STAT40970 – Machine Learning & A.I. (Online)

Assignment 2

Deadline - ~~Tuesday 15th April 2025 at 17:00~~ Monday 21st April 2025 at 17:00

Exercise 1

The following R `keras` chunk of code is used to define a convolutional neural network applied to the classification of image data and trained on a sample of N = 15725 training instances.

```
# model definition
model <- keras_model_sequential() %>%
  #
  # convolutional layers
  layer_conv_2d(filters = 64, kernel_size = c(6,6), activation = "relu",
                input_shape = c(256, 256, 3), name = "conv_1") %>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_conv_2d(filters = 128, kernel_size = c(4,4), activation = "relu",
                name = "conv_2") %>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_conv_2d(filters = 128, kernel_size = c(2,2), activation = "relu",
                name = "conv_3") %>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  #
  # fully connected layers
  layer_flatten() %>%
  layer_dense(units = 64, activation = "relu",
              kernel_regularizer = regularizer_l2(0.2),
              name = "dense_1") %>%
  layer_dense(units = 15, activation = "softmax") %>%
  #
  # compile
  compile(
    loss = "categorical_crossentropy",
    metrics = "accuracy",
    optimizer = optimizer_adam()
  )

# training and validation
fit <- model %>% fit(
  x = x_train, y = y_train,
  validation_data = list(x_val, y_val),
  epochs = 200,
  batch_size = 185,
  callbacks = callback_early_stopping(monitor = "val_loss", patience = 20)
)
```

1. Using the information from the code chunk above, extrapolate the following:

- Type and size of the images in the data.
- Depth of the convolutional layers.
- Size of the filters.
- Number of batches processed in each training epoch.
- Activation functions and output activation.
- Regularization (if any).

(10 marks)

2. Compute the following quantities, showing all steps and calculations.:

- The number of parameters of the second convolutional layer.
- The number of parameters of the first dense layer.

(30 marks)

Exercise 2 – Data analysis

Daily and sports activity recognition

Data description

The data are motion sensor measurements of 19 daily and sports activities, each performed by 8 subjects (4 female, 4 male, between the ages 20 and 30) in their own style for 5 minutes. For each subject, five sensor units are used to record the movement on the torso, arms, and legs. Each unit is constituted by 9 sensors: x-y-z accelerometers, x-y-z gyroscopes, and x-y-z magnetometers.

Sensor units are calibrated to acquire data at 25 Hz sampling frequency. The 5-minute signals are divided into 5-second signal segments, so that a total of 480 signal segments are obtained for each activity, thus $480 \times 19 = 9120$ signal segments classified into 19 classes. Each signal segment is divided into 125 sampling instants recorded using $5 \times 9 = 45$ sensors. Hence, each signal segment is represented as a 125×45 matrix, where columns contain the 125 samples of data acquired from one of the sensors of one of the units over a period of 5 seconds, and rows contain data acquired from all of the 45 sensors at a particular sampling instant.

For each signal matrix:

- Columns 1-9 correspond to the sensors in the torso unit.
- Columns 10-18 correspond to the sensors in right arm unit.
- Columns 19-27 correspond to the sensors in the left arm unit.
- Columns 28-36 correspond to the sensors in the right leg unit.
- Columns 37-45 correspond to the sensors in the left leg unit.

For each set of 9 sensors, the first three are accelerometers, the second three are gyroscopes and the last three magnetometers.

More information about the full data is at the following link:

`https://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities`.

The data file `data_assignment_2_activity_recognition.RData` contains training/validation data and test data. The object `x` includes training/validation data regarding 430 signal segments for each activity (total of 8170 signals), and associated activity-sports labels are in the object `y`. The test data are stored in the object `x_test`, which includes 50 signal segments for each activity (total of 950 signals); associated test data labels are in the object `y_test`. The input data are organized in the form of 3-dimensional arrays, in which the first dimension denotes the signal, the second the sampling instants, and the third the sensors, so each signal is described by a vector of $125 \times 45 = 5625$ features.

Task

The task is to build a predictive model to predict the type of daily/sports activity from the movement sensor data.

1. Deploy **at least 3** different deep learning systems characterized by different configurations, hyperparameters, and training settings (architecture, number of hidden units, regularization, kernel size, filter size, optimization, etc.). These deep learning systems can be of the same type, for example 3 different DNNs characterized by different architectures and settings, or of different types, for example 2 DNNs and 1 CNN with different settings. Motivate clearly the choices made in relation to the settings, configurations, and hyperparameteres used to define the different deep learning systems. Compare appropriately the deep learning systems considered, evaluating and discussing their relative merits. Select the best model a predicting the type of daily/sports activity from the movement sensor data. *(50 marks)*
2. Use the test data to evaluate the predictive performance of the best model. Comment on the ability of the model at recognizing the different activities. *(10 marks)*

Guidelines

- If you want, you can use only a subset of the sensors/features of the data in the model. However, you **must clearly motivate your choice** and why some sensors are discarded.
- If you want, you can consider aggregation of some similar activity classes. Also in this case you **must clearly motivate your choice** and why some classes are aggregated.
- You will not be evaluated on the basis the predictive performance of your models, but you would need to show that attempts have been considered to build a system with reasonable performance.

Submission rules and instructions

- Write a short and tidy report and submit it as a single pdf file (approximately max 10-12 pages, code excluded).
- Include the R code used for the data analysis in the report. The report can be produced using R Markdown (or similar tool), with the code included in the main text or as an appendix. **The code must be working and the analysis must be reproducible in all parts**.
- In general, for full marks you **must explain** concisely and clearly **all reasoning**, as well as **show all steps** and **computations** in your answers. Correct answers alone will not achieve full marks.
- For the data analysis task, you must **discuss and motivate the various decisions taken in all stages of the model building process** for full marks.
- For the data analysis task, submitting only code without any output, commentary, or discussion will not receive any marks.
- Multiple submissions before deadline are allowed and only the latest one will be considered for marking.
- **Submission after deadline will incur in penalization as UCD rules**. See "Module details" document under the "General information" tab on Brightspace.
- **Plagiarism is strictly prohibited and will result in severe penalties**. See "Module details" document and "SMS academic integrity protocol" under the "General information" tab on Brightspace to review which actions constitute plagiarism and further information.
- **Any instance of plagiarism** will result in a **zero grade** in this assessment component **for all students involved**.
- By submitting this assignment, you confirm that you have read and understood the regulations and policies regarding assessment and plagiarism outlined here, in the "Module Details" document, the "SMS Academic Integrity Protocol", and related documents. You also agree to abide by all the regulations stated in these documents.