COS40007 Artificial Intelligence for Engineering

Portfolio Assessment 2: "Systematic Approach to Develop ML Model"

Due: by Sunday of Week 4 (30/03/2025 23:59 PM) in Canvas

AIM

The aim of this task is for you to demonstrate your understanding of a systematic approach to developing a machine-learning model. That is, how to perform data pre-processing, class labelling, data separation, feature selection, training and model evaluation.

About the Dataset

Download the provided dataset (**ampc2.zip**) from Canvas.

This dataset contains acceleration data collected using 17 body-worn sensors in different body positions. From 17 sensors, we get 22 body position values in x, y, and z acceleration, so we have 66 columns. 2 data files contain acceleration values, each belonging to a type of meat processing activity (boning and slicing). Each row contains data from 1 frame, and the length of each frame is 1 second. That means 60 frames contain 1 minute of data.

Disclaimer: The dataset used in this task was initially collected for a funded research project by the Australian Meat Processor Corporation. The dataset here is used solely for educational purposes and can only be used to complete activities for this studio. By any means, this dataset is not shareable with others or in any public domain.

The Task Objective

This task aims to develop a classification model that can distinguish between two types of activities: boning and slicing. To get there, we must perform all the ML model development steps we learned in Week 2. For computing features, you can use a similar set of features we used to develop our ML models in Studio 3. If you examine the Studio 3 data headers closely, you will notice that some statistical features have already been computed for you.

For example,
acc_mean_x_right -> Mean value of x-axis acceleration per minute for the right-hand motion sensor.
acc_std_xyz_right -> Standard deviation of xyz (i.e., the root mean square of x, y, and z values) per minute for the right-hand motion sensor.

In this assessment task, you must compute these features to generate training data for your machine learning models.

Step 1: Data Collection

For this assessment task, you do not need to use all columns. You will work with columns as assigned in the table below (including the frame column)

Student number	Column set 1	Column set 2
Ending with 0	Neck (x,y,z)	Head (x,y,z)
Ending with 1	Right Shoulder (x,y,z)	Left Shoulder (x,y,z)
Ending with 2	Right Upper Arm (x,y,z)	Left Upper Arm (x,y,z)
Ending with 3	Right Forearm (x,y,z)	Left Forearm (x,y,z)
Ending with 4	Right Hand (x,y,z)	Left Hand (x,y,z)
Ending with 5	Right Upper Leg (x,y,z)	Left Upper Leg (x,y,z)
Ending with 6	Right Lower Leg (x,y,z)	Left Lower Leg (x,y,z)
Ending with 7	Right Foot (x,y,z)	Left Foot (x,y,z)
Ending with 8	Right Toe (x,y,z)	Left Toe (x,y,z)
Ending with 9	L5 (x,y,z)	T12 (x,y,z)

So, as the data collection step, extract the columns from the two files that you will work with and combine them along with the class value (0 – boning, 1 – slicing)

So, your data file will contain eight columns: -> frame column, six columns from column sets 1 and 2 and a class value (0 for boning and 1 for slicing)

Step 2: Create composite columns

Create some composite data points for your data as follows

1) Root mean square value of x and y
2) Root mean square value of y and z
3) Root mean square value of z and x
4) Root mean square value of x, y and z
5) Roll (180 * atan2(accelY, sqrt(accelX*accelX + accelZ*accelZ))/PI)
6) Pitch (180 * atan2(accelX, sqrt(accelY*accelY + accelZ*accelZ))/PI)

From the above, you will get 12 more columns for two settings. Merge this with your data in Step 1. Now, you will have a data file with 20 columns.
Column 1: Frame
Columns 2-4: Column set 1
Columns 5-7: Columns set 2
Columns 8-13: 6 computed columns in step 2 for column Set 1
Columns 14-19: 6 calculated columns in step 2 for column Set 2
Column 20: class (0 or 1)

Step 3: Data pre-processing and Feature computation

Now, you must write code to create statistical features per minute (1 minute = 60 consecutive frames) from the 18 columns. Using Studio 3, compute the following features for each column 2-19.

1. Mean
2. Standard deviation
3. Min
4. Max
5. Area under the curve (AUC)
6. Peaks (number of peaks)

So, after combining them, you will have a total of 18 × 6 = 108 features to work with.

Step 4: Training

As you did in Studio 3, with your dataset containing 108 features, develop SVM classifiers with the following settings.

1) Train-Test split (70/30)
2) 10-fold cross-validation
3) 1 and 2 with hyperparameter tuning
4) 1 and 2 with hyperparameter tuning and 10 best features
5) 1 and 2 with hyperparameter tuning and 10 principal components

Create a summary table of your outcome, similar to the one in Studio 3, Activity 6.

Now, train the SGD, Random Forest, and MLP classifiers using an original dataset similar to Studio 3 activity seven and generate the outcome table.

Step 5: Model Selection

Answer the following question:

1) Which SVM model will be the best for your problem
2) Which ML model will be the best for your problem?

Submission

Create a folder and place all your data files, including intermediate data files and code, within that folder. Then, create a sharable link to that folder.

The portfolio assessment submission should be a document (in Word or PDF format) containing the following.

- Your name and Student number
- The studio class you attend (for example, you attend Studio 1-1, then write Studio 1-1)
- Summary Table of Studio 3: Activity 6 **[1 mark]**
- Summary Table of Studio 3: Activity 7 **[1 mark]**
- Step 1: Data collection (link to your source code and data) **[1 mark]**
- Step 2: Create composite columns (link of your source code and data) **[1 mark]**
- Step 3: Data pre-processing (link of your source code and data) **[3 marks]**
- Step 4: Training (outcome summary tables) **[2 marks]**
- Step 5: Model Selection (explain in 1-2 lines the reason for your selection) **[1 mark]**

Total **10 marks**