ISE-291: Introduction to Data Science

Term 221

Homework 03 2022

Covers: Topics 6-7 Material **Deadline: 24 November 2022**

Homework Presentation & Submission:

- You must ***attempt only the first two problems (A & B), while the others (C, D & E) are for practice***.
- It is ***group homework***.
- Provide the ***output solution*** for each sub-problem (part). Otherwise, 50% marks will be deducted.
- Every sub-problem (part) should be answered on a ***DIFFERENT CELL*** as given in the template.
- ***EVERY CELL*** should have a problem and part number clearly written in the first line.
- All cells of your homework should be in ***CHRONOLOGICAL order***. One cell per sub-problem.
- Any text should be written as a comment in the ***code cell***. Do NOT modify the code cell into the markdown cell.
- Submit the entire homework as ***ONE single .ipynb*** document.
- ***Do NOT add/delete*** any cell in the given template.

Problem A [50 Marks]: Solve all the questions using Python. Use Pandas, Seaborn, Sklearn, etc., libraries for all the analysis. Consider the data given in CSV file HW3_DataA. Consider the following data description:

Table 1. Data description

Field	Description
Model	A particular version of a camera produced by a specific camera manufacturing company
Release date	The date from which the product is on sale
Max resolution	Highest dimension in pixel
Low resolution	Lowest dimensions in pixel
Effective pixels	Effective pixels are the pixels that are about actually capturing the image data. These are the pixels that are doing the work of capturing a picture.
Zoom wide (*W*)	Zoom wide on the focal length
Zoom tele (*T*)	Zoom tele of the focal length
Normal focus range	The distance of focus
Macro focus range	It is the nearest distance between the camera and the subject such that the camera can take clear and in-focus shots
Storage included	Local camera storage included
Weight (*inc. batteries*)	Weight of the camera, including batteries
Dimensions	Length and width of the camera
Price	Market purchase price of the camera

Do the following tasks (in exact sequence) using the "HW3_DataA" data:

A-1. *[2 marks]:* Read the HW3_DataA and print the number of columns and rows using a single command. Further, point out which columns have missing values. Finally, drop the rows consisting of missing data and again print the number of rows and columns.

A-2. *[4 marks]:* Create a new column named "price_cat" such that its values can be: "low" if the "Price" is less than and equal to 170, "high" if the "Price" is greater than and equal to 350, and remaining all should be "medium". Display 9 random rows of the modified data frame.

A-3. *[4 marks]:* The column "Model" consists of the brand name and model number. For example, "Olympus FE-150", where Olympus is the model's name and FE-150 is the model number. Split the column "Model" into two new columns, "Brand_Name" and "Model_Name", and delete the column "Model" and "Price". Display the first 7 rows of the modified data frame.

A-4. *[5 marks]:* Plot a line chart to depict the number of camera releases each year and set the title as “Number of models released each year” using font size 23. Comment on the trend learned by the obtained line plot.

A-5. *[5 marks]:* Identify the pair-wise correlation between all the numerical variables using a heatmap plot and set the title as “Correlation between Attributes” using font size 17. From the results, identify which pairs have the most positive and negative correlations.

A-6. *[15 marks]:* Assume “price_cat” is an output variable, then perform a Principal Component Analysis (PCA) only for the numerical columns and set the components equals to two.

a) Print the variances captured by each principal component.

b) What are the coefficients of the linear combination of the columns? Moreover, provide the list of three columns having the highest coefficients.

c) Construct a scatter plot using the first two principal components of the data. Next, differentiate the points using “price_cat” column. Comment on the plot.

A-7. *[5 marks]:* Make a table displaying the average of “Weight (inc. batteries)” for “price_cat” and “Release date” columns.

A-8. *[5 marks]:* Provide a table showing the number of cameras for “price_cat” and “Brand_Name” columns. For each price category, which brand has a large number of cameras?

A-9. *[5 marks]:* Answer the following questions:

a) Which technique(s) can be used for performing descriptive analysis/analytics?

b) Which technique(s) can be used for performing exploratory analysis/analytics?

c) Which technique(s) can be used for performing predictive analysis/analytics?

d) Discuss whether it is possible to include the categorical data fields for Principal Components Analysis.

e) What is the use of idxmin() method in cross or pivot tables?

Problem B [50 Marks]: Solve all the questions using Python. Use Pandas, Seaborn, Sklearn, etc., libraries for all the analysis. Consider the data given in CSV files HW3_DataB and HW3_DataC. Consider the following data description:

Table 2. Data description

Field	Description
Id	player ID (index)
current_rating	Current rating score of the player
potential_rating	Potential rating score of the player
height	The height of the player in cm
weight	The weight of the player in kg
preferred_foot	Preferred foot to score (Right, Lift)
skill_moves	Number of skill moves that the player has
value	The value of the player on the market
wage	The yearly wage of the player
ball_control	The accuracy of the ball control
strength	The strength of the player
long_shots_level	The accuracy level of the long shots (Excellent, Good, Fair)

Do the following tasks (in exact sequence):

B-1. *[2 Marks]:* Given Data HW3_DataB. Read the data and display the first three rows of the data. Identify the number of rows and columns. Display the statistical summaries of all the columns.

B-2. *[3 Marks]:* Type Consistency. For each column in HW3_DataB, identify the type for each field based on value. Also, identify the data types in Python. Report any inconsistency.

B-3. *[2 Marks]:* Handling Missing data. First, permanently remove the "id" column and any inconsistent column(s) from the data. Further, check does any column have missing data? If there is, apply appropriate imputation to fill the data.

B-4. *[5 Marks]:* Normalization. Apply the appropriate transformation for each column in the updated HW3_DataB data. Finally, display the summaries of all the columns in the transposed form.

B-5. *[5 Marks]:* Relation plot. Provide the relation plot along with the distribution plot for all numeric variables of updated HW3_DataB data.

B-6. *[5 Marks]:* OLS Regression. The hypothesis is that all the explanatory variables are linearly related to an explained variable named "wage". Use the formula $\theta = (X^T X)^{-1} X^T Y$, calculate the OLS coefficient estimates of updated HW3_DataB data.

B-7. *[5 Marks]:* OLS Regression. Use the sklearn library to calculate the OLS coefficient estimates of updated HW3_DataB data. Take column 'wage' as the output and all other columns as the input. Compare the coefficients obtained in Part B-6 with the above coefficients. Report any differences between the coefficients from Parts B-6 and B-7.

B-8. *[10 Marks]:* Read the data HW3_DataC and perform all the initial operations implemented on the HW3_DataB. Using the above OLS coefficient estimates (obtained from B-6), calculate the MSE for the updated HW3_DataC data.

B-9. *[5 Marks]:* Ridge Regression. It may be possible that the explanatory variables are not really independent. Thus, the coefficients need regularization (penalization). Do the following:

- Do the ridge analysis, taking updated HW3_DataB data as the training data. Use 6-fold cross-validation and pick the best value of alpha from 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3. Print the best penalty coefficient.
- Using the above coefficient estimates, calculate the MSE for the data given in the updated HW3_DataC data.

B-10. *[5 Marks]:* Lasso Regression. It may be possible that not all the explanatory variables are helpful in predicting the final score. Thus, the coefficients need selection (penalization). Do the following:

- Do the lasso analysis, taking updated HW3_DataB data as training data. Use 7-fold cross-validation and pick the best value of alpha from 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3. Print the best penalty coefficient.
- Using the above coefficient estimates, calculate the MSE for the data given in the updated HW3_DataC data.

B-11. *[3 Marks]:* Regression Analysis. Compare and contrast the coefficient estimates obtained from Parts B-6, B-8, B-9, and B-10.

Problem C: (Practice only. No submission required.)

Consider the following python methods, available in naive Python, or pandas/sklearn libraries:

C-1. pandas.DataFrame.corr

C-2. pandas.DataFrame.concat

C-3. pandas.DataFrame.from records

C-4. pandas.crosstab

C-5. pandas.DataFrame.pivot table()

C-6. matplotlib.pyplot.subplots()

C-7. pandas.DataFrame.idxmax()

C-8. pandas.DataFrame.max()

C-9. sklearn.model selection.train test split()

C-10. sklearn.metrics.mean squared error()

C-11. numpy.linalg.inv()

C-12. numpy.c

C-13. numpy.linspace()

Answer the following questions for each of the above methods:

- State the purpose/usage of the method/attribute.
- List all the argument of the method.
- Classify the arguments as positional or keyword arguments.
- Write the default values for each of the keyword arguments.

Consider the following python class, available in sklearn library:

C-14. sklearn.decomposition.PCA

C-15. sklearn.linear model.LinearRegression

C-16. sklearn.linear model.RidgeCV

C-17. sklearn.linear model.LassoCV

Answer the following questions for the above class:

- List all the methods and properties/attributes.
- Discuss the .fit() method.
- Discuss the .transform() method.
- Discuss the .fit transform() method.

☞ Note: You must use help() function from Python to answer all the above questions.

Problem D: (Practice only. No submission required.)

Explain the following Python codes. Assume df represents an existing pandas’ dataframe, where the columns are C1, C2,.... The columns with odd numbers are categorical, and columns with even numbers are numerical. Also, assume that relevant libraries are imported before executing the following code:

E-1. Code-1:

```
corr = df.corr()
sns.heatmap(corr)
```

E-2. Code-2:

```
print(df[’C1’].values.reshape(-1,1))
print(df[[’C1’]]
```

E-3. Code-3:

```
ndf=pd.concat([df[ 1 ’C1’], df[’C2’], axis=1)
display(ndf.sample(5))
```

E-4. Code-4:

```
plt.figure()
sns.relplot(x=’C2’,y=’C4’,hue=’C1’, palette=[’r’,’b’,’g’,’m’,’c’],
kind=’scatter’,alpha=0.75,height=5, aspect=1,data=df)
plt.show()
```

E-5. Code-5:

```
cat_columns = df.select_dtypes( 1 ’object’).columns.drop(’C1’)
num_columns = df.select_dtypes(exclude=’object’).columns
fig,axes = plt.subplots(len(cat_columns), len(num_columns), figsize=(9,9))
for c,nCol in enumerate(num_columns):
for r,cCol in enumerate(cat_columns):
sns.boxplot(y=cCol,x=nCol,hue=’C1’,data=df, ax=axes[r][c])
plt.show()
```

E-6. Code-6:

```
X = df.iloc[:,:-1].values
y = df.iloc[:, -1].values
print(np.c_[np.ones(len(df.index)), X])
```

E-7. Code-7:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(df[[’C2’,’C4’]])
df[[’C2’,’C4’]] = scaler.transform(df[[’C2’,’C4’]])
df[[’C6’,’C8’]] = scaler.transform(df[[’C6’,’C8’]])
```

Problem E: Consider data given in HW3_DataD.csv. (Practice only. No submission required.)

Table 3. Data description

Field	Description
risk	-3, -2, -1, 0, 1, 2, 3; where 3 implies highest risk.
make	company and model/name of the car
fuel-type	diesel, gas.
aspiration	std, turbo.
num-of-doors	four, two.
body-style	hardtop, wagon, sedan, hatchback, convertible.
drive-wheels	4wd, fwd, rwd.
engine-location	front, rear.
wheel-base	continuous from 86.6 120.9.
length	continuous from 141.1 to 208.1.
width	continuous from 60.3 to 72.3.
height	continuous from 47.8 to 59.8.
curb-weight	continuous from 1488 to 4066.
engine-type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
num-of-cylinders	eight, five, four, six, three, twelve, two.
engine-size	continuous from 61 to 326.
fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
bore	continuous from 2.54 to 3.94.
stroke	continuous from 2.07 to 4.17.
compression-ratio	continuous from 7 to 23.
horsepower	continuous from 48 to 288.
peak-rpm	continuous from 4150 to 6600.
city-mpg	continuous from 13 to 49.
highway-mpg	continuous from 16 to 54.
price	continuous from 5118 to 45400.

Do the following tasks (in exact sequence) using data given in HW3_DataD:

E-1. Given Data. Does any column have missing data? If yes, then drop all the rows that contain any missing values.

E-2. Type Consistency. For each column in HW3_DataD, identify the type of each field and verify that each column in Python is identified correctly. If there is any inconsistency, then resolve it.

E-3. Normalization. For each score column in HW3_DataD, apply the standard scaler such that the mean is zero and the standard deviation is one.

E-4. Correlation Analysis. Identify the top 5 numerical variables that are highly correlated with the "price" column.

E-5. PCA. Do the following:

- Get the first two principal components of the numerical data without considering the "price" column.

- Add the two principal components to the dataframe and rename the components “pc1” and “pc2” respectively.
- Construct a scatter plot using the first two principal components of the data, and differentiate the plot using “price” column. Can the principal components separate “price” column?
- Drop “make”, “pc1,” and “pc2” column from the dataframe, and convert all other non-numerical columns to the numerical columns using one hot encoding.
- Repeat the first three steps using numerical (and encoded) columns as inputs to the PCA.

☞ **Note: Solve all the above questions using Python (not by hand). Use Pandas, Seaborn, SkLearn, etc. libraries for all the above analysis.**

-- THE END --