TASK 4

Open-ended (graded)

You are in the group Qwerty consisting of hongk (hongk@student.ethz.ch (mailto://[u'hongk@student.ethz.ch'])) and kshum (kshum@student.ethz.ch (mailto://[u'kshum@student.ethz.ch'])).

1. READ THE TASK DESCRIPTION

2. SUBMIT SOLUTIONS

3. HAND IN FINAL SOLUTION

1. TASK DESCRIPTION

This is the fourth and final graded task for the Introduction to Machine Learning 2018 class at ETH Zurich.

Multi-class Classification: Your goal is to predict a discrete value *y* in {0, 1, 2, ..., 9} based on a vector *x*.

Potential approaches / tools to consider: Everything taught in class.

DATA DESCRIPTION

Download handout (/+CSCO+1h75676763663A2F2F636562777270672E796E662E7267756D2E7075++/static/task4_s8n2k3nd.zip)

In the handout for this project, you will find the the following files:

- **train_labeled.h5 -** the labeled part of the training set
- **train_unlabeled.h5** - the unlabeled part of the training set
- **test.h5** - the test set (make predictions based on this file)
- **sample.csv** - a sample submission file in the correct format

The training data is contained in two files: *train_labeled.h5* and *train_unlabeled.h5,* both in hdf5 format (https://sslvpn.ethz.ch/+CSCO+0h75676763663A2F2F6A6A6A2E75717374656268632E626574++/HDF5/). Each entry in *train_labeled.h5* is one data instance indexed by an Id which consists of one integer for y and 128 doubles for the vector x1-x128. Each entry in *train_unlabeled.h5* is one data instance indexed by an Id which consists of 128 doubles for the vector x1-x128. The test set file (test.h5) has the same structure as train_labeled.h5 except that the column for y is omitted.

To load the data in Python you may use the pandas package, e.g.:

```
import pandas as pd
train_labeled = pd.read_hdf("train_labeled.h5", "train")
train_unlabeled = pd.read_hdf("train_unlabeled.h5", "train")
test = pd.read_hdf("test.h5", "test")
```

For your convenience, we further provide a sample submission file (sample.csv) which looks as follows:

```
Id,y
30000,2
30001,6
30002,4
30003,2
...
```

Note that, for each prediction, you need to include the same sample id (in the Id column) as specified in test.h5.

SUBMISSION FORMAT

For every data instance in the test set, submission files should contain two columns: *Id* and *y* where *y* should be an integer with the class Id of your prediction.

The file should contain a header and have the following format:

```
Id,y
30000,2
30001,6
30002,4
30003,2
...
```

Please keep in mind that, as a group, you have a limited number of submissions as stated on the submissions page.

EVALUATION

Submissions are evaluated by categorisation accuracy, i.e., the fraction of correct predictions. Random guessing should produce an evaluation score of around 0.1.

How to compute it in Python:

```
from sklearn.metrics import accuracy_score
acc = accuracy_score(y, y_pred)
```

GRADING

We provide you with **one test set** for which you have to compute predictions. We have partitioned this test set into two parts and use it to compute a *public* and a *private* score for each submission. You only receive feedback about your performance on the public part in the form of the public score, while the private leaderboard remains secret. The purpose of this division is to prevent overfitting to the public score. You are motivated to make sure your models will generalize well even to the private part of the test set.

When handing in the task, you need to select which of your submissions will get graded and provide a short description of your approach. We will then compare your selected submission to three baselines (easy, medium and hard). Your final grade depends on both the public score and the private score (weighted equally) and on a properly-written description of your approach. The following **non-binding** guidance provides you with an idea on what is expected to obtain a certain grade: If you hand in a properly-written description and your handed-in submission performs better than the easy baseline, you may expect a grade exceeding a 4. If it further beats the medium baseline, you may expect that the grade will exceed a 5. If in addition your submission performs equal to or better than the hard baseline, you may expect a 6. If you do not hand in a properly-written description of your approach, you may obtain zero points regardless of how well your submission performs.

⚠ Make sure that you properly hand in the task, otherwise you may obtain zero points for this task.

FREQUENTLY ASKED QUESTIONS

WHICH PROGRAMMING LANGUAGE AM I SUPPOSED TO USE? WHAT TOOLS AM I ALLOWED TO USE?

You are free to choose any programming language and use any software library.

CAN YOU HELP ME SOLVE THE TASK? CAN YOU GIVE ME A HINT?

As the tasks are a graded part of the class, **we cannot help you solve them**. However, feel free to ask general questions about the course material during or after the exercise sessions.

CAN YOU GIVE ME A DEADLINE EXTENSION?

⚠ We do not grant any deadline extensions!

CAN I POST ON PIAZZA AS SOON AS HAVE A QUESTION?

This is highly discouraged. Instead,

- Read the details of the task thoroughly.
- Review the frequently asked questions.
- If there is another team that solved the task, spend more time thinking.
- Discuss it with your team-mates.

If you still consider that you should contact the TAs, you can post a **private** question on Piazza. Remember that collaboration with other teams is prohibited.

WHEN WILL I RECEIVE THE PRIVATE SCORES? AND THE PROJECT GRADES?

Before the exam, you will obtain an email with all your private scores. We do not release scores before that. The same email will also contain your project grade.