ML TA hours

HW1 Review

2024.10.1

HW1 Common mistake - KFold

data seems to be sequential

	Youtuber	subscribers	video views	category	uploads	Abbreviation	channel_type_rank	video_views_for_the_last_30_days	lowest_monthly_earnings	highe
0	T-Series	245000000	2.280000e+11	Music	20082	IN	1.0	2.258000e+09	564600.0	
1	YouTube Movies	170000000	0.000000e+00	Film & Animation	1	US	7423.0	1.200000e+01	0.0	
2	MrBeast	166000000	2.836884e+10	Entertainment	741	US	1.0	1.348000e+09	337000.0	
3	Cocomelon - Nursery Rhymes	162000000	1.640000e+11	Education	966	US	1.0	1.975000e+09	493800.0	
4	SET India	159000000	1.480000e+11	Shows	116536	IN	2.0	1.824000e+09	455900.0	
...	...	...	...	...	...	...	...	...	...	
803	Migos ATL	12400000	6.993406e+09	Music	99	US	171.0	4.941200e+07	12400.0	
804	Natan por Aï¿	12300000	9.029610e+09	Sports	1200	BR	172.0	5.525130e+08	138100.0	
805	Free Fire India Official	12300000	1.674410e+09	People & Blogs	1500	IN	69.0	6.473500e+07	16200.0	
806	RobTopGames	12300000	3.741235e+08	Gaming	39	SE	69.0	3.871000e+06	968.0	
807	Make Joke Of	12300000	2.129774e+09	Comedy	62	IN	44.0	2.400000e+07	6000.0	

808 rows × 21 columns

HW1 Common mistake - KFold

```
class sklearn.model_selection.KFold(n_splits=5, *, shuffle=False, random_state=None)
```

shuffle : *bool, default=False*

Whether to shuffle the data before splitting into batches. Note that the samples within each split will not be shuffled.

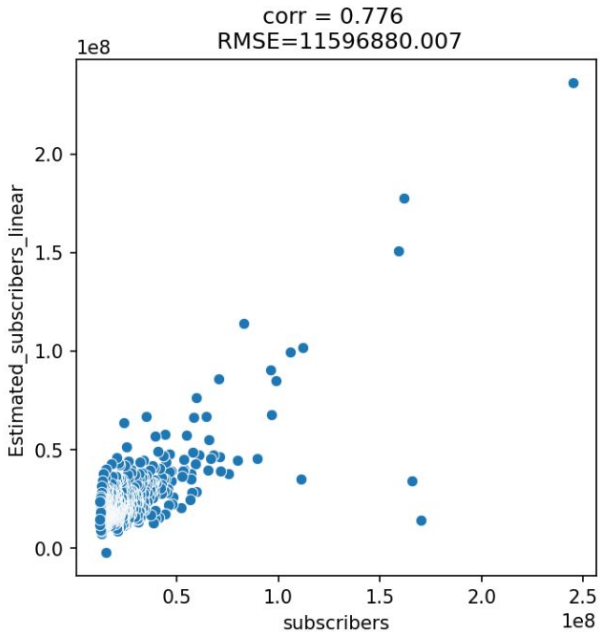
```
# Create CV folds
num_folds = 5
kf = KFold(n_splits=num_folds, random_state=0, shuffle=True)
kfold_indices = {}

for i, (train_index, test_index) in enumerate(kf.split(X)):
    kfold_indices[f"fold_{i}"] = {'train': train_index, 'test': test_index}
```

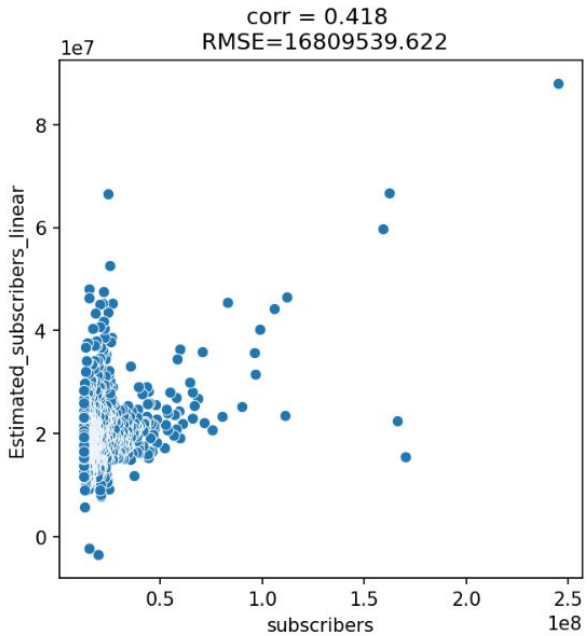
ensure that each fold has a mixture of data

HW1 Common mistake - KFold

shuffle



No shuffle



HW1 Well done！

- Clearly formatted
- Do more experiments:
  - feature selection
  - tuning hyperparameters

If you do extra work, please record it clearly in your homework.

ML TA hours

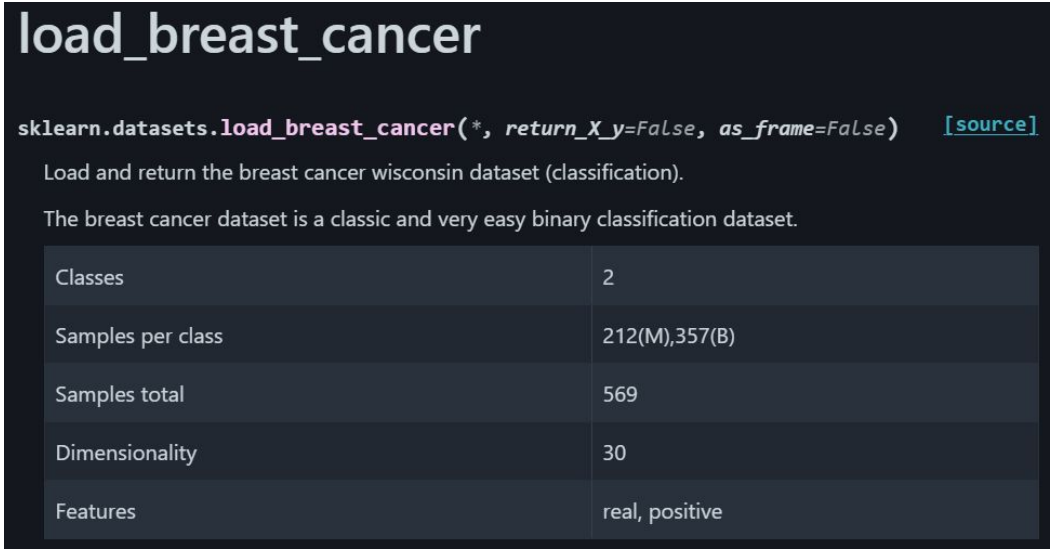
HW3

Colab experiment : Breast cancer classification

2024.10.1

Task description

- In this exercise, we load the Breast cancer wisconsin dataset for classification.
- Binary classification:
  - benign
  - malignant



load_breast_cancer

`sklearn.datasets.load_breast_cancer(*, return_X_y=False, as_frame=False)` [source]

Load and return the breast cancer wisconsin dataset (classification).

The breast cancer dataset is a classic and very easy binary classification dataset.

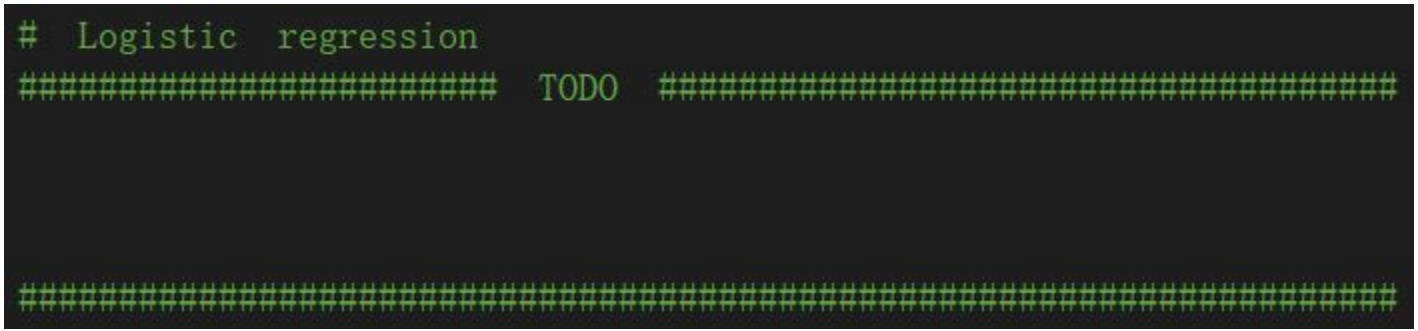
Classes	2
Samples per class	212(M),357(B)
Samples total	569
Dimensionality	30
Features	real, positive

Task description

- Your task is to build and evaluate three models:
  - Logistic Regression
  - SVM
  - Decision Tree

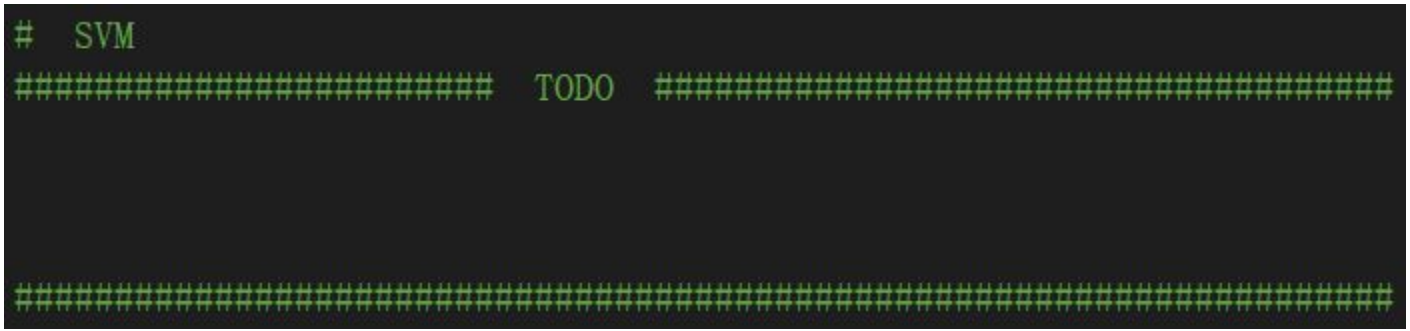
TODO - Logistic regression [ref]

- Make sure to include the following components:
  - StandardScaler() : scaling your features before applying model
  - GridSearchCV() : to search for the optimal hyperparameters
  - Error_rate : calculates the classification error rate on the test data
    - metrics: zero_one_loss



TODO - SVM [ref]

- Make sure to include the following components:
  - StandardScaler() : scaling your features before applying model
  - GridSearchCV() : to search for the optimal hyperparameters
  - Error_rate : calculates the classification error on the test data
    - metrics: zero_one_loss



TODO - Decision tree [ref]

- Make sure to include the following components:
  - GridSearchCV() : to search for the optimal hyperparameters
  - Error_rate : calculates the classification error on the test data
    - metrics: zero_one_loss

```
# Decision tree
########################## TODO ######################################



##############################################################################
```

TODO - Show the results

- report the **mean** and **standard deviation** of the error rates over 5 folds for each method

```
######################  TODO  ###################################
print(f"The  error  rate  over  5  folds  in  CV:")


#################################################################
```

Baseline results

```
The error rate over 5 folds in CV:
Logistic Regression: mean = 0.0298, std = 0.0142
SVM: mean = 0.0263, std = 0.0136
Decision Tree: mean = 0.0632, std = 0.0237
```

Summarize

- Build and evaluate three models
  - Logistic Regression
  - SVM
  - Decision Tree
- Show the results (mean and std)
- Description of the methodology
- Conclusions and discussions

NOTE

- After executing your code, download the .ipynb file and submit it to NTU COOL
- Submitted file name: student ID_week2_colab_homework.ipynb
- HW3 Deadline : 2024/10/7 23:59

- If there are any questions

  Email: r12945048@ntu.edu.tw (add “[ML HW3]” to the beginning of the title.)