Level3

採用したモデル: sklearn.svm.LinearSVC

クラス定義:

sklearn.svm.LinearSVC(penalty='l2', loss='squared_hinge', *, dual=True, tol=0.0001, C=1.0 , multi_class='ovr', fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0 , random_state=None, max_iter=1000)[source]

ハイパーパラメータの説明

C: float 型。Default = 1.0。正則化パラメータ。正則化の強さは C に反比例する。C は必ず正の値でなければならない。

max_iter: int 型。Default = 1000。実行する反復処理の最大数。

Level4

Code

```
import pandas as pd
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")
#print(df)
X = df[["OG", "ABV", "pH", "IBU"]]
Y = df["style"]
print("X.shape = ", X.shape) shape -> (行数, 列数)をタプルで取得
print("Y.shape = ", Y.shape)
print(X.loc[0]) loc -> 行列と列名で位置を指定
print(Y[0])
X.shape = (150, 4)
Y.shape = (150,)
OG
        11.5
ABV
         4.2
         4.2
рΗ
```

```
IBU 14.0 Name: 0, dtype: float64 Premium Lager """" 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() print(model.fit(X, Y)) print(model.predict(X)) print(f'C = \{c\}: scores=\{scores\}, average=\{average:.3f\}')
```

Code の説明。

今回の機械学習では、Flow chart を参考にして sklearn の LinearSVC を採択。評価には cross_val_score を用いた。

対象となるデータセットは beer.csv であり、数値データ(量的変数)である、"OG", "ABV", "pH", "IBU"を X とし、分類クラスに当たる"style"を Y とした。ハイパーパラメータである C を[0.5, 1.0, 1,5]でチューニングし、150 の data を 5 分割検定により training set, testing set に分け学習を行った。scores には 0~1 の値が入り、精度が高ければ高いほど、1 に近づく。5 分割検定したデータ(scores)の平均値を average という形で算出した。

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 Lage

'Premium Lager' 'IPA' 'Light Lager' |

学習評価

C = 1.0 のとき

分類結果

LinearSVC()

['Premium Lager' 'Premium Lage

'Premium Lager' 'IPA' 'Premium Lager' 'Light Lager']

学習評価

C = 1.5 のとき

分類結果

LinearSVC(C=1.5)

['Premium Lager' 'Premium Lage

'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager' 'Premium Lager' 'IPA' 'Light Lager' 'Light Lager']

学習評価

C = 1.5: scores=[0.93333333 1. 0.96666667 1. 1.], average=0.980

Option

実行毎に average が異なる値を出すことより、さらに 10 回試行して C が 0.5,1.0,1.5 のときの average の平均をとって average の精度を上げ、C の値が 3 つのうちどの場合において一番精度が高いか調べてみた。

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| C=0.5 | 0.993 | 0.987 | 0.993 | 0.987 | 0.993 | 0.987 | 0.993 | 0.993 | 0.987 | 0.987 |
| C=1.0 | 0.987 | 0,987 | 0.987 | 0.993 | 0.987 | 0.980 | 0.993 | 0.980 | 0.993 | 0.987 |
| C=1.5 | 0.980 | 0,987 | 0.987 | 0.967 | 0.987 | 0.993 | 0.993 | 0.987 | 0.993 | 1.000 |

従って、下の表より、C = 0.5 のときが一番精度が高い結果となった。

| | Average of aver |
|---------|-----------------|
| C = 0.5 | 0.990 |
| C = 1.0 | 0.9874 |
| C = 1.5 | 0.9874 |