
빅데이터 분석
및 처리 기술
Crayon

지난 시간에는

6 빅데이터 회귀 분석

- 머신러닝의 개념과 기본 절차
- Scikit-Learn API 사용법
- 회귀 분석 심화

이번 시간에는

7 빅데이터 분류 분석

- 빅데이터 분류 분석의 절차
- 초모수와 모델 검증 방법
- 분류 심화 : 나이브 베이즈 기법

이번 시간에는

7 빅데이터 분류 분석

학습 목표

분류 분석의 개념과 절차에 대해 이해한다.
초모수와 모델 검증 방법을 이해한다.
나이브 베이즈 기법을 활용한 분류 심화 과정을 실습한다.

이번 시간에는

7 빅데이터 분류 분석

- 분류 분석의 개념
- 분류 분석의 기본 절차

지도학습 (Supervised learning)

데이터의 측정된 특징과 데이터와 관련된 레이블 사이의 관계를 모델링

비지도 학습(Unsupervised learning)

레이블을 참조하지 않고 데이터 세트의 특징을 모델링

이번 시간에는

7 빅데이터 분류 분석

- 분류 분석의 개념
- 분류 분석의 기본 절차

I 분류 분석의 개념

1. 데이터의 특징과 레이블

"지도학습"

- 데이터의 특징과 레이블 사이의 관계 모델링

기온	습도	불쾌도
20	60	쾌적
30	80	불쾌
30	70	불쾌
35	75	?

I 분류 분석의 개념

1. 데이터의 특징과 레이블

"지도학습"

- 데이터의 특징과 레이블 사이의 관계 모델링

기온	습도	불쾌도
20	60	쾌적
30	80	불쾌
30	70	불쾌
35	75	?

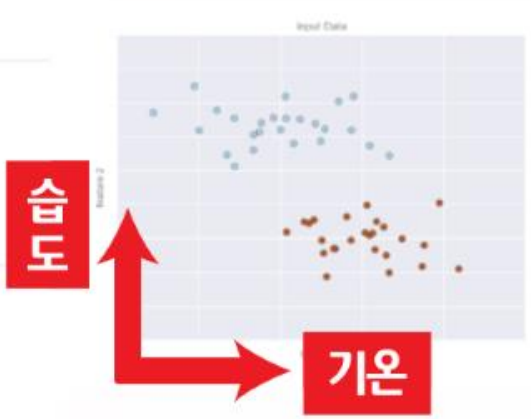
이산적인
범주형 레이블

I 분류 분석의 개념

2. 분류

"이산적인 레이블 예측하기"

기온	습도	불쾌도
20	60	쾌적
30	80	불쾌
30	70	불쾌
35	75	?



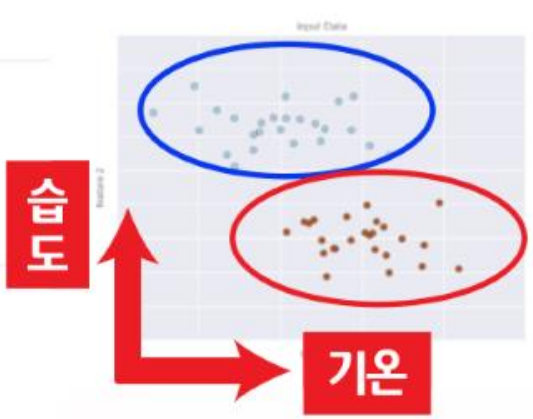
2차원 데이터

I 분류 분석의 개념

2. 분류

"이산적인 레이블 예측하기"

기온	습도	불쾌도
20	60	쾌적
30	80	불쾌
30	70	불쾌
35	75	?



2차원 데이터

파란색 : 쾌적, 빨간색 : 불쾌

I 분류 분석의 개념

2. 분류

"이산적인 레이블 예측하기"

기온	습도	불쾌도
20	60	?
30	80	
30	70	
35	75	?

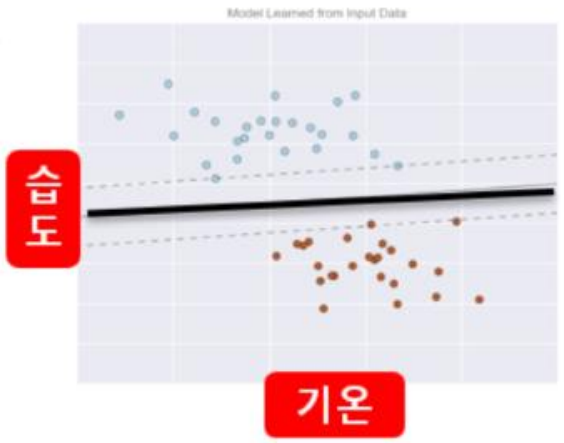
레이블이 없다고 가정했을 때

기존에 조사한 데이터를 사용해서 새로운 데이터의 레이블을 결정하는 모델을 만들 수 있다면?

I 분류 분석의 개념

2. 분류

"모델과 모수"



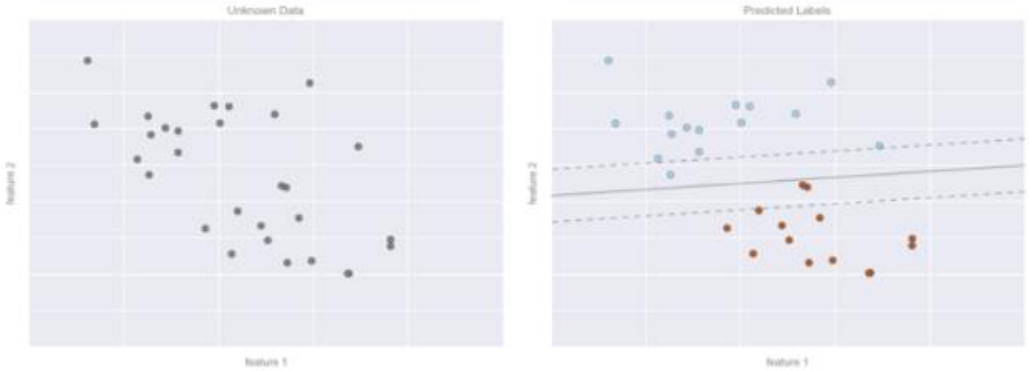
- 하나의 직선이 레이블의 클래스를 구분한다.
- 분류 모델의 정확도는 직선이 어느 위치, 어느 방향에서 그려지는가에 따라 달라진다.
- 모델의 모수 : 직선의 위치와 방향을 설명할 수 있는 특정 숫자
- 모수를 찾는 과정 : 적합(fitting)

I 분류 분석의 개념

2. 분류

"모델 적용"

- 학습한 모델을 이용하여 레이블이 없는 새 데이터를 분류하는 것

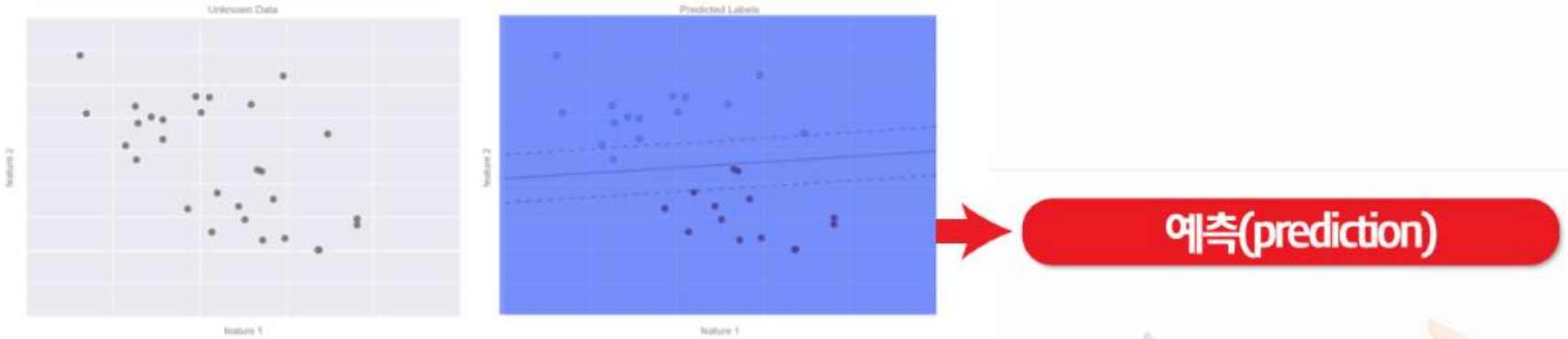


I 분류 분석의 개념

2. 분류

"모델 적용"

- 학습한 모델을 이용하여 레이블이 없는 새 데이터를 분류하는 것

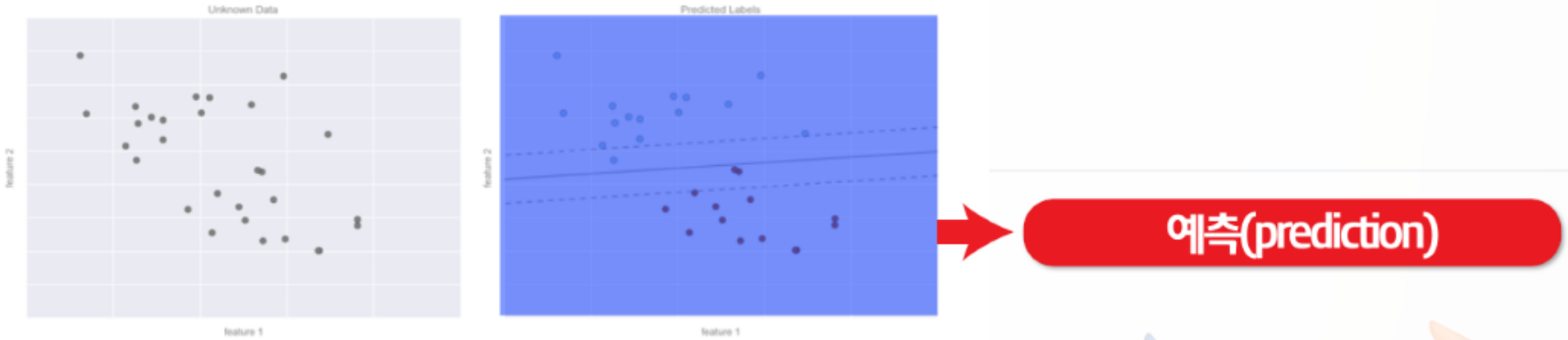


I 분류 분석의 개념

2. 분류

"모델 적용"

- 학습한 모델을 이용하여 레이블이 없는 새 데이터를 분류하는 것



더 큰 차원, 더 큰 데이터 세트로 확장 가능

분류 분석은 머신러닝에서 유용한 도구 중 하나

II 분류 분석의 절차

1. Scikit Learn API 사용 단계

1. Scikit Learn API에서 적절한 추정기 클래스를 임포트 해서 사용하고자 하는 모델을 선택한다.
2. 클래스로부터 인스턴스를 생성하고 초모수(Hyper-parameter)를 설정한다.
3. 데이터를 특징 배열과 대상 배열로 준비한다.
4. 모델 인스턴스의 fit() 메소드를 호출해서 데이터를 학습한다.
5. 정확도를 확인하고 새로운 데이터에 모델을 적용한다.

II 분류 분석의 절차

2. 분류 분석의 예제

"붓꽃 데이터"

```
import seaborn as sns
iris = sns.load_dataset('iris')
iris.head(4)
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa

II 분류 분석의 절차

2. 분류 분석의 예제

"붓꽃 데이터"

```
import seaborn as sns
iris = sns.load_dataset('iris')
iris.head(4)
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa

II 분류 분석의 절차

2. 분류 분석의 예제

"붓꽃 데이터"

```
import seaborn as sns
iris = sns.load_dataset('iris')
iris.head(4)
```

붓꽃의 외형 데이터로부터

붓꽃의 품종을 예측하는 모델 만들기

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 클래스 선택"

- from sklearn.naive_bayes import GaussianNB

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 클래스 선택"

- from sklearn.naive_bayes import GaussianNB

- 가우스 나이브 베이즈 생성 모델
- 처리 속도가 빠르고 초모수를 선택할 필요가 없기 때문에 기본 분류기로 사용하기에 적합

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 인스턴스화"

- model = GaussianNB ()

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 인스턴스화"

- model = GaussianNB ()

II 분류 분석의 절차

2. 분류 분석의 예제

"데이터 분할"

```
from sklearn.model_selection import train_test_split
Xtrain, Xtest, ytrain, ytest = train_test_split(X_iris, y_iris, test_size=0.25,
random_state=1)
```

II 분류 분석의 절차

2. 분류 분석의 예제

```
from sklearn.model_selection import train_test_split
Xtrain, Xtest, ytrain, ytest = train_test_split(X_iris, y_iris, test_size=0.25,
random_state=1)
```

II 분류 분석의 절차

2. 분류 분석의 예제

"데이터에 모델 적합"

- model.fit(Xtrain, ytrain)

II 분류 분석의 절차

2. 분류 분석의 예제

"데이터에 모델 적합"

- model.fit(Xtrain, ytrain)

(특징 행렬, 대상 배열)

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 적용"

- y_model = model.predict(Xtest)

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 적용"

- y_model = model.predict(Xtest)

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 정확도 확인"

- from sklearn.metrics import accuracy_score
  accuracy_score(ytest, y_model);

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 정확도 확인"

- from sklearn.metrics import accuracy_score
  accuracy_score(ytest, y_model);

두 배열 사이에 서로 일치하는 요소의 비율 계산

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 정확도 확인"

- from sklearn.metrics import accuracy_score
  accuracy_score(ytest, y_model);

두 배열 사이에 서로 일치하는 요소의 비율 계산

(붓꽃의 품종, 예측한 레이블)

II 분류 분석의 절차

2. 분류 분석의 예제

"모델 정확도 확인"

- from sklearn.metrics import accuracy_score
  accuracy_score(ytest, y_model);

두 배열 사이에 서로 일치하는 요소의 비율 계산

(붓꽃의 품종, 예측한 레이블)

정확도 97%

이번 시간에는

1 빅데이터 분류 분석 절차

- 분류 분석의 개념
- 분류 분석의 기본 절차

이번 시간에는

1 빅데이터 분류 분석 절차

- 지도 학습
  - 회귀 분석
  - 분류 분석
    - 모델 불러오기
    - 인스턴스 생성
    - 초모수 설정
    - 모델 적합
    - 모델 평가
    - 모델 적용

이번 시간에는

실습 참고 자료

Colab 노트북 파일
Scikit-Learn 공식 사이트 자료
→ https://scikit-learn.org/stable/user_guide.html

다음 시간에는

2 초모수와 모델 검증

- 잘못된 방식의 모델 검증
- 데이터 분할 검증 방법
- 교차 검증 방법