Chapter. 14

이건 꼭 알아야 해: 지도학습 모델의 핵심 개념

지도학습 개요

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

| 지도 학습

- 컴퓨터에게 입력과 출력을 주고, **입력과 출력 간 관계를 학습**하여 새로운 입력에 대해 적절한 출력을 내도록 하는 기계학습의 한 분야

- 입력을 특징 (feature) 혹은 특징 벡터 (feature vector)라고 하며, 출력을 라벨 (label)이라고 함

특징 벡터	라벨
$\boldsymbol{x}^{(1)}$	$y^{(1)}$
$\boldsymbol{x}^{(2)}$	$y^{(2)}$
$\boldsymbol{x}^{(3)}$	$y^{(3)}$
⋮	⋮
$\boldsymbol{x}^{(n)}$	$y^{(n)}$

학습 데이터

학습 → $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ Q(\boldsymbol{\theta})$

파라미터 추정
(Q: 비용 함수)

반영 → $f\left(\boldsymbol{x}|\hat{\boldsymbol{\theta}}\right)$

학습된 모델

입력 → $\boldsymbol{x}^{(n+1)}$

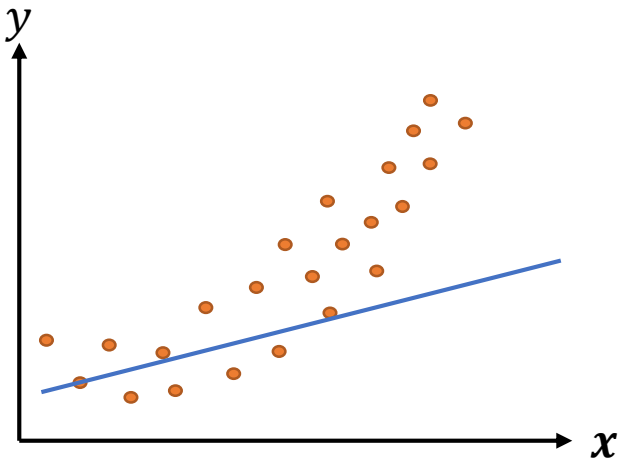
예측 → $\hat{y} = f\left(\boldsymbol{x}^{(n+1)}|\hat{\boldsymbol{\theta}}\right)$

- 라벨이 범주형 변수면 분류라고 하며, 연속형 변수면 예측 혹은 회귀라고 함

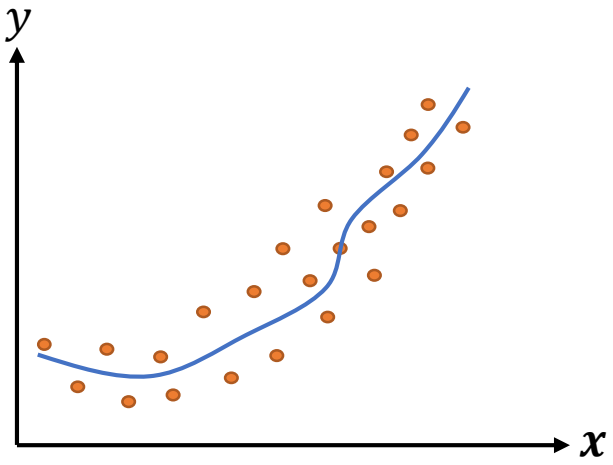
| 과적합

- 지도학습 모델은 학습 데이터를 분류하고 예측하는 수준으로, 학습에 사용되지 않은 데이터도 정확히 분류하고 예측하리라 기대하며, 이러한 기대가 충족되는 경우 **일반화**되었다고 함

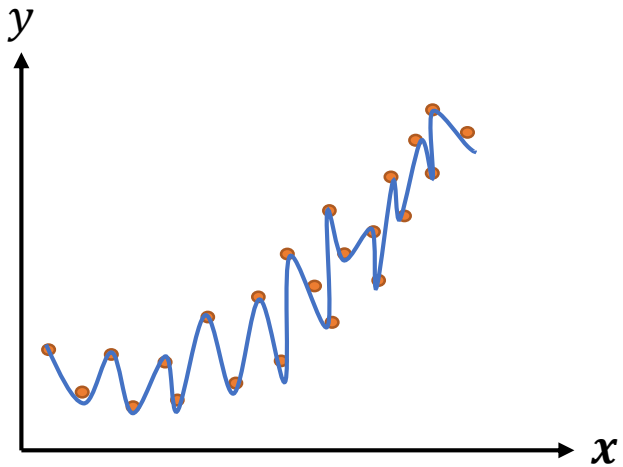
- 모델이 너무 복잡해서 학습 데이터에 대해서만 정확히 분류하고 예측하는 모델을 **과적합**되었다고 하며, 반대로 너무 단순해서 어떠한 데이터에 대해서도 부적합한 모델을 과소적합되었다고 함



과소적합



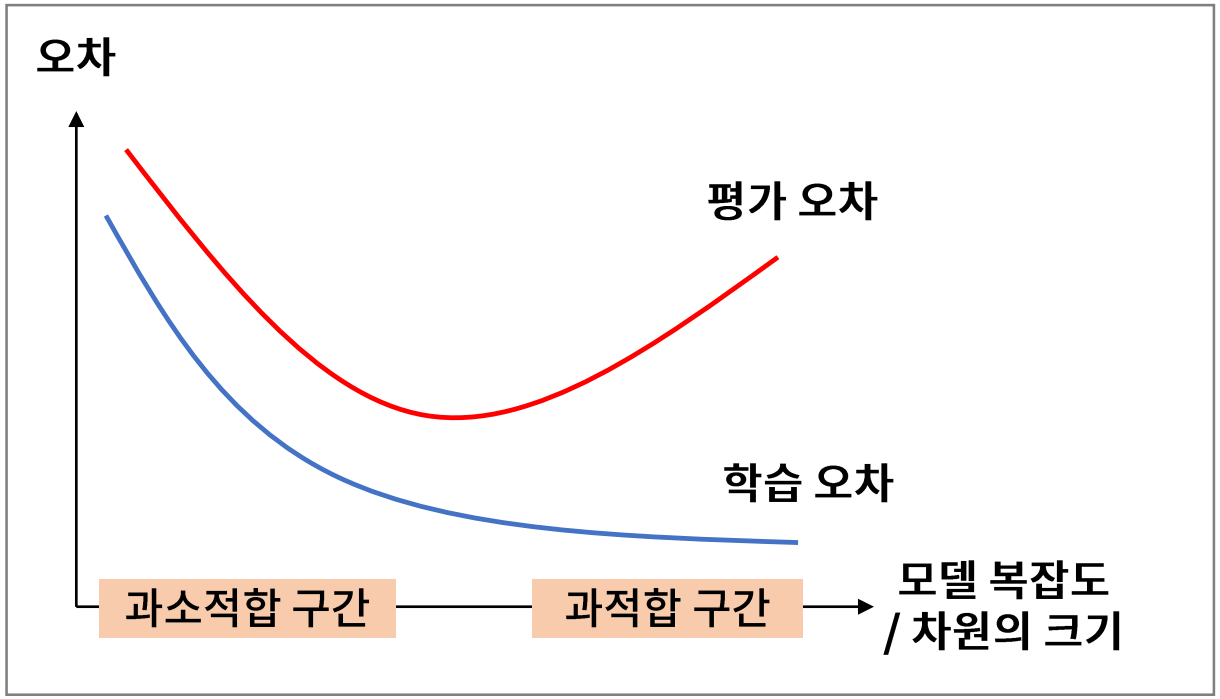
적정적합



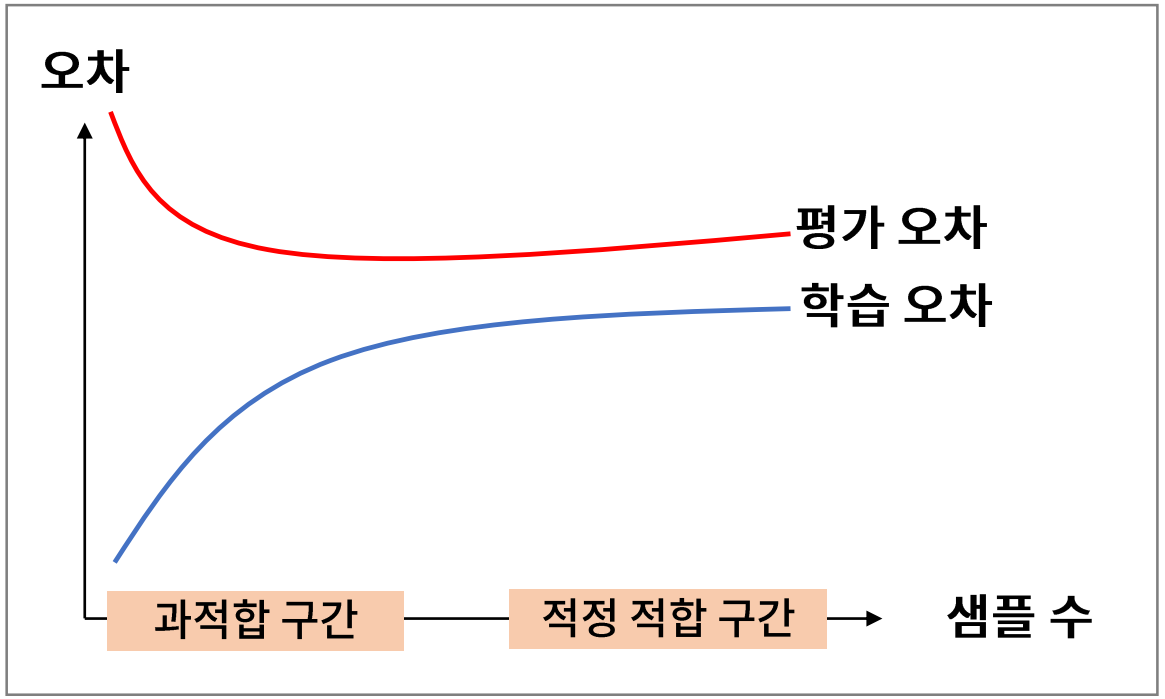
과적합

| 과적합 (계속)

- 과적합과 과소적합에 영향을 끼치는 주요 인자로는 **모델의 복잡도, 샘플 수, 차원의 크기** 등이 있음



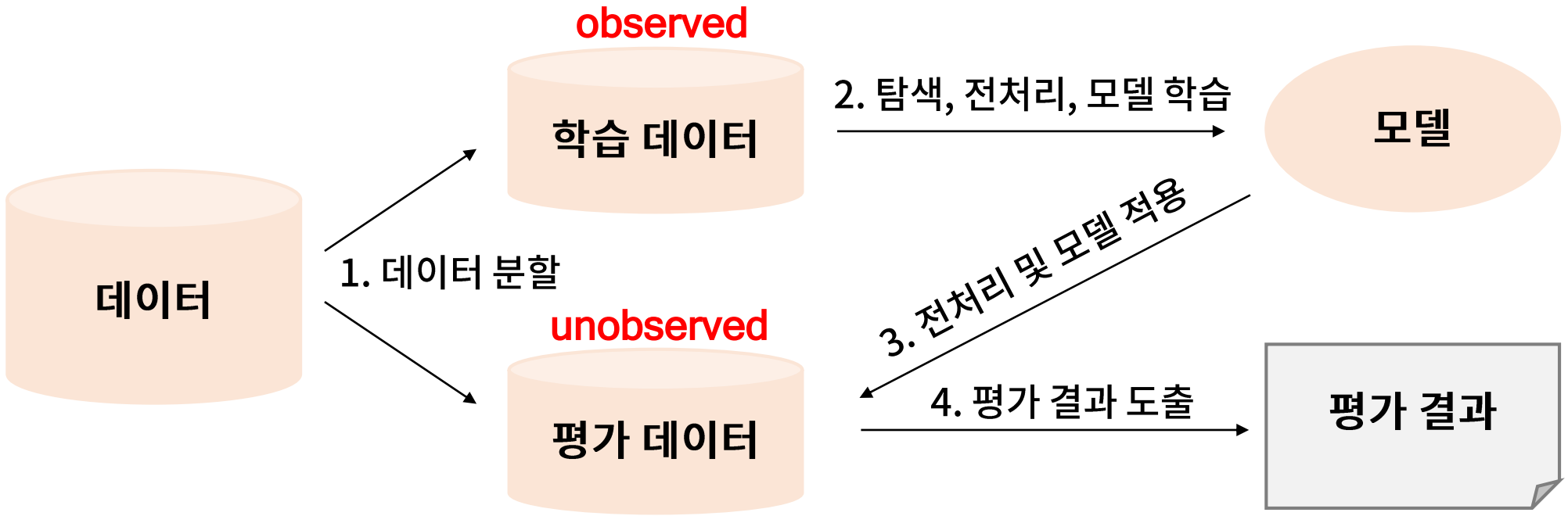
모델의 복잡도 및 차원의 크기와 과적합 간 관계



샘플 수와 과적합 간 관계

I 데이터 분할

- **과적합된 모델을 좋게 평가하는 것을 방지**하기 위해서, 데이터를 학습 데이터와 평가 데이터로 분할함



- 학습 데이터와 평가 데이터가 지나치게 유사하거나 특정 패턴을 갖지 않도록 분할해야 함

I 파라미터와 하이퍼 파라미터

- 하이퍼 파라미터(hyper parameter)는 일종의 **사용자 옵션**으로, 모델 성능에 직접적으로 영향을 끼치므로 자세한 **데이터 탐색 결과를 바탕으로 선택**해야 함

구분	파라미터	하이퍼 파라미터
정의	모델 내부에서 결정되는 변수	파라미터에 영향을 주는 파라미터
예시	신경망의 가중치, SVM의 가중치	신경망의 은닉층 구조, SVM의 커널
추정	비용 함수를 최소화하는 값으로 미리 정의된 컴퓨터 연산을 통해 추정	사용자가 직접 설정하며, 최적의 설정 방법은 없고, 휴리스틱한 방법이나 경험에 의한 설정이 대부분임

I 이진 분류 모델 평가: 혼동 행렬

- 이진 분류: 클래스 변수의 상태 공간이 크기가 2인 분류

- 혼동 행렬: 분류 모델을 평가하는데 사용하는 표
  - ➢ Positive class: 분석의 관심 대상 (보통 1로 설정)
  - ➢ Negative class: 분석 관심 대상 외 (보통 0이나 -1로 설정)

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

I 이진 분류 모델 평가: 대표적인 지표

- 각 지표의 한계 때문에, 가능한 **여러 지표를 사용**하여 모델을 평가해야 함

지표	수식	의미	사용시 주의사항
정확도	$Acc = \frac{TP+TN}{TP+FN+FP+TN}$	모든 샘플 가운데 정확히 분류한 샘플의 비율	클래스 불균형 문제가 있는 데이터에 대해 매우 취약함
정밀도	$Pre = \frac{TP}{TP+FP}$	긍정이라고 분류한 샘플 가운데, 실제 긍정인 샘플의 비율로, 이 수치가 높을수록 사람의 공수가 줄어드는 경향이 있음	긍정이라고 분류하는 샘플 수가 작으면 작을수록 정밀도가 높아짐
재현율	$Rec = \frac{TP}{TP+FN}$	실제 긍정 샘플 가운데, 긍정이라고 분류한 샘플의 비율로, 검출력과 관련이 있음	모든 샘플을 긍정이라고 분류하면 재현율이 높아짐
F1-score	$F_1 = 2 \times \frac{Pre \times Rec}{Pre+Rec}$	정밀도와 재현율의 조화 평균	해석이 까다로우며, 정확도처럼 해석하는 경우가 잦음

I 다중 분류 모델 평가

- 다중 분류: 클래스 변수의 상태 공간이 크기가 3이상인 분류
- **각 클래스를 긍정으로 간주**하여 평가 지표를 계산한 뒤, 이들의 산술 평균이나 가중 평균으로 평가

클래스 A를 긍정으로 간주

		예측		
		A	B	C
실제	A	AA	AB	AC
	B	BA	BB	BC
	C	CA	CB	CC

클래스 B를 긍정으로 간주

		예측		
		A	B	C
실제	A	AA	AB	AC
	B	BA	BB	BC
	C	CA	CB	CC

클래스 C를 긍정으로 간주

		예측		
		A	B	C
실제	A	AA	AB	AC
	B	BA	BB	BC
	C	CA	CB	CC

범례

TP	FN
FP	TN

$Score_A$ $Score_B$ $Score_C$

$Score$

I 예측 모델 평가

- 대표적인 예측 모델 평가 지표로 루트 평균 제곱 오차(root mean squared error; RMSE)와 평균 절대 오차 (mean absolute error; MAE)가 있으며, 두 지표 모두 값이 작을수록 좋음
  - 루트 평균 제곱 오차: $\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}}$
  - 평균 절대 오차: $\frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n}$
- RMSE와 MAE를 정확히 평가하려면, 해당 분야의 **도메인 지식**이나 **클래스 변수의 스케일**을 고려해야 함
  - 예를 들어, 코스피 지수를 예측하는 모델의 MAE가 1010이라면 무의미한 수준의 모델이지만, 전세계 인구 수를 예측하는 모델의 MAE가 1010이라면 매우 우수한 모델임

Chapter. 14

이건 꼭 알아야 해: 지도학습 모델의 핵심 개념

모델 개발 프로세스

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

책에서나 볼 수 있는 프로세스

I 문제 정의

- 전체 프로세스 가운데 가장 중요한 단계로, 명확한 목적 의식을 가지고 프로세스를 시작해야 함

- 이 단계에서 수행하는 활동은 다음과 같음
  - 과업 종류 결정 (분류, 예측 등)
  - **클래스 정의**
  - **도메인 지식 기반의 특징 정의**
  - 사용 데이터 정의

I 데이터 수집

- 문제 정의에서 정의한 데이터를 수집하는 단계로, 크롤링, 센서 활용, 로그 활용 등으로 데이터를 수집
- 기업 내 구축된 DB에서 SQL을 통해 추출하는 경우가 가장 많으며, 이때는 클래스를 중심으로 수집

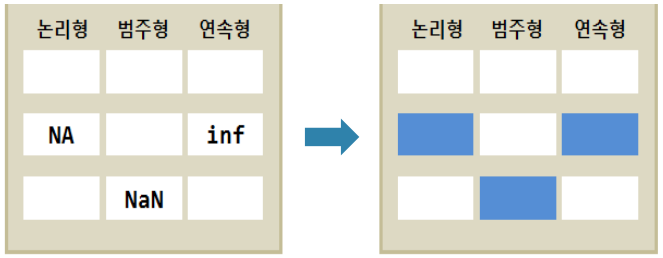
I 데이터 탐색

- 데이터가 어떻게 생겼는지를 확인하여, **프로세스를 구체화**하는 단계

- 데이터 탐색 단계에서 변수별 분포, 변수 간 상관성, 이상치와 결측치, 변수 개수, 클래스 변수 분포 등을 확인하며, 이 **탐색 결과는 데이터 전처리 및 모델 선택에 크게 영향**을 미침
  - 데이터 크기 확인 → 과적합 가능성 확인 및 특징 선택 고려
  - 특징별 기술 통계 → 특징 변환 고려 및 이상치 제거 고려
  - 특징 간 상관성 → 특징 삭제 및 주성분 분석 고려
  - 결측치 분포 → 결측치 제거 및 추정 고려
  - 변수 개수 → 차원 축소 기법 고려
  - 클래스 변수 분포 → 비용민감 모델 및 재샘플링 고려

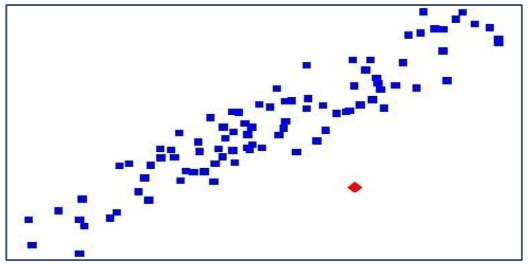
I 데이터 전처리

- 원활한 모델링을 위해 **데이터를 가공**하는 단계로, 여기서 수행하는 대표적인 작업은 다음과 같음

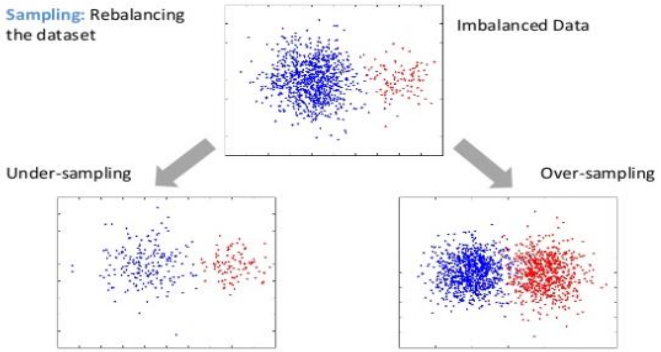


결측 값 처리

데이터 통합



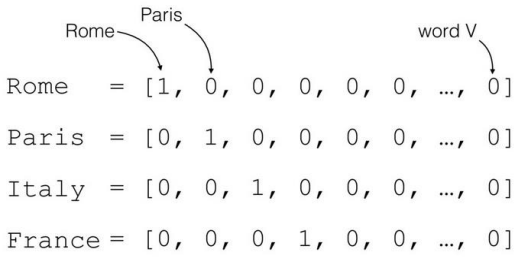
이상치 제거



재샘플링

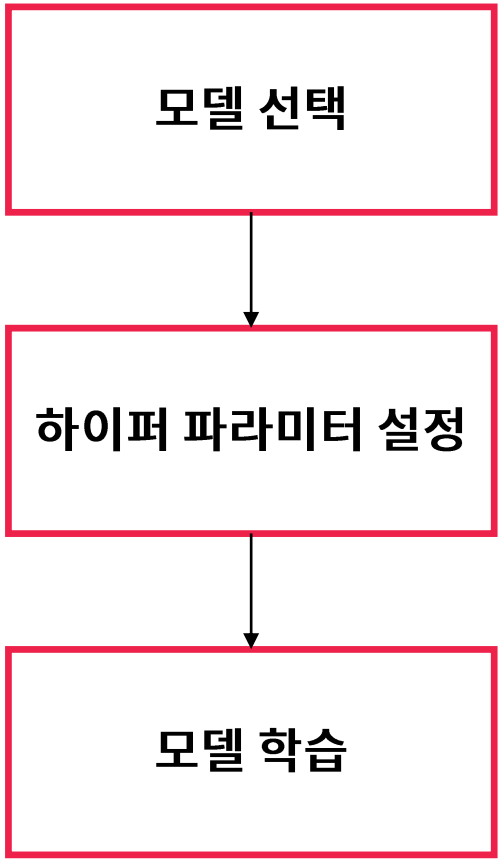
X_1 X_4 X_5 X_8
X_2 X_3 X_6 X_7 X_9

특징 선택



더미 변수 생성

I 모델링



- 데이터 특성, 성능, 설명력 등을 기준으로 모델 선택
- 예1) 설명력이 중요한 경우 의사결정나무 혹은 베이지안 네트워크 사용
- 예2) 이진 텍스트 분류를 하는 경우 나이브베이즈 혹은 서포트 벡터 머신을 사용

- 모델의 성능을 결정짓는 하이퍼 파라미터를 설정
- 최적의 하이퍼 파라미터 설정은 굉장히 어려움

- 모델에 포함된 파라미터를 추정
- 이미 잘 개발되어 있는 모듈/패키지가 있기에 전혀 어렵지 않음

I 모델 평가

- 분류 모델의 대표적인 지표
  - 정확도 (accuracy): 전체 샘플 가운데 제대로 예측한 샘플의 비율
  - 정밀도 (precision): 긍정 클래스라고 예측한 샘플 가운데 실제 긍정 클래스 샘플의 비율
  - 재현율 (recall): 실제 긍정 클래스 샘플 가운데 제대로 예측된 샘플의 비율
  - F1 점수 (F1-score): 정밀도와 재현율의 조화 평균

- 예측 모델의 대표적인 지표
  - 평균 제곱 오차 (mean squared error)
  - 평균 절대 오차 (mean absolute error)
  - 평균 절대 퍼센트 오차 (mean absolute percentage error)

- 잘못된 평가를 피하기 위해, **둘 이상의 평가 지표를 쓰는 것이 바람직**함

l 결과보고서 작성

- 지금까지의 분석 결과를 바탕으로 보고서를 작성하는 단계

- 결과보고서의 통일된 구성은 없지만, 일반적으로 다음과 같이 구성됨
  1. 분석 목적
  2. 데이터 탐색 및 전처리
  3. 분석 방법
  4. 분석 결과 및 **활용 방안**

I 부적절한 문제 정의

- 부적절한 문제 정의의 가장 흔한 유형으로는 (1) 구체적이지 않은 문제 정의, (2) 부적절한 특징 정의, (3) 수집 불가한 데이터 정의가 있음

- 타이어 A사 사례
  - 문제 정의: “타이어 생산 공정에서 발생하는 데이터를 바탕으로 공정을 최적화하고 싶다”
  - 제공 데이터: 생산 공정에서 발생한 거의 모든 데이터

- 카드 A사 사례
  - 문제 정의: “일별 콜수요를 예측하고 싶다”
  - 특징: (1) 상담원 수, (2) 상담원의 역량

I 부적절한 데이터 수집

- 부적절한 데이터 수집의 가장 흔한 유형으로 (1) 측정 오류 등으로 수집한 데이터가 실제 상황을 반영하지 못하는 경우, (2) 해결하고자 하는 문제와 무관한 데이터를 수집한 경우, (3) 특정 이벤트가 데이터에 누락된 경우가 있음

- 자동차 시트 A사 사례
  - 배경: 자동차 시트 폼을 보관하는 과정에서 창고의 온도 및 습도 등에 따라, 폼이 수축하기도 이완하기도 함
  - 문제 정의: 창고의 환경에 따른 폼의 수축 및 이완 정도를 예측하고, 그 정도가 심할 것이라 판단되면 환경을 제어
  - 수집 데이터: 창고의 환경 조건과 폼의 수축/이완 정도 데이터 (수집 시기: 2015년 6월 ~ 8월)

I 부적절한 데이터 탐색

- 피드백 루프를 발생시키는 핵심 원인이 부적절한 데이터 탐색 혹은 데이터 탐색 생략임
- 데이터 탐색을 제대로 하지 않으면, **적절한 모델 선택 및 전처리를 할 수 없어**, 모델 평가 단계에서 좋은 성능을 내는 것이 거의 불가능함
- 자동차 A사 사례
  - 문제 정의: 서비스 센터로 등록되는 클레임 가운데 안전 관련 클레임을 자동 검출하는 시스템 개발
  - 부적절한 탐색 내용: 클레임에서 오탈자 및 비표준어가 굉장히 많았지만, 빈발 단어의 분포 등만 확인함
  - 결과: 스펠링 교정 등의 전처리를 생략해서, 검출 성능이 매우 낮은 모델이 학습되어 다시 탐색 단계로 되돌아감

부적절한 데이터 전처리

- 데이터 전처리는 크게 모델 개발을 위해 필수적인 전처리와 모델 성능 향상을 위한 전처리로 구분됨
- 보통 모델 성능 향상을 위한 전처리를 생략해서 이전 단계로 되돌아가는 경우가 가장 흔함

| 부적절한 모델링 및 모델 평가

- 모델링에서는 주로 부적절한 모델 및 파라미터 선택으로 잘못되는 경우가 대부분이며, 모델링 자체가 잘못되는 경우는 매우 드묾

- 모델 평가는 적절하지 않은 지표를 사용해서 잘못되는 경우가 대부분이며, 대표적인 사례로 단일 지표만 써서 부적절한 모델을 우수한 모델이라고 판단하는 경우가 있음

실제 / 예측	Positive	Negative
Positive	0	1
Negative	0	9999

⇒ 정확도 99.99%의 모델

⇒ 재현율 0%의 모델

Chapter. 14

이건 꼭 알아야 해: 지도학습 모델의 핵심 개념

주요 모델의 구조 및 특성

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

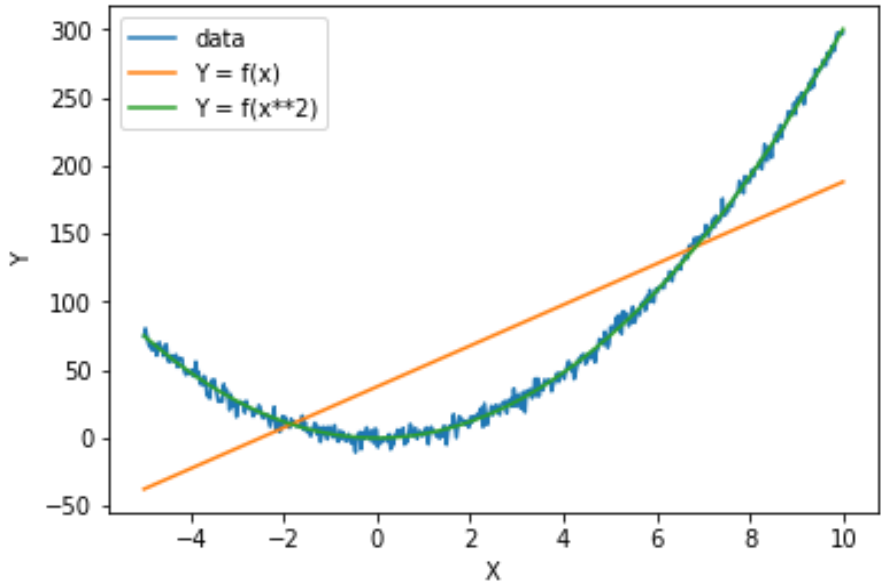
강사. 안길승

I 선형 회귀 모델과 정규화 회귀 모델

- 모델 구조
  - $f\left(\boldsymbol{x}^{(i)}\right) = w_1 x_1^{(i)} + w_2 x_2^{(i)} + \cdots + w_d x_d^{(i)} + b$
  - $\boldsymbol{w} = (w_1, w_2, \cdots, w_d)^T$: 가중치 벡터 (계수)
  - $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_d^{(i)}\right)$: 샘플 i $(i = 1, 2, \cdots, n)$
  - b: 절편 (bias)
- 비용 함수: 오차제곱합
  - 선형 회귀 모델: $\sum_{i=1}^{n} \left(y^{(i)} - f\left(\boldsymbol{x}^{(i)}\right)\right)^2$
  - 정규화 회귀 모델 (Ridge): $\sum_{i=1}^{n} \left(y^{(i)} - f\left(\boldsymbol{x}^{(i)}\right)\right)^2 + \lambda \sum_{j=1}^{d} w_j^2$

I 선형 회귀 모델과 정규화 회귀 모델

- 특징과 라벨 간 비선형 관계가 무시될 수 있으므로, 특징 변환이 필요



➢ 데이터 (파란색 선)의 분포를 보면, x와 y는 2차식 관계라는 것을 대략적으로 추정할 수 있음

➢ x를 x^2로 변환한다면 과소적합된 주황색 모델이 아닌 적절하게 적합된 파란색 모델을 학습할 수 있음

- 특징 간 스케일 차이에 크게 영향을 받아, 예측 모델링을 할 때 스케일링이 필요함

- λ와 max_iter에 따라 과적합 정도가 직접 결정됨

I 로지스틱 회귀 모델

- **모델 구조**
  - $\Pr\left(y^{(i)}=1\right)=\frac{1}{1+\exp\left(-w_1x_1^{(i)}-w_2x_2^{(i)}-\cdots-w_dx_d^{(i)}-b\right)}$
  - $\hat{y}^{(i)}=\begin{cases}1, \Pr\left(y^{(i)}=1\right)\geq c\\ 0, \quad \text{otherwise}\end{cases}$, 여기서 $0<c<1$는 cut-off value
  - $\boldsymbol{w}=(w_1,w_2,\cdots,w_d)^T$: 가중치 벡터 (계수)
  - b: 절편 (bias)

- **비용 함수: 크로스 엔트로피**
  - $\sum_{i=1}^{n}\left(-y_i\log\left(\Pr\left(y^{(i)}=1\right)\right)-(1-y_i)\log\left(1-\Pr\left(y^{(i)}=1\right)\right)\right)$

I 로지스틱 회귀 모델

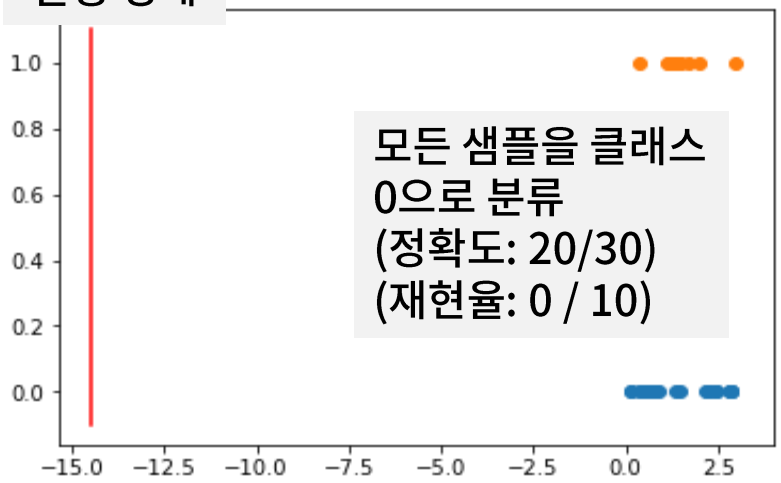
- 특징의 구간 별로 라벨의 분포가 달라지는 경우, 적절한 구간을 나타낼 수 있도록 **특징 변환**이 필요함

	$y=0$	$y=1$
$x \le 1$	9	1
$1 < x \le 2$	2	8
$2 \le x$	9	1

$$\tilde{x} = \begin{cases} 1, 1 < x \le 2 \\ 0, \text{otherwise} \end{cases}$$

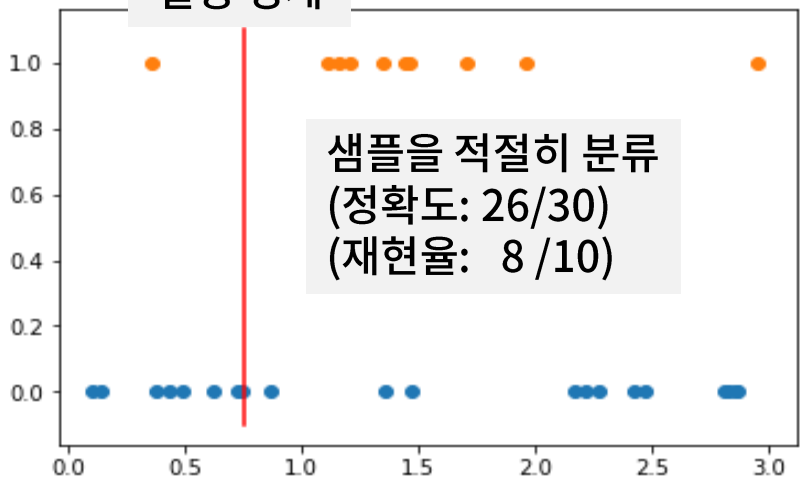
	$y=0$	$y=1$
$\tilde{x}=0$	18	2
$\tilde{x}=1$	2	8

결정 경계



모든 샘플을 클래스 0으로 분류
(정확도: 20/30)
(재현율: 0 / 10)

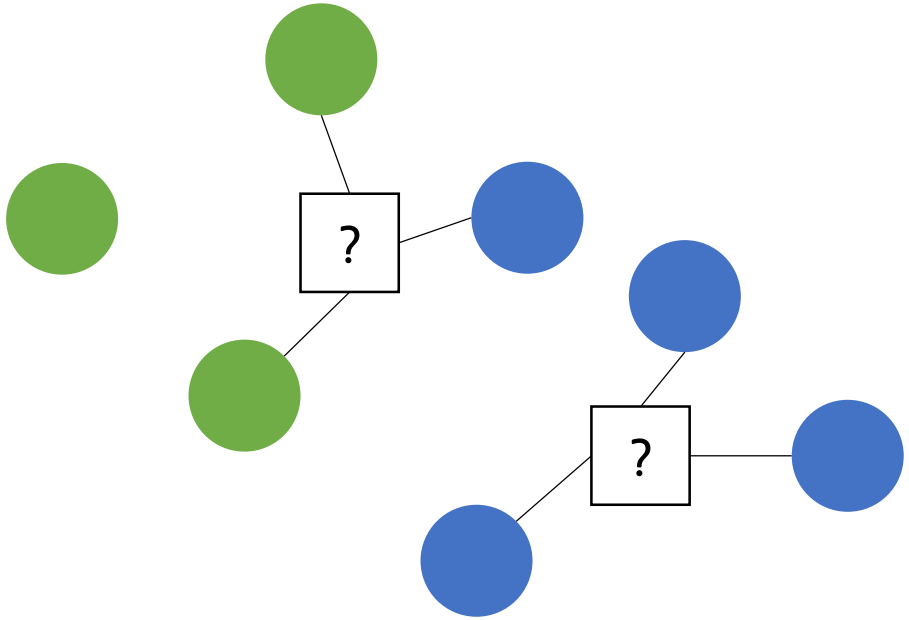
결정 경계



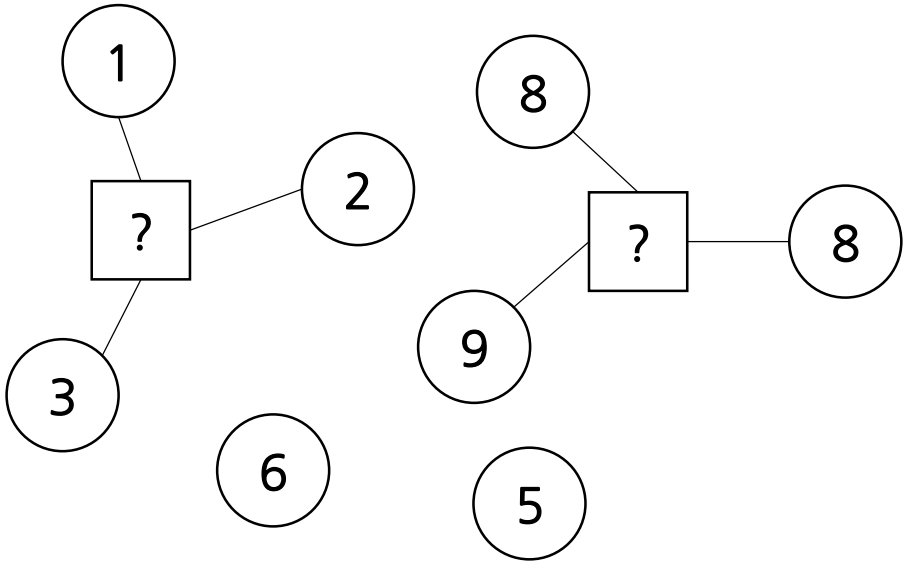
샘플을 적절히 분류
(정확도: 26/30)
(재현율: 8 /10)

I k-최근접 이웃 (k-Nearest Neighbors; kNN)

- 모델 구조
  - $\hat{y}^{(i)} = mode(\{y^{(j)} | j \in N_i\})$ (분류: 샘플 i의 k개 이웃 샘플들의 클래스 최빈값으로 분류)
  - $\hat{y}^{(i)} = \frac{\sum_{j \in N_i} y^{(j)}}{k}$ (예측: 샘플 i의 k개 이웃 샘플들의 클래스 평균으로 예측)



분류를 위한 kNN



예측을 위한 kNN

I k-최근접 이웃 (k-Nearest Neighbors; kNN)

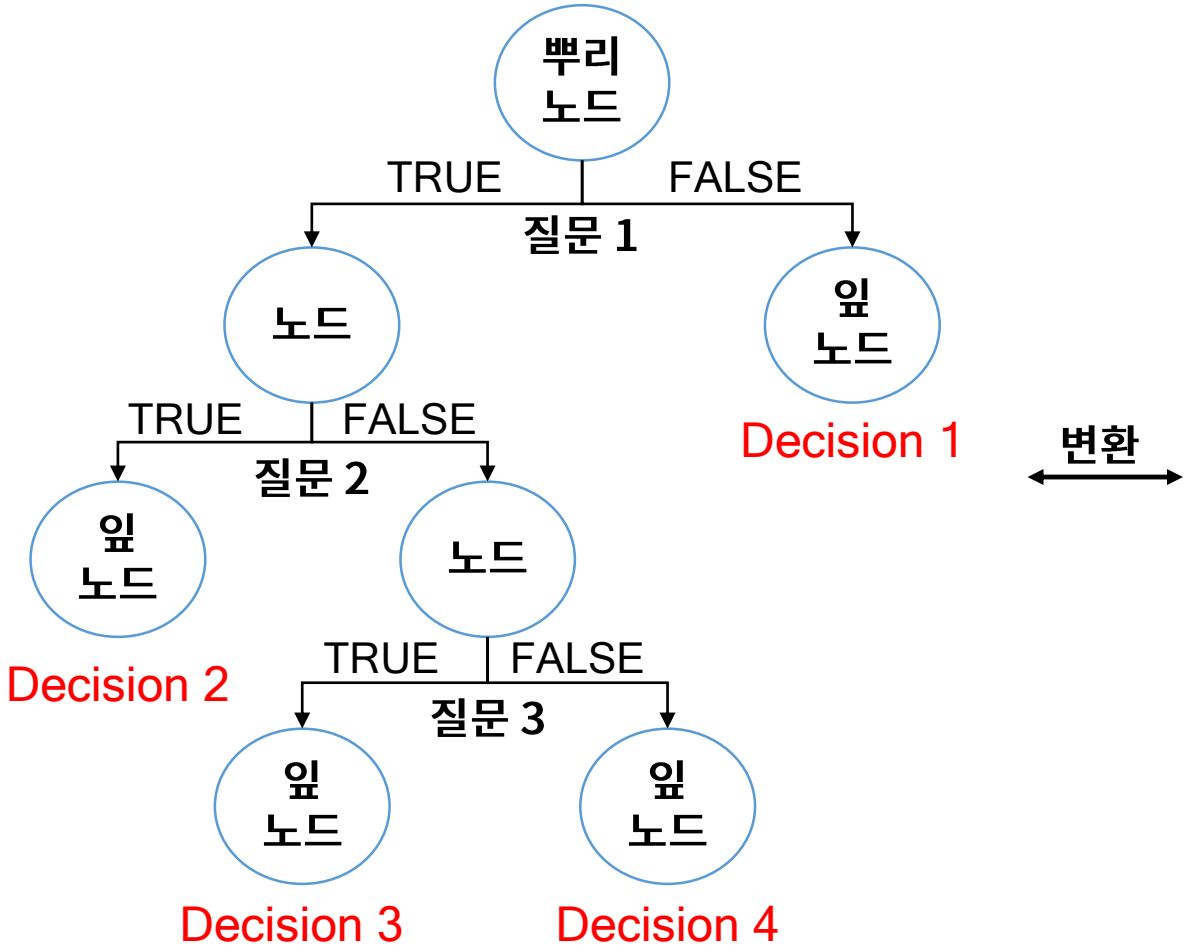
- 주요 파라미터와 설정 방법
  - ➢ 이웃 수 (k): 홀수로 설정하며, 특징 수 대비 샘플 수가 적은 경우에는 k를 작게 설정하는 것이 바람직함
  - ➢ 거리 및 유사도 척도
    - ✓ 모든 변수가 서열형 혹은 정수인 경우: 맨하탄 거리
    - ✓ 방향성이 중요한 경우 (예: 상품 추천 시스템): 코사인 유사도
    - ✓ 모든 변수가 이진형이면서 희소하지 않은 경우: 매칭 유사도
    - ✓ 모든 변수가 이진형이면서 희소한 경우: 자카드 유사도
    - ✓ 그 외: 유클리디안 거리

I k-최근접 이웃 (k-Nearest Neighbors; kNN)

- 특징 추출이 어려우나 유사도 및 거리 계산만 가능한 경우 (예: 시퀀스 데이터)에 주로 활용
- 모든 특징이 연속형이고 샘플 수가 많지 않은 경우에 좋은 성능을 보인다고 알려져 있음
- 특징 간 스케일 차이에 크게 영향을 받아, 스케일링이 반드시 필요함 (코사인 유사도를 사용하는 경우 제외)
- 거리 및 유사도 계산에 문제가 없다면, 별다른 특징 변환이 필요하지 않음

I 의사결정나무 (Decision tree)

- 모델 구조



변환

```
IF 질문 1:
        IF 질문 2:
                Decision 2
        ELSE:
                IF 질문 3:
                        Decision 3
                ELSE:
                        Decision 4
ELSE:
        Decision 1
```

나무 구조

규칙 집합

I 의사결정나무 (Decision tree)

- 예측 과정을 잘 설명할 수 있다는 장점 덕분에 많은 프로젝트에서 활용
  - A 보험사: 고객의 이탈 여부를 예측하고, 그 원인을 파악해달라
  - B 밸브사: 밸브의 불량이 발생하는 공정 상의 원인을 파악해달라
  - C 홈쇼핑사: 방송 조건에 따라 예측한 상품 매출액 기준으로 방송 편성표를 추천해주고, 그 근거를 설명해달라

- 선형 분류기라는 한계로 예측력이 좋은 편에 속하지는 못하나, 최근 각광받고 있는 앙상블 모델(예: XGBoost, LightGBM)의 기본 모형으로 사용됨

I 의사결정나무 (Decision tree)

- 주요 파라미터
  - max_depth: 최대 깊이로 그 크기가 클수록 모델이 복잡해짐
  - min_samples_leaf: 잎 노드에 있어야 하는 최소 샘플 수로, 그 크기가 작을수록 모델이 복잡해짐

I 나이브 베이즈 (Naive Bayes)

- **모델 구조**
  - 베이즈 정리를 사용하고 **특징 간 독립**을 가정하여 사후 확률 $\Pr(y|\boldsymbol{x})$을 계산
  - $\Pr(y|\boldsymbol{x}) \propto \Pr(y) \times \prod_{j=1}^{d} \Pr(x_j \mid y)$
  - 가능도 $\Pr(x_j \mid y)$은 조건부 분포를 가정하여 추정함
    - 이진형 변수: 베르누이 분포
    - 범주형 변수: 다항 분포
    - 연속형 변수: 가우시안 분포

- **모델 특성**
  - 특징 간 독립 가정이 실제로는 굉장히 비현실적이므로, 일반적으로 높은 성능을 기대하긴 어려움
  - 설정한 분포에 따라 성능 차이가 크므로, 특징의 타입이 서로 같은 경우에 사용하기 바람직함
  - 특징이 매우 많고 그 타입이 같은 문제 (예: 이진형 텍스트 분류)에 주로 사용됨

I 나이브 베이즈 (Naive Bayes)

- 모델 특성
  - ➢ 특징 간 독립 가정이 실제로는 굉장히 비현실적이므로, 일반적으로 높은 성능을 기대하긴 어려움
  - ➢ 설정한 분포에 따라 성능 차이가 크므로, 특징의 타입이 서로 같은 경우에 사용하기 바람직함
  - ➢ 특징이 매우 많고 그 타입이 같은 문제 (예: 이진형 텍스트 분류)에 주로 사용됨

I 서포트 벡터 머신 (Support Vector Machine; SVM)

- 모델 구조
  - $w_1 x_1^{(i)} + w_2 x_2^{(i)} + \cdots + w_d x_d^{(i)} + b = 0$
  - $\boldsymbol{w} = (w_1, w_2, \cdots, w_d)^T$: 가중치 벡터 (계수)
  - b: 절편 (bias)

- 최적화 모델
  - 목적식: $\|\boldsymbol{w}\| + C \times \sum_{i=1}^{n} \xi_i$
  - 제약식: $y^{(i)}\left(\boldsymbol{w}\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi_i, \forall i$

- 오차를 최소화하면서 동시에 마진을 최대화하는 분류 모델로, 커널 트릭을 활용하여 저차원 공간을 고차원 공간으로 매핑함

- 마진의 개념을 회귀에 활용한 모델을 서포트 벡터 회귀 (Support Vector Regression)이라 함

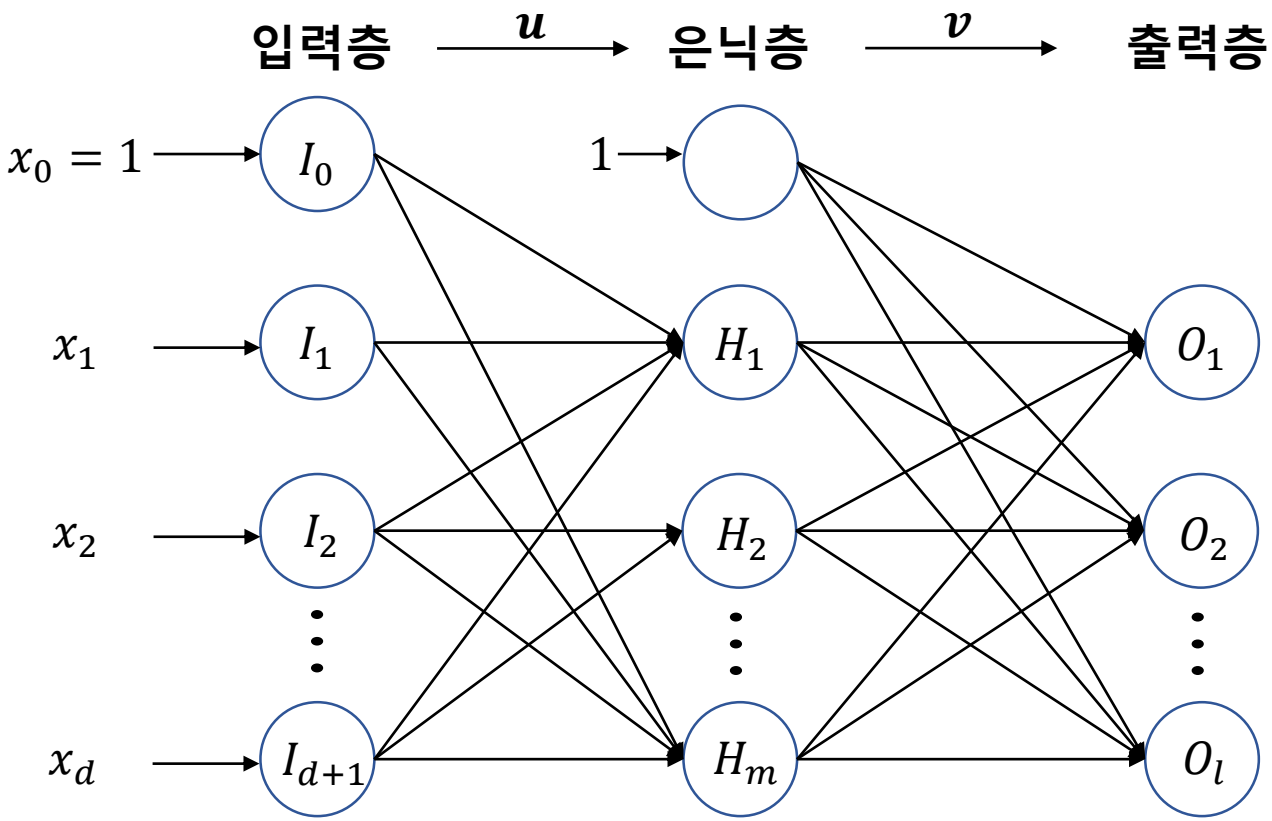
I 서포트 벡터 머신 (Support Vector Machine; SVM)

- 주요 파라미터
  - ➢ kernel: 통상적으로 이진 변수가 많으면 linear 커널이, 연속 변수가 많으면 rbf 커널이 잘 맞는다고 알려져 있음
  - ➢ C: 오차 패널티에 대한 계수로, 이 값이 작을수록 마진 최대화에 클수록 학습 오차 최소화에 신경을 쓰며, 보통 10^n 범위에서 튜닝함
  - ➢ γ: rbf 커널의 파라미터로, 크면 클수록 데이터의 모양을 잘 잡아내지만 오차가 커질 위험이 있으며, C가 증가하면 γ도 증가하게 튜닝하는 것이 일반적임

- 파라미터 튜닝이 까다로운 모델이지만, 튜닝만 잘하면 좋은 성능을 보장하는 모델임

I 신경망 (Neural Network)

- 모델 구조



- 입력 노드: 입력 값을 받는 역할을 수행
- 은닉 노드 및 출력 노드: 입력 노드 (은닉 노드) 혹은 다른 은닉 노드로부터 들어온 값들을 가중합하고 활성 함수를 적용하여 출력을 냄

I 신경망 (Neural Network)

- 초기 가중치에 크게 영향을 받는 모델로, 세밀하게 random_state와 max_iter 값을 조정해야 함
- 은닉 노드가 하나 추가되면 그에 따라 하나 이상의 가중치가 추가되어, 복잡도가 크게 증가할 수 있음
- 모든 변수 타입이 연속형인 경우에 성능이 잘 나오는 것으로 알려져 있으며, 은닉 층 구조에 따른 복잡도 조절이 파라미터 튜닝에서 고려해야 할 가장 중요한 요소임
- 최근 딥러닝의 발전으로 크게 주목받는 모델이지만, 특정 주제(예: 시계열 예측, 이미지 분류, 객체 탐지 등)를 제외하고는 깊은 층의 신경망은 과적합으로 인한 성능 이슈가 자주 발생함

I 트리 기반의 앙상블 모델

- 최근 의사결정나무를 기본 모형으로 하는 앙상블 모형이 캐글 등에서 자주 사용되며, 좋은 성능을 보임
- 랜덤 포레스트: 배깅 (bagging) 방식으로 여러 트리를 학습하여 결합한 모델
- XGboost & LightGBM: 부스팅 방식으로 여러 트리를 순차적으로 학습하여 결합한 모델
- 랜덤 포레스트를 사용할 때는 트리의 개수와 나무의 최대 깊이를 조정해야 하며, XGboost와 LightGBM을 사용할 때는 트리의 개수, 나무의 최대 깊이, 학습률을 조정해야 함
  - ➢ 트리의 개수: 통상적으로 트리의 개수가 많으면 많을수록 좋은 성능을 내지만, 어느 수준 이상에서는 거의 큰 차이를 보이지 않음
  - ➢ 나무의 최대 깊이: 4이하로 설정해주는 것이 과적합을 피할 수 있어, 바람직함
  - ➢ 학습률: 이 값은 작으면 작을수록 과소적합 위험이 있으며, 크면 클수록 과적합 위험이 있음. 통상적으로 0.1로 설정함

Chapter.

이건 꼭 알아야 해: 지도학습 모델의 핵심 개념

감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승