〈데이터분석과기계학습 1주차〉

강의 소개 및 개론

인공지능융합공학부 데이터사이언스전공

곽찬희

강의소개 - 데이터 분석과 기계학습

- 강의 목표
  - ✓ 기계 학습의 원리를 정확히 이해하고, 실제 데이터 분석에 적용할 수 있다
  - ✓ 머신러닝 + 데이터분석
- 대상: 데이터사이언스 전공
  - ✓ 단, 복수전공, 부전공, 연계전공은 수강 가능
- 난이도
  - ✓ 파이썬에 익숙함
  - ✓ 데이터 분석 관련 과목 수강함
  - ✓ 머신러닝 기본 지식 있음

질문이 있을 땐

- Ecampus 질문 게시판
  - ✓ 쪽지 확인 어려움 (알림 안 뜸)
- E-mail (chk @ kangnam . ac . Kr)
  - ✓ 메일로 질문을 보낼 때엔 다음과 같은 규칙을 준수해주시기 바랍니다.
  (아래 내용 복사해서 쓰세요. 형식을 갖추지 않은 메일은 답장하지 않습니다.)

  제목: [과목이름] 질문요지 간단히
  내용:
  안녕하세요,
  저는 ***수업을 수강하는 **학과 ***입니다.
  이러이러한 질문이 있어서 메일 드렸습니다.

  감사합니다.
  ***드림

평가기준

- 프로젝트 진행 90%
  - ✓ 1차: 20%
  - ✓ 2차: 30%
  - ✓ 3차: 40%
  - ✓ 미제출 시 0점
  - ✓ 늦은 제출은 받지 않습니다 (시간을 충분히 드립니다).
- 출석 10%
  - ✓ 출석 미달 시 F이므로 출석에 유의!
  - ✓ 수강 후 출석체크가 되었는지 반드시 확인 (당일에는 반영이 안될 수도 있습니다.)
  - ✓ 지각(수강시간 미달) -1점, 결석(수강 안 함) -3점

프로젝트

- 1차 내용
  - ✓ 분석 대상 선정 및 분석 계획 수립
  - ✓ 주제 선정
  - ✓ 이 분석이 가치를 가지는 이유 설명
  - ✓ 데이터 수집 계획 (수집이 완료되었다면 수집된 데이터 설명)
  - ✓ PPT1장 발표 (5분)

프로젝트

- 2차
  - ✓ 분석 진행 및 데이터분석
  - ✓ 이 단계에서는 완벽하지 않아도 다양한 시도가 중요
  - ✓ 무엇이 가장 좋은 길일지 고민하기
  - ✓ 코드 발표 (with 시각화. 10분. Notebook 형태로)

프로젝트

- 3차
  - ✓ Hyperparameter 최적화
  - ✓ Model Validation
  - ✓ 보고서(PPT 혹은 Notebook)로 만들기
  - ✓ 발표! (10분)

프로젝트 기타 사항

- 개인 혹은 2인 팀을 구성해서 진행할 수 있음
- 주제는 자유롭게 선택함
- 머신러닝/딥러닝 요소와 데이터 분석 요소 모두 포함되어야 함
- 각 단계별 자료를 github에 올려야 함
- 발표

성적

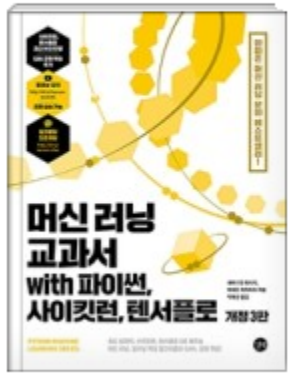
- 성적최대비율
  - ✓ A 50 %
  - ✓ B 50 %
  - ✓ 학교 정책에 따라 다를 수 있음

- 이러면 성적이 당연히 안 좋겠죠?
  - ✓ 과제를 내지 않거나,
  - ✓ 과제를 대애애애애충 내거나
  - ✓ 출석이 매우 미달이거나…
- 상대평가

교재

- **머신러닝 교과서 - 세바스챤 라시카, 바히드 미자리리 저 (길벗)**
  - ✓ **교재가 아니더라도 정말 알찬 책이니, 한번쯤 공부하면 좋겠습니다.**

머신 러닝 교과서 with 파이썬, 사이킷런, 텐서플로 최신 넘파이, 사이킷런, 텐서플로 2로 배우는 머신 러닝, 딥러닝 핵심 알고



★★★★★ 0.0 | 네티즌리뷰 1건

저자 세바스찬 라시카, 바히드 미자리리 | 역자 박해선 | 길벗 | 2021.03.31

페이지 868 | ISBN ? 9791165215187

도서 39,600원 ~~44,000원~~ -10%

e북 31,680원 ~~35,200원~~ -10%

구매혜택 상세보기 > | ♡ 3

바로구매

예스24 N Pay 1%	39,600원	구매
인터넷 교보문고 N Pay 1%	39,600원	구매
알라딘 N Pay 1%	39,600원	구매
인터파크 도서 N Pay 6%	39,600원	구매
영풍문고 N Pay 6%	39,600원	구매
도서11번가	39,600원	구매
커넥츠북 N Pay 1%	39,600원	구매
e북 예스24 N Pay 1%	31,680원	구매
e북 알라딘 N Pay 1%	31,680원	구매
e북 인터넷 교보문고 N Pay 1%	31,680원	구매
e북 리디북스 N Pay 1%	35,200원	구매
e북 네이버 시리즈	35,200원	구매

실습환경구성

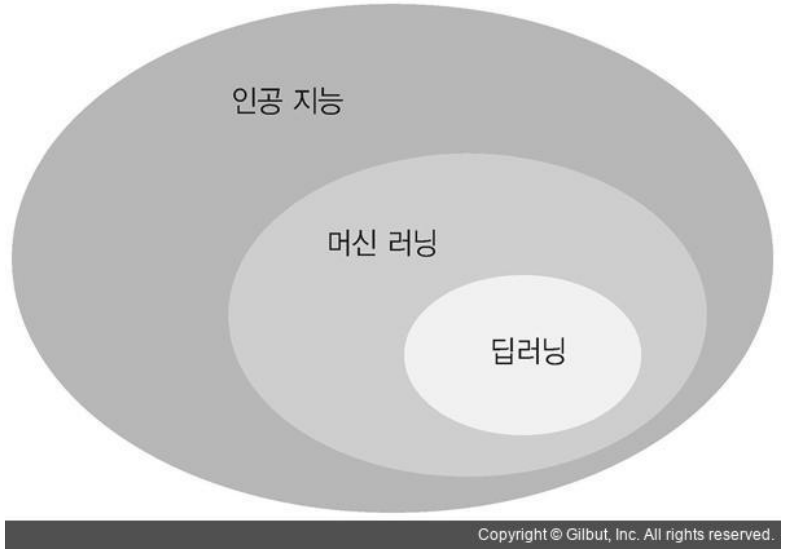
- 실습은 제 컴퓨터 기준으로…
  - ✓ Windows 10 + Anaconda + Chrome + JupyterLab + Google Colab (Optional)

- Chrome 이 기본 브라우저가 아니면 실행이 되지 않을 수 있습니다.

1. 컴퓨터는 데이터에서 배운다

인공지능, 머신러닝, 딥러닝…

- 인공지능: 지능을 인공적으로 만듦
- 머신러닝: 데이터에서 지식을 배움(학습)
- 딥러닝: 인공 신경망(Artificial Neural Network) 이 깊은 단계까지 복합적으로 연결됨



머신러닝의 종류

- 머신러닝의 세 가지 종류
  - ✓ 지도 학습 (Supervised Learning)
  - ✓ 비지도 학습 (Unsupervised Learning)
  - ✓ 강화 학습 (Reinforced Learning)

그림 1-1 머신 러닝의 세 가지 학습 종류

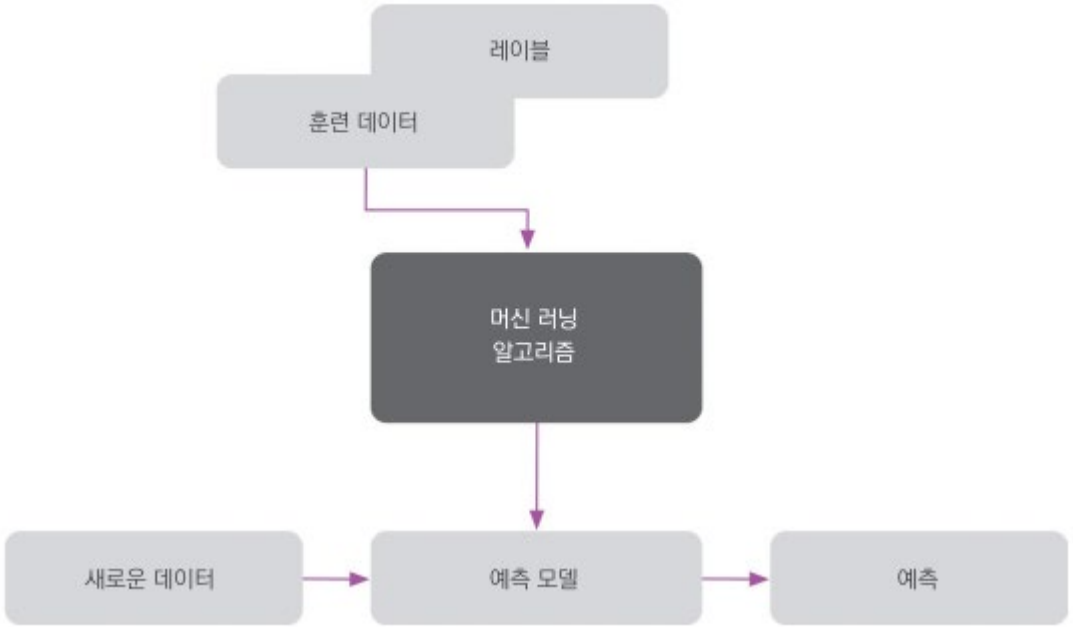
지도 학습의 특징

- 지도 학습
  - ✓ 레이블(Label)된 훈련 데이터를 학습한 모델이 미래 데이터에 대해 예측하는 것

- Label?
  - ✓ Label의 존재에 따라 지도/비지도 결정
  - ✓ 정답표

- 분류 (Classification)와 회귀(Regression)로 나뉨

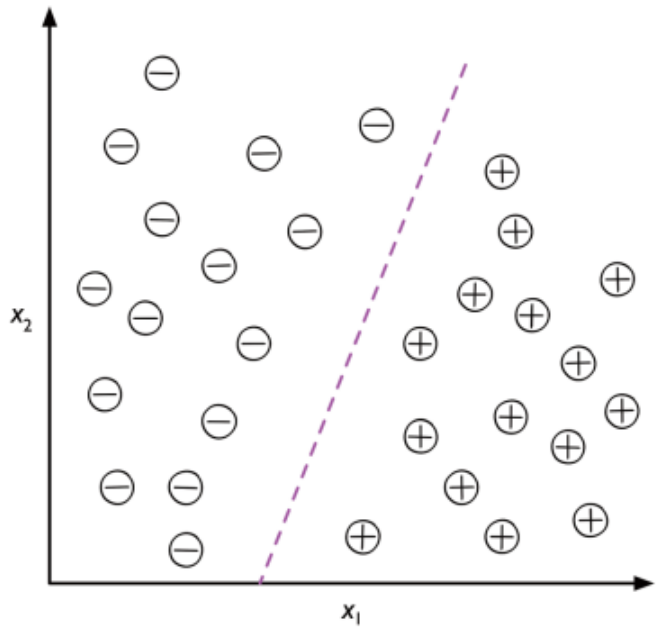
그림 1-2 지도 학습



지도 학습 1 - 분류: 클래스 레이블 예측

- 분류란?
  - ✓ 과거의 관측을 근거로 새로운 샘플(데이터, 사례 등)의 범주형 클래스 레이블을 예측
- 분류의 종류
  - ✓ 이진 분류 (binary classification): 0 or 1 / 홀 or 짝 / 흑 or 백 / 짜 or 짬 / 부먹 or 찍먹…
  - ✓ 다중 분류 (multiclass classification): 학년 (1, 2, 3, 4), 군대 계급 …
- 결정 경계 (decision boundary)
  - ✓ 클래스를 구분하는 경계

그림 1-3 두 개의 클래스를 구분하는 결정 경계

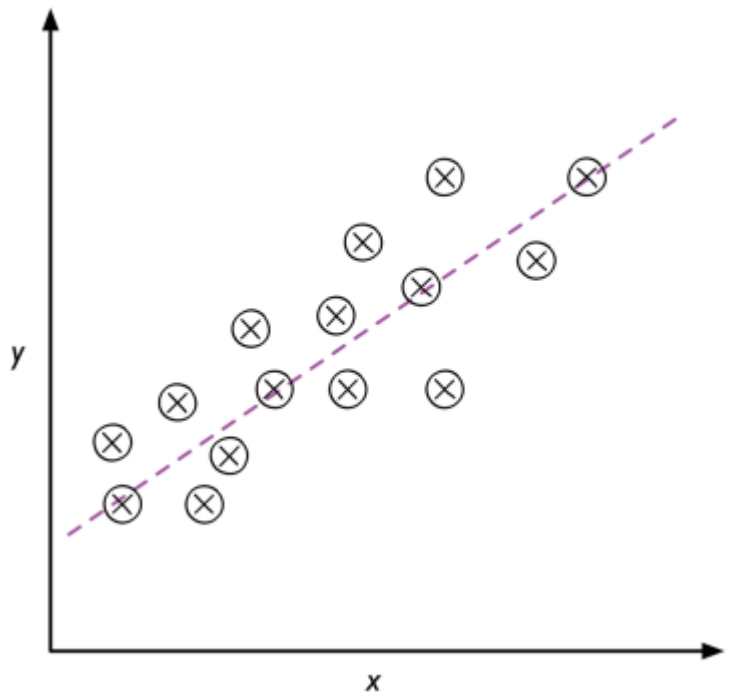


지도 학습 2 - 회귀: 연속적인 출력 값 예측

- 회귀 (regression) 란?
  - ✓ 예측 변수 (predictor variable, 또는 설명변수 explanatory variable, 또는 입력 input) 와 연속적인 반응 변수 (response variable , 또는 출력 outcome, 타깃 target)가 주어졌을 때 출력 값을 예측하는 변수 사이의 관계를 찾음
  - ✓ 예) 키, 얼굴

- 선형회귀의 예
  - ✓ 입력 x 와 타깃 y 가 주어졌을 때,
    직선과 점들 사이 거리가 최소가 되는 직선

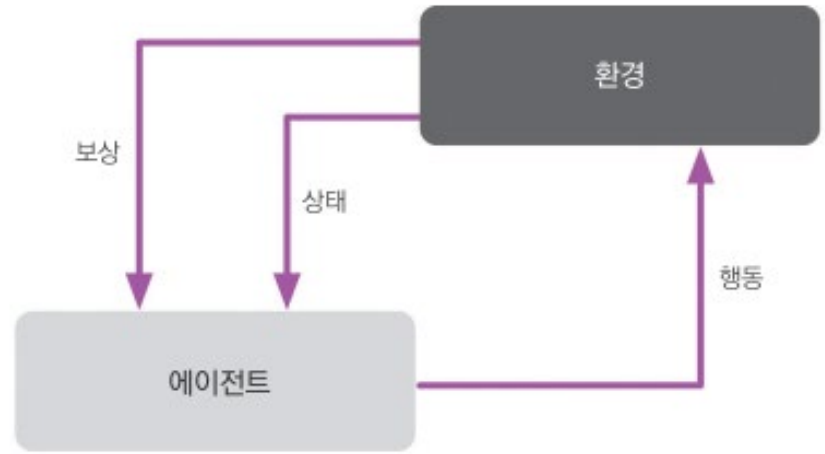
그림 1-4 선형 회귀의 예



강화 학습

- 강화 학습 (reinforced learning)
  - ✓ 환경과 상호 작용하여 시스템 (에이전트, agent)의 성능을 향상시키는 것이 목적
  - ✓ 보상 (reward) 함수로 행동이 얼마나 좋은지 판단

그림 1-5 강화 학습

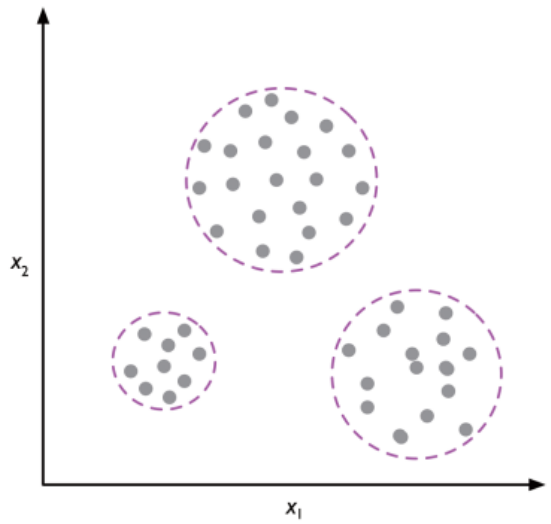


비지도학습

- 비지도 학습?
  - ✓ 레이블되지 않거나 구조를 알 수 없는 데이터를 학습하여 정보나 지식을 추출

- 군집 (Clustering, 또는 비지도 분류)
  - ✓ 사전 정보 없이 쌓여 있는 그룹 정보를 의미 있는 서브그룹 (subgroup) 또는 클러스터 (cluster)로 조직하는 탐색적 데이터 분석 기법

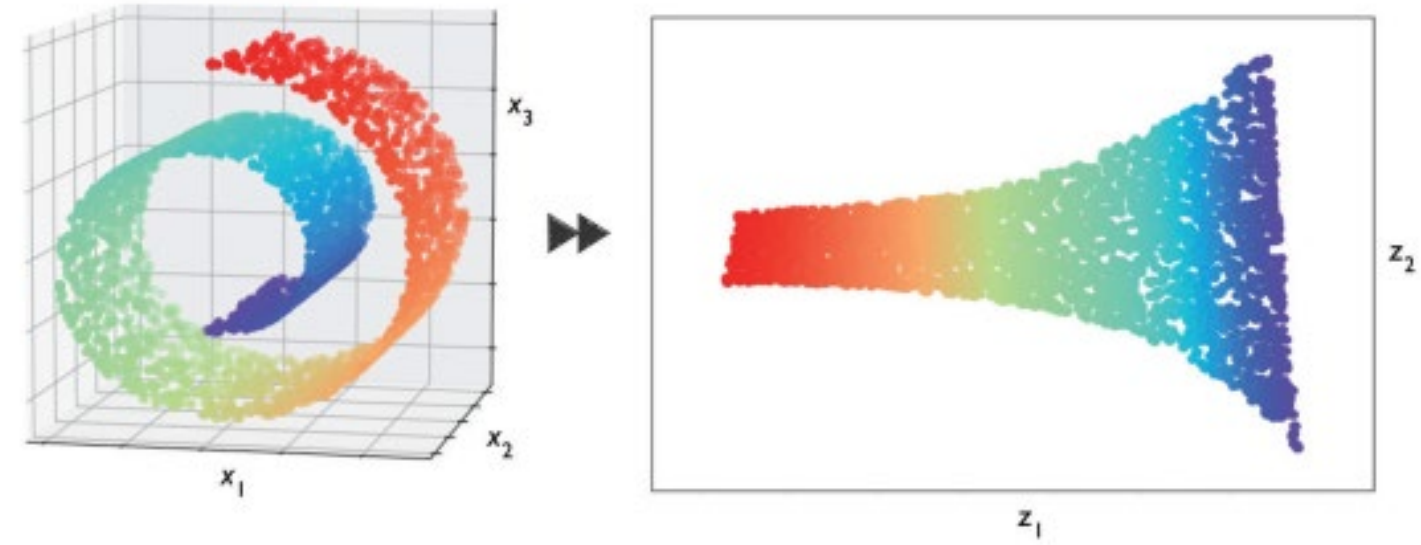
그림 1-6 군집의 예



비지도학습 - 차원 축소: 데이터 압축

- 차원 축소 (dimensionality reduction)
  - ✓ 고차원의 데이터를 정보가 유지되는 선에서 더 작은 차원의 부분 공간 (subspace) 로 변환

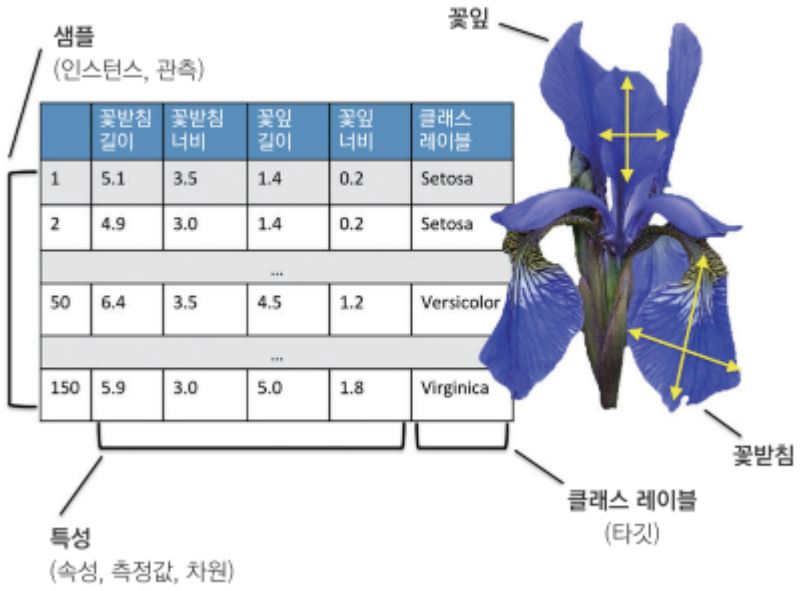
그림 1-7 차원 축소의 예



기본 용어와 표기법

- 데이터셋은 다음으로 구성
  - ✓ 행(row)
  - ✓ 열(column) 혹은 특성(feature)

그림 1-8 붓꽃 데이터셋



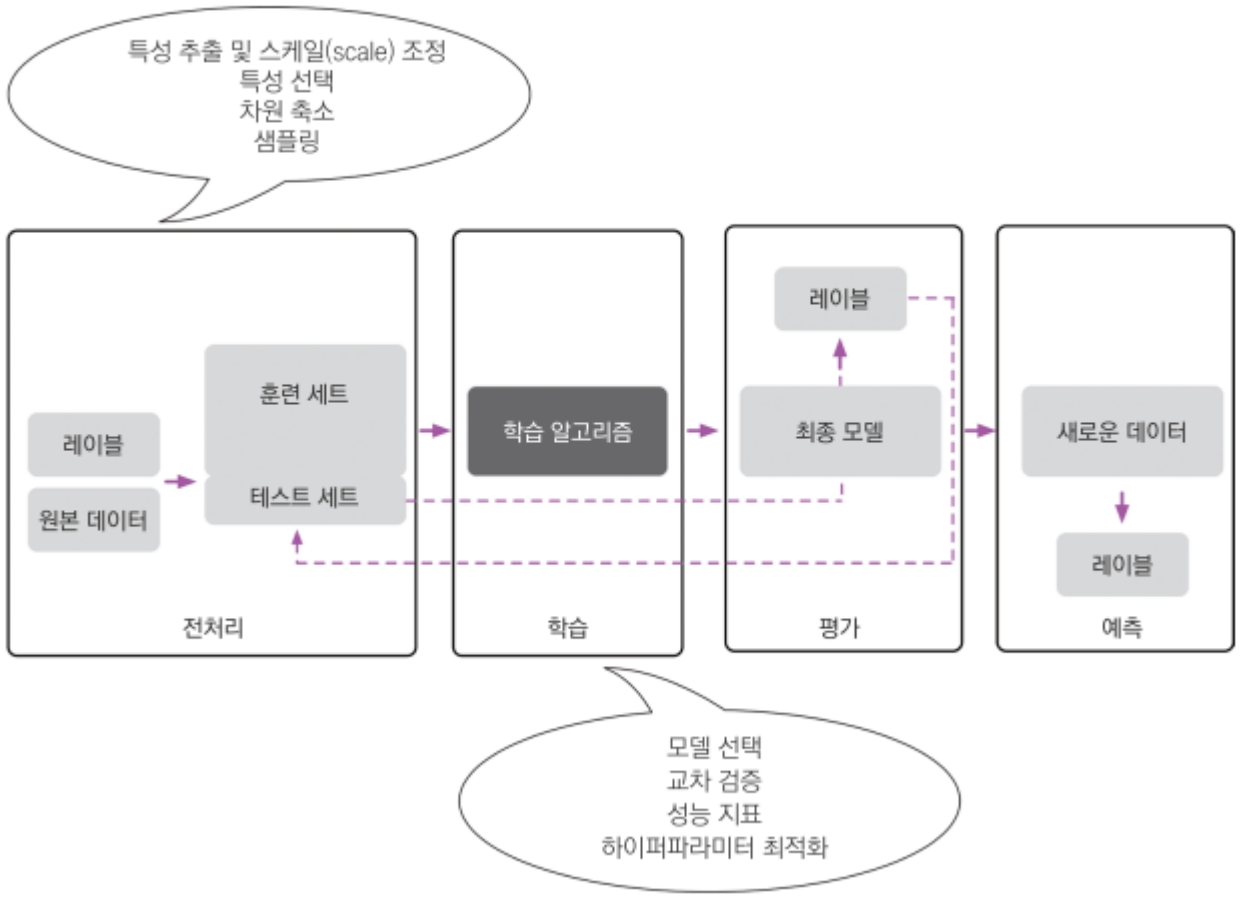
	꽃받침 길이	꽃받침 너비	꽃잎 길이	꽃잎 너비	클래스 레이블
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

머신 러닝 용어

- 훈련 샘플: 데이터셋을 나타내는 테이블 행
- 훈련: 모델 피팅(model fitting) 혹은 파라미터 추정(parameter estimation)
- 특성(x): 예측변수, 변수, 입력, 속성, 공변량
- 타깃(y): 결과, 출력, 반응 변수, 종속 변수, 레이블,
- 손실함수(loss function): 비용함수(cost function).

머신러닝 시스템 구축 로드맵

그림 1-9 머신 러닝의 작업 흐름

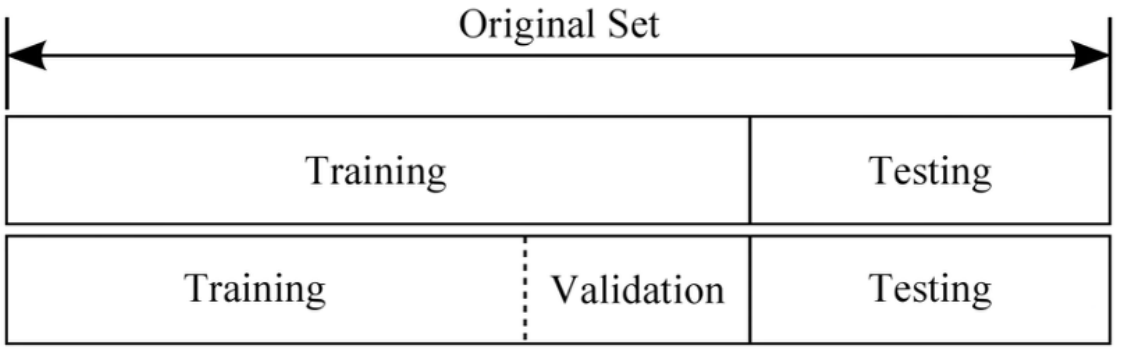


1. 전처리

- 데이터가 머신러닝 알고리즘에 사용될 수 있도록 정리하고 정제하는 작업
  - ✓ 특성을 [0, 1] 범위로 변환
  - ✓ 표준 정규 분포 (standard normal distribution) 로 변환
  - ✓ 중복된 정보를 갖는 경우 차원 축소 시행
  - ✓ 훈련셋과 테스트셋을 나눔

2. 예측 모델 훈련과 선택

- 여러 알고리즘을 비교할 척도가 필요
  - ✓ 정확도 (accuracy) 가 대표적
- 교차 검증 기법 (cross validation)
  - ✓ 훈련셋을 다시 훈련셋과 검증셋으로 나눔



- 하이퍼파라미터(hypterparameter) 튜닝
  - ✓ 알고리즘의 세부 조건을 바꿔가며 최적의 모델 도출

3. 모델을 평가하고 본 적 없는 샘플로 예측

- 이전에 본 (사용한) 데이터를 이용한 변환(Transformation), 차원 축소(Dimension Reduction), 특성 추출(Feature Extraction) 등을 새로운(테스트) 데이터에도 동일하게 적용해야 함

- 그렇지 않으면 과도하게 긍정적인 결과 도출

- Sklearn 에서 fit_transform 과 transform 의 차이