Assignment 2

April 9, 2021

Question 1: 10 points

Use KNeighborsClassifier to build a classifier for the MNIST dataset that achieves over 97% accuracy on the test set. You might need to use a grid search to find good hyperparameters.

Question 2: 30 points

- Write a function that can shift a MNIST image in any direction (left, right, up, or down) by one pixel.
- Then, for each image in the training set, create four shifted copies (one per direction) and add them to the training set.
- Finally, train your best model on this expanded training set and measure its accuracy on the test set.
- You should observe that your model performs even better now!
- This technique of artificially growing the training set is called **data augmentation** or **training set expansion**.

Question 3: 40

- If you do not know Linux, study my video lessons, if you do not have an invite, let me know ASAP.
- Study `midway.pdf` from previous lesson.
  - Run `l2_n2.ipynb` on midway2 compute node. You might have to modify it to download data from the login node.
  - Show what part of code you ran on login node and what you changed in the notebook to run on compute node.
  - Provide screenshots proving that you did it.

- Study `skyway.pdf` and run the same notebook on Skyway
  - Using jupyter interactively. Provide screenshots.
  - Submitting a job in batch: convert notebook to python, instead of plotting pictures, save them. Provide a batch file and the output files.
  - To submit Q3 use Linux' `tar zcvf` command to pack all the relevant files into a single compressed file with `.tar.gz` extension.

Requirements

These are the general rules for submitting the homework not only for this assignment.

1. Submit jupyter notebook(s) in ipynb and html formats: html format allows the grader to easily read it in canvas while ipynb allows to execute it if something is not clear from html.
2. The notebook should be well formatted:
   - Use markdown to break it into sections, bullet points, etc.
     - See `l1_n3.ipynb` for example how to use it.
     - If you double click on markdown cell, you can see the code.
     - To show the formatted text, execute the corresponding markdown cell.
     - Start the notebook with a markdown title cell of the form: '# Assignment X, Jane Doe, date'.
     - When answering a particular question from the assignment, start the answer with markdown cell '## Question Y' that indicates what question you are answering.
3. Name your notebook (and the corresponding html file) according to the following template: `Assignment_X_Jane_Doe.ipynb`
4. If you are submitting several notebooks, append to the name `'_partZ'` and explain in canvas what each notebook is about.
5. There should not be any failed cells in the notebook. Every cell should work and have the results of the execution.
6. The notebook should contain only what is needed to answer a particular question and nothing else. More is as bad as less.
7. All the plots, if needed for the assignment, should be of production quality with readable labels, titles, etc. Also, try to find the most approriate type of plot to clearly demonstrate your statement.
8. Due date: 04/16/2021