Deep Learning KU (708.220) WS23

Assignment 2: Training Neural Networks

We will use TensorFlow and Keras to train a neural network for a regression task on the California Housing Dataset[1]. Our goal is to predict the values of the houses based on various predictive variables. Provided file `california-housing-dataset.pkl` contains the training and test datasets which you can load by executing:

```
import pickle
dict = pickle.load(open('california-housing-dataset.pkl', 'rb'))
x_train, y_train = dict['x_train'], dict['y_train']
x_test, y_test = dict['x_test'], dict['y_test']
```

There are 15,480 training and 5,160 test examples with 8 predictive input features and 1 target variable. Input features consist of the spatial locations of the districts that data is collected from (latitude, longitude), demographic information in the districts (income, population, house occupancy), and general information regarding the houses (number of rooms, number of bedrooms, age of the house). Since these statistics are measured per district, the features correspond to averages or medians across "block groups" (a geographical unit with a population of 600 to 3,000 people).

The input variables are provided in the following order:

- MedInc : median income in block group
- HouseAge : median age of a house within a block
- AveRooms : average number of rooms per household
- AveBedrms : average number of bedrooms per household
- Population : block group population
- AveOccup : average number of household members
- Latitude : a measure of how far north a house is
- Longitude : a measure of how far west a house is

The target variable is the *median house value* for California districts, expressed in hundreds of thousands of dollars ($\$100,000$).

For all the tasks below, create the appropriate code and discuss your experimentation and reasoning process, findings and choices in your report.

[1] `https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html`

Task details:

a) (3 pts) : Get familiar with the dataset. Construct a validation set consisting of samples from the training data, which will be used during the model selection process. You will use the test set only for final evaluations. Investigate the feature distributions, and normalize the data if it is necessary.

b) (8 pts) : Design your neural network architecture for the regression task. Explain your choices for the output layer and the error function that you will use. Minimize the error by using mini-batches of suitable size. Test different architectures with varying numbers of hidden units and hidden layers. Compare these choices and report training and validation set errors in a table.

c) (5 pts) : Investigate and compare different optimization procedures such as stochastic gradient descent (SGD), momentum SGD, and ADAM. Accordingly, try a number of learning rates and also try out adapting the learning rate during training by scheduling. Provide a table where training and validation set errors of various optimization hyper-parameters are compared.

d) (4 pts) : Clearly summarize your final model once your architecture choices are fixed. Provide a plot where the evolution of the training and validation set errors during training are shown throughout iterations. Perform a final training with this model on the whole training set. Report and comment on the final test error. Provide a scatter plot in which you compare model predictions with their ground truth values (on the test set).

e) (5 pts) : Now assume that we want to use a similar architecture for the binary classification problem of determining if the median house value is below or over $\$200,000$. Explain which parts of the architecture and the training pipeline you would need to change, and which test set evaluation metrics should be investigated in this case. Explain the reasons for these differences.

Implement these changes to the architecture and the training pipeline you had in part (d), and train a single model for this binary classification task using the whole training set. Note that you will also need to redefine your target variables by simply executing the following lines in the order:

```
y_train[y_train<2], y_test[y_test<2] = 0, 0
y_train[y_train>=2], y_test[y_test>=2] = 1, 1
```

Evaluate your model on the test set, report and comment on its performance.

Total: 25 points

Present your results clearly and structured. Submit your commented code (a single .py file) and a report (.pdf) at TeachCenter. Do **not** provide one zip file with code and PDF, but submit them separately. Your code should be directly executable (assuming that the provided files are present at the working directory).

Assignment details:

- *Assignment issued:* October 25th, 2023, 08:00
- *Deadline:* November 22nd, 2023, 08:00
- *Solution submission:* Upload to TeachCenter a PDF (report) and a single .py (code) file, separately as explained above (i.e., **not** in a zip file).
- *Rules:* **Groups of up to two students are allowed for this task.** In the report, indicate your group partner on top of the first page (write "Group partner: ⟨First name⟩, ⟨Last Name⟩, ⟨Matrikelnumber⟩"). It is sufficient if only one of the group members submits the group's solutions. Copying of solutions or reports from other students is strictly forbidden.