하기 때문



서포트 벡터(Support Vector)

결정 경계와 가장 가까이 있는 데이터 포인트들

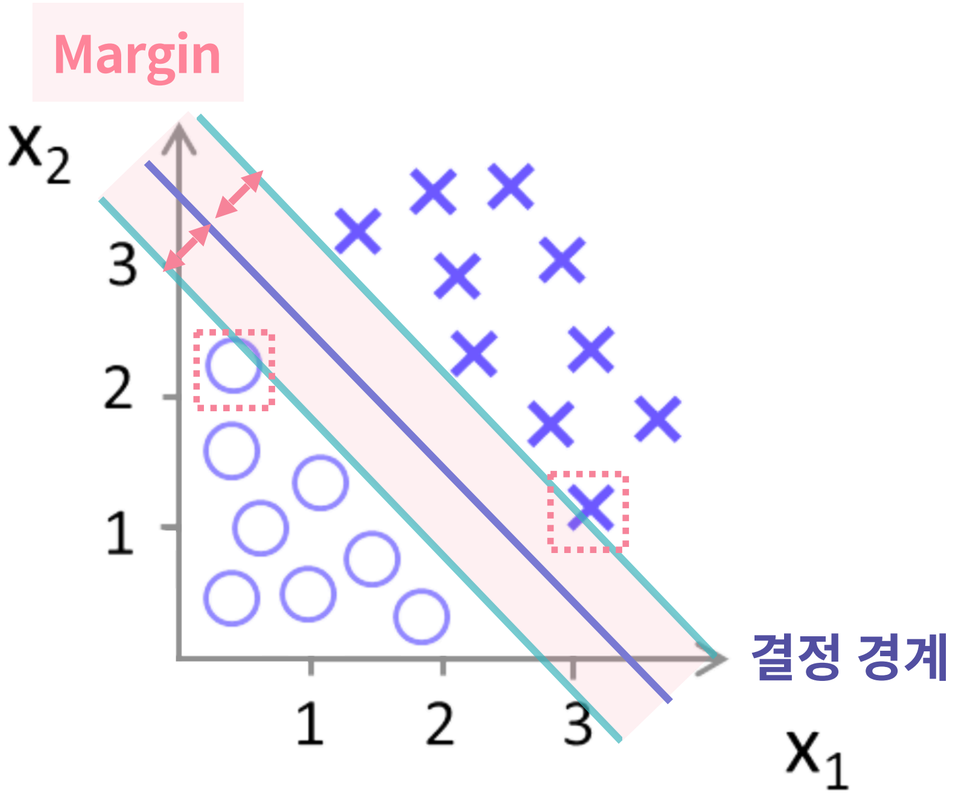


결정경계 여유? Margin

클래스를 분류하는 기준선에
여유(Margin)를 둘 수 있다

여유(Margin)
= 결정 경계와 서포트 벡터 사이의 거리

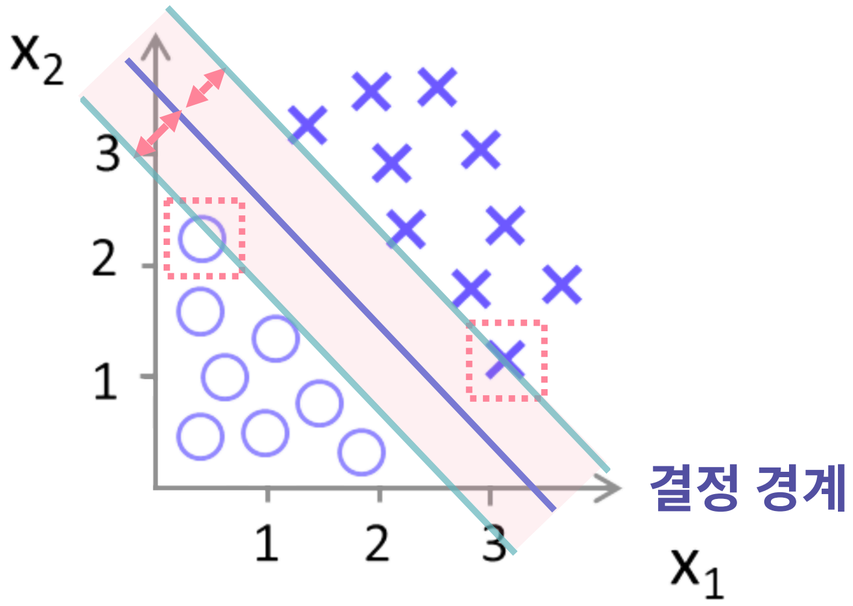
Margin을 최대화 하는 결정 경계를 찾음



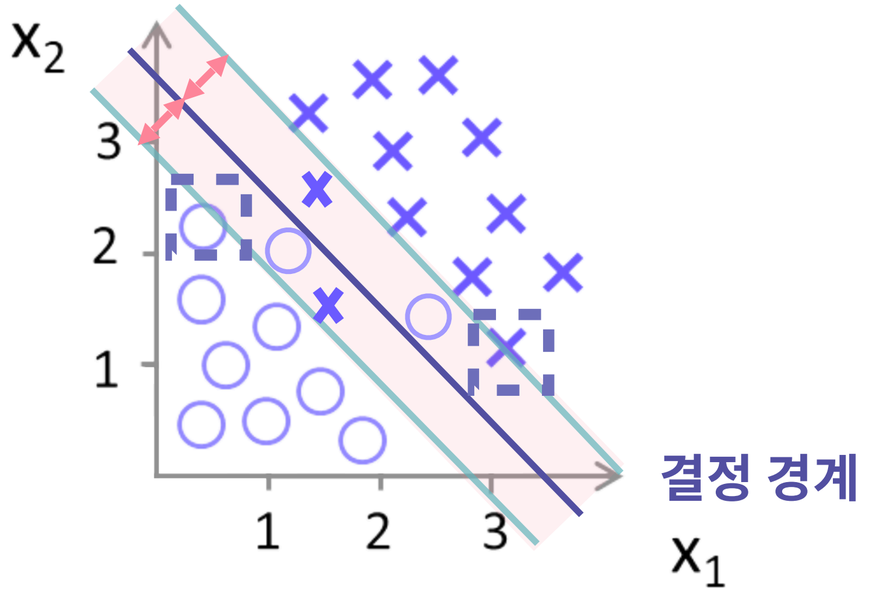
Hard Margin vs Soft Margin

이상치(Outlier) 허용 범위에 따라 Hard Margin과 Soft Margin으로 구분됨

Hard Margin



Soft Margin



02 SVM(Support Vector Machine)

SVM 특징

- 선형 분류와 비선형 분류 모두 가능
- 고차원 데이터에서도 높은 성능의 결과를 도출
- 회귀에도 적용 가능

03

나이브 베이즈 분류

문제 정의와 해결 방안

- 문제 정의

10만개의 메일 중
스팸 메일과 정상 메일을 분류하고 싶다면?

*__메일__ : 독립 사건으로 가정하는 텍스트 데이터

- 해결 방안

나이브 베이즈 분류 알고리즘 활용

나이브 베이즈 분류(Naïve Bayes Classification)

각 특징들이 독립적 즉, 서로 영향을 미치지 않을 것이라는 가정 설정
베이즈 정리(Bayes Rule)를 활용한 확률 통계학적 분류 알고리즘

베이즈 정리

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

$P(A \mid B)$: 사건 B가 발생했을 때, A도 같이 발생했을 확률

베이즈 정리

맑은 날 비가 오지 않을 확률은?

P(비가 안옴|맑은 날)

$$= \frac{P(\text{맑은 날}|\text{비가 안옴}) * P(\text{비가 안옴})}{P(\text{맑은 날})}$$

	비가 옴	비가 안옴	
맑은 날	2	8	10
흐린 날	5	5	10
	7	13	20

베이즈 정리

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

문제 적용

스팸 메일 분류

- 1. 스팸 메일과 정상 메일의 단어를 체크
- 2. 새로운 메일의 단어들에 대한 확률로 스팸 메일을 구분

$P(\text{스팸} \mid \text{단어}1, \text{단어}2, \text{단어}3\ldots) > P(\text{정상} \mid \text{단어}1, \text{단어}2, \text{단어}3\ldots)$ 이면 스팸

03 나이브 베이즈 분류

나이브 베이즈 분류 원리

베이즈 정리를 활용하여 입력값이 해당 클래스에 속할 확률을 계산하여 분류

스팸 메일 분류 예시

- $P(\text{스팸} \mid \text{단어}1, \text{단어}2, \text{단어}3\ldots) > P(\text{정상} \mid \text{단어}1, \text{단어}2, \text{단어}3\ldots)$ 이면 스팸

나이브 베이즈 분류 특징

- 각 특징들이 독립이라면 다른 분류 방식에 비해 결과가 좋고, 학습 데이터도 적게 필요
- 각 특징들이 독립이 아니라면 즉, 특징들이 서로 영향을 미치면 분류 결과 신뢰성 하락
- 학습 데이터에 없는 범주의 데이터일 경우 정상적 예측 불가능

04

KNN(K-Nearest Neighbor)

04 KNN(K-Nearest Neighbor)

문제 정의와 해결 방안

- 문제 정의

고객이 평가한 영화 평점 데이터를 기준으로
기존 보유 고객을 분류한 이후 새로 유입된 고객을
기준에 따라 분류하고자 하는 경우

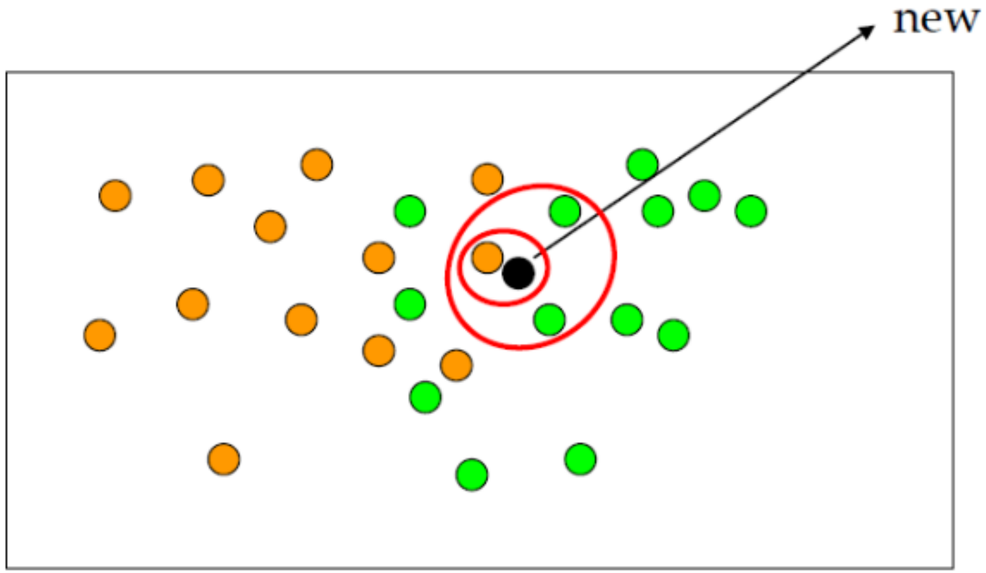
- 해결 방안

KNN(k-Nearest Neighbor) 알고리즘

KNN(K-Nearest Neighbor)

기존 데이터 가운데
가장 가까운 k개 이웃의 정보로
새로운 데이터를 예측하는 방법론

유사한 특성을 가진 데이터는
유사 범주에 속하는 경향이 있다는 가정 하에 분류

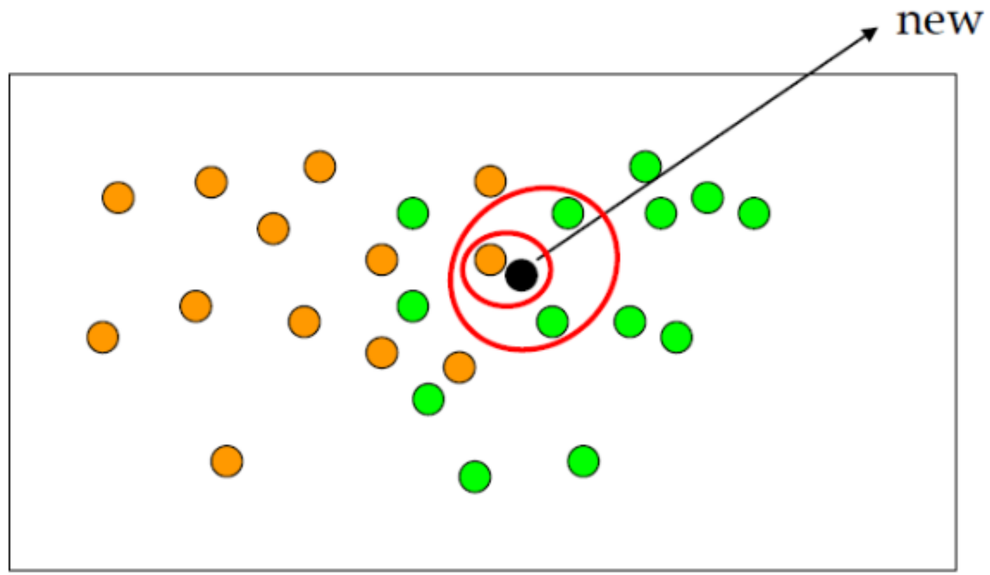


04 KNN(K-Nearest Neighbor)

KNN(K-Nearest Neighbor) 원리

설정된 K값에 따라
가까운 거리 내의 이웃의 수에 따라 분류

새로운 고객 데이터(검정색)이 들어왔을 때
만약
K=1 이면 **주황색 클래스**로 분류
K=3 이면 **초록색 클래스**로 분류



04 KNN(K-Nearest Neighbor)

KNN(K-Nearest Neighbor) 특징

- 직관적이며 복잡하지 않은 알고리즘, 결과 해석이 쉬움
- K값 결정에 따라 성능이 크게 좌우됨
- 딱히 학습이랄 것이 없는 Lazy Model

05

분류 알고리즘 평가 지표(1)

05 분류 알고리즘 평가 지표(1)

혼동 행렬(Confusion Matrix)

분류 모델의 성능을 평가하기 위함

True Positive, True Negative, False Positive, False Negative

		예측	
		Positive	**Negative**
실제	**Positive**	True Positive	False Negative
	Nagative	False Positive	True Negative

05 분류 알고리즘 평가 지표(1)

True Positive, True Negative

True Positive : 실제 **Positive** 인 값을 **Positive** 라고 예측 (정답)
True Negative : 실제 **Negative** 인 값을 **Negative** 라고 예측(정답)

False Positive : 실제 **Negative** 인 값을 **Positive** 라고 예측 (오답) - **1형 오류**
False Negative : 실제 **Positive** 인 값을 **Negative** 라고 예측 (오답) - **2형 오류**

05 분류 알고리즘 평가 지표(1)

예시

전체 100개 항공기 관련 정보를 활용하여 지연 여부 예측을 실시했을 때 결과

실제 결과	예측 결과
지연 O : **20**개 지연 X : **80**개	지연 O : **60**개 지연 X : **40**개

05 분류 알고리즘 평가 지표(1)

TP, FN, FP, TN

		예측	
		Positive	**Negative**
실제	**Positive**	True Positive : 20개	False Negative : 0개
	Negative	False Positive : 40개	True Negative : 40개

05 분류 알고리즘 평가 지표(1)

정확도(Accuracy)

전체 데이터 중에서 제대로 분류된 데이터의 비율로,
모델이 얼마나 정확하게 분류하는지를 나타냄

일반적으로 분류 모델의 주요 평가 방법으로 사용됨

그러나, 클래스 비율이 **불균형**할 경우
평가 지표의 신뢰성을 잃음

$$Accuracy = \frac{TP + TN}{P + N}$$

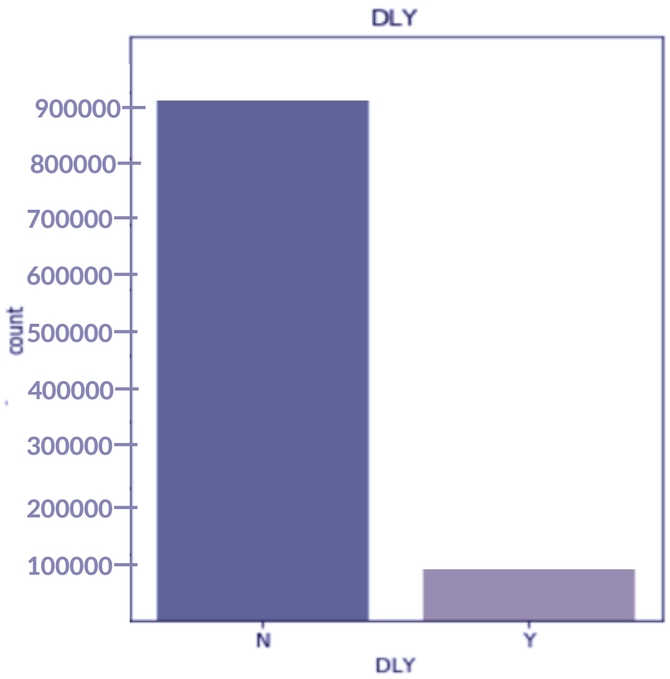
05 분류 알고리즘 평가 지표(1)

불균형한 클래스에서의 정확도

만약 전체 100만개 항공 데이터 중
90만개가 정상 운행,
10만개만이 지연 운행인 데이터를 예측하고자 할 때,

분류 모델 A가 전체 결과가 모두 지연되지 않았다고 예측할 경우

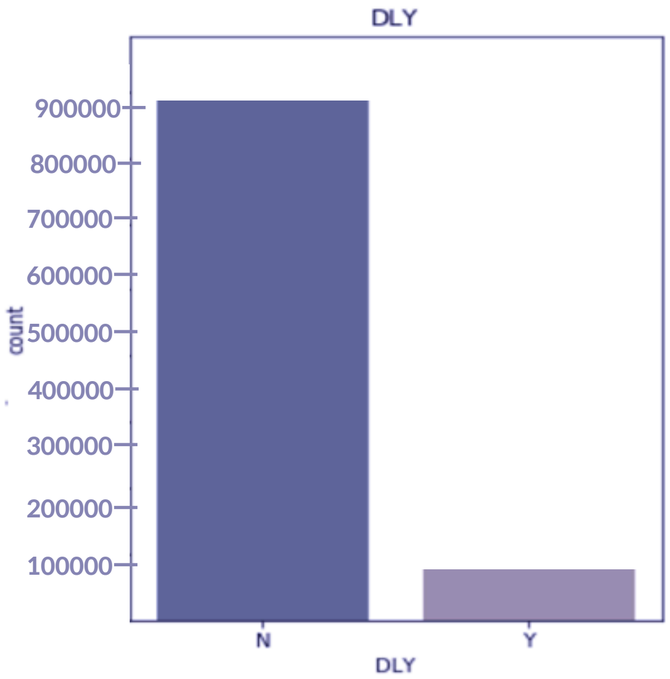
$$\text{정확도} = \frac{0\text{개} + 90\text{만개}}{100\text{만개}} = 90\%$$



05 분류 알고리즘 평가 지표(1)

불균형한 클래스에서의 정확도

$$\text{정확도} = \frac{0\text{개} + 90\text{만개}}{100\text{만개}} = 90\%$$



수치 상으로는 맞으나,
해당 모델에 대한 정확도가 90% 라고 단정하기에는 위험함
다양한 평가 지표 고려 필요성

06

분류 알고리즘 평가 지표(2)

정밀도(Precision)

모델이 Positive라고 분류한 데이터 중에서
실제로 Positive인 데이터의 비율

$$Precision = \frac{TP}{TP + FP}$$

Negative가 중요한 경우

즉, 실제로 Negative인 데이터를
Positive라고 판단하면 안되는 경우 사용되는 지표

06 분류 알고리즘 평가 지표(2)

Negative가 중요한 경우

스팸 메일 판결을 위한 분류 문제

해당 메일이 스팸일 경우 **Positive,**
스팸이 아닐 경우 즉, 일반 메일일 경우 **Negative**

일반 메일을 **스팸 메일(Positive)로 잘못 예측**했을 경우
중요한 메일을 전달받지 못하는 상황이 발생할 수 있음

06 분류 알고리즘 평가 지표(2)

재현율(Recall, TPR)

실제로 Positive인 데이터 중에서
모델이 Positive로 분류한 데이터의 비율

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Positive가 중요한 경우

즉, 실제로 Positive인 데이터를
Negative라고 판단하면 안되는 경우 사용되는 지표

06 분류 알고리즘 평가 지표(2)

Positive가 중요한 경우

악성 종양 여부 판결을 위한 검사

악성 종양일 경우 **Positive,**
악성 종양이 아닐 경우 즉, 양성 종양일 경우 **Negative**

악성 종양(Positive)을 **양성 종양(Negative)으로 잘못 예측**했을 경우
제 때 치료를 받지 못하게 되어 생명이 위급해질 수 있음

06 분류 알고리즘 평가 지표(2)

FPR(False Positive Rate)

실제로 Negative인 데이터 중에서
모델이 Positive로 분류한 데이터의 비율

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

06 분류 알고리즘 평가 지표(2)

FPR 지표와 비정상 사용자 검출 예시

- 게임에서 비정상 사용자 검출 시 **FPR이 높다.**
  = 정상 사용자를 **비정상 사용자**로 검출하는 경우가 많다.
- 이 때 비정상 사용자에 대해서 계정정지 등 페널티를 부여할 경우 선의의 사용자가 피해를 입게 될 가능성이 높음

→ 이는 곧 게임에 대한 충성도를 떨어트리는 계기가 될 수 있음

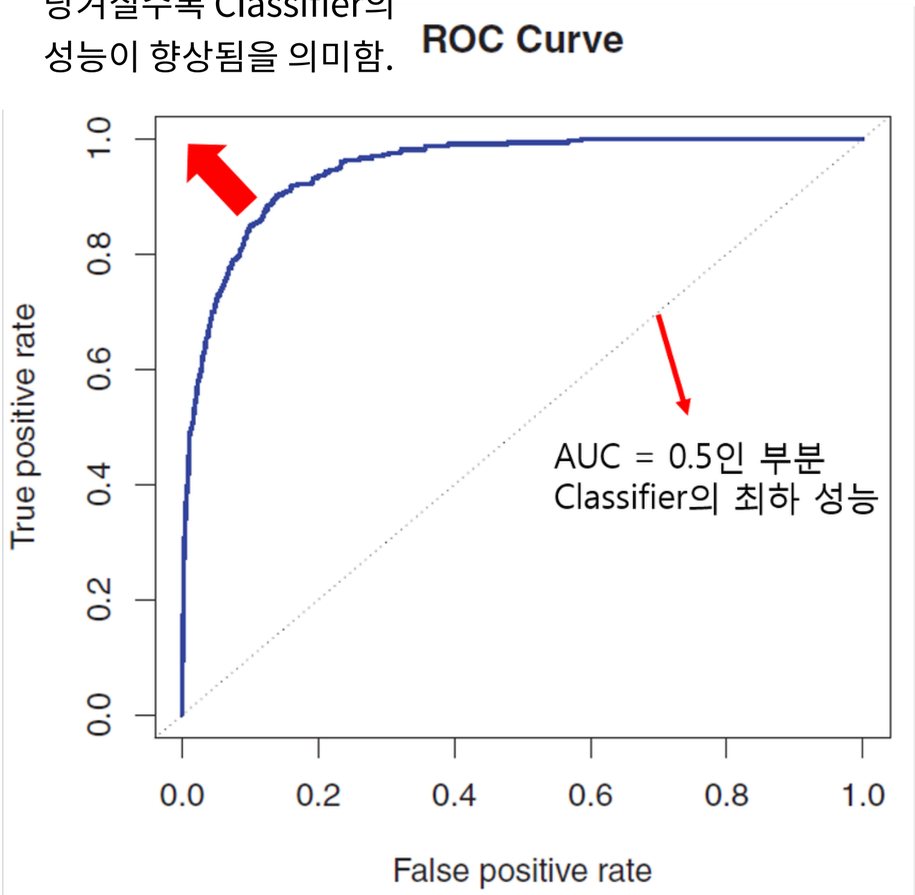
06 분류 알고리즘 평가 지표(2)

ROC Curve와 AUC

x축을 **False Positive Rate**
y축을 **Recall(True Positive Rate)**
로 두고 시각화한 그래프

ROC Curve 아래 면적인
AUC(Area Under Curve)를
이용해 모델의 성능을 평가

화살표 쪽으로 커브가
당겨질수록 Classifier의
성능이 향상됨을 의미함.



06 분류 알고리즘 평가 지표(2)

평가 지표 선정 방법

상황에 따라 선정해야 하는 평가 지표가 다르므로
다양한 평가 지표를 적용하여 **결과 비교**해보기

Contact

TEL

070-4633-2015

WEB

https://elice.io

E-MAIL

contact@elice.io