DATA.ML.200 Pattern Recognition and Machine Learning

Exercise Set 1: Level test (no exercise sessions)

1. *Install Python.*

   You should know how to do that. You may check the course Moodle page for quick instructions for Anaconda installation. You may use your favourite IDE or code editor (PyCharm, VS Code, Spyder, Emacs etc), but make sure you return exercise answers in the requested format.

2. *TeenMagi-2022 Kaggle competition.*

   This competition is our official course competition for which course points are assigned. You should register as a Kaggle user, for instructions see:

   - `www.kaggle.com`

   After registration go to TeenMagi competition page at

   - `https://www.kaggle.com/c/teenmagi-2022/`

   read the instructions and download the data files.

3. **`python`** *Scikit nearest neighbor classifier. (10 pts)*

   The main Python library for machine learning oriented coding is the community project Scikit-learn. See its Web page for description and documentation

   - `https://scikit-learn.org`

   In this exercise we will use Scikit-learn implementation of the *Nearest Neighbor* classification rule. You implemented the same functions yourself if you participated the introduction course (DATA.ML.100), but now you can use ready implementation. Read the NN manual page and see the code examples

   - `https://scikit-learn.org/stable/modules/neighbors.html#classification`

   In principle, all functionality is embedded into three standard sklean function calls

   ```
   clf = neighbors.KNeighborsClassifier(n_neighbors=1,
      algorithm='kd_tree')
   clf.fit(x_tr, y_tr)
   y_pred = clf.predict(x_val)
   ```

   for which you must prepare the competition data. For example, x_tr must be a row matrix where each 64-dimensional row is one training sample and y_tr contains the corresponding class label.

   You should make a code that classifies all validation samples and stores the results in a CSV file expeted by the Kaggle submission system:

```
Id,Class
1,522
2,330
3,6
...
...
48234,468
48235,816
48236,806
48237,522
48238,683
```

Find out suitable parameters for the NN classifier and submit your final results to the Kaggle competition.

Note: The sample images are $8 \times 8 \times 3$, but since they are gray level the three color channels are all same. Therefore you should use them as $8 \times 8 \times 1$, i.e. when you convert them to vectors the vector length should be 64 values per sample.

Note: NN can be super slow, so better to start by using only a small portion of training samples!