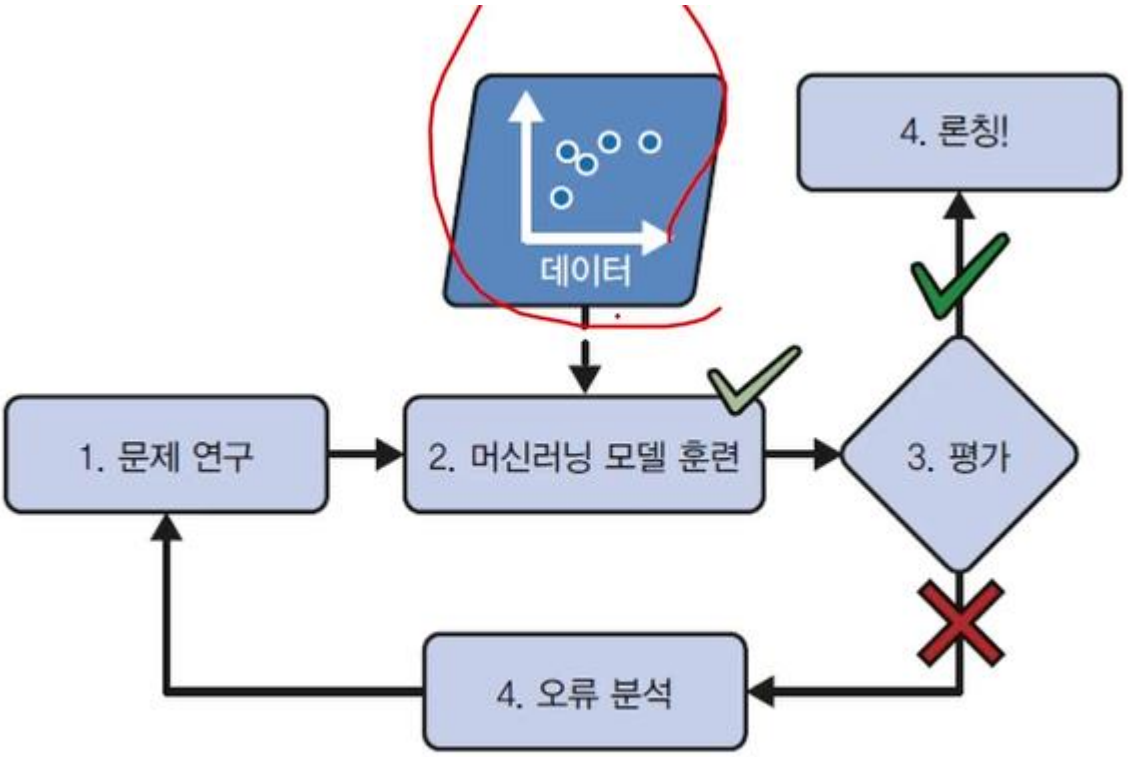
기계학습 2025-1 중간고사 20241519 조예성

데이터에서 학습하도록 컴퓨터를 프로그래밍 하는 과학의 명칭: 머신러닝

- 대표 예시 1개: 스팸 메일 분류

머신러닝 기본 용어 3개: 훈련 세트, 훈련 사례, 모델



머신러닝 접근 방법

머신러닝은 자동으로 (변화에) 적응한다

대용량의 데이터를 분석하여 숨겨진 패턴을 발견하는 기법: 데이터 마이닝

머신러닝 훈련 지도방식: 지도 / 비지도 / 강화학습 / 준지도 / 자기지도

레이블이 있는 지도방식: 지도

레이블이 없는 학습방식: 비지도

데이터 없이 행동으로 습득: 강화학습

시스템이 학습하는 데 사용하는 샘플: 훈련 세트

각 훈련 데이터: 훈련 사례 / 셀플

머신러닝 작업 전처리: 시각화, 차원 축소, 특성 추출

탐지 등 3개: 이상치 탐지, 특이치 탐지, 연관 규칙 학습

한번에 모아서: 배치 학습

실시간으로 업데이트: 온라인 학습

나누어서 하는 학습: 외부 메모리 학습

데이터에 따라 적응하는 속도: 학습률

시스템이 훈련 샘플을 기억함으로써 학습: 사례 기반 학습

샘플들의 모델을 만들어 예측: 모델 기반 학습

라이브러리 sklearn

모델이 훈련 데이터에만 잘 맞지만 일반성이 떨어지는 경우: 과대적합

반대로: 과소적합

과대적합 방지를 위하여: 규제

검증을 하기 위하여 (훈련 데이터)와 검증 (데이터로 분류함)

프로젝트 진행
큰 그림 -> 데이터 구하기 -> 데이터 탐색/시각화 -> 데이터 준비 ->
선택훈련 -> 모델 조정 -> 솔루션 제시 -> 유지보수

데이터에 대한 간단한 설명하는 메소드: df.info()

테스트 학습 분할: train_test_split()

상관관계 메소드: corr()

누락된 데이터 관리: 해당 구역 제거, 전체 특성 삭제, 중간값 대체

중간값으로 대체하는 클래스: SimpleImputer

문자로 되어있는 카테고리형 변수를 숫자로 변경하는 클래스: OneHotEncoder() 클래스

입력값이 고정 포인트에서 멀어질수록 출력 값이 지수적으로 감소하는 함수: 방사 기저 함수

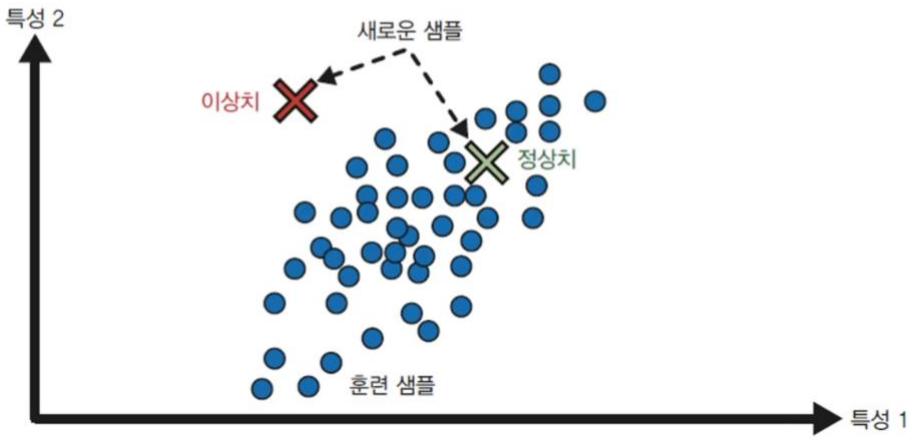
성능 향상 위하여 값 미세조정: GridSearchCV

가장 좋은 하이퍼파라미터 값 구하기: .cv_results

시간 소요 문제 -> 랜덤화: RandomizedSearchCV

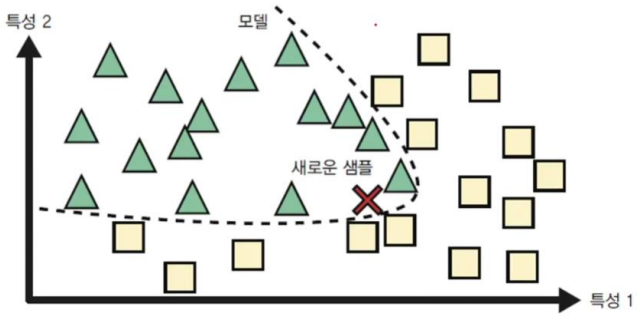
강화 학습은 (환경)을 관찰해서 (행동) 을 실행하고 그 결과로 (보상) 또는 (벌점) 부과

데이터를 나누어서 하는 학습: 외부 메모리 학습



강화학습의 학습하는 시스템을 칭하는 명칭: 에이전트

(사례 기반 학습)



(모델 기반 학습)

```
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.linear_model import LinearRegression

# 데이터를 다운로드하고 준비합니다.
data_root = "https://github.com/ageron/data/raw/main/"
lifesat = pd.read_csv(data_root + "lifesat/lifesat.csv")
X = lifesat[["GDP per capita (USD)"]].values
y = lifesat[["Life satisfaction"]].values

# 데이터를 그래프로 나타냅니다.
lifesat.plot(kind='scatter', grid=True,
             x="GDP per capita (USD)", y="Life satisfaction")
plt.axis([23_500, 62_500, 4, 9])
plt.show()

# 선형 모델을 선택합니다.
model = LinearRegression()

# 모델을 훈련합니다.
model.fit(X, y)

# 키프로스에 대해 예측을 만듭니다.
X_new = [[37_655.2]]       # 2020년 키프로스 1인당 GDP
print(model.predict(X_new)) # 출력: [[6.30165767]]
```

최근접 이웃 모델 구성: model = KneighborsRegressor(n_neighbors = 3)

성능 측정 지표 대표 2종류: RMSE, MAE

```
>>> housing.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

가장 먼저 수정이 필요한 컬럼은?

데이터에 관한 간략한 설명을 보여주는 메소드: info()

각 카테고리마다 몇 개가 있는지 보여주는 메소드: value_counts()

숫자형 특성의 요약 정보를 보여주는 메소드: describe()

특성간 표준 상관관계를 보여주는 메소드 corr

n개의 중복되지 않은 서브셋을 랜덤으로 분할하는 메소드: k-폴드 교차 검증

모든 A/B 쌍에 대해 클래스 A의 샘플이 클래스 B로 분류된 횟수를 세는 것: 오차 행렬

- 만드는 함수: confusion_matrix()

```
ray([[53892,  687],
     [ 1891, 3530]])
```

True Negative	False Positive
False Negative	True Positive

오차 행렬로 만들 수 있는 것: 정밀도, 재현율

양성 예측의 정확도: 정밀도

분류기가 정확하게 감지한 양성 샘플의 비율: 재현율

거짓 양성 비율에 대한 진짜 양성 비율의 곡선: ROC 곡선

완벽한 분류기는 ROC의 AUC가 (1) 이 됨

매우 강력하고 선형 비선형 회귀 이상치 탐색에도 사용할 수 있는 다목적 머신러닝 모델: SVM

모든 데이터가 확실하게 분류되어야 하는 분류: 하드 마진 분류

어느정도 오류는 유연: 소프트 마진 분류

소프트함을 정하기 위한 하이퍼파라미터 약어: c

비선형 SVM 분류: PolynomialFeatures

비선형 데이터를 처리할 수 있는 커널 함수: 가우스 RBF

```
 = make_pipeline(StandardScaler(),
                 SVC(kernel="rbf", gamma=5, C=0.001))
```

서포트 벡터로 분류하는 약어: SVC

서포트 벡터로 수치 예측: SVR

객관식 15개

단답/주관 3개

서술형 2개