COS40007 Artificial Intelligence for Engineering

Portfolio Assessment-2: "Systematic approach to develop ML model"

Due: by Sunday of Week 4 (18/08/2024 23:59 PM) in Canvas

Aim

The aim of this task is for you to demonstrate your understanding of systematic approach to develop 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)

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,z acceleration, so we have total 66 columns. There are 2 data files that contains acceleration values and each belong to a type of meat processing activity (boning and slicing). Each row contains data from 1 frame and length of each frame is 1 second. That's mean 60 frames contains 1 minute data.

Disclaimer: The dataset used in this task was originally collected for a funded research project by Australian Meat processor Corporation. The dataset here is used solely for educational purposes and can be only used for completing activities of this studio. By any mean this dataset is not shareable to others or any public domain.

Objective of the task

Now the objective of this task is to develop a classification model that can distinguish between 2 types of activities boning and slicing. To get there, we need to perform all the steps of machine learning model development that we learned in Week 2. For computing features you can use the concept of similar set of features that we used to develop our ML models in Studio 3. If you closely look at the studio 3 data headers you will see that some statistical features were already computed for you.

For example,
acc_mean_x_right -> mean value of x axis acceleration per minute for right hand motion sensor
acc_std_xyz_right-> standard deviation of xyz (with is basically root mean square of x y and z values) per minute for right hand motion sensor

In this assessment task you need to compute such features by yourself to generate training data of your ML models.

Step 1: Data collection

For this assessment task you don't need to use all columns. You will work with columns as assigned in the below table (including 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 data collection step extract the columns from the 2 files that you will work with and combine them along with class value (0- boning, 1 – slicing)

So, your data file will contain 8 columns -> frame column, 6 columns from column set 1 and 2 and 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)

So, you will get 6 more columns from the above. Merge this with your data in Step 1. Now you will have data file with 14 columns.

Column1: Frame

Column 2-4: Column set 1
Column 5-7: Columns set 2
Column 8-13: 6 computed column in step 2
Column 14: class (0 or 1)

Step 3: Data pre-processing and Feature computation

Now you will need to write code to create statistical features per minute (1 minute = consecutive 60 frames) from the 13 columns. Like we use in Studio 3 compute the following features for each column 2-13.

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 total (12X6=72 features to work with)

Step 4: Training

Like you did in Studio 3, with your dataset containing 72 features develop SVM classifiers with the following settings

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

Create the summary table of your outcome similar to Studio 3 Activity 6.

Now, train SGD, RandomForest and MLP classifier with original dataset similar to Studio 3 activity 7 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 of your data file (including intermediate data file) and code in that folder. Then create a sharable link of that folder

The portfolios assessment submission should be a document (word or pdf) with 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 of 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 line the reason of your selection) **[1 mark]**

Total **10 marks**