Dr. Eick

COSC 6335*“Data Mining”*

ProblemSet3 Fall 2022

Deep Learning and Dimensionality Reduction

Deadlines; Nov. 22, 2022

Last updated: Nov 10 at 9.30pm

Task 4: Variational Auto-encoders and Dimensionality Reduction

Individual Task

Fourth Draft

**Learning Objectives**:

1. Learn to use deep learning models such as VAE
2. Learn to use dimensionality reduction
3. Learn differn tools to create different deep learning models
4. Learning how to interpret quality of models

**Datasets**: **are available in Teams/Files/Dataset folder**

In the project we will use the MNIST computer vision digit dataset. You can use following blog to see how to download MNIST dataset (2) in python. Reference 1 contains the link of the dataset:

**Task 4 Subtasks:**

1. Use VAE auto-encoder model to reduce the dimensionality of the digit dataset. You can use the model given in reference [2]. You need to optimize the model such a way that you will get minimum loss. During the optimization procudure consider the layers are fixed, and the number of nodes in each layer are the only learnable parameters. Now develop a function that will find the set of paramters θ for which the loss is minimum.

Optimal\_model(θ) = arg min L(D)

Where, 200 <=h dimension<=600 and 10 <= z dimnesion <= 40

Here, D represents your dataset and L is the loss function.

[You can find simple VAE architecture in this reference [2]. A suggestion of dimensionality reduction tutorial using auto-encoder is available here [3] . A detaild lecture on the topic available in referece [4]]

1. Use the decoder part of the VAE and use it gernerate character images. Include some examples of digits you created and and describe how you created them in your report.

[Sample example available in reference 2]

1. Develop a classification models to recognize the digits in the dataset using the CNN classifier. Verify the model using K-fold cross validation.

[You can use tutorial 3 to see how to build CNN model to recognize digits.]

* 1. For subtask 4:
		1. Architecture of the classifier model
		2. A plot that will show the compartive accuracies
		3. Your interpreation of the results
1. Complete any one from the below 2. You can get upto 10 additional points for doing both.
	1. Develop two classifiers using any model such as SVM, Tree or neural network models; Train one model using the main dataset and second model on top of the extracted features obtained from the VAE encoder.
2. Comprate both models and sssess the benefits of using demensiality reduction for the dataset at hand based on the accuracy you get; discuss the issues of using dimensionality reduction for this dataset.
	1. As an alternative approach, namely PCA, to reduce dimentionality of your dataset and use them as input to your second classifier model (same model used on top of using the dimensionality reduced features from VAE) and compare the results against that second model. [Reduced dimensionality need to be same for all methods]

[Reference 5 contains tutorials ]

**References:**

1. [**https://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits**](https://archive.ics.uci.edu/ml/datasets/optical%2Brecognition%2Bof%2Bhandwritten%2Bdigits)
2. [**https://github.com/dataflowr/notebooks/blob/master/HW3/VAE\_clustering\_empty.ipynb**](https://github.com/dataflowr/notebooks/blob/master/HW3/VAE_clustering_empty.ipynb)
3. [**https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-from-scratch-for-mnist-handwritten-digit-classification/**](https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-from-scratch-for-mnist-handwritten-digit-classification/)
4. [**https://dataflowr.github.io/website/modules/6-convolutional-neural-network/**](https://dataflowr.github.io/website/modules/6-convolutional-neural-network/)
5. [**https://www.analyticsvidhya.com/blog/2021/06/dimensionality-reduction-using-autoencoders-in-python/**](https://www.analyticsvidhya.com/blog/2021/06/dimensionality-reduction-using-autoencoders-in-python/)
6. [**https://www.kaggle.com/code/arthurtok/interactive-intro-to-dimensionality-reduction/notebook**](https://www.kaggle.com/code/arthurtok/interactive-intro-to-dimensionality-reduction/notebook)

**Deliverables for Task 4:**

1. A report
2. Code and Readme
3. The report must contains following iteams:
	1. For subtask 1:
		1. Properly written algorithm with explanantion for the algorithm used to find the proper architecture
		2. Different loss updates visualized for different parameter settings by the architecture
		3. Optimal architecture
		4. Loss diagram for the optimal model
	2. For subtask 2:
		1. An algorithm discussing how the outputs are generated from the decoder
		2. Sample inputs and outputs
	3. For subtask 3:
		1. Model Architecture
		2. Train-Test-Validation graph
		3. A discussion that will explain the outputs (Show overfitting/ underfitting)
	4. For subtask 4.a:
		1. Discussion of the classifier model
		2. A plot that will show the compartive accuracies
		3. Your interpreation of the results
	5. For subtask 4.b:
		1. Discussion about the PCA model
		2. A plot that will show the compartive accuracies
		3. Your interpreation of the results