

Chapter. 21

많다고 좋은게 아니다: 차원의 저주 문제

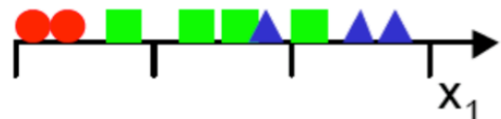
| 문제 정의 및 해결 방안

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

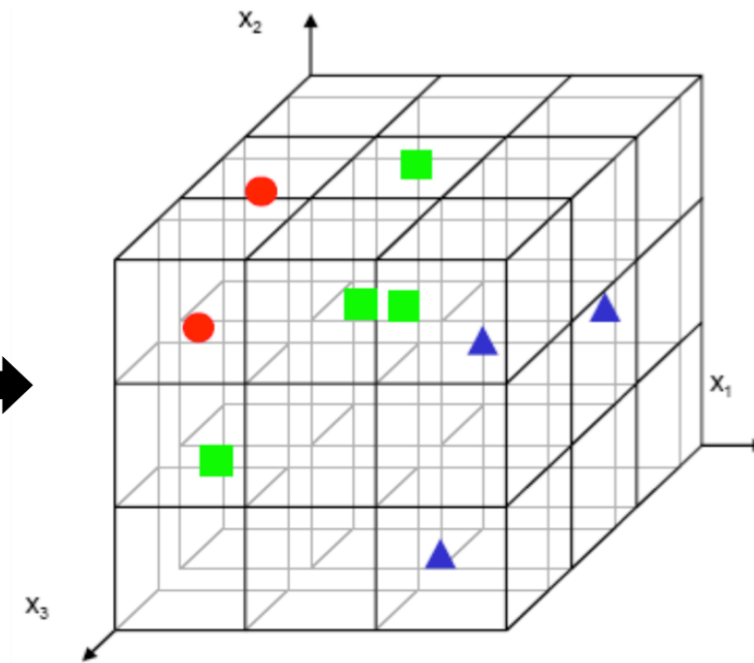
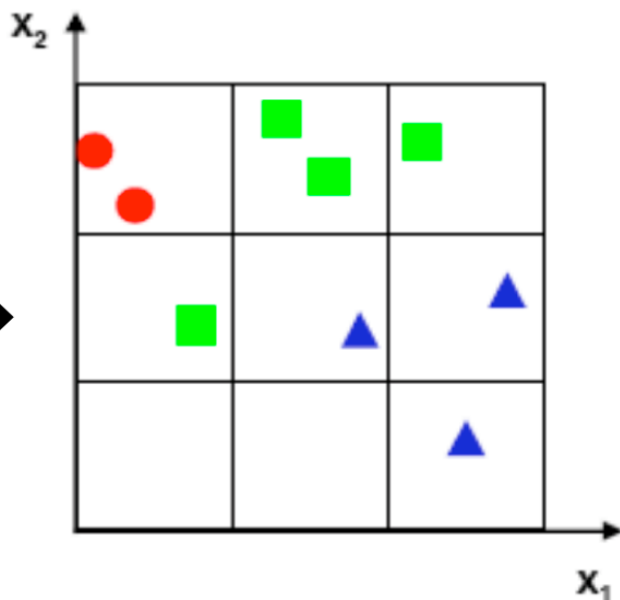
강사. 안길승

I 문제 정의

- 차원이 증가함에 따라 필요한 **데이터의 양**과 **시간 복잡도**가 기하급수적으로 **증가**하는 문제를 의미함



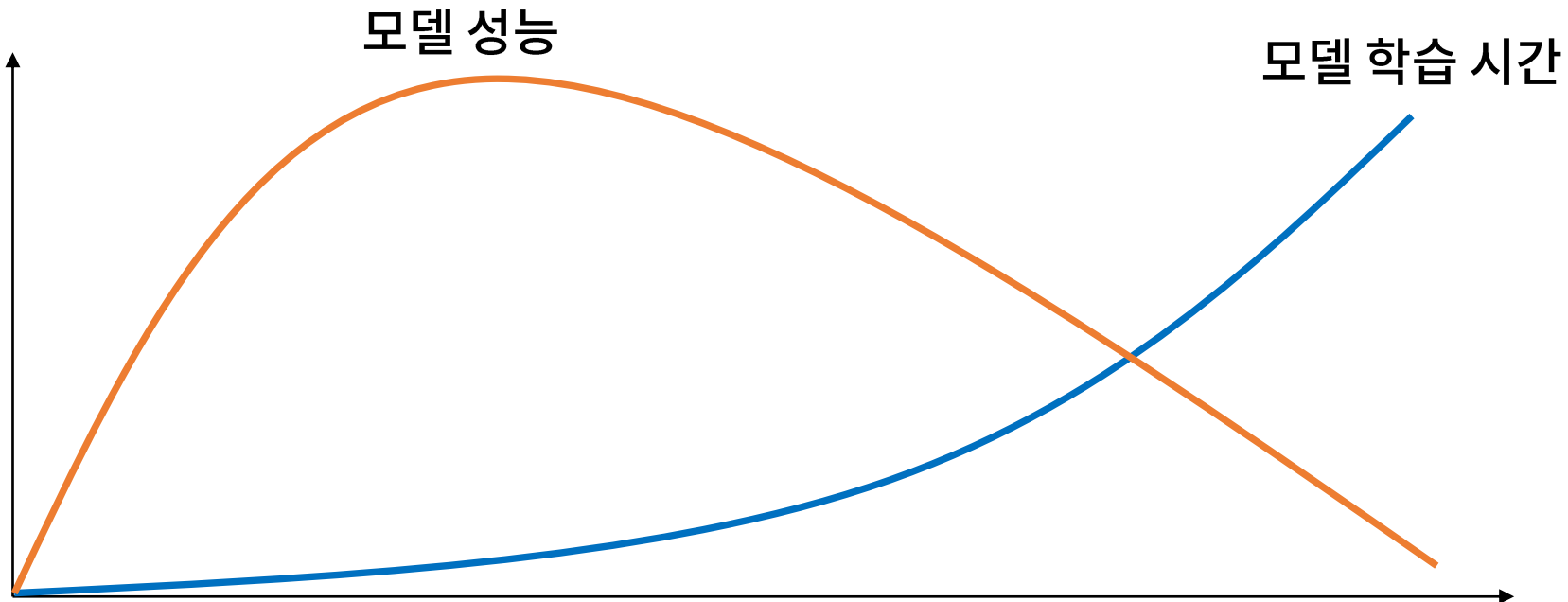
빈 공간 없음



빈 공간 발생
(과적합 가능성)

I 차원을 줄여야 하는 이유

- 차원이 증가함에 따라 모델 학습 시간이 정비례하게 증가함
- 차원이 증가하면 **각 결정 공간에 포함되는 샘플 수가 적어져**, 과적합으로 이어져 성능 저하가 발생할 수 있음



Chapter. 21

많다고 좋은게 아니다: 차원의 저주 문제

| 특징 선택

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

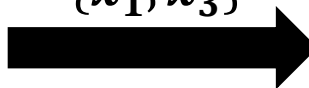
강사. 안길승

I 개요

- 분류 및 예측에 **효과적인 특징만 선택**하여 차원을 축소하는 방법
- n 개의 특징으로 구성된 특징 집합 $\{x_1, x_2, \dots, x_n\}$ 에서 $m \ll n$ 개의 특징을 선택하여, 새로운 특징 집합 $\{x'_1, x'_2, \dots, x'_m\} \subset \{x_1, x_2, \dots, x_n\}$ 을 구성하는 방법

x_1	x_2	x_3	x_4	y
$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	$x_4^{(1)}$	$y^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$	$x_4^{(2)}$	$y^{(2)}$
$x_1^{(3)}$	$x_2^{(3)}$	$x_3^{(3)}$	$x_4^{(3)}$	$y^{(3)}$
$x_1^{(3)}$	$x_2^{(4)}$	$x_3^{(4)}$	$x_4^{(4)}$	$y^{(4)}$
\vdots	\vdots	\vdots	\vdots	\vdots
$x_1^{(N)}$	$x_2^{(N)}$	$x_3^{(N)}$	$x_4^{(N)}$	$y^{(N)}$

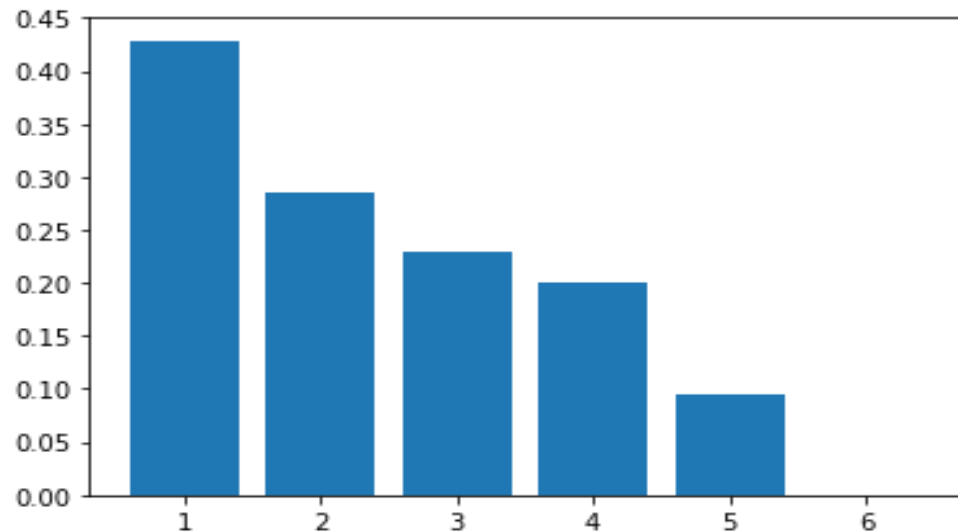
특징 선택
 $\{x_1, x_3\}$



x_1	x_3	y
$x_1^{(1)}$	$x_3^{(1)}$	$y^{(1)}$
$x_1^{(2)}$	$x_3^{(2)}$	$y^{(2)}$
$x_1^{(3)}$	$x_3^{(3)}$	$y^{(3)}$
$x_1^{(3)}$	$x_3^{(4)}$	$y^{(4)}$
\vdots	\vdots	\vdots
$x_1^{(N)}$	$x_3^{(N)}$	$y^{(N)}$

I 적용 대상

- 특징 선택은 특징이 많은 데이터에만 적용해야 한다? **그렇지 않다!**
- (예제) 특징이 7개인 데이터에 대해, 모든 특징 집합을 비교해보기
 - 모든 특징을 사용했을 때의 성능 (f1-score): 0.3333
 - 특징을 두 개만 쓰는 경우에 가장 좋은 성능을 보였음 ({At1, At6}: 0.5714)



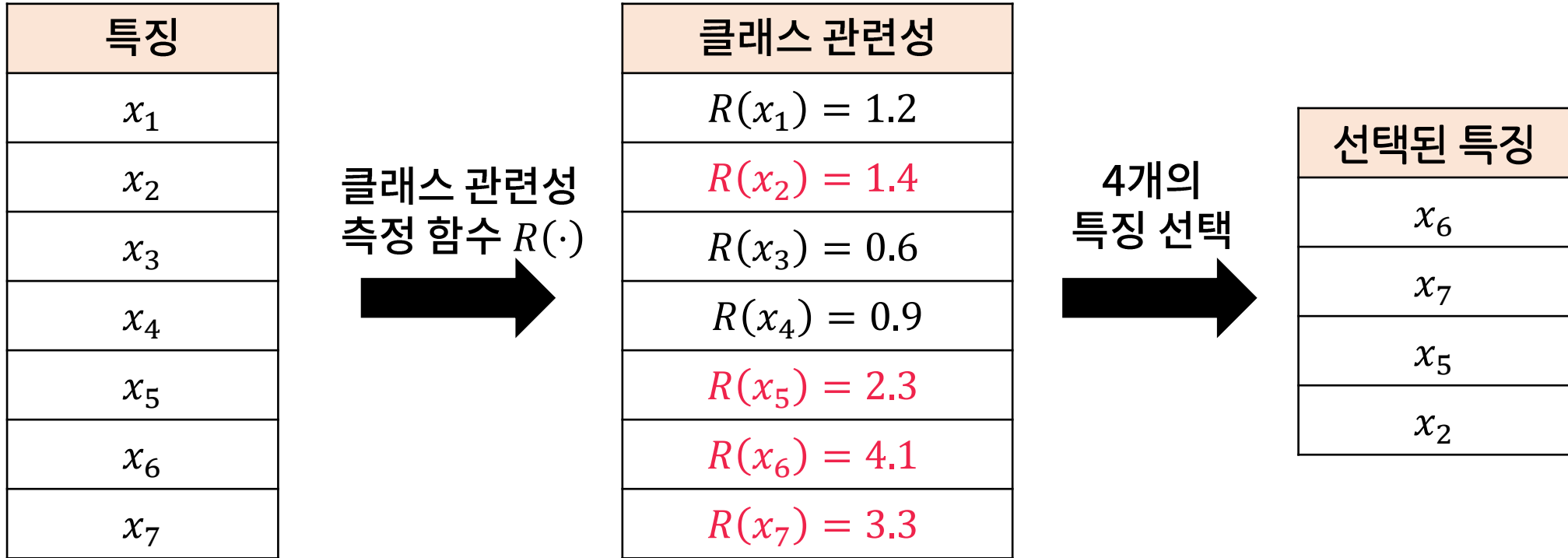
모든 특징을 썼을 때보다 성능이 좋았던 특징 집합의 비율

I 주먹구구식 특징 선택

- 선택 가능한 모든 특징 집합을 비교/평가하여 **가장 좋은 특징 집합**을 선택하는 방법
- 그러나 특징 개수가 n 개라면, $2^n - 1$ 번의 모형 학습이 필요하므로, 현실적으로 적용 불가능함
 - **1초에 1억 번의 모형**을 학습할 수 있는 슈퍼 컴퓨터가, **1000개의 특징**이 있는 데이터에 대해, 이 방법을 적용하여 가장 좋은 특징 집합을 선택하는데 소요되는 시간은 **400조년**

I 필터링 기반의 특징 선택

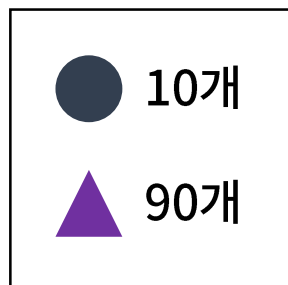
- 특징과 라벨이 얼마나 관련이 있는지를 나타내는 **클래스 관련성이 높은 특징**을 우선 선택하는 방법



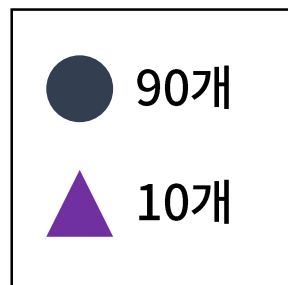
I 클래스 관련성

- 클래스 관련성 (class relevance)란, 한 특징이 클래스를 얼마나 잘 설명하는지를 나타내는 척도로, 상관계수, 카이제곱 통계량, 상호정보량 등의 특징과 라벨 간 독립성을 나타내는 통계량을 사용하여 측정
- 즉, 클래스 관련성이 높은 특징은 분류 및 예측에 도움이 되는 특징이며, 그렇지 않은 특징은 도움이 되지 않는 특징임
- (예시) 범주형 특징 - 분류

클래스 관련성이 높은 이진형 특징 x_1

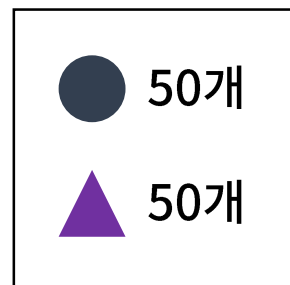


$x_1 = 1$ 인 샘플의
클래스 변수의 분포

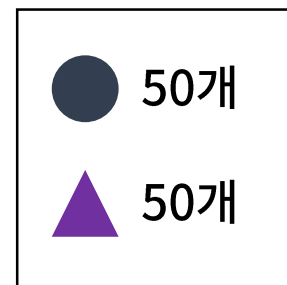


$x_1 = 0$ 인 샘플의
클래스 변수의 분포

클래스 관련성이 낮은 이진형 특징 x_2



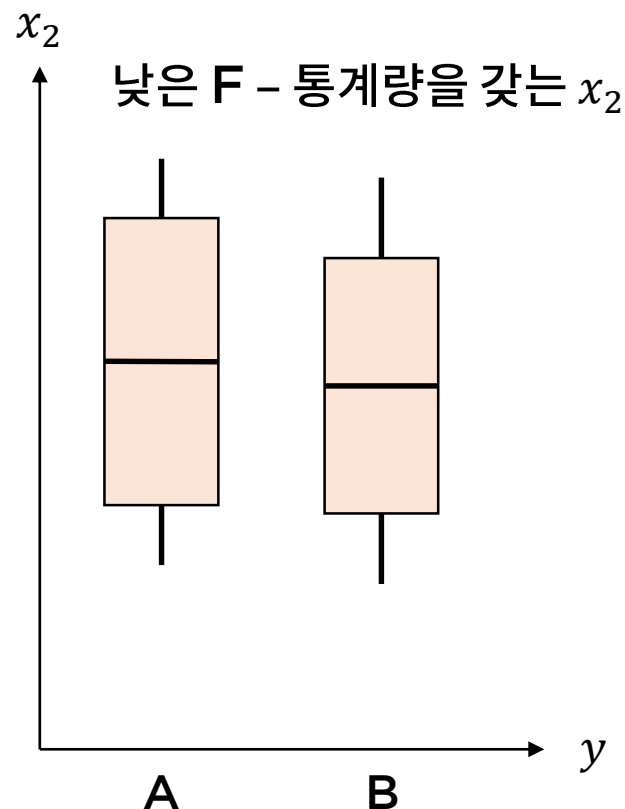
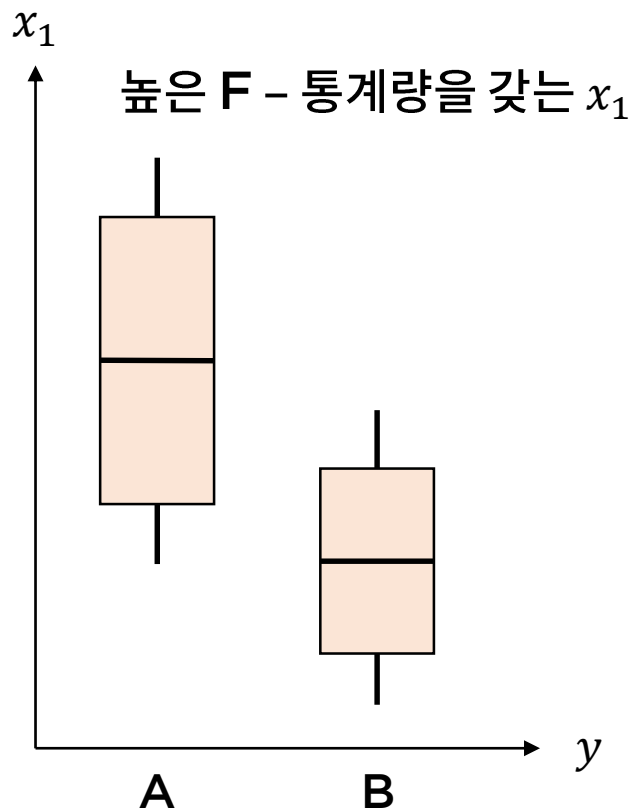
$x_2 = 1$ 인 샘플의
클래스 변수의 분포



$x_2 = 0$ 인 샘플의
클래스 변수의 분포

I 클래스 관련성 척도 예시: F - 통계량

- F - 통계량은 ANOVA에서 사용하는 통계량으로, 집단 간 평균 차이가 있는지를 측정하기 위한 통계량



- 지도학습에서 집단이란 특징 값 혹은 클래스 값을 기준으로 나뉜 샘플 집합을 의미
- x_1 의 값에 따라 y 분류가 어느정도 가능하지만, x_2 는 그렇지 않음

I 클래스 관련성 척도 분류

- 클래스 관련성 척도는 특징과 라벨의 유형에 따라 선택함

통계량	특징 유형	라벨 유형	관련 함수
카이제곱 통계량	이진형	이진형 (분류)	chi2
상호 정보량	이진형	이진형 (분류)	mutual_info_classif
	연속형		
	이진형	연속형 (예측)	mutual_info_regression
	연속형		
F - 통계량	연속형	이진형 (분류)	f_classif
	연속형	연속형 (예측)	f_regression

I 관련 함수: `sklearn.feature_selection.SelectKBest`

- 주요 입력
 - `scoring_func`: 클래스 관련성 측정 함수 (예: `chi2`, `mutual_info_classif`, `f_regression` 등)
 - `k`: 선택하는 특징 개수
- 주요 메서드
 - `.fit`, `.transform`, `.fit_transform`: 특징을 선택하는데 사용하는 메서드
 - `.get_support()`: 선택된 특징의 인덱스를 반환
- 주요 속성: `scoring_func(X, Y)`의 결과물과 같음
 - `scores_`: `scoring_func`으로 측정한 특징별 점수
 - `pvalues_`: `scoring_func`으로 측정한 특징별 p-value
(1에 가까울수록 독립적이며, 0에 가까울수록 관련성이 높음)

Chapter.

많다고 좋은게 아니다: 차원의 저주 문제

| 감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승