Data Science HW # 4

Buy Less, Choose Well! Fashion Item Classification

HW # 4: Buy Less, Choose Well! Fashion Item Classification

- Kaggle In-class
- Link: https://www.kaggle.com/c/2018falldatascience-hw4
- Deadline: 17/12/2018 11:59 PM

HW # 4: Buy Less, Choose Well! Fashion Item Classification

- Task is to preditct whether the images are belong to which kind of categories. The images data have already transform into pixel value from 0(white) to 255(black) in grayscale.
- The data are consisting of a training set of 60,000 examples and a test set of 10,000 examples(50% for private and 50 % for public).

Data

- Therer are 784 attributes for one instance, which represesnt the grayscale of the 28x28 image.
- Pixel: from 0(white) to 255(Black)

Training data

label	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel9	pixel10	pixel11	pixel12	pixel13	pixel14	pixel15	pixel16	pixel17	pixel18	pixel19	pixel20	pixel21	pixel22	pixel23	pixel24	pixel25	pixel26
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	5	0	0	0	105	92	101	107	100	132	0	0	2	4	0	1	0	0	0
0	0	0	0	1	2	0	0	0	0	0	114	183	112	55	23	72	102	165	160	28	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	46	0	21	68	0	0	0	0	0	0	0	0	0	0
4	0	0	0	5	4	5	5	3	5	6	2	0	0	0	0	0	0	0	2	3	7	4	4	5	5	0
4	0	0	0	0	0	0	0	0	0	0	0	159	161	143	180	142	21	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	3	2	0	0	60	234	215	229	223	231	107	0	0	0	0	4	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	141	45	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	41	162	167	84	30	38	94	177	176	26	0	0	0	1	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	18	143	174	176	150	20	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	1	1	0	0	0	0	49	218	208	185	200	202	85	11	0	0	0	3	0	0	0	0
2	0	0	0	0	0	0	0	0	16	43	71	55	29	17	18	33	59	88	50	28	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	34	109	9	0	0	0	80	137	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	38	155	0	0	0	0	0	0	108	94	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	25	111	107	12	171	141	0	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	5	52	67	32	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	1	32	88	149	153	89	44	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	1	0	0	0	109	65	42	47	92	91	0	0	0	3	0	0	0	0	0

Testing data

	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel9	pixel10	pixel11	pixel12	pixel13	pixel14	pixel15	pixel16	pixel17	pixel18	pixel19	pixel20	pixel21	p
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	31	116	96	132	125	110	134	117	91	87	23	0	
3	0	0	0	0	0	0	0	0	0	47	168	107	59	52	50	55	65	78	193	104	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	0	0	0	0	0	0	0	11	50	41	23	34	29	32	18	20	31	32	45	50	0	
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Evaluation

- The competition will take 50% of the test data to calculate the accuracy
- Final rank will show on E3 after the competition.
- The evaluation for this competition is Mean F1-Score.

$$F1 = 2\frac{p \cdot r}{p + r} \quad \text{where } p = \frac{tp}{tp + fp}, \quad r = \frac{tp}{tp + fn}$$

Submission files

- The maximum number of daily submissions is 10.
- The file should be CSV file contains two columns:

ID: the index

label: between 0 and 9

	id	label
1		
2	0	0
3	1	0
4	2	0
5	3	0
6	4	0
7	5	0
8	6	0
9	7	0
10	8	0
11	9	0

Grading policy

Below baseline (0.8): 0

top 10%: 100

top 25%: 90

top 50%: 80

top 75%: 75

Others: 70

Requirements

Please archive your code as zip NOT RAR NOT 7Z , testing result and submit on E3.

Deadline: 10/12/2018 11:59 PM

Submission folder (your team name on Kaggle) should contain 3 files:

- [Student ID].py

- answer.csv

- readme

Details

- folder名字是你在kaggle上的隊伍名
- 裡面的主程式請命名為 學號.py
- 主程式請寫成可以讀取兩個檔案的格式(train.csv/ test.csv)，這兩個檔案的內容就是kaggle上的檔案。你的程式要用train.csv的資料，train出一個model，然後用test.csv做預測並寫入answer.csv，寫入的格式就按照交到kaggle上的格式一樣。
- 若用了特殊的packages請寫在readme中。

Contact Information

If you have any questions, please email 許家銘

email : ming2242.cm06@g2.nctu.edu.tw

or feel free to come 716 during TA hours