

# データサイエンス概論 第一回レポート

201821636 村松 直哉

平成 30 年 11 月 19 日

このレポートにはアダルトな項目を含んでいます。18 歳未満の方は読まないでください。

## 1 データセット

本レポートでは、Sexualitics<sup>\*1</sup>の xHamster<sup>\*2</sup>に関するデータセット<sup>\*3</sup>を利用する。

データセットの構造を表 1 に示す。ただし表が大きくなりすぎるため、一部を省略している。

## 2 重回帰分析

### 2.1 利用したデータ

今回の分析では、説明変数として「nb\_views（その動画の視聴回数）」、「nb\_comments（動画に対して投稿されたコメント数）」、「runtime（動画の時間（秒））」を利用した。目的変数として「nb\_votes（動画の平均評価値）」を利用した。この値は、動画に対して 1 ユーザーにつき、良かったか悪かったかどちらかを投票できる。nb\_votes は、良かった時を 1、悪かった時を -1 として、全投票の合計値である。

データ数はそれぞれ 786121 個ある。

### 2.2 重回帰分析結果

重回帰分析に用いたプログラムのソースコードを 1 に示す。プログラミング言語は Python を利用した。各要素が nb\_votes に与える影響を詳しく知りたかったため、説明変数はすべて標準化している。

ソースコード 1 重回帰分析のプログラム

```
1 import os
2 import numpy as np
3 import pandas as pd
4 import matplotlib.pyplot as plt
5 %matplotlib inline
6
7 DATASET_DIR = '/data'
8
9 # === データセットの読み込み ===
10 path = os.path.join(DATASET_DIR, 'xhamster.csv')
11 xhamster_dataset = pd.read_csv(path, sep=",")
12 xhamster_dataset.info()
13 xhamster_dataset.head() # データを表示
14
15 # === 重回帰分析 ===
```

---

\*1 <http://sexualitics.github.io/>

\*2 <https://xhamster.com/>

\*3 <http://pornstudies.sexualitics.org/data/xhamster.csv.tar.gz>

```

16 from sklearn import linear_model
17 clf = linear_model.LinearRegression()
18
19 # データの標準化
20 dataset = xhamster_dataset[['nb_views', 'nb_comments', 'runtime', 'nb_votes']]
21 dataset = dataset.apply(lambda x: (x - np.mean(x)) / (np.max(x) - np.min(x)))
22 print(dataset.head())
23 print('')
24
25 # 説明変数
26 xs = dataset[['nb_views', 'nb_comments', 'runtime']]
27 X = xs.values
28 # 目的変数
29 Y = dataset['nb_votes'].values
30
31 X = np.nan_to_num(X)
32 Y = np.nan_to_num(Y)
33 print('X_shape: {}'.format(X.shape))
34 print('Y_shape: {}'.format(Y.shape))
35 print('')
36
37 # 予測モデルを作成
38 clf.fit(X, Y)
39
40 # 偏回帰係数
41 print(pd.DataFrame({
42     "Name": xs.columns,
43     "Coefficients": clf.coef_
44 }).sort_values(by='Coefficients'))
45
46 # 切片 (誤差)
47 print(clf.intercept_)

```

出力結果を以下に示す.

```

nb_views  nb_comments  runtime  nb_votes
0 -0.007029    -0.000810 -0.000202 -0.003661
1 -0.008997         NaN -0.000240 -0.007272
2 -0.008380    -0.009860 -0.000186 -0.006806
3 -0.007574     0.003715  0.000474 -0.002380
4 -0.005144    -0.006466  0.000234 -0.003079

```

X shape: (786121, 3)

Y shape: (786121,)

```

Coefficients      Name
2    -0.008247      runtime
1     0.332681  nb_comments
0     0.431295    nb_views
-1.1260282263925007e-20

```

したがって、最終的に得られる回帰式は以下ようになる。ただし、 $x_1$  は nb\_views,  $x_2$  は nb\_comments,  $x_3$  は runtime,  $y$  は nb\_votes を示す。

$$y = -1.13 \times 10^{-20} + 0.43x_1 + 0.33x_2 - 0.008x_3 \quad (1)$$

この結果から nb\_views が, nb\_votes に最も大きな影響を与えているのがわかる. 一方で視聴回数 (nb\_views) が多い場合, 投票する人が増えるため, 各作品を平等に評価できない可能性がある. 同様に, コメント数 (nb\_comments) も nb\_views に相関すると考えられる.

以上のことから, 説明変数同士の相関の影響が現れている可能性がある. 説明変数同士の回帰分析を行うことで, nb\_votes に対する影響の詳細を知ることができる.

またタグ情報も nb\_votes や nb\_views に対して, 影響を及ぼしていることが考えられる. 各タグに対する集計を取ることで, 人気のタグがわかると予想される.

## 2.3 データ行列

重回帰分析を理解するために, 説明変数のデータ行列をソースコード 2 により求めた.

ソースコード 2 データ行列のプログラム

```
1 import os
2 import numpy as np
3 import pandas as pd
4 import matplotlib.pyplot as plt
5 %matplotlib inline
6
7 DATASET_DIR = '/data'
8
9 # === データセットの読み込み ===
10 path = os.path.join(DATASET_DIR, 'xhamster.csv')
11 xhamster_dataset = pd.read_csv(path, sep=",")
12 xhamster_dataset.info()
13 xhamster_dataset.head() # データを表示
14
15 # === データ行列 ===
16 # 説明変数
17 dataset = xhamster_dataset[['nb_views', 'nb_comments', 'runtime']]
18 dataset = dataset.apply(lambda x: (x - np.mean(x)) / (np.max(x) - np.min(x)))
19 X = dataset[['nb_views', 'nb_comments', 'runtime']]
20 data_num = X.shape[0]
21 X = np.concatenate((np.ones([data_num, 1], dtype=np.float32), X), axis=1)
22 X = np.nan_to_num(X)
23
24 data_matrix = np.dot(X.T, X)
25
26 print(data_matrix)
```

出力結果を以下に示す.

```
[[ 7.86121000e+05 -3.05006020e-13 -1.42080792e-13 -6.22852464e-15]
 [-3.05006020e-13  3.43732358e+02  1.48922987e+02  1.23231432e+00]
 [-1.42080792e-13  1.48922987e+02  2.05637159e+02  4.96063643e-01]
 [-6.22852464e-15  1.23231432e+00  4.96063643e-01  1.17774672e+00]]
```

表 1 xHamster データセット

id	upload_date	title	channels	description	nb_views	nb_votes	nb_comments	runtime	uploader
378466	2010-06-29	girl riding black cock	['BBW', (省略)]	(省略)	17262	65	11	120	(省略)
478576	2010-11-07	masturbation	['Masturbation']	(省略)	953	3	NA	15	(省略)
287146	2010-02-12	sexy horny booty dance	['Babes', (省略)]	(省略)	6060	11	3	163	(省略)
1583074	2012-11-18	its rather big	['Amateur', (省略)]	NA	64413	55	4	146	(省略)
1246688	2012-06-01	dildo	['Men']	NA	247	1	NA	51	(省略)
1450532	2012-09-14	tribute to me	['Men']	(省略)	428	18	128	146	(省略)
1450531	2012-09-14	(省略)	['Men']	NA	199	1	1	57	(省略)
1450537	2012-09-14	(省略)	['Double Penetration', (省略)]	NA	33481	194	14	1137	(省略)
1450536	2012-09-14	x036	['BDSM', (省略)]	Part 3/3	104039	72	7	619	(省略)
1450535	2012-09-14	boy and mature woman part1	['Matures', (省略)]	NA	400167	171	3	240	(省略)
1450534	2012-09-12	(省略)	['Amateur', (省略)]	(省略)	39276	37	4	780	(省略)
968454	2011-12-22	(省略)	['Brunettes', (省略)]	NA	20864	40	3	404	(省略)
1246681	2012-06-01	amante tetona 01	['Amateur', (省略)]	Amateur	8921	16	1	127	(省略)
1246682	2012-06-01	british slut victoria	['Amateur', (省略)]	NA	180901	125	14	1505	(省略)
968457	2011-12-22	my cock	['Men']	my cock	868	5	3	173	(省略)
968450	2011-12-22	masturbation in black dress	['Men']	(省略)	845	4	4	362	(省略)
968451	2011-12-22	cuckold cleans it up 2	['Amateur', (省略)]	(省略)	353342	497	48	747	(省略)
778288	2011-07-29	(省略)	['Amateur', (省略)]	(省略)	310578	269	24	1059	(省略)
1246687	2012-05-30	school time orgy	['Big Boobs', (省略)]	(省略)	50202	74	8	185	(省略)
978809	2011-12-31	sexy gir	['Hardcore']	NA	1234	16	5	284	(省略)
978802	2012-01-03	denise young bbw threesomes	['BBW', (省略)]	bbw porno	63186	150	16	1090	(省略)
978806	2012-01-03	blonde with pigtails fucking 2	['Amateur', (省略)]	NA	3775	8	1	199	(省略)
978807	2011-12-31	up bra	['Amateur', (省略)]	NA	337309	182	8	96	(省略)
1698377	2013-01-21	strokin i	['Men']	BBC	87	1	1	23	(省略)
1698375	2013-01-21	amateur cellphone sex	['Amateur', 'Hardcore']	NA	6954	18	2	174	(省略)
89377	2008-12-06	fisting	中略 ['Amateur']	NA	122575	206	3	289	NA