Pass Task 2

Solve the following set of problems using Python and submit the code file with the extension .ipynb in Olympus as part of your pass activity.

1. Load the "housing.csv" dataset and follow this link for the data description (features and target variable). Split the dataset in train and test set (use your choice of splitting). Train a linear regression model and report the performance (use your choice of at least 4 performance metric).

2. Apply PCA on the dataset and select the first three principal components. Split the dataset into train and test using the same method used in Q1. Compare the performance of this model with the performance obtained in Q1. Explain the outcome.

3. Load "IRIS " datasets from SKlearn and follow this link for the data description (features and target variable). Apply PCA on the dataset and select the first three principal components. Split the dataset in train and test set (use your choice of splitting). Train a logistic regression model and report the performance (use your choice of at least 4 performance metric).
4. Apply L1 or L2 regulariser on the logistic regression model developed using the same train and test data used in Q3 and calculate performance of the new model. Compare performance of this model with the performance reported in Q3. Explain the outcome.

5. Load "digits" datasets from SKlearn and print the dimension of the dataset. Apply PCA on the dataset and select first three components. Print the dimension of modified dataset and visualise the data using appropriate plotting tool/s.

6. Classify the digit classes available in the dataset (use the modified dataset) using SVM with RBF kernel. Select appropriate data splitting approach and performance metrics. Report the performances and the used model hyper-parameters.

7. Load "diabetes" datasets from SKlearn and print the dimension of the dataset. Apply tSNE method to reduce dimension and select first three components. Plot the selected components using appropriate visualisation technique.

8. Create a model for detecting diabetes using SVM with a poly kernel. Select appropriate data splitting approach and performance metrics. Report the performances and the used model hyper-parameters.

9. Based on the model hyper-parameters used in Q6 and Q8, share your understanding of hyper-parameters tuning in ML model development.

Assessment feedback

The results with comments will be released within 5 business days from the due date.

Referencing

You must correctly use the Harvard method in this assessment. See the Deakin referencing guide.

Academic integrity, plagiarism and collusion

Plagiarism and collusion constitute extremely serious breaches of academic integrity. They are forms of cheating, and severe penalties are associated with them, including cancellation of marks for a specific assignment, for a specific unit or even exclusion from the course. If you are ever in doubt about how to properly use and cite a source of information refer to the referencing site above.

Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task.

Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose, with the intent of obtaining an advantage in submitting an assignment or other work.

Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions: https://www.deakin.edu.au/students/study-support/referencing/academic-integrity.