Homework3 Question 1

17/11/2023

59.5/60 Points

Attempt 2

Review Feedback
17/11/2023

Attempt 2 Score:
59.5/60

Add Comment

Anonymous Grading: **no**

Unlimited Attempts Allowed

21/10/2023

Details

This assignment is designed to give you practical programming experience with dimensionality reduction, unsupervised learning and neural networks. Please carefully read all the instructions below. Do not hesitate to use Slack and Inscribe Q&A community (right-hand side menu on Canvas) to ask questions.

You can complete this assignment individually or in a group (up to 4 members). To select your group (even if you work individually) go to *People* section in Canvas and select *Groups*. There, you can join the desired group *(please do not create your own groups, but join one of the groups that we created)*. Please communicate with your colleagues before forming the groups.

Your assignment should be submitted by uploading your code (in the form of a **Jupyter Notebook (.ipynb) AND pdf copy of the files –** so we can make comments directly on the file) to Canvas.

Be sure to run the file before committing so that we can directly see your results (please double-check the file you uploaded to make sure it is the right file and that all the results are visible). Please mention all the resources that were used to solve the problem (e.g., websites, books, research papers, other people, etc.). To complete the assignment, you can use any Python (or R) package that you want, but we recommend using Scikit-Learn and Tensorflow or Pytorch.

Question

In this problem, you will use Rock dataset located here: https://osf.io/d6b9y/ (https://osf.io/d6b9y/) .

Extra tips:

- If you are using Colab, make sure to select a GPU (even if you're not using a GPU via Edit->Notebook settings, it will give you more memory; no need for the paid version of Colab).
- You can convert images to grayscale from RGB and complete the assignment using only grayscale images.
- Feel free to downsize images if you are facing issues with the memory/running time

<
(https://iu.instructure.com/courses/2165858/modules/items/30965979)

N
Attempt
(https://iu.instructure.com/c

(https://osf.io/cvwu9/wiki/Data%20File%20Descriptions/)

Answer the questions below directly in your Jupyter Notebook. Please answer questions in order and use Markdown cells to clearly indicate which question you are answering.

1. Apply PCA to the images from folder '360 Rocks'. How many components do you need to preserve 95% of the variance? **[3 points]**
2. Plot 10 images of your choice in the original form (without PCA) and then plot their reconstruction (projection in the original space) after you kept 95% of variance using PCA. **[3 points]**
3. Each of the images belongs to one of three rock categories. The category is indicated by the first letter in the filename (I, M and S). We will now try to see if the visualization can help us identify different clusters.
   A. Use PCA to reduce dimensionality to only 2 dimensions. How much of the variance is explained with the first two principal components? **[2 points].**
   B. Plot a 2D scatter plot of the images spanned by the first two principal components. Each image will be represented with a dot. Make the color of the dot correspond to the image category (so you will have three different colors). Then add some rock images to the visualization to better understand what features in the images are accounting for the majority of variance in the data (your visualization should look similar to the one after line 71 in this file **https://github.com/ageron/handson-ml3/blob/main/08_dimensionality_reduction.ipynb** (https://github.com/ageron/handson-ml3/blob/main/08_dimensionality_reduction.ipynb) but with images of rocks instead of MNIST digits). Repeat the process and create the same type of plots for t-SNE, LLE and MDS. **[6 points]**
   C. Which of the visualizations do you prefer?**[1 point]**
4. Now let's see if these dimensionality reduction techniques can give us similar features to those that humans use to judge the images. File mds_360.txt contains 8 features for each of the images (rankings are in the same order as the images in '360 Rocks' folder. Run PCA, t-SNE, LLE and MDS to reduce the dimensionality of the images to 8. Then, compare those image embeddings with the ones from humans that are in the mds_360.txt file. Use Procrustes analysis to do the comparison (here is one example of how to do that `mtx1, mtx2, disparity = procrustes(matrix_with_human_data, matrix_with_pca_embeddings_data)`. Here `matrix_with_human_data` and `matrix_with_pca_embeddings_data` should be 360 by 8. `disparity` will tell you the difference in the data. Report `disparity` for each of the four dimensionality reduction methods. Compute the correlation coefficient between each dimension of `mtx1` and `mtx2` for each of the four methods - display results in a table. **[7 points]**
5. Cluster the 360 images using K-Means.
   A. To speed up the algorithm, use PCA to reduce the dimensionality of the dataset to two. Determine the number of clusters using one of the techniques we discussed in class. **[4 points]**
   B. Visualize the clusters in a similar way to the visualization after line 28 here:

&lt;
(https://iu.instructure.com/courses/2165858/modules/items/30965979)
(https://iu.instructure.com/c

color each dot based on the clusters it belongs to using the labels taken from the filename as in question 3 (I, M and S). **[4 points]**

6. Cluster the 360 images using EM.
    - A. Same as in the previous question, to speed up the algorithm, use PCA to reduce the dimensionality of the dataset to two. Determine the number of clusters using one of the techniques we discussed in class. **[4 points]**
    - B. Visualize the clusters in a similar way to the visualization after line 28 here: **https://github.com/ageron/handson-ml3/blob/main/09_unsupervised_learning.ipynb** **(https://github.com/ageron/handson-ml3/blob/main/09_unsupervised_learning.ipynb)** , but color each dot based on the clusters it belongs to using the labels taken from the filename as in question 3 (I, M and S). **[4 points]**
    - C. Use the model to generate 20 new rocks (using the sample() method), and visualize them in the original image space (since you used PCA, you will need to use its inverse_transform() method). **[4 points]**
7. Build a feedforward neural network (using dense and/or CNN layers) with a few hidden layers (we suggest using Keras (within Tensorflow) or Pytorch). Train the network to classify on 360 rock images using rock name as the label - the category is indicated by the first letter in the filename (I, M and S). Use images from '120 Rocks' folder as your validation data. Choose the number of neurons you find appropriate and efficient (so you have enough time to run it), but make the last layer before the softmax should consist of 8 neurons. The hidden layers should have ReLU activation function. Train the network for multiple epochs until it converges (if the process is too slow, tweak the learning rate and consider simplifying your network). We will not deduct points based on the simplicity of your network, but we expect you to have performance that is above chance performance that could be obtained with an untrained network - in other words, we expect to see train and validation loss decrease and accuracy increase throughout the training. We recommend using Colab (the free version should be totally fine), but make sure to run it with a GPU to speed up the training - to add a GPU on Colab go to Edit->Notebook settings).
    - A. Report the training time (use code to do this). **[1 point]**
    - B. Plot training and validation loss and accuracy as a function of training epochs. **[13 points]**
    - C. How many parameters does the network have? How many of those parameters are bias parameters? **[1 points]**
    - D. Compare the activity of neurons in the next to the last layer (the one with 8 neurons) with the human data. (to get human data use mds_360.txt and mds_120.txt files). Similar to before, use Procrustes analysis to do the comparison. For training and validation data (separately), report `disparity` and compute the correlation coefficient between each dimension of `mtx1` and `mtx2`. Display results in a table. **[3 points]**

View Rubric

S

<

(https://iu.instructure.com/courses/2165858/modules/items/30965979)

(https://iu.instructure.com/c

HW3 Rubric

Criteria	Ratings		Pts
Question -1	**3 to >0 pts** **Full Marks** PCA (2) Number of Component required to preserve 95 % variance (1) ▲	**0 pts** **No Marks**	3 / 3 pts
Question- 2	**3 to >0 pts** **Full Marks** 10 images in original from (1.5) Their reconstruction (1.5) ▲	**0 pts** **No Marks**	3 / 3 pts
Question- 3 A	**2 to >0 pts** **Full Marks** PCA with 2 dimensions (1) Amount of Variance preserved with these 2 components (1) ▲	**0 pts** **No Marks**	2 / 2 pts
Question-3 B	**6 pts** **Full Marks** Scatter Plots of Components with some Rock Images a) t-SNE (2) b) LLE (2) c) MDS (2) ▲	**0 pts** **No Marks**	6 / 6 pts
Question 3-C	**1 to >0 pts** **Full Marks** Discussion on the visualizations (preferred or not) (1) ▲	**0 pts** **No Marks**	1 / 1 pts
Question-4	**7 to >0 pts** **Full Marks** PCA (1.75), t-SNE (1.75), LLE	**0 pts** **No Marks**	7 / 7 pts

<

(https://iu.instructure.com/courses/2165858/modules/items/30965979) (https://iu.instructure.com/c

HW3 Rubric

Criteria	Ratings		Pts
	has been reduced to 8 (0.25) Disparity has been reported after Procrustes Analysis (1) Correlation Table has been Reported. (0.5) ▲		
Question 5 A view longer description	**4 to >0 pts** **Full Marks** K-means is implemented correctly with the PCA reduced data set (2) Selection of the number of clusters using one of the techniques (2) ▲	**0 pts** **No Marks**	4 / 4 pts
Question -5 B	**4 to >0 pts** **Full Marks** Visualization: Boundaries can be inferred with centroids of each cluster (3), Dots are color mapped according to labels (1) ▲	**0 pts** **No Marks**	4 / 4 pts
Question 6 A view longer description	**4 to >0 pts** **Full Marks** EM has been implemented correctly (2), Evaluation for the number of clusters (2) ▲	**0 pts** **No Marks**	4 / 4 pts
Question 6 B	**4 to >0 pts** **Full Marks** Visualization: Boundaries can be inferred with centroids of each cluster (3), Dots are color mapped according to labels (1) ▲	**0 pts** **No Marks**	4 / 4 pts

<

(https://iu.instructure.com/courses/2165858/modules/items/30965979) (https://iu.instructure.com/c

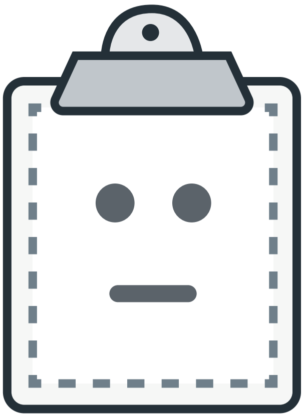
HW3 Rubric

Criteria	Ratings		Pts
	Full Marks Generation of 20 rock images using sample method with visualization - (4) ▲	**No Marks**	
Question 7 A view longer description	**1 to >0 pts** **Full Marks** Training Time has been Reported ▲	**0 pts** **No Marks**	1 / 1 pts
Question 7 B	**13 to >0 pts** **Full Marks** Sequential Model has been implemented correctly with right number of neurons -5 Validation Data has been incorporated -1 Accuracy is increasing with epochs -2 Plots of val and training loss via training epochs -5 ▲	**0 pts** **No Marks**	13 / 13 pts
Question 7 C	**1 to >0 pts** **Full Marks** Total Number of parameters (0.75) Number of Bias Parameters (0.25) ▲	**0 pts** **No Marks**	1 / 1 pts
Question 7 D	**3 to >0 pts** **Full Marks** Disparity and Correlation tables have been computed. For training data (1.5) For Validation Data (1.5) ▲	**0 pts** **No Marks**	2.5 / 3 pts
			Total Points: 59.5

<

(https://iu.instructure.com/courses/2165858/modules/items/30965979) (https://iu.instructure.com/c

File Name	Size	
AML_Assig...nal.ipynb	6.22 MB	✓
AML_Assig...final.pdf	4.16 MB	✓



Preview Unavailable

AML_Assignment3_Q1_final.ipynb

Download (https://iu.instructure.com/files/165347260/download?download_frd=1&verifier=sETaew28zaMPssoO5TpWGP7bXB0nBQhXTRBA6cTH)

< (https://iu.instructure.com/courses/2165858/modules/items/30965979) (https://iu.instructure.com/c