BLG 454E

Homework 3

December 21, 2022

Due date is **8th January 2023, 23:59!**

kamard@itu.edu.tr
korkmazmer@itu.edu.tr

Homework Policy

- Use comments whenever necessary to explain your code.
- You will use Python as the programming language. However, you are NOT allowed to use built-in functions from Python libraries. Following are libraries that are NOT allowed:
  - scipy, Bio, Theano, Tensorflow, Keras, PyTorch and any other similar machine learning modules.

  Following are libraries that are allowed:
  - pandas, numpy, math, random, matplotlib, csv, etc... Basically anything that is not related to machine learning. When you are in doubt, simply ask.
- IMPORTANT: This is an individual assignment! You are expected to act according to Student Code of Conduct, which forbids all ways of cheating and plagiarism. It is okay to discuss the homework with others, but it is strictly forbidden to use all or parts of code from other students' codes or online sources and let others do all or part of your homework.
- Only electronic submissions through Ninova will be accepted. You only need to submit your Jupyter Notebook file. Any comments and discussions should be included in the same file.
- If you have any questions, you can contact us. Do NOT hesitate to send an e-mail if you are confused.

For easy reference, a mathematical notation summary table is also attached in the last page of this homework.

Jupyter Notebook Installation

You need to have Python and Pip installed in your computer to install Jupyter Notebook.

Windows

On command prompt (cmd.exe with admin mode):

```
C:\**path**> python –m pip install jupyter
```

After changing your current folder to the folder which you want to work on (see ’cd’ command):

```
C:\**your_working_folder**> jupyter notebook
```

Then it will be launched on your default browser

Ubuntu/Linux/Unix/Mac

On Terminal:

```
$ pip install notebook
```

Then launch with:

```
$ jupyter notebook
```

For more information: https://jupyter.org/install

Problem Description

One application of neural networks is autoencoders, in which we can learn different representations of data. As illustrated in Figure 1, the autoencoder takes some input, encodes it and creates a hidden space representation from it, then decodes it and tries to generate the input data from the latent space representation. If the dimension of the hidden space representation is lower than the dimension of the input, the autoencoder performs dimensionality reduction on the data. The output on the other hand, always has the same dimensions as the input.

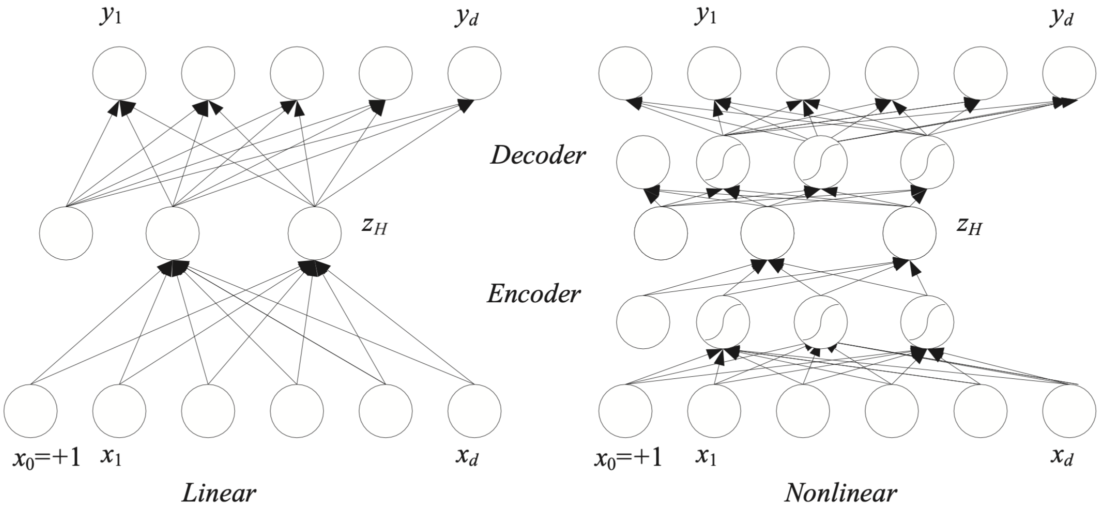

Figure 11.19 In the autoencoder, there are as many outputs as there are inputs and the desired outputs are the inputs. When the number of hidden units is less than the number of inputs, the MLP is trained to find the best coding of the inputs on the hidden units, performing dimensionality reduction. On the left, the first layer acts as an encoder and the second layer acts as the decoder. On the right, if the encoder and decoder are multilayer perceptrons with sigmoid hidden units, the network performs nonlinear dimensionality reduction.

Figure 1: The architecture of an autoencoder.

Thus, the job of an autoencoder is to represent the data in the hidden layer, then use that representation to regenerate the input data. This is why the MLP is pushed to discover the most accurate representation of the data in the latent space[1].

[1] For further information, please check out Chapter 11 of Introduction to Machine Learning (Third Edition) by Ethem Alpaydin.

Homework 3

Task Description

Use the optdigit dataset[2] and train a MLP with 64 inputs, two hidden units, and 64 outputs. Reduce the dimensionality of the dataset using the outputs of hidden units. The output of the hidden units for each example will be your new feature vector for that example. Plot the reduced dataset in two dimensional space (Hidden1 vs Hidden2) as given in Figure 2 taken from Alpaydin's book. Please note that the output of the autoencoder should be identical to the inputs. Conduct the MLP training using the train set of the optdigits dataset.

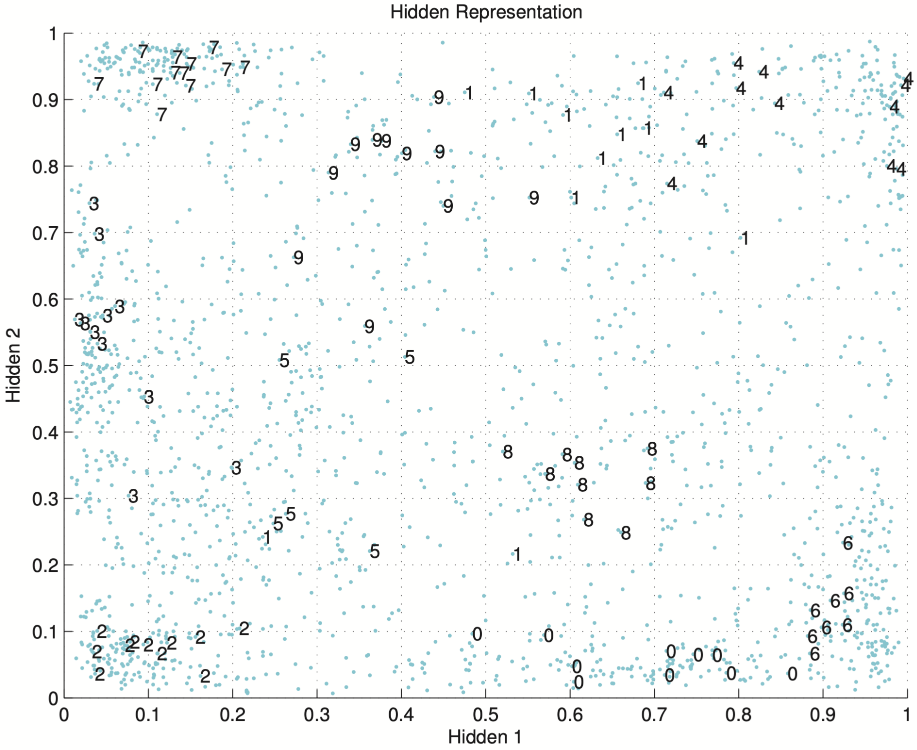

Figure 11.18 Optdigits data plotted in the space of the two hidden units of an MLP trained for classification. Only the labels of one hundred data points are shown. This MLP with sixty-four inputs, two hidden units, and ten outputs has 80 percent accuracy. Because of the sigmoid, hidden unit values are between 0 and 1 and classes are clustered around the corners. This plot can be compared with the plots in chapter 6, which are drawn using other dimensionality reduction methods on the same dataset.

Figure 2

[2] For more information on the dataset, please visit the Optdigits dataset website.

Dataset

You are provided 3 files for this assignment:

- **digits.tra**: This file contains the set of training data. It consists of 3823 instances. There are a total of 64 attributes per instance (and one label). The attributes take values in the range 0,...,16 and the labels take values in the range 0,...,9. **Recall that autoencoders are unsupervised and you will not be needing the labels!**
- **digits.tes**: This file contains 1797 instances of test data.
- **digits.names**: This file contains detailed descriptions of the dataset.

You can use the following code snippet to load the data. Note that the last column contains labels.

```
#Loading the Data
training_df = pd.read_csv(’data/optdigits.tra’,header=None)
X_training, y_training = training_df.loc[:,0:63], training_df.loc[:,64]

testing_df = pd.read_csv(’data/optdigits.tes’,header=None)
X_testing,  y_testing  = testing_df.loc[:,0:63],  testing_df.loc[:,64]
```

Deliverables

You are responsible for turning in your code, and a report for this assignment.

- **Code:**

  Please provide clear code with comments. All code must be in your notebook. Include explanations throughout the notebook on what each section of the code does. Include intermediate outputs in the notebook. Leave copies of the generated plots in the notebook in appropriate places as well.

- **Report:**

  Report the details of your approach in your notebook. Provide a plot of the reduced datasets (refer to Figure 2). Also plot the loss throughout the training procedure. Compare the training loss with the test loss. All discussion should be included in your notebook.

IMPORTANT NOTE: Use of machine learning frameworks is strictly prohibited.
Note: Include any references you used.