

KNN의 적용

Classification

기계 학습의 일반적인 실습 순서

- 데이터셋 불러오기
 - seaborn 라이브러리 사용, 유명한 데이터 셋 대부분 지원 (예. Iris)
- 데이터셋 카테고리의 실수화
 - setosa, versicolor, virginica → "0", "1", "2"
- 데이터 분할
 - 학습데이터와 테스트 데이터로 나누기
- (옵션) 입력데이터의 표준화
- 모형 추정 혹은 사례중심학습
- 결과 분석
 - Confusion matrix 로 확인

Iris 데이터셋 불러오기

- Iris 데이터셋이란?

- 데이터명 : IRIS (아이리스, 붓꽃 데이터)
- 레코드수 : 150개
- 필드개수 : 5개
- 데이터설명 : 아이리스(붓꽃) 데이터에 대한 데이터. 꽃잎의 각 부분의 너비와 길이 등을 측정한 데이터이며 150개의 레코드로 구성되어 있음.

- 필드의 이해 :

총 6개의 필드로 구성되어있음. caseno는 단지 순서를 표시하므로 분석에서 제외. 2번째부터 5번째의 4개의 필드는 **입력 변수(X)**로 사용되고, 맨 아래의 Species 속성이 **목표(종속) 변수(Y)**로 사용된다.



	caseno	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
1	1	5.1	3.5	1.4	.2	setosa
2	2	4.9	3.0	1.4	.2	setosa
3	3	4.7	3.2	1.3	.2	setosa

Iris 데이터셋 불러오기



```
1 # 3장 KNN
2 import seaborn as sns # seaborn을 불러오고 sns로 축약함.
3 iris=sns.load_dataset('iris') # iris라는 변수명으로 Iris data를 download함.
4 print(iris.head()) # 최초의 5개의 관측치를 print
```



	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
4	5.0	3.6	1.4	0.2	setosa



```
1 print(iris.shape) # iris data의 행과 열의 수
2
3 X = iris.drop('species', axis=1) # 'species'열을 drop하고 input x를 정의함.
4 print(X.shape)
5
6 y=iris['species'] # 'species'열을 label y를 정의함.
```

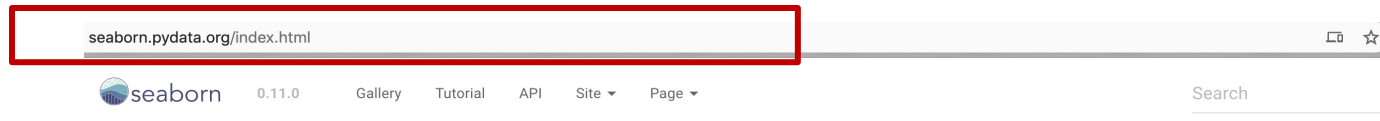


```
(150, 5)
(150, 4)
```

Iris 데이터셋 불러오기

Seaborn 라이브러리란?

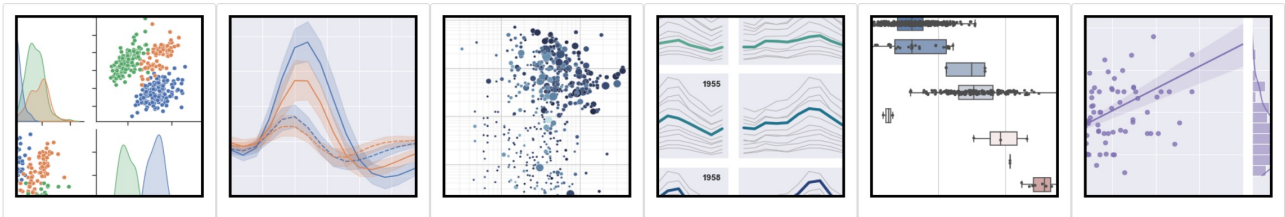
- 파이썬에서 데이터 시각화를 담당하는 모듈
- 유익한 통계 그래픽을 그리기 위한 고급 인터페이스를 제공
- 파이썬 사용자들이 약간의 변수와 파라미터 조정으로 쉽게 그래프를 표현해 볼 수 있게 해주는 도구



seaborn.pydata.org/index.html

seaborn 0.11.0 Gallery Tutorial API Site Page Search

seaborn: statistical data visualization



Seaborn is a Python data visualization library based on [matplotlib](#). It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the [introductory notes](#). Visit the [installation page](#) to see how you can download the package and get started with it. You can browse the [example gallery](#) to see what you can do with seaborn, and then check out the [tutorial](#) and [API reference](#) to find out how.

To see the code or report a bug, please visit the [GitHub repository](#). General support questions are most at home on [stackoverflow](#) or [discourse](#), which have dedicated channels for seaborn.

Contents

- [Introduction](#)
- [Release notes](#)
- [Installing](#)
- [Example gallery](#)
- [Tutorial](#)
- [API reference](#)

Features

- Relational: [API](#) | [Tutorial](#)
- Distribution: [API](#) | [Tutorial](#)
- Categorical: [API](#) | [Tutorial](#)
- Regression: [API](#) | [Tutorial](#)
- Multiples: [API](#) | [Tutorial](#)
- Style: [API](#) | [Tutorial](#)
- Color: [API](#) | [Tutorial](#)

카테고리의 실수화

```
[3] 1 from sklearn.preprocessing import LabelEncoder # LabelEncoder() method를 불러옴
    2 import numpy as np # numpy를 불러옴
    3 classle=LabelEncoder()
    4 y=classle.fit_transform(iris['species'].values) # species 열의 문자열은 categorical 값으로 전환
    5 print('species labels:', np.unique(y)) # 중복되는 y 값을 하나로 정리하여 print
```

```
☞ species labels: [0 1 2]
```

```
▶ 1 yo=classle.inverse_transform(y) # 원래의 species 문자열로 전환
   2 print('species:', np.unique(yo))
```

```
☞ species: ['setosa' 'versicolor' 'virginica']
```

- [주의] DictVectorizer 클래스 vs LabelEncoder 클래스
 - One-hot encoding vs 범주형 라벨

데이터 분할

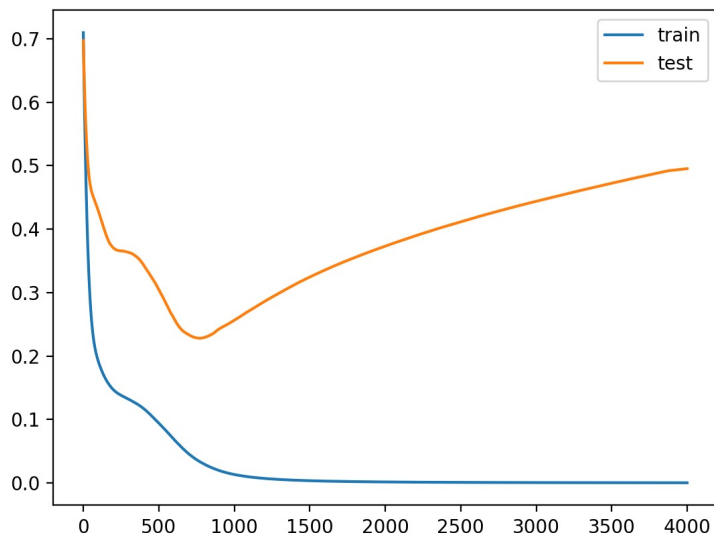


- 데이터 분할이란?

- 학습 데이터(train)와 시험 데이터(test)가 서로 겹치지 않도록 나누는 것

- 데이터 분할의 목적

- 학습데이터로 자료를 학습시키고 학습에 전혀 사용하지 않은 시험데이터에 적용하여 학습 결과의 일반화(generalization)가 가능한지 알아보기 위함



모델이 과적합되었다면, validation 셋으로 검증 시 **예측율이나 오차율이 떨어지는 현상**을 확인할 수 있으며, 이런 현상이 나타나면 **학습을 종료**

데이터 분할

```
1
2 from sklearn.model_selection import train_test_split #Scikit-Learn 의 model_selection library를 train_test_split로 명명
3 X_train,X_test,y_train,y_test=train_test_split(X,y, test_size=0.3, random_state=1, stratify=y) # x와 y의 data를 각각 30%, 70%의 비율
4 print(X_train.shape)
5 print(X_test.shape)
6 print(y_train.shape)
7 print(y_test.shape)
```

```
↳ (105, 4)
   (45, 4)
   (105,)
   (45,)
```

▪ train_test_split() 함수의 인자 설명

옵션 값 설명

- **test_size**: 테스트 셋 구성의 비율을 나타냅니다. train_size의 옵션과 반대 관계에 있는 옵션 값이며, 주로 test_size를 지정해 줍니다. 0.2는 전체 데이터 셋의 20%를 test (validation) 셋으로 지정하겠다는 의미입니다. **default 값은 0.25** 입니다.
- **shuffle**: **default=True** 입니다. split을 해주기 이전에 섞을건지 여부입니다. 보통은 default 값으로 놔둡니다.
- **stratify**: **default=None** 입니다. classification을 다룰 때 매우 중요한 옵션값입니다. stratify 값을 target으로 지정해주면 각각의 **class 비율(ratio)**을 **train / validation에 유지**해 줍니다. (한 쪽에 **쏟려서 분배되는 것을 방지**합니다) 만약 이 옵션을 지정해 주지 않고 classification 문제를 다룬다면, 성능의 차이가 많이 날 수 있습니다.
- **random_state**: 세트를 섞을 때 해당 int 값을 보고 섞으며, 하이퍼 파라미터를 튜닝시 이 값을 고정해두고 튜닝해야 매번 데이터셋이 변경되는 것을 방지할 수 있습니다.

모형 추정 및 사례중심 학습



```
1 # KNN 의 적용
2 from sklearn.neighbors import KNeighborsClassifier #KNN 불러오기
3 knn=KNeighborsClassifier(n_neighbors=5,p=2) #5개의 인접한이웃, 거리측정기준:유클리드
4 #knn.fit(X_train_std,y_train) #모델 fitting과정
5 knn.fit(X_train,y_train) #모델 fitting과정
```

```
[> KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                        metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                        weights='uniform')
```

```
[17] 1 #y_train_pred=knn.predict(X_train_std) #train data의 y값 예측치
      2 y_train_pred=knn.predict(X_train) #train data의 y값 예측치
      3 y_test_pred=knn.predict(X_test_std) #모델을 적용한 test data의 y값 예측치
      4 y_test_pred=knn.predict(X_test) #모델을 적용한 test data의 y값 예측치
      5 print('Misclassified training samples: %d' %(y_train!=y_train_pred).sum()) #오분류 데이터 갯수 확인
      6 print('Misclassified test samples: %d' %(y_test!=y_test_pred).sum()) #오분류 데이터 갯수 확인
```

```
[> Misclassified training samples: 2
     Misclassified test samples: 1
```

```
[18] 1 from sklearn.metrics import accuracy_score #정확도 계산을 위한 모듈 import
      2 print(accuracy_score(y_test,y_test_pred)) # 45개 test sample중 42개가 정확하게 분류됨.
```

```
[> 0.9777777777777777
```

결과 분석

- 성능 평가
 - 분류 문제는 회귀 분석과 달리 다양한 성능 평가 기준(metric)이 필요함
 - 평가 방법마다 장단점이 존재함
- 싸이킷런에서 제공하는 분류 성능 평가 방법
 - `confusion_matrix(y_true, y_pred)`
 - `accuracy_score(y_true, y_pred)`
 - `precision_score(y_true, y_pred)`
 - `recall_score(y_true, y_pred)`
 - `fbeta_score(y_true, y_pred, beta)`
 - `f1_score(y_true, y_pred)`
 - `roc_curve`
 - `auc`

결과 분석

	예측 클래스 0	예측 클래스 1	예측 클래스 2
정답 클래스 0	정답 클래스가 0, 예측 클래스가 0인 표본의 수	정답 클래스가 0, 예측 클래스가 1인 표본의 수	정답 클래스가 0, 예측 클래스가 2인 표본의 수
정답 클래스 1	정답 클래스가 1, 예측 클래스가 0인 표본의 수	정답 클래스가 1, 예측 클래스가 1인 표본의 수	정답 클래스가 1, 예측 클래스가 2인 표본의 수
정답 클래스 2	정답 클래스가 2, 예측 클래스가 0인 표본의 수	정답 클래스가 2, 예측 클래스가 1인 표본의 수	정답 클래스가 2, 예측 클래스가 2인 표본의 수

- **혼합 행렬 (confusion matrix):** 타겟의 원래 클래스와 모형이 예측한 클래스가 일치하는지는 갯수로 센 결과를 표나 나타낸 것

```
1 from sklearn.metrics import confusion_matrix# 오분류표 작성을 위한 모듈 import
2 conf=confusion_matrix(y_true=y_test,y_pred=y_test_pred) # 대각원소가 각각 setosa, versicolor, virginica를 정확하게 분류한 갯수.
3 print(conf)
4 # setosa는 모두 정확하게 분류되었고 versicolor는 15개 중 2개가 virginica로 오분류 되었으며 virginica는 15개 중 1개가 versicolor로 오분류됨.
```

```
[[15  0  0]
 [ 0 13  2]
 [ 0  1 14]]
```

(옵션) 입력데이터의 표준화

■ 표준화

- 특성 자료의 측정 단위(Scaling)에 의해 영향 받지 않도록 하는 과정
- 사이킷런의 **StandardScaler 클래스**를 호출하여 사용
- 시험 데이터(test data)의 표준화는 학습 데이터(train data)에서 구한 특성 변수의 평균과 표준편차를 이용함
- 표준화로 인해 데이터의 분포인 통계적 특성이 깨지면 머신러닝의 학습 저하를 가져옴

```
[7] 1 from sklearn.preprocessing import StandardScaler #Scikit-Learn 의 model_selection library를 train_test_split로 명명
    2 sc=StandardScaler()
    3 sc.fit(X_train)
    4 X_train_std=sc.transform(X_train) # training data의 표준화
    5 X_test_std=sc.transform(X_test) # test data의 표준화
    6
    7 #표준화된 data의 확인
    8 print(X_train.head()) # X_train data 최초 5개의 관측치
    9 X_train_std[1:5,] # X_train_std data 최초 5개의 관측치
```

```
sepal_length  sepal_width  petal_length  petal_width
33           5.5         4.2          1.4          0.2
20           5.4         3.4          1.7          0.2
115          6.4         3.2          5.3          2.3
124          6.7         3.3          5.7          2.1
35           5.0         3.2          1.2          0.2
array([[ -0.55053619,  0.76918392, -1.16537974, -1.30728421],
       [ 0.65376173,  0.30368356,  0.84243039,  1.44587881],
       [ 1.0150511 ,  0.53643374,  1.0655204 ,  1.18367281],
       [-1.03225536,  0.30368356, -1.44424226, -1.30728421]])
```

(옵션) 입력데이터의 표준화

```
1 # KNN 의 적용
2 from sklearn.neighbors import KNeighborsClassifier #KNN 불러오기
3 knn=KNeighborsClassifier(n_neighbors=5,p=2) #5개의 인접한이웃, 거리측정기준:유클리드
4 knn.fit(X_train_std,y_train) #모델 fitting과정
5 #knn.fit(X_train,y_train) #모델 fitting과정
```

```
↳ KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                        metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                        weights='uniform')
```

```
[11] 1 y_train_pred=knn.predict(X_train_std) #train data의 y값 예측치
      2 #y_train_pred=knn.predict(X_train) #train data의 y값 예측치
      3 y_test_pred=knn.predict(X_test_std) #모델을 적용한 test data의 y값 예측치
      4 #y_test_pred=knn.predict(X_test) #모델을 적용한 test data의 y값 예측치
      5 print('Misclassified training samples: %d' %(y_train!=y_train_pred).sum()) #오분류 데이터 갯수 확인
      6 print('Misclassified test samples: %d' %(y_test!=y_test_pred).sum()) #오분류 데이터 갯수 확인
```

```
↳ Misclassified training samples: 4
    Misclassified test samples: 3
```

```
[12] 1 from sklearn.metrics import accuracy_score #정확도 계산을 위한 모듈 import
      2 print(accuracy_score(y_test,y_test_pred)) # 45개 test sample중 42개가 정확하게 분류됨.
```

```
↳ 0.9333333333333333
```

- 표준화로 인해 정확도가 97.8 ➔ 93.3 으로 떨어진 사례
- 표준화 여부는 시험 데이터(test data)의 정밀도(accuracy)를 점검 하여 결정함