ML Applications - Vehicle Motion State Classification

Lab assignment #4 (worth 10%; deadline: Friday, November 29 at 11:59pm)

Learning objectives

After completing this lab assignment students should be able to:

- Understand and implement ML model training and hyperparameter tuning using grid search with cross validation
- Apply Support Vector Machine for classification of vehicle motion state using onboard sensor data
- Appreciate significance of underlying process behind data, the potential error sources in training datasets, and explore ways to train models that are resilient despite presence of errors in data

Preliminary Task – (25 Marks)

Similarly to previous assignments, each group must fill in the **certificate of work**, which is in the given notebook of this lab assignment (5 marks). Each group is responsible for maintaining a clean Git repository like the Bookkeeping tasks from Assignment 1, E.g., clear project structure, consistent version control, include only necessary files etc. (10 marks). Make sure the code is readable, E.g., consistent naming conventions, proper commenting, clear code structure, consistent Indentation and Spacing etc. (10 marks)

Background

In this assignment, machine learning is applied to the vehicle motion state classification problem. The motion states of a travelling vehicle is to be identified using data recorded from an inertial measurement unit (IMU) sensor onboard the vehicle. Supervised learning algorithms must be trained to detect three vehicle states: rest, straight driving, and turning. Labeled inertial sensor data, derived from real-world driving dataset have been provided [1] for model training. Overall, three datasets are provided:

1. *UrbanNav-HK_Whampoa-20210521_sensors.csv (**training and tuning**)*
2. *UrbanNav-HK_TST-20210517_sensors.csv (**first independent testing**)*
3. *2020-03-14-16-45-35.csv (**second independent testing**)*

The first dataset is to be used for model training and hyperparameter tuning using grid search with cross validation (GridSearchCV). The second and the third datasets are to be used as independent test datasets.

Each dataset consists of 33 columns: the first 32 columns contain *features* derived from inertial sensor data, while the last column contains the *target* column or data label. The machine learning algorithm must use the features from the first 32 columns to predict the target column. The detailed explanation behind the feature engineering and data labeling is provided in the video lecture accompanying the assignment.

It is highly recommended to watch this video. The video also provides important background information regarding IMU sensors, which is necessary for answering questions in this assignment.

Task 1 – (33 Marks)

In the first section, a Support Vector Machine classifier (SVC) model must be trained using *UrbanNav-HK_Whampoa-20210521_sensors.csv* training dataset. All 32 features provided must be used.

The section starts with importing and exploring the datasets used for the entire assignment. Next GridSearchCV must be performed: the training data must be separated into features and target, further train/test split must be performed, the grid search parameters are to be defined, and the search must be performed to find the best model parameters using the train split. The results of training and cross validation are to be explored and analyzed using the accuracy score, confusion matrix, and classification report.

Finally, the trained SVC model must be evaluated using the independent test datasets: *UrbanNav-HK_TST-20210517_sensors.csv,* and *2020-03-14-16-45-35.csv*. The test datasets, which were loaded at the start, must also be separated into features and target. The model then must be evaluated on these datasets using the accuracy score, confusion matrix, and the classification report.

***Detailed instructions are provided in the assignment Colab notebook*.**

Task 2 – (12 Marks)

In the second section, another SVC model must be trained also using the *UrbanNav-HK_Whampoa-20210521_sensors.csv.* However, only the following selected features must be used in model training:

- *ax_var* - windowed variance of accelerometer readings in *x* direction
- *ay_var* - windowed variance of accelerometer readings in *y* direction
- *az_sum* - windowed sum of accelerometer readings in *z* direction
- *a_sum* - windowed sum of L2-norm of accelerometer readings
- *gz_var* - windowed variance of gyroscope readings in *z* direction

The steps from *Task 1* must then be repeated using the subset of features mentioned above. The model performance must be evaluated and then also compared to the performance achieved in *Task 1*.

***Detailed instructions are provided in the assignment Colab notebook*.**

Task 3 – (10 Marks)

Groups of 3 will be graded out of 80 and will not be awarded marks for this section

In the third section, a final SVC model must be trained also using the *UrbanNav-HK_Whampoa-20210521_sensors.csv* dataset. However, only a *subset of the 32 features* may be used. The *subset of features to use must be selected by the group* members. The reasoning behind the feature selection must be clearly presented and non-trivial. Use of ML tools which help in feature selection, such as Principal Component Analysis (PCA), is recommended but not required.

The same steps from *Tasks 1* and *2* are repeated using the group's selected features. The model performance must be evaluated and compared to the performance achieved in *Tasks 1* & *2*. Possible reasoning and insight for the performance observed must be provided.

***Detailed instructions are provided in the assignment Colab notebook*.**

Task 4 – (10 Marks)

Groups of 3 will be graded out of 80 and will not be awarded marks for this section

In section 4, dimensionality reduction of the feature set must be first performed using *Principal Component Analysis* (PCA). An SVC model will be trained using grid-search with cross validation, during which the PCA's hyperparameter will also be tuned. The PCA and SVC combined classification pipeline must be trained using the *UrbanNav-HK_Whampoa-20210521_sensors.csv* dataset.

The same steps from previous tasks are repeated. The performance must be compared to the performance achieved in *Tasks 1, 2,* and *3*. Possible reasoning and insight for the performance observed must be provided.

***Detailed instructions are provided in the assignment Colab notebook*.**

Task 5 – (10 Marks)

In the final section, a *classification algorithm chosen by the group* must be trained using the selected subset of features provided in *Task 2*. The classification algorithm must be trained using the *UrbanNav-HK_Whampoa-20210521_sensors.csv* dataset. The reasoning behind the model selected must be clearly presented and non-trivial. Algorithms presented in course lectures which are listed below maybe selected.

- Random Forest
- K-Nearest Neighbors
- Gradient Boosting Machines
- Neural Networks

The same steps from previous tasks are repeated using the group's selected classification algorithm. The model performance must be compared to the performance achieved in *Task 2*. Possible reasoning and insight for the performance observed must be provided.

***Detailed instructions are provided in the assignment Jupyter notebook*.**

Submission Materials

All the listed materials **must** be submitted in the *Lab Assignment 4 D2L* dropbox by the deadline. **Late** and **incomplete** submissions will result in a deduction of marks.

- A zip file of the `Lab_Assignment4` folder, which should include the Assignment 4*.ipynb* file
- A link to the group GitHub repository
- **NO REPORT** is required for this assignment

As stated from "*Additional Lab Information*", please remember that:

- Each member is **required** to make at least one commit and push to the group repository. Failure to do so will result in a 1% **individual** course grade deduction.
- The D2L submission dropbox will close 24 hours after the deadline. All submissions in this 24-hour window are subject to a 20% lab grade deduction.
- There must not be any commits to the `Lab_Assignment4` folder in your GitHub repository after the deadline, or there will be a 20% lab grade deduction.

References

[1] L.-T. Hsu *et al.*, "Hong Kong UrbanNav: An Open-Source Multisensory Dataset for Benchmarking Urban Navigation Algorithms," *Navig. J. Inst. Navig.*, vol. 70, no. 4, Dec. 2023, doi: 10.33012/navi.602.