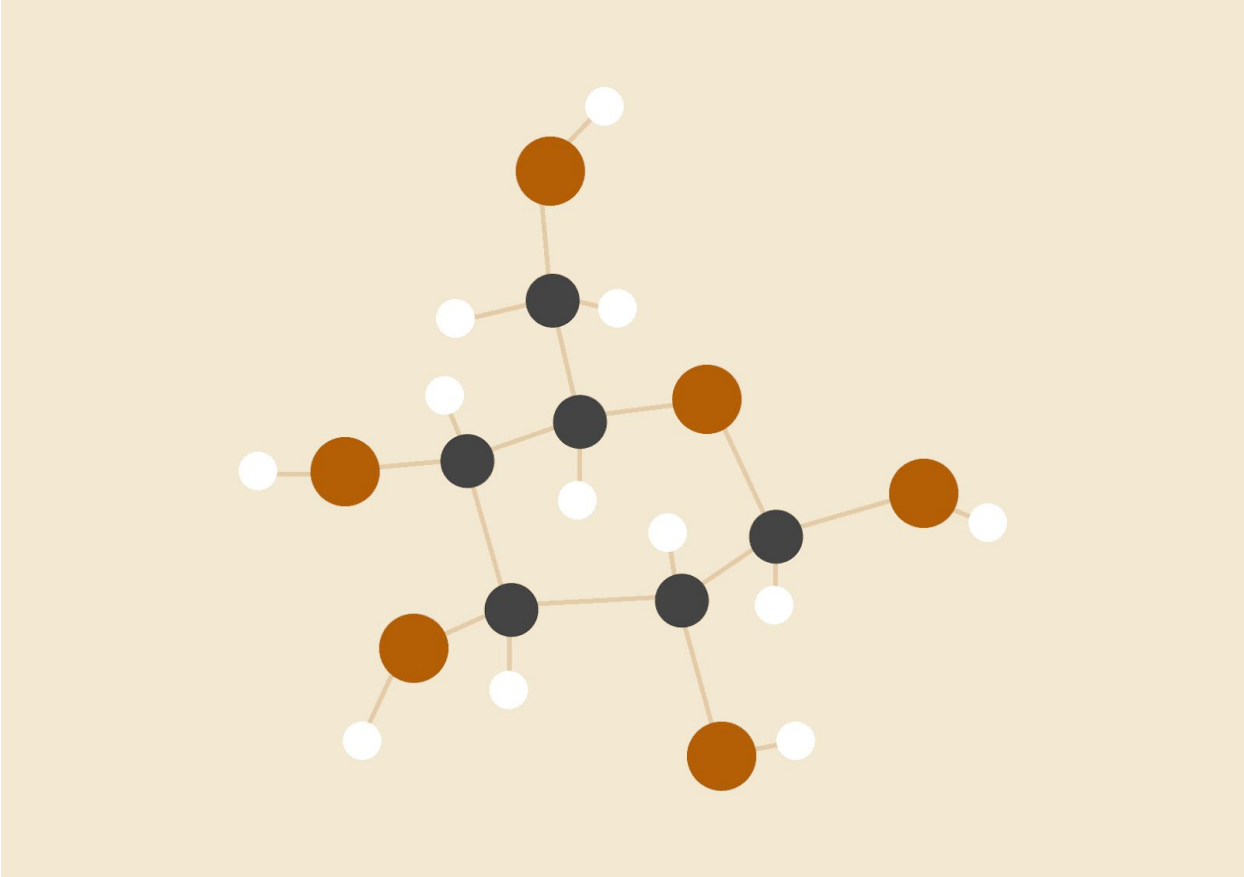
LAB 03 REPORT

MTH00051 - Applied Mathematics and Statistics



Nguyen Cong Phuc - 1753085

31.08.2020

Ho Chi Minh University Of Science

LIBRARIES

1. Data Manipulation
   Pandas, numpy.
2. Building Model
   Statsmodels.
3. Features Manipulation, Features Preprocessing, Scoring
   Scikit-learn.

HELPERS

1. SMWrapper(BaseEstimator, RegressorMixin)
   Wrapper for statsmodels regressors to use in scikit-learn functions.

RESULTS

a) Full Attributes

	coefficient
fixed acidity	0.005925
volatile acidity	-1.108038
citric acid	-0.263046
residual sugar	0.015322
chlorides	-1.730503
free sulfur dioxide	0.003801
total sulfur dioxide	-0.003899
density	4.338588
pH	-0.458535
sulphates	0.729719
alcohol	0.308859

Building model: model_full = sm.OLS(y,X).fit()
Compute average r2 score of 5 folds: mf_score = np.mean(cross_val_score(SMWrapper(sm.OLS), X, y, scoring='r2', cv=5))

b) Sorted By Mean Of All Fold R2 Score
Alcohol is the best single attribute score.

	Attributes	r2 Score
1	alcohol	0.163181
2	volatile acidity	0.028825
3	total sulfur dioxide	-0.073043
4	citric acid	-0.084692
5	chlorides	-0.129122
6	fixed acidity	-0.132844
7	pH	-0.139813
8	free sulfur dioxide	-0.143103
9	residual sugar	-0.147915
10	density	-0.162868
11	sulphates	-0.176188

Loop through each attributes and compute the score:
for col in X.columns:

```
att_score.append(np.mean(cross_val_score(SMWrapper(sm.OLS), X[col], y, scoring='r2', cv=5)))
```

Make it a dataframe and sorted by the score:
df_score = pd.DataFrame(zip(X.columns, att_score),columns=['Attributes','r2 Score']).sort_values(by='r2 Score', ascending= False).reset_index(drop=True)

c)

Sorted by Fold1

	r2 Score fold1	r2 Score fold2	r2 Score fold3	r2 Score fold4	Attributes
1	0.063166	0.197865	0.133434	0.226652	alcohol
2	-0.011582	-0.000635	0.048694	-0.014407	volatile acidity
3	-0.093504	0.001409	-0.143423	-0.223374	total sulfur dioxide
4	-0.144679	0.019359	-0.074247	-0.239200	citric acid
5	-0.155578	-0.007559	-0.164458	-0.358526	fixed acidity
6	-0.213676	-0.050337	-0.148017	-0.276457	chlorides
7	-0.215880	-0.045690	-0.139860	-0.310643	pH
8	-0.217667	-0.047050	-0.159717	-0.315076	free sulfur dioxide
9	-0.219697	-0.053777	-0.156278	-0.310569	residual sugar
10	-0.328743	-0.314689	-0.059804	-0.226508	density
11	-0.520102	0.034172	-0.054130	-0.281518	sulphates

Sorted by Fold2

	r2 Score fold1	r2 Score fold2	r2 Score fold3	r2 Score fold4	Attributes
1	0.063166	0.197865	0.133434	0.226652	alcohol
2	-0.520102	0.034172	-0.054130	-0.281518	sulphates
3	-0.144679	0.019359	-0.074247	-0.239200	citric acid
4	-0.093504	0.001409	-0.143423	-0.223374	total sulfur dioxide
5	-0.011582	-0.000635	0.048694	-0.014407	volatile acidity
6	-0.155578	-0.007559	-0.164458	-0.358526	fixed acidity
7	-0.215880	-0.045690	-0.139860	-0.310643	pH
8	-0.217667	-0.047050	-0.159717	-0.315076	free sulfur dioxide
9	-0.213676	-0.050337	-0.148017	-0.276457	chlorides
10	-0.219697	-0.053777	-0.156278	-0.310569	residual sugar
11	-0.328743	-0.314689	-0.059804	-0.226508	density

Sorted by Fold3

	r2 Score fold1	r2 Score fold2	r2 Score fold3	r2 Score fold4	Attributes
1	0.063166	0.197865	0.133434	0.226652	alcohol
2	-0.011582	-0.000635	0.048694	-0.014407	volatile acidity
3	-0.520102	0.034172	-0.054130	-0.281518	sulphates
4	-0.328743	-0.314689	-0.059804	-0.226508	density
5	-0.144679	0.019359	-0.074247	-0.239200	citric acid
6	-0.215880	-0.045690	-0.139860	-0.310643	pH
7	-0.093504	0.001409	-0.143423	-0.223374	total sulfur dioxide
8	-0.213676	-0.050337	-0.148017	-0.276457	chlorides
9	-0.219697	-0.053777	-0.156278	-0.310569	residual sugar
10	-0.217667	-0.047050	-0.159717	-0.315076	free sulfur dioxide
11	-0.155578	-0.007559	-0.164458	-0.358526	fixed acidity

Sorted by Fold4

	r2 Score fold1	r2 Score fold2	r2 Score fold3	r2 Score fold4	Attributes
1	0.063166	0.197865	0.133434	0.226652	alcohol
2	-0.011582	-0.000635	0.048694	-0.014407	volatile acidity
3	-0.093504	0.001409	-0.143423	-0.223374	total sulfur dioxide
4	-0.328743	-0.314689	-0.059804	-0.226508	density
5	-0.144679	0.019359	-0.074247	-0.239200	citric acid
6	-0.213676	-0.050337	-0.148017	-0.276457	chlorides
7	-0.520102	0.034172	-0.054130	-0.281518	sulphates
8	-0.219697	-0.053777	-0.156278	-0.310569	residual sugar
9	-0.215880	-0.045690	-0.139860	-0.310643	pH
10	-0.217667	-0.047050	-0.159717	-0.315076	free sulfur dioxide
11	-0.155578	-0.007559	-0.164458	-0.358526	fixed acidity

d) Improving Model

Full Attributes CV Score: 0.2604353464190211

Improved Final CV Score: 0.325033020959744

Steps:

1) Make the polynomial with max degree = 2

```
poly = PolynomialFeatures(degree= 2)
poly_x = poly.fit_transform(X)

Rename x0..x10 to the original attributes name
cols = poly.get_feature_names()
cols[0] = 'bias'
for i in range(len(cols)):
        cols[i] = cols[i].replace(' ',' . ').replace('x10','alcohol')
        for j in range(len(X.columns)):
                cols[i] = cols[i].replace(f'x{j}',X.columns[j])
```

2) Set the p value threshold (in this case, 0.05)

```
threshold = 0.05
```

3) Recurrence compute and remove attributes which have p value equal or greater than the threshold p value.

```
Finding current max p value and its index
max_pval = np.max(df_pval['Pvalues'])
max_pval_index = np.argmax(df_pval['Pvalues'])

Add to drop list
drop_att = df_pval.iloc[max_pval_index]['Atrributes']
drop_list.append(drop_att)

Building model without dropped attributes
X2_new = X2.drop(drop_list, axis=1)
model = sm.OLS(y,X2_new).fit()
model_score = np.mean(cross_val_score(SMWrapper(sm.OLS), X2_new, y,
scoring='r2', cv=5))
```

4) Final model with the optimize R2 score

	coefficient
fixed acidity	-28.546312
alcohol	26.380443
fixed acidity^2	-0.034238
fixed acidity . volatile acidity	-0.131926
fixed acidity . chlorides	-1.287134
fixed acidity . density	29.421214
volatile acidity . total sulfur dioxide	0.010700
volatile acidity . pH	-0.988443
volatile acidity . alcohol	0.276158
citric acid . density	6.683647
citric acid . pH	-3.493100
citric acid . alcohol	0.440354
chlorides . density	9.553580
free sulfur dioxide . total sulfur dioxide	-0.000175
free sulfur dioxide . density	0.036245
free sulfur dioxide . sulphates	-0.027197
total sulfur dioxide^2	0.000043
total sulfur dioxide . density	-0.012626
density^2	-18.477886
density . pH	9.831975
density . alcohol	-25.986988
pH^2	-1.328206
sulphates^2	-0.939161
sulphates . alcohol	0.287921
alcohol^2	-0.031657

REFERENCES

1. https://numpy.org/
2. https://pandas.pydata.org/pandas-docs/stable/index.html
3. https://www.statsmodels.org/stable/index.html
4. https://scikit-learn.org/stable/index.html