HCM-UNIVERSITY OF SCIENCE

Applied Mathematics and Statistics

Project 03: Linear Regression

20127056 – Võ Duy Nhân

1/8/2022

Lab teacher: Phan Thị Phương Uyên – Nguyễn Văn Huy

Index

Index 1

1. Libraries and reasons to use them 2
2. Functions description 2
   a. Method fit(self, X, y): 2
   b. Method get_params(self): 3
   c. Method predict (self, X): 3
   d. rmse (y, y_hat) 3
3. Models predicting life expectancy using OLS Linear Regression 4
   a. Models using 10 attributes to predict life expectancy. 4
      - Idea explanation 4
      - Result 4
   b. Models using best attributes to predict life expectancy. 5
      - *k_fold Cross Validation (k=5)* [*] 6
      - Result 6
      - Retrain best_feature_model with best attribute: Schooling 7
      - Comment 7
   c. Best models to predict life expectancy. 7
      - Model 1 7
      - Model 2 8
      - Model 3 8
      - Result after train 3 model: 9
      - Retrain with the best model: y = w1 * HIV/AIDS + w2 * Income composition of resources + w3 * Schooling 9
      - Comment: 9
4. Work cited. 10
5. Reference. 10

1. Libraries and reasons to use them

- pandas: Used to read data from file *train.csv* and *test.csv* provided.
- numpy: Used to convert data frames to numpy arrays and manipulate data arrays with numpy functions provided.
- math: Used to calculate the square root of a number.

2. Functions description

```
class OLSLinearRegression:

    def fit(self, X, y):

        X_pinv = np.linalg.inv(X.T @ X) @ X.T    #
np.linalg.pinv(X)

        self.w = X_pinv @ y

        return self

    def get_params(self):

        return self.w

    def predict(self, X):

        return np.sum(self.w.ravel() * X, axis=1)
```

a. Method fit(self, X, y):

i. Input: Array X and y. X, y has 2 dimension, X.shape[0] = y.shape[0] (array X, y has the same number of rows), y.shape[1] = 1

ii. Output: The instance of the class with instance variable w, an array.

iii. Applying OLS Linear Regression, $w = (X^T X)^{-1} X^T Y$

iv. Step 1: X_pinv is set to $(X^T X)^{-1} X^T$, with X.T : transpose of matrix, np.linalg.inv(): calculate the inverse of a matrix, Numpy operator @: matrix multiplication.

v. Step 2: Instance variable w is set to X_pinv @ y. w.shape = (X.shape[1] , 1)

vi. Step 3: Return self, instance of the class.

b. Method get_params(self):

i. Input: Nothing
ii. Output: The instance variable w
iii. This method simply return the instance variable w

c. Method predict (self, X):

i. Input: array X, X has 2 dimensions and X.shape[1] = w.shape[0], X has the same dimension value as X in *fit* method.
ii. Output: An array has 2 dimension like array y in *fit* method
iii. Step 1: Turn self.w into a 1-D array by numpy ravel() method.
iv. Step 2: Multiply self.w.ravel() with X. Here * operator means multiplying each element of self.w.ravel() with each element in every row of X, then function numpy.sum will add all of that into 1 number. So in return we have a new array with shape = (X.shape[0],1)

In summary, this class uses 2 arrays X,y to train the model using OLS Linear Regression and get the parameters w, then use them to predict another X data.

d. rmse (y, y_hat)

```
def rmse(y, y_hat):
    se = np.mean((y.ravel() - y_hat.ravel())**2)
    return math.sqrt(np.mean(se))
```

- Input: 2 numpy array y, y_hat. (y, y_hat has the same shape = (n, 1), n >0)
- Output: Distance between 2 vector y.ravel(), y_hat.ravel()
- Step 1: Turn 2 array y, y_hat into 1_D array by function *.ravel()*
- Step 2: Subtract 2 array above. (*)
- Step 3: Create new array includes square each element of result (*) by ** 2 (exponent 2) (**)
- Step 4: Calculate mean of n element in array (**) by *numpy.mean* (***)
- Step 5: Return square root of (***) by *math.sqrt()*

3. Models predicting life expectancy using OLS Linear Regression

a. Models using 10 attributes to predict life expectancy.

```
(X, y) = (np.array(X_train),np.array(y_train).reshape(-1,1))

lr = OLSLinearRegression().fit(X,y)

y_test_predict = lr.predict(np.array(X_test))

print(lr.get_params())
```

- **Idea explanation**
  - Teacher has provided *X_train*, *y_train* from file *train.csv*. (*X_train.shape* = (1085,10) , *y_train.shape* = (1085,))
  - Step 1: Convert data frame *X_train*, *y_train* to array *X*, *y* using function *numpy.array*. At the same time, reshape *y* to (-1,1) to match the shape format in method *fit()* of class *OLSLinearRegression*
  - Step 2: Create an *OLSLinearRegression* ***lr*** by calling *OLSLinearRegression().fit (X, y).* Now we may have parameters *w. (w.shape = (10, 1))*
  - Step 3: Predict *y_test_predict* by calling method *.predict()* or ***lr*** with *np.array(X_test)* passed in

The process above train X, y to get parameters w and then predict y_test_predict from X_test

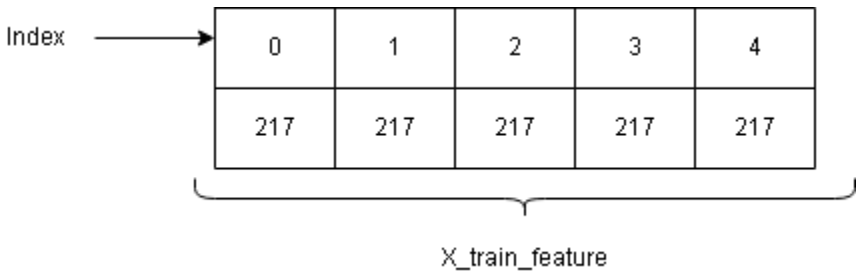
- **Result**
  - Parameters w: [0.0151013627, 0.0902199807, 0.0429218175, 0.139289117, -0.567332827, -0.000100765115, 0.740713438, 0.190935798, 24.5059736, 2.39351661]
  - RMSE = 7.064046430584466
  - Life expectancy = 0.0151013627 *Adult Morality + 0.0902199807 * BMI + 0.0429218175 * Polio + 0.139289117 * Diphtheria - 0.567332827 * HIV/AIDS - 0.000100765115 * GDP + 0.740713438 * Thinness age 10 - 19 + 0.190935798 * Thinness age 5 - 9 + 24.5059736 * Income composition of resources + 2.39351661 * Schooling

b. Models using best attributes to predict life expectancy.

```
# Phần code cho yêu cầu 1b
# Tìm ra đặc trưng tốt nhất
# In ra các kết quả cross-validation như yêu cầu

data = np.hstack((np.array(X_train).copy(),np.array(y_train).copy().reshape(-1,1) ))
np.random.shuffle(data)

X_train_clone = data[:, : -1]
y_train_clone = data[:, -1]

    
RMSE_list = np.ones(10)
for i in range(10):
    X_train_feature = X_train_clone[:,i]
    y_train_feature = y_train_clone

    RMSE = 0
    for j in range(5):
        X_val = X_train_feature[j * 217 : (j+1) * 217]
        y_val = y_train_feature[j * 217 : (j+1) * 217]

        Xtrain_kfold = np.concatenate ((X_train_feature[0 : j * 217],  X_train_feature[(j+1) * 217 : 217 * 5]))
        ytrain_kfold = np.concatenate ((y_train_feature[0 : j * 217], y_train_feature[(j+1) * 217 : 217 * 5]))


        lr = OLSLinearRegression().fit(Xtrain_kfold.reshape(-1,1),ytrain_kfold.reshape(-1,1))

        y_val_predict = lr.predict(X_val.reshape(-1,1))

        RMSE += rmse (y_val, y_val_predict)

    RMSE_list[i] = RMSE / 5

print (RMSE_list)

```

- Line 5 → 6: We create a copy of *X_train, y_train* because any changes would not affect the original array. We create a new array *data* by stacking 2 array copies of *X_train, y_train* horizontally by *numpy.hstack.* Shuffle *data* by function *numpy.random.shuffle()*
- Line 8→9: Create 2 arrays *X_train_clone, y_train_clone* by getting from *data* with slicing index. (*data[: , : -1]* : get from the 1st to 9nd column , *data[: , -1]* : get the last column.)
- Line 12 : Prepare a RMSE_list.shape = (10,) to store 10 rmse values of 10 models trained with every single attribute out of 10.
- Line 13 → 32: I performed *k_fold Cross Validation (k=5)* on each attribute out of 10 and store rmse values of each model to list. For each time (0 <= i <= 9), I prepare *X_train_feature = X_train_clone[: , i]* (get column i th in matrix X_train_clone), *y_train_featrue = y_train_clone*

- ***k_fold Cross Validation (k=5)*** **[*]**



  - We now have *X_train_feature* (1085 rows, 1 column) and y_train_feature (1085 rows, 1 column)
  - Step 1: We divide *X_train_feature* set into 5 parts (1 part has 217 rows, 1 column) and *y_train_feature* set into 5 parts (1 part has 217 rows, 1 column). Each part of *y_train_feature* must match each part of *X_train_feature* as the original data.
  - Step 2: Every time, we take 1 part of *X_train_feature* and 1 part of *y_train_feature* corresponding as *X_val*, *y_val* and the rest part would be *Xtrain_kfold*, *ytrain_kfold* $(\text{line } 18 \rightarrow 23)$
  - Step 3: We train with *Xtrain_kfold*, *ytrain_kfold*, predict with *X_val* and measure rmse with *y_val*, *y_val_predict* $(\text{line } 26 \rightarrow 30)$
  - Step 4: After 5 times performing training and predicting with 5 part of *X_train_feature*, *y_train_feature*, we sum up rmse value then divide by 5 (line 32)
- **Result**

  After 10 times performing *k_fold Cross Validation* with 10 model of 10 attribute

No	Model with 1 attribute	RMSE
1	Adult Mortality	46.22512429
2	BMI	27.90470836
3	Polio	17.99305181

4	Diphtheria	15.98564663
5	HIV/AIDS	67.1001855
6	GDP	60.18481154
7	Thinness age 10 - 19	51.81247783
8	Thinness age 5 - 9	51.73478913
9	Income composition of resources	13.18601473
10	Schooling	11.77609994

- **Retrain best_feature_model with best attribute: Schooling**
  - w = 5.5573994
  - RMSE = 10.260950391655376
  - Life expectancy = 5.5573994 * Schooling
- **Comment**

  Schooling can be interpreted as literacy. Literacy has a link with life expectancy through a range of socioeconomic factors. People with poor literacy are likely to be unemployed and have low income and so have low health behaviors, which lead to shorter life expectancy. Literacy is linked to life expectancy through health. That explains well attribute schooling is the best attribute out of 10. (Gilbert, Lisa, et al. "Life and Expectancy, An evidence review exploring the link between literacy and life expectancy in England through health and socioeconomic factors." *National Literacy Trust*, February 2018, p. 3.)

c. Best models to predict life expectancy.

- **Model 1**

$$y = w1 * Polio + w2 * Diphtheria + w3 * HIV/AIDS$$

As we can see, life expectancy relates pretty much to health. One person can live longer due to better health conditions, which is affected by health care facilities. Out of 10 attributes, *Polio, Diphtheria, HIV/AIDS* are those factors I believe represent the health care condition one person can get. Therefore, model 1 determines that *y (life expectancy)* has a linear link with *Polio, Diphtheria* and *HIV/AIDS*. As quote said:

"…The age we live depends primarily on where and how we live. People die earlier in countries that are badly affected by hunger and armed conflict and that have only poor health care facilities.Here, infectious diseases can rapidly become fatal, and women also die much more frequently during childbirth…" (Reiff, Susanne. "Life expectancy is increasing globally." *Alumniportal Deutschland*, 2017,)

- **Model 2**

$$y = w1 * HIV/AIDS + w2 * Income\ composition\ of\ resources + w3 * Schooling$$

"…There was an inverse and significant linear relationship between life expectancy and young age dependency rate, total fertility rate, child mortality rate, and a positive relation with HDI, adult literacy, contraceptive prevalence rate, HIV incidence rate, and TB incidence rate…" (Girum, Tadele, et al. "Determinants of life expectancy in low and medium human development index countries." *ResearchGate*, September 2018, 219, 220, 221, 222, 223, 224, 225.)

As quoted above in the article, life expectancy has a significant linear relationship with HDI, adult literacy, and HIV incidence rate. HDI can be understood as Income composition of resources (as defined in project 3 in lab). Adult literacy can be related to schooling (number of school years). HIV incidence rate can be described well by HIV/AIDS attribute. Besides, there are more factors affecting life expectancy as quoted, but out of 10 factors data provided, I see those 3 above are the most suitable for the model . This explains my model 2.

- **Model 3**

$$y = w * GDP^{2}$$

A research article named *"Relationship between GDP, Life Expectancy and Growth Rate of G7 Countries",* published in June 2019 discussed particularly the link between life expectancy and GDP. Quote saying:

"…Schnabel & Eilers (2009) explored that life expectancy has a nonlinear influence on wealth. They followed research on Preston's study, in which life expectancy and GDP had a curvilinear relationship…" (Shafi, Rafia, and Samreen Fatima. "Relationship between GDP, Life Expectancy and Growth Rate of G7 Countries." *International Journal of Sciences*, no. 8, 2019, 75,76,77,78,79.). And here is another finding in the article :

"... Table 3 represents the correlation coefficients between GDP and Population Growth Rate, between GDP and Life Expectancy, and between Population Growth Rate and Life

Expectancy…The correlation coefficients between GDP and Life Expectancy are highly positive for all seven countries, showing a strong bonding between the two variables… " (Shafi, Rafia, and Samreen Fatima. "Relationship between GDP, Life Expectancy and Growth Rate of G7 Countries." *International Journal of Sciences*, no. 8, 2019, 75,76,77,78,79.)

Population growth rate is not considered in the data provided. Therefore, GDP is the factor seen to affect expectancy. Furthermore, GDP should be squared because as the quote said *"life expectancy and GDP had a curvilinear relationship".* This explains my model 3.

- **Result after train 3 model:**

No	Model	RMSE
1	y = w1 * Polio + w2 * Diphtheria + w3 * HIV/AIDS	15.04847921
2	y = w1 * HIV/AIDS + w2 * Income composition of resources + w3 * Schooling	11.2842458
3	$y = w * GDP^2$	66.1290711

- **Retrain with the best model: y = w1 * HIV/AIDS + w2 * Income composition of resources + w3 * Schooling**
    - w = [0.0269354523, 35.1354088, 3.72006160]
    - RMSE= 9.617618622409767
    - Life expectancy = 0.0269354523 * HIV/AIDS + 35.1354088 * Income composition of resources + 3.72006160 * Schooling
- **Comment:**

Depending on RMSE value model 2 gave, we can see that life expectancy depends on many factors. Model 1 and 3 just describe the correlation between life expectancy with 1 or 2 factors, which does not explain well how life expectancy depends on. Furthermore, there is absolutely a linear relationship between life expectancy and HDI, HIV/AIDS incidence rate, adult literacy,...(said in Girum, Tadele, et al. "Determinants of life

expectancy in low and medium human development index countries." *ResearchGate*, September 2018, 219, 220, 221, 222, 223, 224, 225.)

4. Work cited.

[1] Girum, Tadele, et al. "Determinants of life expectancy in low and medium human development index countries." *ResearchGate*, September 2018, 219, 220, 221, 222, 223, 224, 225.

[2] Reiff, Susanne. "Life expectancy is increasing globally." *Alumniportal Deutschland*, 2017,https://www.alumniportal-deutschland.org/en/global-goals/sdg-03-health/increasing-life-expectancy-age-ageing/. Accessed 31 July 2022.

[3] Shafi, Rafia, and Samreen Fatima. "Relationship between GDP, Life Expectancy and Growth Rate of G7 Countries." *International Journal of Sciences*, no. 8, 2019, 75,76,77,78,79.

[4] Gilbert, Lisa, et al. "Life and Expectancy, An evidence review exploring the link between literacy and life expectancy in England through health and socioeconomic factors." *National Literacy Trust*, February 2018, p. 3.

5. Reference.

[1] Lab 04 and Project 03 lab at courses.ctda.hcmus.edu.vn