COM4509/6509 MLAI - Assignment Part 2 Brief

Deadline: Friday, December 3, 2021 at 15:00 hrs

Please READ the whole assignment first, before starting to work on it.

Scope: Sessions 6 to 8; Number of marks available for Part 2: 25

How and what to submit

1. A **Jupyter Notebook** with the code in all the cells executed, outputs displayed, and code documented. ***We expect to see all the answers in the notebook. Yes, you need to create one.***

2. Name your Notebook as **COM4509-6509_Assignment_Part2_Username_XXXXXX.ipynb** where XXXXXX is your username such as abc18de.

3. Upload a .zip file to Blackboard before the deadline that contains two Jupyter Notebooks, **one for Part 1 and one for Part 2** (COM4509-6509_Assignment_Part1_Username_XXXXXX.ipynb and COM4509-6509_Assignment_Part2_Username_XXXXXX.ipynb)

4. **NO DATA UPLOAD**: Please do not upload the data files used in this Notebook. We have a copy of the **data** already. Instead, please use a relative file path in your code (data files under folder ‘data’), as in the lab notebook so that we can run your code smoothly when needed. So ‘./data/’, instead of ‘/User/username/myfiles/mlai/assignment1/’

Assessment Criteria

1) Being able to build complete, reproducible machine learning pipelines from loading data to evaluating prediction performance.
2) Being able to design different machine learning models to compare/optimise prediction performance.
3) Being able to perform exploratory data analysis to gain insights.

Late submissions

We follow Department's guidelines about late submissions, i.e., a deduction of 5% of the mark each working day the work is late after the deadline. NO late submission will be marked one week after the deadline because we will release a solution by then. Please read this link if you are taking COM4509 or read this link if you are taking COM6509.

Use of unfair means

"Any form of unfair means is treated as a serious academic offence and action may be taken under the Discipline Regulations." (from the students Handbook). Please carefully read this link on what constitutes Unfair Means if not sure, for COM4509. If you are taking COM6509, please read this link if you are not sure what is Unfair means. If you still have questions, please ask your Personal tutor or the Lecturers.

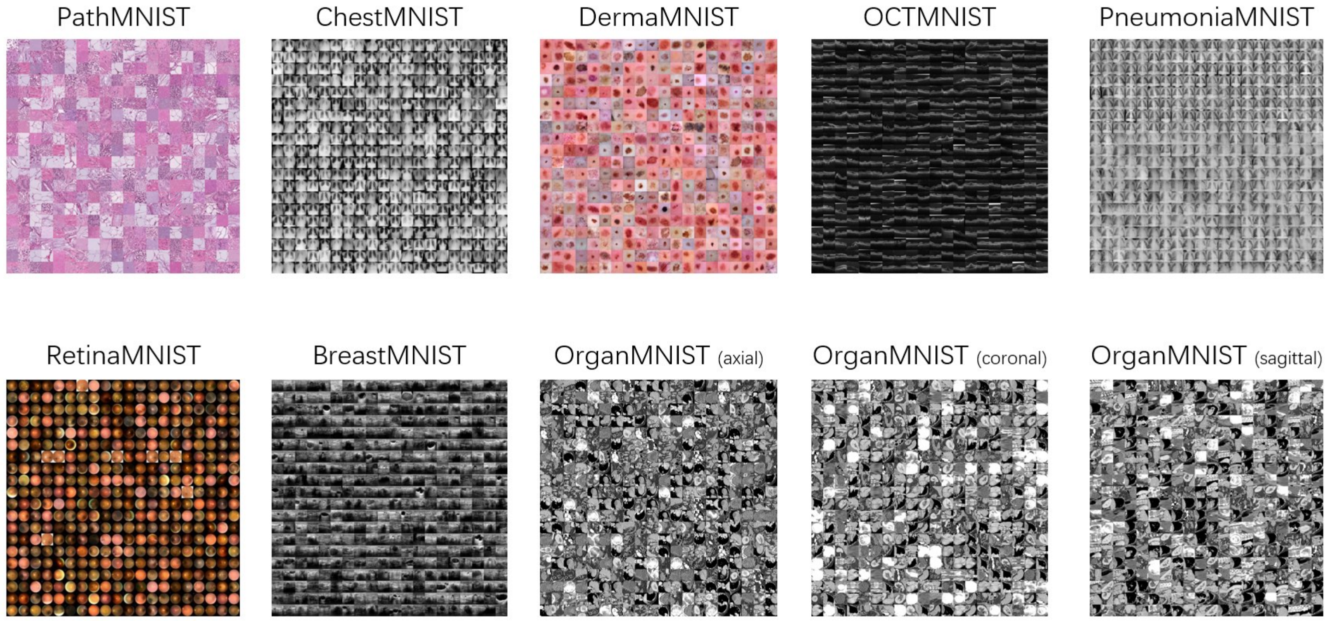
A. Reproducibility & readability

Whenever there is randomness in the computation, you MUST set a random seed for reproducibility. Use your UCard number XXXXXXXXX (or the digits in your registration number if you do not have one) as the random seed throughout this assignment.

Answers for each question should be clearly indicated in your notebook, e.g., including question numbers below in bold such as **B2.1a**. All code should be clearly documented and explained.

Note: You will make several design choices (e.g. hyperparameters) in this assignment. There are no “standard answers”. You are encouraged to explore several design choices to settle down with good/best ones, if time permits.

B. Logistic Regression on BreastMNIST [9 marks]



The above shows the **first version** of the MedMNIST, published in ISBI21 (2010.14925.pdf (arxiv.org)). As taken from the paper: BreastMNIST is based on a dataset of 780 breast ultrasound images. It is categorized into 3 classes: normal, benign and malignant originally but in BreastMNIST, the task is simplified into binary classification by combining normal and benign as positive, and classifying them against malignant as negative. The source dataset with a ratio of 7 : 1 : 2 into training, validation and test set. The source images of 1 × 500 × 500 are resized into 1 × 28 × 28.

We aim to train a **L2-regularised** logistic regression model to classify the two classes in BreastMNIST using the standard train/validation/test split with decent performance, i.e. much better than the chance level at worst.

B1 Data loading and inspection [3 mark]

Follow instructions at https://github.com/MedMNIST/MedMNIST to download and load the data. Display at least ten images for each class, i.e. at least 20 images, from the training set. Display at least ten images for each class from the validation set, and display at least ten images for each class from the test set.

B2 Logistic regression [4 marks]

Keep a record of the three metrics M1 to M3 below for the two models below:

M1) Training accuracy: the prediction accuracy of a trained model on the training dataset.

M2) Validation accuracy: the prediction accuracy of a trained model on the validation dataset.

M3) Testing accuracy: the prediction accuracy of a trained model on the test dataset.

a) Using the built-in logistic regression functions in scikit-learn, train a logistic regression model with **L2 regularisation** on the training set, use the validation set to choose a good regularisation parameter (a hyperparameter) from at least three choices, and test the chosen model on the test set. Report the three metrics M1 to M3 [2 marks]

b) Using **PyTorch** (see Question 5 of Lab 6), train a logistic regression model with L2 regularisation on the training set, use the validation set to choose a good regularisation parameter (a hyperparameter) from at least three choices, and test the chosen model on the test set. Report the three metrics M1 to M3 [2 marks]

B3 Performance comparison (2 marks)

a) Summarise each of the three metrics from the two models in B2 using one or more bar graphs. [1 mark]

b) Describe at least two observations interesting to you. [1 mark]

C. Convolutional Neural Networks on OCTMNIST [8 marks]

OCTMNIST is based on a prior dataset of 109,309 valid optical coherence tomography (OCT) images for retinal diseases, with 4 different types, leading to a multi-class classification task. The source training set is split with a ratio of 9 : 1 into training and validation sets, and uses its source validation set as the test set. The source images are single channel, and their sizes are (384−1, 536)×(277−512), which are center-cropped and resized to 1 × 28 × 28.

C1 Data loading and inspection [2 mark]

Follow instructions at https://github.com/MedMNIST/MedMNIST to download and load the data. Display at least ten images for each class, i.e. at least 40 images, from the training set.

C2 Convolutional neural networks [4 marks]

Keep a record of the four metrics M1 to M4 below for the two models below:

M1) Training accuracy: the prediction accuracy of a trained model on the training dataset.

M2) Validation accuracy: the prediction accuracy of a trained model on the validation dataset.

M3) Testing accuracy: the prediction accuracy of a trained model on the test dataset.

M4) Training time: the time taken to train the model (i.e. to learn/estimate the learnable parameters) on the training dataset.

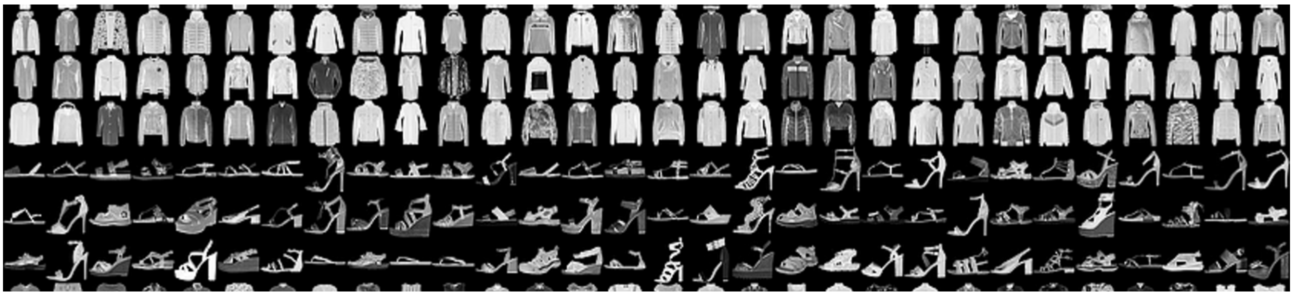
This question asks you to design convolutional neural networks (CNNs). Only the number of convolutional (Conv) layers and the number of fully connected (FC) layers will be specified below. You are free to design other aspects of the network. For example, you can use other types of operation (e.g. padding), layers (e.g. pooling, or preprocessing (e.g. augmentation), and you choose the number of units/neurons in each layer. Likewise, you may choose the number of epochs and many other settings according to your accessible computational power.

a) Design a CNN with two Conv layers and two FC layers. Train the model on the training set, use the validation set to choose the best design among at least three different choices, and test the chosen model on the test set. Report the four metrics M1 to M4 [2 marks]
b) Design a CNN with three Conv layers and three FC layers. Train the model on the training set, use the validation set to choose the best design among at least three different choices, and test the chosen model on the test set. Report the four metrics M1 to M4 [2 marks]

C3 Performance comparison (2 marks)

c) Summarise each of the four metrics from the two models in B2 using one or more bar graphs. [1 mark]
d) Describe at least two observations interesting to you. [1 mark]

D. Unsupervised learning on Fashion-MNIST [8 marks]



Fashion-MNIST is a dataset of Zalando's article images, with examples shown above. It consists of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes: *0=T-shirt/top; 1=Trouser; 2=Pullover; 3=Dress; 4=Coat; 5=Sandal; 6=Shirt; 7=Sneaker; 8=Bag; 9=Ankle boot*.

Choose any two out of the 10 classes and use only the **test data for these two chosen classes** to complete tasks in this section. It will be better to finish reading the remaining part of this section before choosing the two classes. Again, you may choose any two and there is no “correct” answer about which two to choose but some choices may make your studies below more interesting than others.

Use the PyTorch API for Fashion-MNIST to load the training/test data of Fashion-MNIST. You may refer to similar procedures in Lab 7 for CIFAR-10.

D1. Dimensionality reduction and clustering [7 marks]

a) Apply PCA to all images of these two chosen classes. Visualise the top 5 eigenvectors as images and display them in the order of descending corresponding values (the one corresponding to the largest eigenvalue first). [1 marks]
b) Use the top 30 PCs to reconstruct 10 images, with 5 from each class (any 5 images are fine from each class). Show these 10 **pairs** of reconstructed and original images. [1 marks]

c) Visualise the two-dimensional PCA representations of all data points in a 2D plane (i.e. using the top two PCs). Use different colours/markers for the two classes for better visualisation (*Hint: You need to use the class labels here for visualisation*). [1 marks]

d) Use *spectral clustering* to cluster all data points as represented by the top two PCs (clustering of two-dimensional vectors, where each vector has two values, PC1 and PC2). Visualise the two clusters with different colours/markers in 2D. [2 marks].

e) Design a new autoencoder with five Conv2d layers and five ConvTranspose2d layers. You are free to choose the activation functions and settings such as stride and padding. Train this new autoencoder on all images of these two chosen classes for at least 20 epochs. Plot the loss against the epoch. [2 marks]

D2 Observation [1 marks]

Describe at least two observations interesting to you from D1 above.

The END of Assignment

FAQs