UNIVERSITI MALAYA
UNIVERSITI MALAYA

UJIAN IJAZAH SARJANA MUDA KEJURUTERAAN
TEST FOR THE DEGREE OF BACHELOR OF ENGINEERING

SESI AKADEMIK 2023/2024 : SEMESTER II
ACADEMIC SESSION 2023/2024 : SEMESTER II

KIG4068 : Pembelejaran Mesin
Machine Learning

April/Mei 2024

Masa: 2 jam
April/May 2024 *Time: 2 hours*

ARAHAN KEPADA CALON:
INSTRUCTIONS TO CANDIDATES:

Calon dikehendaki menjawab semua soalan.
Candidate is required to answer all questions.

Ini adalah peperiksaan Buku Terbuka. Calon dibenarkan merujuk kepada helaian sokongan/nota kuliah/buku/internet.
This is an Open Book Examination. Candidates are allowed to refer to support sheet/lecture notes/books/internet.

Peringatan: Kertas soalan ini **TIDAK DIBENARKAN** dibawa keluar dari Dewan Peperiksaan.
*Reminder: This question paper is **NOT ALLOWED** to be taken out from the Examination Hall.*

(Kertas soalan ini mengandungi 2 soalan dalam 5 halaman yang dicetak)
(*This question paper consists of 2 questions on 5 printed pages)*

Soalan 1
Question 1

Anda diberikan dataset ('data_Q1.csv') yang mengandungi pelbagai ciri (Feature 1 hingga Feature 11) dan pembolehubah sasaran kategorikal (Target). Tugas anda adalah untuk melakukan pemprosesan data dan membina saluran pra pemprosesan untuk dataset ini. Anda boleh memuat turun dataset dari SPECTRUM>Midterm Test Datasets.
You are provided with a dataset ('data_Q1.csv') containing various features (Feature 1 to Feature 11) and a categorical target variable (Target). Your task is to perform data processing and construct a preprocessing pipeline for this dataset. You can download the dataset from SPECTRUM>Midterm Test Datasets.

a) Muatkan dataset ke dalam struktur data yang sesuai (contohnya, pandas DataFrame) dan paparkan beberapa baris pertama untuk memahami strukturnya.
Periksa jenis data setiap lajur dan semak adakah terdapat nilai yang hilang.
Penjelasan/Komen: Berikan penjelasan atau komen mengenai setiap ciri, termasuk apa yang anda perhatikan dalam histogram, plot taburan, atau visualisasi lain. Selain itu, huraikan sebarang corak atau korelasi yang dikenal pasti dan bincangkan bagaimana ini mungkin mempengaruhi keputusan pra pemprosesan anda.
Load the dataset into a suitable data structure (e.g., pandas DataFrame) and display the first few rows to understand its structure.
Examine the data types of each column and check for any missing values.
Explanation/Comment: Provide explanations or comments on each feature, including what you observe in histograms, scatter plots, or any other visualizations. Additionally, describe any patterns or correlations you identify and discuss how these may influence your preprocessing decisions.

(*2* markah/*marks*)

b) Berdasarkan pemerhatian dari langkah a, tentukan langkah-langkah pra pemprosesan seperti menangani nilai yang hilang, kod pembolehubah kategorikal, dan penskalaan ciri-ciri berangka.
Penjelasan/Komen: Terangkan alasan/penaakulan anda di sebalik setiap langkah pra pemprosesan. Sebagai contoh, mengapa anda memilih kaedah imputasi tertentu untuk menangani nilai yang hilang? Bagaimana anda memutuskan kaedah penyandian untuk pembolehubah kategorikal? Bincangkan implikasi penskalaan ciri-ciri berangka.
Based on your observations from step a, decide on preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features.
Explanation/Comment: Explain your reasoning behind each preprocessing step. For example, why did you choose a particular imputation method for handling missing values? How did you decide on the encoding method for categorical variables? Discuss the implications of scaling numerical features.

(4markah/*marks*)

c) Bina saluran pra pemprosesan yang mengintegrasikan langkah-langkah pra pemprosesan yang ditakrifkan dalam langkah b.
Penjelasan/Komen: Berikan penjelasan atau komen mengenai reka bentuk saluran anda. Bincangkan bagaimana setiap langkah pra pemprosesan menyumbang kepada pemprosesan data secara keseluruhan. Komen tentang sebarang cabaran yang anda hadapi dan bagaimana anda menangani mereka.
Kombinasi Ciri Tambahan untuk Penciptaan Saluran:

I. Feature Combination 1:
Attribute Name: Ratio_Feature1_2
Penerangan: Cipta Feature baru yang mewakili nisbah Feature1 kepada Feature2.
Formula: Ratio_Feature1_2 = Feature1 / Feature2

II. Feature Combination 2:
Nama Ciri: Sum_Feature3_4_5
Penerangan: Cipta Feature baru yang mewakili jumlah Feature3, Feature4, dan Feature5.
Formula: Sum_Feature3_4_5 = Feature3 + Feature4 + Feature5

III. Feature Combination 3:
Attribute Name: Multiplication_Feature6_7
Penerangan: Cipta Feature baru yang mewakili perkalian Feature6 dan Feature7.
Formula: Multiplication_Feature6_7 = Feature6 * Feature7

Construct a pipeline that integrates the preprocessing steps defined in step b.
Explanation/Comment: Provide explanations or comments on the design of your pipeline. Discuss how each preprocessing step contributes to the overall data processing. Comment on any challenges you encountered and how you addressed them.
Additional Feature Combinations for Pipeline Creation:

I. Feature Combination 1:
Attribute Name: Ratio_Feature1_2
Description: Create a new feature representing the ratio of Feature1 to Feature2.
Formula: Ratio_Feature1_2 = Feature1 / Feature2

II. Feature Combination 2:
Attribute Name: Sum_Feature3_4_5
Description: Create a new feature representing the summation of Feature3, Feature4, and Feature5.
Formula: Sum_Feature3_4_5 = Feature3 + Feature4 + Feature5

III. Feature Combination 3:
Attribute Name: Multiplication_Feature6_7
Description: Create a new feature representing the multiplication of Feature6 and Feature7.
*Formula: Multiplication_Feature6_7 = Feature6 * Feature7*

(*4* markah/*marks*)

Soalan 2
Question 2

Anda diberikan dengan dataset regresi ('data_Q2.csv') yang mengandungi n sampel dan m ciri, yang boleh dimuat turun dari SPECTRUM>Midterm Test Datasets. Tugas anda adalah untuk menangani kelebihan penyesuaian dengan mengguna teknik-teknik peminggiran Ridge, Lasso, dan ElasticNet serta menganalisis keberkesanan mereka dalam meningkatkan generalisasi model.
You are provided with a regression dataset ('data_Q2.csv') containing n samples and m features, which can be downloaded from SPECTRUM>Midterm Test Datasets. Your task is to address overfitting by applying Ridge, Lasso, and ElasticNet regularization techniques and analyzing their effectiveness in improving model generalization.

a) Muatkan dataset regresi yang disediakan dan bahagikan ia kepada set latihan dan ujian menggunakan kaedah yang sesuai, memastikan saiz ujian adalah 20%. Standardkan ciri-ciri untuk memastikan bahawa mereka mempunyai min 0 dan sisihan piawai 1.
Load the provided regression dataset and split it into training and testing sets using an appropriate method, ensuring a test size of 20%. Standardize the features to ensure that they have a mean of 0 and a standard deviation of 1.

(*1* markah/*marks*)

b) Latih model Regresi Linear standard menggunakan dataset latihan dan nilai prestasinya diuji pada dataset ujian menggunakan Mean Squared Error (MSE) dan R-squared (R^2) sebagai metrik penilaian.
Train a standard Linear Regression model using training dataset and evaluate its performance on the testing dataset using Mean Squared Error (MSE) and R-squared (R^2) as evaluation metrics.

(*1* markah/*marks*)

c) Lakukan penyesuaian hiperparameter untuk setiap teknik peminggiran menggunakan carian grid cross-validation:
   - Untuk peminggiran Ridge, optimalkan parameter alpha dalam julat [0.01, 0.1, 1.0, 10.0].
   - Untuk peminggiran Lasso, optimalkan parameter alpha dalam julat [0.01, 0.1, 1.0, 10.0].
   - Untuk peminggiran ElasticNet, optimalkan kedua-dua parameter alpha dalam julat [0.01, 0.1, 1.0, 10.0] dan l1_ratio dalam julat [0.1, 0.5, 0.9].

   Perform hyperparameter optimization for each regularization technique using cross-validation grid search:
   - *For Ridge regression, optimize the alpha parameter in the range [0.01, 0.1, 1.0, 10.0].*
   - *For Lasso regression, optimize the alpha parameter in the range [0.01, 0.1, 1.0, 10.0].*
   - *For ElasticNet regression, optimize both the alpha parameter in the range [0.01, 0.1, 1.0, 10.0] and l1_ratio parameter in the range [0.1, 0.5, 0.9].*

(*3* markah/*marks*)

d) Latih setiap teknik peminggiran menggunakan dataset latihan dengan hiperparameter yang dioptimumkan yang diperoleh daripada carian grid cross-validation, dan nilai prestasi setiap model diuji pada dataset ujian menggunakan Mean Squared Error (MSE) dan R-squared (R^2) sebagai metrik penilaian.
Train each regularization technique using the training dataset with the optimized hyperparameters obtained from the cross-validation grid search, and evaluate the performance of each model on the testing dataset using Mean Squared Error (MSE) and R-squared (R^2) as evaluation metrics.

(*2* markah/*marks*)

e) Bandingkan dan bincangkan hasil yang diperoleh daripada teknik-teknik peminggiran Ridge, Lasso, dan ElasticNet, termasuk:
   - Prestasi pada dataset latihan (MSE dan R-squared)
   - Prestasi pada dataset ujian (MSE dan R-squared)
   - Kompleksiti model dan pemilihan ciri.

   Compare and discuss the results obtained from Ridge, Lasso, and ElasticNet regularization techniques, including:

   - *Performance on the training dataset (MSE and R-squared)*
   - *Performance on the testing dataset (MSE and R-squared)*
   - *Model complexity and feature selection.*

(*2* markah/*marks*)

f) Berikan pandangan tentang teknik peminggiran yang mungkin lebih berkesan dalam menangani kelebihan penyesuaian dalam dataset ini dan terangkan alasan/penaakulan anda.

   Provide insights into which regularization technique might be more effective in addressing overfitting in this dataset and explain your reasoning.

(*1* markah/*marks*)

TAMAT
END