Practical Exam Data Analytics 1: 12.06.2024

Hints and Instructions:

This exam includes 4 problems.

1. Write your code in a file named Forename_Surename.R, or Student-ID.R. Submit your .ipynb codes in separate files.

2. Submit your .R and .ipynb files in the "Examination" section on Moodle.

3. Use # or ## to insert comments. Your comments or description of your code could have some points in case your script has errors.

4. You can use the following material:

- Your computer.
- Slides from course.
- Internet searches and resources.

5. You **can not** use the following material:

- ChatGPT or any Generative AI
- Chatting tools

Download the following CSV file, and use that for the following questions:

dataset.csv:
https://www.dropbox.com/scl/fi/i887jbzpfgmqljzrhfwvu/dataset.csv?rlkey=k5ae1jftetl7mgqms92b848pk&st=i6ylumlp&dl=0

sample.csv:
https://www.dropbox.com/scl/fi/e2bskyc92zxbtpieusi25/sample.csv?rlkey=46ail9zyyjgx9upg81qfi892y&st=fjn9js7g&dl=0

mydata.csv:
https://www.dropbox.com/scl/fi/bbqkx6l17wpwdcmbpymft/mydata.csv?rlkey=ajk1lugh7omgv5gqtgaie8pbu&st=v0ccpsiy&dl=0

data_1.csv:
https://www.dropbox.com/scl/fi/hwwoex6texqk7ptz95ngm/data_1.csv?rlkey=lo7oaqfsspjoeux6a191oj8lt&st=mrft1mpy&dl=0

data_2_test.csv:
https://www.dropbox.com/scl/fi/cilzrajdqrqqoqx06zfov/data_2_test.csv?rlkey=po0dacqufhpj1x1jtoq6wzmqj&st=kywf1ox1&dl=0

data_2_train.csv:
https://www.dropbox.com/scl/fi/7mp1u7o30zkm1acbmbutf/data_2_train.csv?rlkey=3v34tnu2va4gyg2o7cdm2b3t0&st=2beylf3a&dl=0

Hi, Ashwith. When you submit this form, the owner will see your name and email address.

1.
    1. Load dataset.csv file and perform feature selection using Boruta to identify important features for predicting V2 using all other variables in the dataset and except for V1.
    2. Get non-rejected features from Boruta.
    3. Build random forest model with Boruta-selected features.
    4. Print the random forest model.
    5. Define the control parameters for cross-validation.
    6. Train the logistic regression model with cross-validation.
    7. Print the model with details and the MAE.
    8. Load the sample.csv file and display the first few rows of the dataset.
    9. Generate the kNN distance plot for k = 5 and explain the significance of the plot.
    10. Based on the kNN distance plot and the elbow method, determine a suitable value for eps.
    11. Perform DBSCAN clustering with minPts = 5 and eps you got from the previous step.
    12. Visualize the clustering result using fviz_cluster and on second and third column of your data.
    13. Perform hierarchical clustering
    14. Cut Dendrogram to 5 subtrees to Define Clusters
    15. Visualize the dendrogram with the identified clusters
    16. Draw rectangles around the clusters on the dendrogram (Hint: k=5)

Enter your answer

2.

1. Load necessary libraries and the mydata.csv file
2. Change the rating column to 1 (if rating > 5) or 0
3. Convert the rating column to a factor type
4. Split the dataframe into train and test sets with an 80% ratio for training and 20% for testing: (Trying to predict "rating")
5. Build the random forest model and make predictions for the logistic regression model
6. Build the logistic regression model and make predictions for the random forest model
7. Create ROC curves for both models
8. Compare the ROC results of both models by plotting the ROC results of both models
9. Calculate AUC for logistic regression and random forest models and print the results

Enter your answer

3. import numpy as np

   import pandas as pd
   from scipy.io.arff import loadarff
   from sklearn.cluster import KMeans

   df = pd.read_csv("./data/data_1.csv")
   X = df.values

   Question:

   - X is your dataset with features
   - Please define the optimal number of clusters
   - Consider the optimal number of clusters should be between [4, 12]
   - Use silhouette Score as evaluation metric, and plot it.

- You can use KMeans as cluseter

Enter your answer

4. import numpy as np
   import pandas as pd
   import matplotlib.pyplot as plt
   from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
   from statsmodels.tsa.arima.model import ARIMA

   df = pd.read_csv('./data/data_2_train.csv')
   X = df.values

   df_test = pd.read_csv('./data/data_2_test.csv')
   X_test = df_test.values

   Question:

   - Build a ARIMA model based on X (data_2_train.csv)
   - Please demonstrate how do you select:
     - the differencing d
     - autoregressive term p
     - and monving average term q
   - Please use visual inspection and ADF test to determine the stationarity of time series
   - Once you crate your model, please plot the residuals of your ARIMA model and the density of residuals
   - Furthermore, please load the test data (data_2_test.csv) and plot your forecast and test data on a same plot, including the train data as well

   Enter your answer

Microsoft 365

This content is created by the owner of the form. The data you submit will be sent to the form owner. Microsoft is not responsible for the privacy or security practices of its customers, including those of this form owner. Never give out your password.

Microsoft Forms | AI-Powered surveys, quizzes and polls Create my own form

Privacy and cookies | Terms of use