Datamining rep2　　205713B 朝比奈太郎　2022/05/30

Level1

preprocess_numerical.ipynb を参考に、何か一つ以上の特徴に対して前処理を施せ。なお、選択理由についても検討すること。

・選択した前処理 -> min-max scalering

・選んだ理由

Min-max 法はデータの最大値と最小値の範囲が明確な場合に適した手法である。また、min-max 法は外れ値に敏感であるが、今回の対象のデータでは目立った外れ値がないことから min-max 法を採用した。

・Min-max 法の定義

$$x_i' = \frac{x\_i - \min(x)}{\max(x) - \min\,(x)}$$

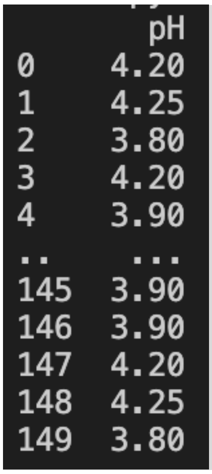
(i = 1 ~ n)

データ全体を正規化するには、1 番目から n 番目までの各データを 1 つ 1 つスケーリングしていく必要がある。

Level2

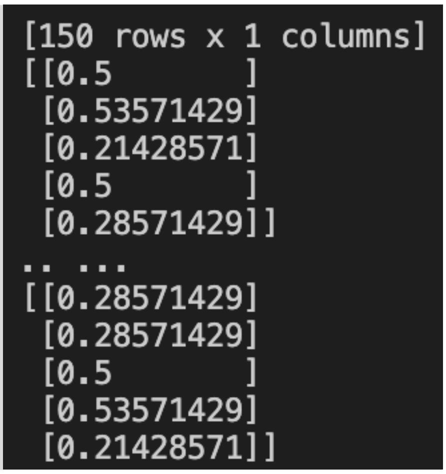
特徴ベクトルの 1 つである、[pH]に対して前処理を施した。

処理前



```
        pH
0     4.20
1     4.25
2     3.80
3     4.20
4     3.90
..     ...
145   3.90
146   3.90
147   4.20
148   4.25
149   3.80
```

処理後



```
[150 rows x 1 columns]
[[0.5       ]
 [0.53571429]
 [0.21428571]
 [0.5       ]
 [0.28571429]]
.. ...
[[0.28571429]
 [0.28571429]
 [0.5       ]
 [0.53571429]
 [0.21428571]]
```

Level3

前処理後のデータを用い、分類タスクを実行せよ。課題レポート 1 の結果（前処理なしでの結果）と比較し、考察せよ。

Source code

```
import pandas as pd
from sklearn import preprocessing
from sklearn.svm import LinearSVC
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold

df = pd.read_csv("beer.csv")
temp = pd.DataFrame(df["pH"])
#min-max 法
min_max_sacler = preprocessing.MinMaxScaler()
temp_minmax = min_max_sacler.fit_transform(temp)
df["pH"] = temp_minmax #pH の値を min-max 法で求めた値に置き換える
#print(df["pH"]) 置き換えられた pH の値を出力

X = df[["OG", "ABV", "pH", "IBU"]]
Y = df["style"]
Cs = [0.5, 1.0, 1.5] #ハイパーパラメータ
k_folds = 5 #5 分割検定

for c in Cs:
    model = LinearSVC(C=c)
    scores = cross_val_score(model, X, Y, cv=KFold(n_splits=k_folds,
shuffle=True))
    #KFold(n_splits(分割個数), shuffle(シャッフル), random_state(乱数シード))
    average = scores.mean()
    model.fit(X, Y)
    print(model.predict(X))
    print(f'C = {c}: scores={scores}, average={average:.3f}')
```

分類結果と学習評価

C = 0.5 のとき

分類結果

```
LinearSVC(C=0.5)
['Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'Premium Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Premium Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Premium Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Premium Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager']
```

学習評価

```
C = 0.5: scores=[1.         0.93333333 1.         0.93333333 1.        ],
average=0.973
```

C = 1.0 のとき

分類結果

```
LinearSVC()
['Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'Premium Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Premium Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Premium Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager']
```

学習評価

```
C = 1.0: scores=[0.96666667 0.96666667 1.         0.93333333 1.        ],
average=0.973
```

C = 1.5 のとき

分類結果

```
LinearSVC(C=1.5)
['Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager'
 'Premium Lager' 'Premium Lager' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA'
 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'IPA' 'Premium Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Premium Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Premium Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager' 'Light Lager'
 'Light Lager' 'Light Lager' 'Light Lager']
```

学習評価

```
C = 1.5: scores=[1.         0.96666667 1.         0.96666667 0.96666667],
average=0.980
```

前回の結果(前処理なし)の場合と比較して考察せよ。

前回の学習評価は

	Average
C = 0.5	0.990
C = 1.0	0.9874
C = 1.5	0.9874

これに比べて今回の学習評価は

	Average of aver
C = 0.5	0.973
C = 1.0	0.973
C = 1.5	0.980

となった。

Min-max 法で前処理を行った後のデータの方が前処理なしのデータに比べ、5 分割検定において精度が低いという結果が出た。この要因として、前処理の方法が適していなかった。サンプル数が少ないために評価にばらつきが出た。といった 2 つが挙げられる。Min-max 法と似た手法で標準化という前処理方法もある。Min-max 法と標準化の主な使い分けは、最大値および最小値が決まっているか。また、外れ値があるかどうかで使い分けをすることが多い。今回のデータにおいて私は最大値および最小値が決まっており、目立った外れ値がないため min-max 法を採用した。しかし、実際には機械学習を行う上での外れ値が存在したため精度が低くなったのではないかと考える。従って今後機械学習を行なっていく上で学習コストが不問である場合は積極的に標準化を使っていきたいと思う。