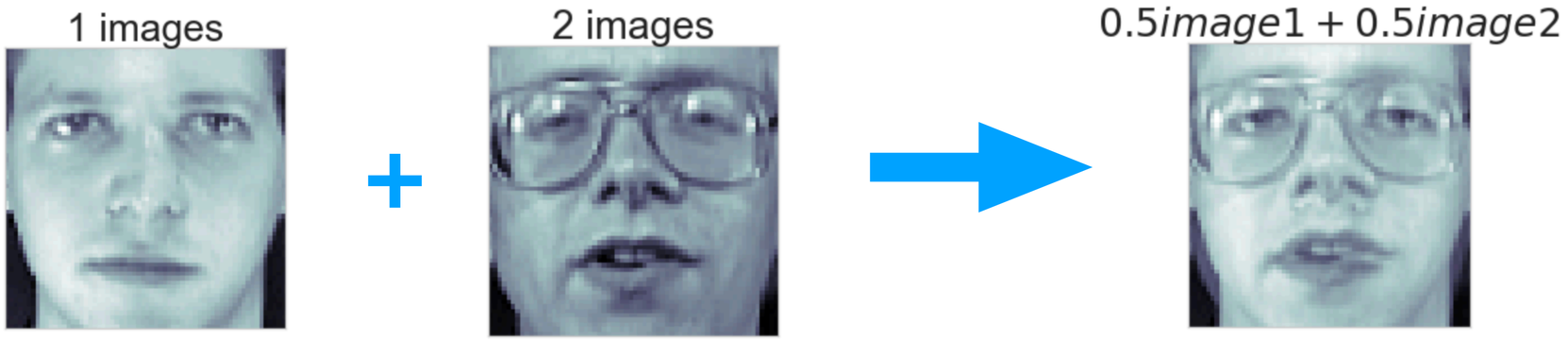
Practice - Mathematics and MachineLearning

Table of contents

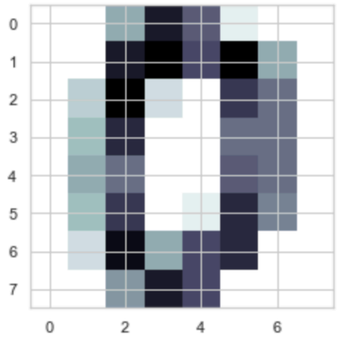
Practice 1. Linear Combination with Image data (Morphing)

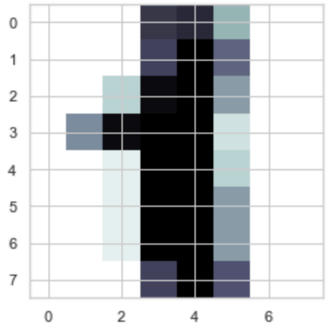
Image data의 가중평균(선형결합)으로 새로운 이미지로 변형(Morphing)



Practice 2. Cosine similarity with MNIST digit image

‘0’과 ‘8’의 코사인 유사도가 더 큼을 확인



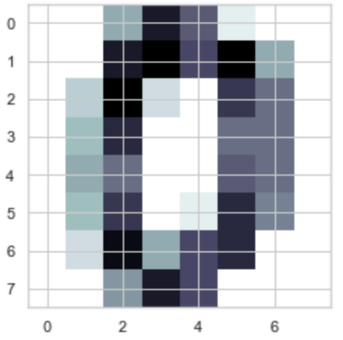


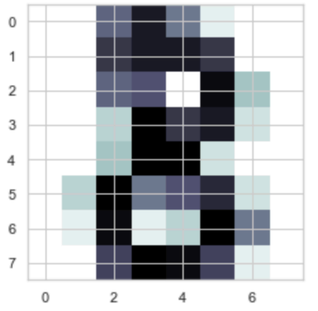
```
((_0.T@_1)/(np.linalg.norm(_0)*np.linalg.norm(_1)))[0][0]
```

```
0.5191023426414686
```

Cosine similarity

<





```
((_0.T@_8)/(np.linalg.norm(_0)*np.linalg.norm(_8)))[0][0]
```

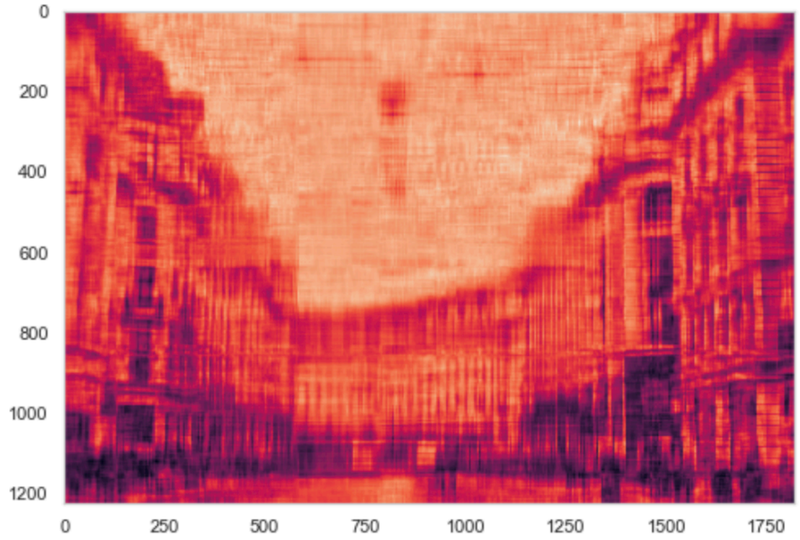
```
0.7515122122359871
```

Practice 3. Approximation of Image with Singular Value Decomposition

이미지 데이터를 Rank-1 Matrix로 분해(SVD)후 선형결합해 근사

$$A_{image} = U\Sigma V^T \approx \sum_{i=1}^{18} u_i \sigma_i v_i^T$$

```
plt.imshow(np.real(S[0]*U[:,:1]@VT[:1,:])
           +np.real(S[1]*U[:,1:2]@VT[1:2,:])
           +np.real(S[2]*U[:,2:3]@VT[2:3,:])
           +np.real(S[3]*U[:,3:4]@VT[3:4,:])
           +np.real(S[4]*U[:,4:5]@VT[4:5,:])
           +np.real(S[5]*U[:,5:6]@VT[5:6,:])
           +np.real(S[6]*U[:,6:7]@VT[6:7,:])
           +np.real(S[7]*U[:,7:8]@VT[7:8,:])
           +np.real(S[8]*U[:,8:9]@VT[8:9,:])
           +np.real(S[9]*U[:,9:10]@VT[9:10,:])
           +np.real(S[10]*U[:,10:11]@VT[10:11,:])
           +np.real(S[11]*U[:,11:12]@VT[11:12,:])
           +np.real(S[12]*U[:,12:13]@VT[12:13,:])
           +np.real(S[13]*U[:,13:14]@VT[13:14,:])
           +np.real(S[14]*U[:,14:15]@VT[14:15,:])
           +np.real(S[15]*U[:,15:16]@VT[15:16,:])
           +np.real(S[16]*U[:,16:17]@VT[16:17,:])
           +np.real(S[17]*U[:,17:18]@VT[17:18,:]))
plt.show()
```



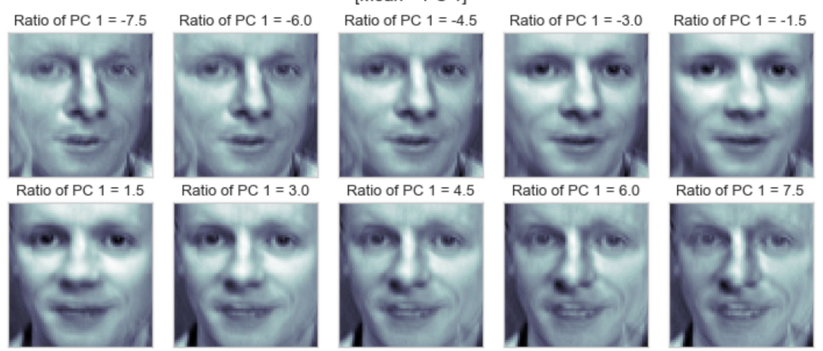
Practice 4. Principal Component Analysis (PCA)

주성분 분석을 통해 주성분의 역할 확인

Original image data

PCA

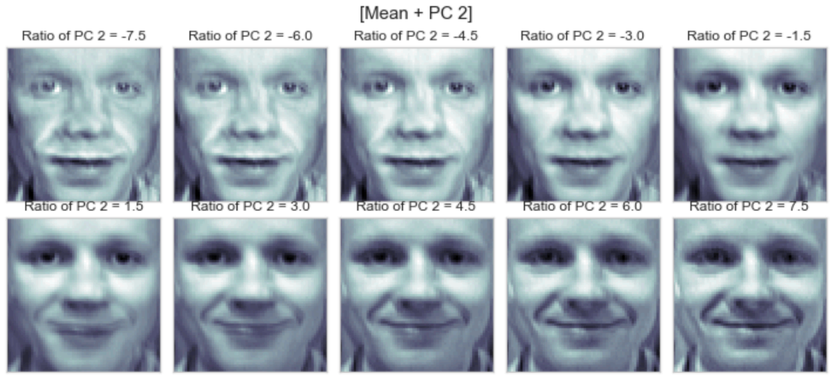
[평균값 + 주성분 1]



주성분 1의 역할 : 얼굴의 좌우 회전

PCA

[평균값 + 주성분 2]



주성분 2의 역할 : 얼굴의 표정 정보

Practice 5. Regression Analysis

[회귀분석 대상 데이터 : Boston House dataset (from scikit-learn)]

1. Target data : MEDV, 1978년 보스턴 주택 가격 (506개 타운의 주택 가격 중앙값 (단위 1,000 달러))

2. Feature data(Columns) :

CRIM: 범죄율
INDUS: 비소매상업지역 면적 비율
NOX: 일산화질소 농도
RM: 주택당 방 수
LSTAT: 인구 중 하위 계층 비율
B: 인구 중 흑인 비율
PTRATIO: 학생/교사 비율
ZN: 25,000 평방피트를 초과 거주지역 비율
CHAS: 찰스강의 경계에 위치한 경우는 1, 아니면 0
AGE: 1940년 이전에 건축된 주택의 비율
RAD: 방사형 고속도로까지의 거리
DIS: 직업센터의 거리
TAX: 재산세율

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67	22.4
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88	11.9

506 rows × 14 columns

3. shape : (506,13 + 1)

Practice 5. Regression Analysis

1. OLS regression analysis(Basic model)

$$MEDV = Const + w_1 CRIM + w_2 ZN + \ldots + w_{13} LSTAT$$

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Wed, 15 Jul 2020	Prob (F-statistic):	6.72e-135
Time:	18:18:10	Log-Likelihood:	-1498.8
No. Observations:	506	AIC:	3026.
Df Residuals:	492	BIC:	3085.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	36.4595	5.103	7.144	0.000	26.432	46.487
CRIM	-0.1080	0.033	-3.287	0.001	-0.173	-0.043
ZN	0.0464	0.014	3.382	0.001	0.019	0.073
INDUS	0.0206	0.061	0.334	0.738	-0.100	0.141
CHAS	2.6867	0.862	3.118	0.002	0.994	4.380
NOX	-17.7666	3.820	-4.651	0.000	-25.272	-10.262
RM	3.8099	0.418	9.116	0.000	2.989	4.631
AGE	0.0007	0.013	0.052	0.958	-0.025	0.027
DIS	-1.4756	0.199	-7.398	0.000	-1.867	-1.084
RAD	0.3060	0.066	4.613	0.000	0.176	0.436
TAX	-0.0123	0.004	-3.280	0.001	-0.020	-0.005
PTRATIO	-0.9527	0.131	-7.283	0.000	-1.210	-0.696
B	0.0093	0.003	3.467	0.001	0.004	0.015
LSTAT	-0.5248	0.051	-10.347	0.000	-0.624	-0.425

Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	Cond. No.	1.51e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.51e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

R-squared : 0.74
Condition number : 1.51e+04

[분석 결과]
1) 매우 높은 조건수를 확인
- 다중공선성 확인을 위해 추가적인 분석 필요
- 주요 변수들의 scaling으로 조건수 조정이 필요

Scaling

Lower Cond No.

2. OLS Regression Analysis(model with scaled features)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Wed, 15 Jul 2020	Prob (F-statistic):	6.72e-135
Time:	18:20:02	Log-Likelihood:	-1498.8
No. Observations:	506	AIC:	3026.
Df Residuals:	492	BIC:	3085.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	22.3470	0.219	101.943	0.000	21.916	22.778
scale(CRIM)	-0.9281	0.282	-3.287	0.001	-1.483	-0.373
scale(ZN)	1.0816	0.320	3.382	0.001	0.453	1.710
scale(INDUS)	0.1409	0.421	0.334	0.738	-0.687	0.969
scale(NOX)	-2.0567	0.442	-4.651	0.000	-2.926	-1.188
scale(RM)	2.6742	0.293	9.116	0.000	2.098	3.251
scale(AGE)	0.0195	0.371	0.052	0.958	-0.710	0.749
scale(DIS)	-3.1040	0.420	-7.398	0.000	-3.928	-2.280
scale(RAD)	2.6622	0.577	4.613	0.000	1.528	3.796
scale(TAX)	-2.0768	0.633	-3.280	0.001	-3.321	-0.833
scale(PTRATIO)	-2.0606	0.283	-7.283	0.000	-2.617	-1.505
scale(B)	0.8493	0.245	3.467	0.001	0.368	1.331
scale(LSTAT)	-3.7436	0.362	-10.347	0.000	-4.454	-3.033
CHAS	2.6867	0.862	3.118	0.002	0.994	4.380

Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	Cond. No.	10.6

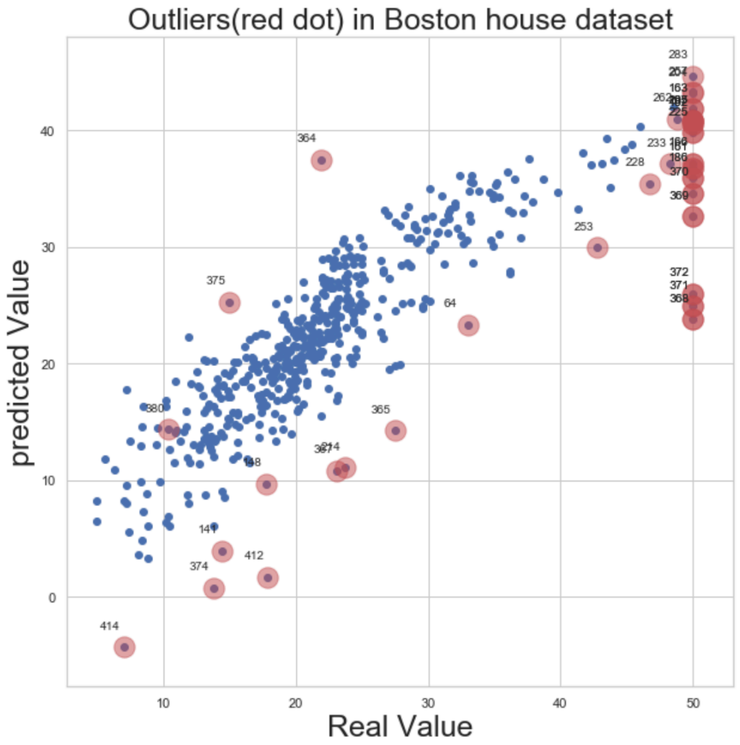
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

R-squared : 0.74
Condition number : 10.6

[분석 결과]
1) 스케일링으로 조건수의 조정을 확인
- R-sqaured 값은 유지

Practice 5. Regression Analysis

3. OLS Regression Analysis(model with outlier removed according to ‘Fox recommendation’)



Outliers Removed

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.812
Method:	Least Squares	F-statistic:	126.9
Date:	Wed, 15 Jul 2020	Prob (F-statistic):	9.10e-127
Time:	18:38:44	Log-Likelihood:	-957.69
No. Observations:	380	AIC:	1943.
Df Residuals:	366	BIC:	1999.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	21.4555	0.163	131.785	0.000	21.135	21.776
scale(CRIM)	-0.4351	0.236	-1.847	0.066	-0.898	0.028
scale(ZN)	0.9480	0.237	3.992	0.000	0.481	1.415
scale(INDUS)	-0.1444	0.294	-0.492	0.623	-0.722	0.433
scale(NOX)	-1.2558	0.328	-3.830	0.000	-1.901	-0.611
scale(RM)	2.6007	0.227	11.468	0.000	2.155	3.047
scale(AGE)	-0.7655	0.293	-2.616	0.009	-1.341	-0.190
scale(DIS)	-2.3313	0.314	-7.417	0.000	-2.949	-1.713
scale(RAD)	1.6630	0.393	4.228	0.000	0.889	2.436
scale(TAX)	-1.6769	0.406	-4.132	0.000	-2.475	-0.879
scale(PTRATIO)	-1.4405	0.196	-7.337	0.000	-1.827	-1.054
scale(B)	0.9273	0.188	4.939	0.000	0.558	1.297
scale(LSTAT)	-2.2301	0.299	-7.455	0.000	-2.818	-1.642
CHAS	1.1045	0.669	1.652	0.099	-0.210	2.419

Omnibus:	21.697	Durbin-Watson:	1.262
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28.733
Skew:	0.465	Prob(JB):	5.76e-07
Kurtosis:	3.975	Cond. No.	10.7

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The Number of Outliers : 55

R-squared : 0.818
Condition number : 10.7

[분석 결과]

1) Outlier의 제거를 통해 선형회귀 모델의 R-squared 상승(0.07 +) 확인
- 조건수는 유지

Practice 5. Regression Analysis

4. OLS Regression Analysis(model with Polynomial model multicollinearity controlled)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.872
Model:	OLS	Adj. R-squared:	0.868
Method:	Least Squares	F-statistic:	199.9
Date:	Wed, 15 Jul 2020	Prob (F-statistic):	1.56e-185
Time:	18:47:46	Log-Likelihood:	317.45
No. Observations:	456	AIC:	-602.9
Df Residuals:	440	BIC:	-536.9
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.0338	0.007	433.880	0.000	3.020	3.048
scale(CRIM)	-0.3471	0.044	-7.976	0.000	-0.433	-0.262
scale(I(CRIM ** 2))	0.3075	0.071	4.331	0.000	0.168	0.447
scale(ZN)	-0.0465	0.022	-2.110	0.035	-0.090	-0.003
scale(I(ZN ** 2))	0.0440	0.020	2.206	0.028	0.005	0.083
scale(INDUS)	0.0037	0.012	0.323	0.747	-0.019	0.026
scale(NOX)	-0.0652	0.013	-5.001	0.000	-0.091	-0.040
scale(RM)	0.0999	0.011	9.195	0.000	0.079	0.121
scale(AGE)	-0.0273	0.011	-2.438	0.015	-0.049	-0.005
scale(np.log(DIS))	-0.1008	0.014	-7.368	0.000	-0.128	-0.074
scale(RAD)	0.1634	0.020	8.106	0.000	0.124	0.203
scale(TAX)	-0.0934	0.018	-5.153	0.000	-0.129	-0.058
scale(np.log(PTRATIO))	-0.0699	0.008	-8.872	0.000	-0.085	-0.054
scale(B)	0.0492	0.007	6.699	0.000	0.035	0.064
scale(np.log(LSTAT))	-0.1487	0.013	-11.074	0.000	-0.175	-0.122
CHAS	0.0659	0.026	2.580	0.010	0.016	0.116

Omnibus:	28.653	Durbin-Watson:	1.309
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43.266
Skew:	0.465	Prob(JB):	4.03e-10
Kurtosis:	4.188	Cond. No.	35.2

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

R-squared : 0.872
Condition number : 35.2

[분석 결과]
1) 조건수가 다소 상승
- Polynomial 모델로서 다중공선성이 원인일 가능성이 높음

	VIF Factor	features
0	1.061624	CHAS
1	1.338325	scale(B)
2	1.478553	Intercept
3	1.780320	scale(np.log(PTRATIO))
4	2.596496	scale(RM)
5	3.748931	scale(AGE)
6	3.807459	scale(INDUS)
7	4.682812	scale(np.log(LSTAT))
8	5.071802	scale(NOX)
9	5.215025	scale(np.log(DIS))
10	9.107858	scale(TAX)
11	10.218588	scale(I(CRIM ** 2))
12	11.254736	scale(RAD)
13	11.751869	scale(I(ZN ** 2))
14	14.646056	scale(ZN)
15	21.260182	scale(CRIM)

각 feature별 다중공선성 영향 정도 확인
: 15개 중 6개 변수 선택해 모델 재구성

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.836
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	380.7
Date:	Sun, 19 Jul 2020	Prob (F-statistic):	1.42e-172
Time:	21:42:51	Log-Likelihood:	260.52
No. Observations:	456	AIC:	-507.0
Df Residuals:	449	BIC:	-478.2
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.0192	0.007	445.252	0.000	3.006	3.033
CHAS	0.0884	0.028	3.141	0.002	0.033	0.144
scale(B)	0.0558	0.008	6.989	0.000	0.040	0.072
scale(CRIM)	-0.1179	0.013	-9.120	0.000	-0.143	-0.092
scale(np.log(PTRATIO))	-0.0508	0.007	-6.936	0.000	-0.065	-0.036
scale(RM)	0.1153	0.011	10.828	0.000	0.094	0.136
scale(np.log(LSTAT))	-0.1570	0.011	-14.179	0.000	-0.179	-0.135

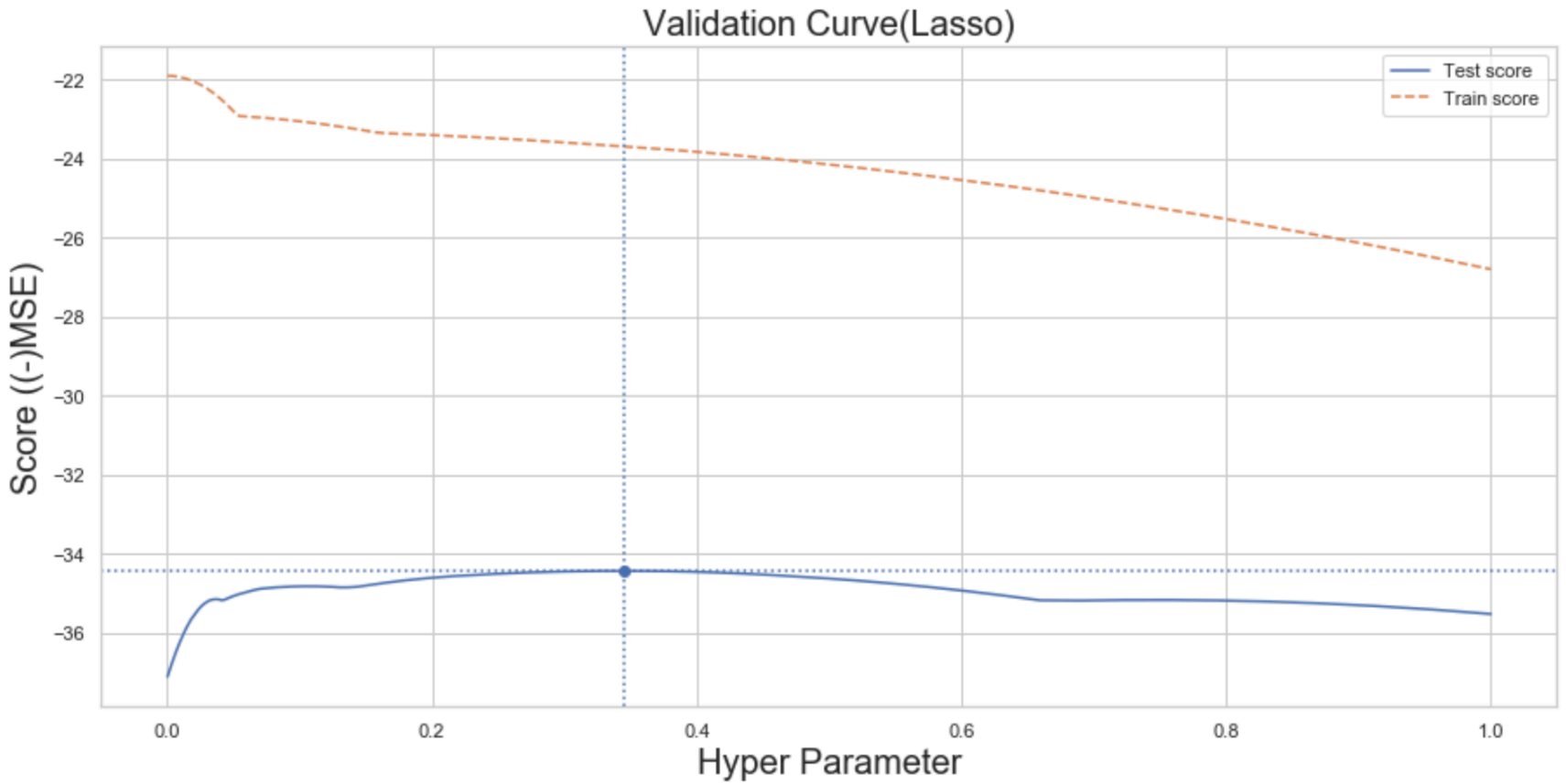
Omnibus:	29.141	Durbin-Watson:	1.113
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.637
Skew:	0.483	Prob(JB):	5.51e-10
Kurtosis:	4.145	Cond. No.	5.91

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

R-squared : 0.836
Condition number : 5.91

[분석 결과]
1) 조건수 하락 및 성능 유지
- 절반도 안되는 변수로 성능은 유지하며,
조건수는 낮춰 보다 안정적인 성능 도출 가능함을 확인

5. OLS Regression Analysis(Regularized model (Lasso))



Hyper Parameter 가 약 '0.345' 일 때, MSE 가 약 '-34.44' 로 최적의 Test 성능을 보임을 확인

Practice 6. Machine Learning models (Breast Cancer dataset)

1) Scikit-learn Dataset Practice

Breast Cancer dataset

:Number of Instances: 569

:Number of Attributes: 30 numeric, predictive attributes and the class

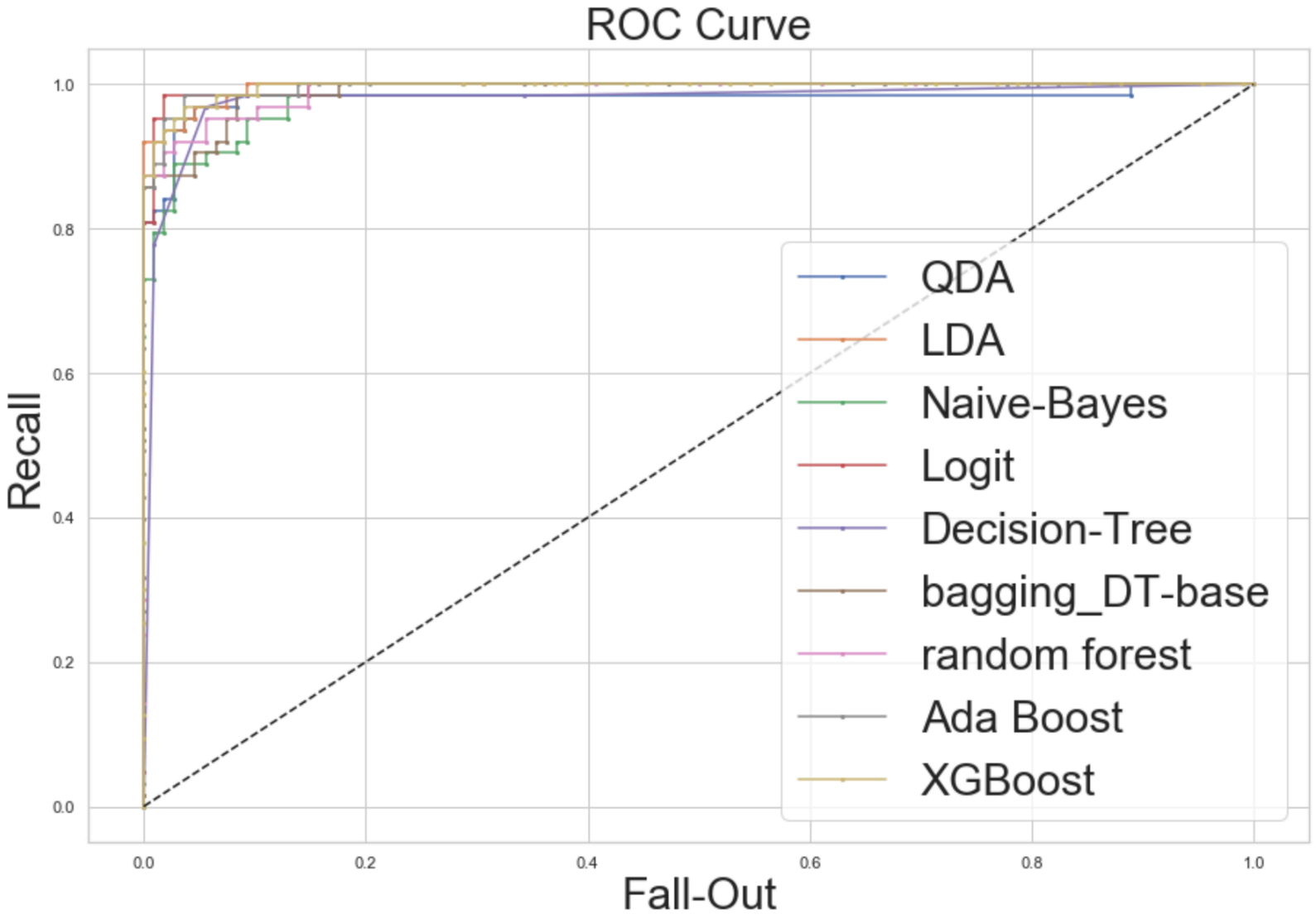
:Attribute Information:
- **radius (mean of distances from center to points on the perimeter)**
- **texture (standard deviation of gray-scale values)**
- **perimeter**
- **area**
- **smoothness (local variation in radius lengths)**
- **compactness (perimeter^2 / area - 1.0)**
- **concavity (severity of concave portions of the contour)**
- **concave points (number of concave portions of the contour)**
- **symmetry**
- **fractal dimension ("coastline approximation" - 1)**

- **class:**
  - **WDBC-Malignant (음성)**
  - **WDBC-Benign (양성)**

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	con
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.16220	
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.12380	
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.14440	
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.20980	
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.13740	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	2027.0	0.14100	
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0	0.11660	
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0	0.11390	
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0	0.16500	
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6	0.08996	

569 rows × 31 columns

Practice 6. Machine Learning models (Breast Cancer dataset)



AUC (Area Under the Curve)

Decision Tree : 0.9957
Logistic Regression : 0.9957
LDA : 0.9957
XGBoost : 0.9956
AdaBoost : 0.9951
Random Forest : 0.9909

Practice 7. Machine Learning models (Titanic Dataset)

2) Titanic Dataset

:Number of Instances : 1,300

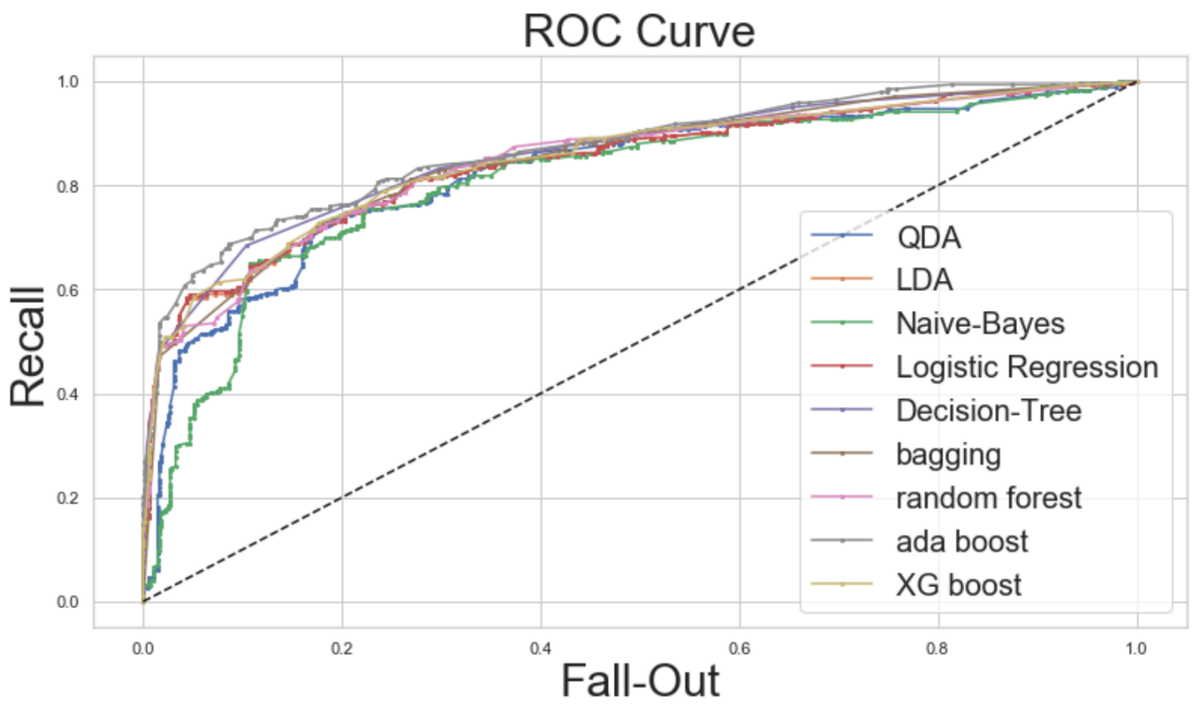
:Number of Attributes : 11 numeric, categorical attributes and the class

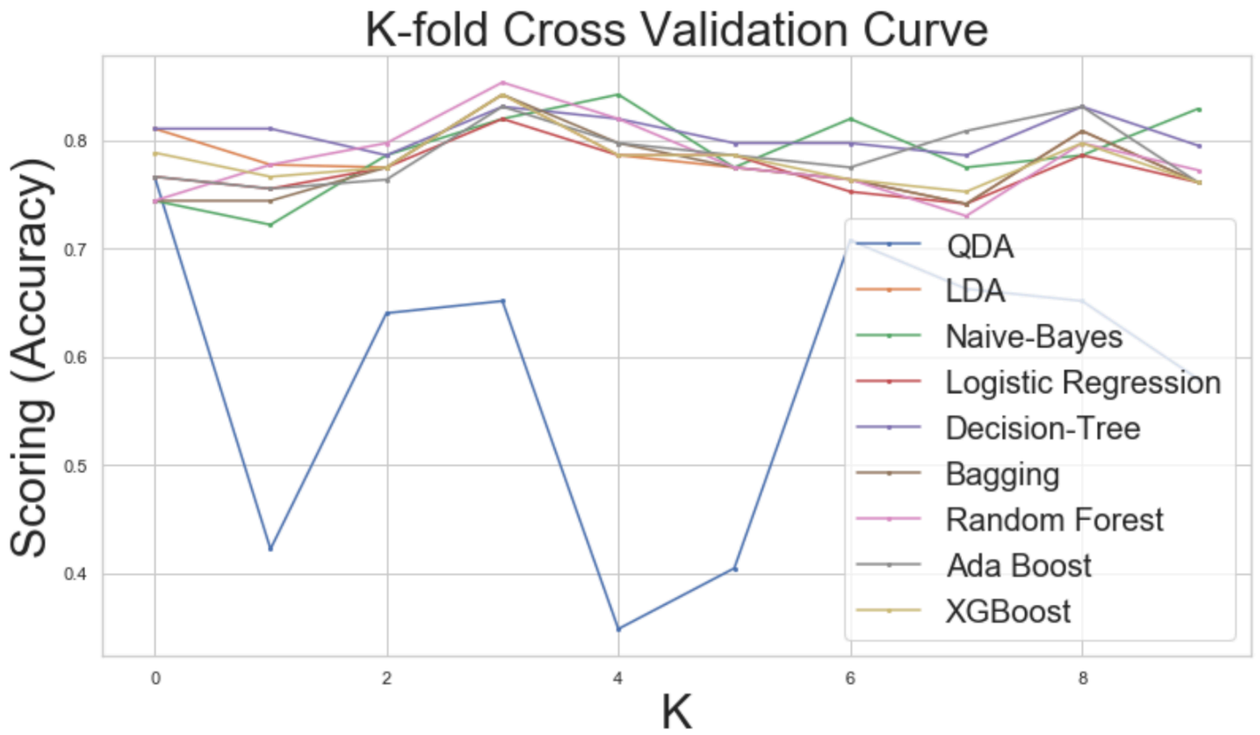
- **Survival** - 생존 여부. 0이면 사망, 1이면 생존한 것으로 간주
- **Pclass** - 티켓 등급. 1등석(1), 2등석(2), 3등석(3)이 있으며, 1등석일수록 좋고 3등석일수록 좋지 않음
- **Sex** - 성별. 남자(male)와 여자(female)이 있음
- **Age** - 나이. 틈틈히 빈 값이 존재하며, 소수점 값도 존재함
- **SibSp** - 해당 승객과 같이 탑승한 형재/자매(siblings)와 배우자(spouses)의 총 인원 수
- **Parch** - 해당 승객과 같이 탑승한 부모(parents)와 자식(children)의 총 인원 수
- **Ticket** - 티켓 번호. 다양한 텍스트(문자열)로 구성되어 있음
- **Fare** - 운임 요금. 소수점으로 구성되어 있음
- **Cabin** - 객실 번호. 많은 빈 값이 존재하며, 다양한 텍스트(문자열)로 구성되어 있음
- **Embarked** - 선착장. C는 셰르부르(Cherbourg)라는 프랑스 지역, Q는 퀸스타운(Queenstown)이라는 영국 지역, S는 사우스햄튼(Southampton)이라는 영국 지역

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...
887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 11 columns

Practice 7. Machine Learning models (Titanic Dataset)





AUC (Area Under the Curve)

AdaBoost : 0.8702
Decision Tree : 0.8591
XGBoost : 0.8495
Bagging : 0.8473
Random Forest : 0.8461

Accuracy Score
(Mean value of 10 fold Cross Validation)

Decision Tree : 0.8069
Naive-Bayes : 0.7903
Ada Boost : 0.7879
Random Forest : 0.7834
XGBoost : 0.7823

Practice 8. Machine Learning models (Naver sentiment movie corpus v1.0)

3) Naver sentiment movie corpus v1.0 Dataset

- **All 200k reviews**
- **150 (train), 50(test) reviews data**
- **100 (positive review, rating 9-10), 100(negative review, rating 1-4) data**

- 영화 리뷰데이터에 대한 Sentiment analysis(긍정 / 부정 판단)

```
3 pprint(data[:10])
```

```
[['9976970', '아 더빙.. 진짜 짜증나네요 목소리', '0'],
 ['3819312', '흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나', '1'],
 ['10265843', '너무재밓었다그래서보는것을추천한다', '0'],
 ['9045019', '교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정', '0'],
 ['6483659',
  '사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다',
  '1'],
 ['5403919', '막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.', '0'],
 ['7797314', '원작의 긴장감을 제대로 살려내지못했다.', '0'],
 ['9443947',
  '별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫겟다 납치.감금만반복반복..이드라마는 가족도없다 '
  '연기못하는사람만모엿네',
  '0'],
 ['7156791', '액션이 없는데도 재미 있는 몇안되는 영화', '1'],
 ['5912145', '왜케 평점이 낮은건데? 꽤 볼만한데.. 헐리우드식 화려함에만 너무 길들여져 있나?', '1']]
```

Practice 8. Machine Learning models (Naver sentiment movie corpus v1.0)

Classification Report (부정 : '0', 긍정 : '1')

1. 전처리 : CountVectorizer, 모델 : NB-multinomial

	precision	recall	f1-score	support
0	0.81	0.84	0.83	24827
1	0.84	0.81	0.82	25173
accuracy			0.83	50000
macro avg	0.83	0.83	0.83	50000
weighted avg	0.83	0.83	0.83	50000

2. 전처리 : CountVectorizer, 모델 : NB-multinomial
KoNlpy 형태소 분석기 활용

	precision	recall	f1-score	support
0	0.85	0.86	0.85	24827
1	0.86	0.85	0.85	25173
accuracy			0.85	50000
macro avg	0.85	0.85	0.85	50000
weighted avg	0.85	0.85	0.85	50000

3. 전처리 : CountVectorizer, 모델 : NB-multinomial
KoNlpy 형태소 분석기 활용
gram 범위 : 1-2 gram

	precision	recall	f1-score	support
0	0.86	0.87	0.87	24827
1	0.87	0.86	0.87	25173
accuracy			0.87	50000
macro avg	0.87	0.87	0.87	50000
weighted avg	0.87	0.87	0.87	50000