Falak_Jain_HW3

April 7, 2022

HW3 - Falak Jain

[1]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold, cross_val_score
```

1.

(i) Yes

(ii) Yes

2. Are the following statements correct?

(1) No

(2) No

3. Subset selection methods:

(i) 2 ^ 19 = 524,288

(ii) [1 + p(p+1)/2] - [(p-k*)(p-k*+1)/2]

4. Regression using K-fold CV

[2]:
```
df_auto = pd.read_csv("auto.csv")
df_auto.head()
```

[2]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70		
4	17.0	8	302.0	140	3449	10.5	70		

```
3       1              amc rebel sst
4       1                ford torino
```

[3]:
```
# Model 1 with displacement and horsepower as predictor variables
# Model 2 with acceleration and weight as predictor variables
kfolds_regression = KFold(n_splits=10, random_state=2, shuffle=True)
regression_model = LinearRegression()
r2_model_1_cv =␣
 ↪cross_val_score(regression_model,df_auto[['displacement','horsepower']],␣
 ↪df_auto['mpg'], cv=kfolds_regression,scoring = 'r2')
r2_model_2_cv =␣
 ↪cross_val_score(regression_model,df_auto[['acceleration','weight']],␣
 ↪df_auto['mpg'], cv=kfolds_regression,scoring = 'r2')
print("Linear Regression: \n")
print("r squared of 10-folds with displacement and horsepower as input:
 ↪",r2_model_1_cv,"(mean r squared:",np.mean(r2_model_1_cv),")\n")
print("r squared of 10-folds with acceleration and weight as input:
 ↪",r2_model_2_cv,"(mean r squared:",np.mean(r2_model_2_cv),")")
```

```
Linear Regression:

r squared of 10-folds with displacement and horsepower as input: [0.63862276
0.67734208 0.6728976  0.72577554 0.64892779 0.67435645
 0.52565632 0.60286615 0.72099218 0.67427306] (mean r squared:
0.6561709931353421 )

r squared of 10-folds with acceleration and weight as input: [0.66708203
0.74825549 0.7557161  0.73255447 0.70727716 0.70700609
 0.56823354 0.6645276  0.70780473 0.67915667] (mean r squared:
0.6937613870721294 )
```

Therefore, we can see that the linear regression model with acceleration and weight has a higher r-squared

5. Cp and AIC criteria for linear regression

- Cp = 1/n * (RSS + 2$d$sigma_hat^2)
- AIC = 1/(n*sigma_hat^2) * (RSS + 2*d*sigma_hat^2)

Why do they give the same ranking for models:

- As we can see, Cp and AIC are directly proportional to each other. It is calculated by fit of large class of models of maximum likelihood. Therefore, similarly to Cp, lowest AIC provides the best model and vice versa. Hence, they give the same ranking for models
- Both criteria are proportional with the AIC term having an additional constant in the denominator, which remains the same for each model

[ ]: