Icon

Description automatically generatedAutoencoders

In this chapter, we will look at autoencoders (including the sparse variant). This chapter is a theoretical one, so it will cover the mathematics and the fundamentals of autoencoders. We will discuss what they are, what the limitations are, the typical use cases, and we will look at some examples. We will start with a general introduction to autoencoders, and we will discuss the role of the activation function in the output layer. We will then discuss the loss functions that are typically used and what is the reconstruction error. We will then look at typical applications as dimensionality reduction, classification, and anomaly detection.

# What is an autoencoder

As we have seen in many previous chapters, neural networks are typically used in a supervised setting. Meaning that for each training observation we will have a label or expected value . The neural network model will then learn the relationship between the input data and the expected values. Now suppose we have only unlabeled observations, meaning we only have our training dataset , made of the observations with

Where in general, as noted previously in the book, with . In general, we can give an intuitive definition of an autoencoder as

Definition: an autoencoder is a type of algorithm that learns to reconstruct the input observations with the lowest error possible*[[1]](#footnote-1)*.

And in a more informative definition we could say

Definition: an autoencoder is a type of algorithm with the main purpose of learning an “informative” representation of the data that then can be used for different applications*[[2]](#footnote-2)*

In their most used form, autoencoders are neural networks, but variants exist (see for example XXXX). In general an autoencoder has the structure visualized in Figure 25.1

# Feed Forward Autoencoders

A Feed-Forward Autoencoder (FFA) is a neural network made of dense layers[[3]](#footnote-3). In Figure (25.1) you can see an example of an FFA.



Figure 25.1: A typical architecture of a Feed-Forward Autoencoder. The number of neurons in the layers at first goes down as we move through the network until it reaches the middle, and then starts to grow again until the last layer has the same number of neurons as the input dimensions.

A typical architecture (although it is no strict requirements) has an odd number of layers and is symmetrical with respect to the middle layer. Typically, the first layer has a number of neurons (the size of the input observation . As we move toward the center of the network, the number of neurons in each layer drops in some measure. The middle layer (remember we have an odd number of layers) usually is the one with the smallest number of neurons. In almost all practical applications, the layers after the middle one are a mirrored version of the layers before the middle one. For example, an autoencoder with 3 layers could have the following numbers of neurons: , and then (supposing we are working on a problem where the input dimension is ). All the layers up to and including the middle one make what is called the **encoder**, and all the layers from and including the middle one (up to the output) make what is called the **decoder**, as you can see depicted in Figure (25.1).

The **encoder** can be written generally as a function

Where is the output of the middle layer in Figure (25.1) when we do a forward pass through the network up to the middle layer for input observation . Note that with . It is relevant to point out that could even be one or two orders of magnitude smaller than . The **decoder (and the output of the network** ) **can be written as a generic function**

If the FFA training is successful, the output will be a good approximation of the input, in other words . What is essential to notice is that the decoder can reconstruct the input by using only a much smaller number () of features than the input observations originally have (). The output of the middle layer are called a *learned representation* of the input observation .

**Note** The **encoder** can reduce the number of dimensions of the input observation () and to create a learned representation (of the input that has a much smaller dimension (. This learned representation is enough for the **decoder** to reconstruct the input accurately (if the autoencoder training was successful as intended).

## Activation Function of the Output Layer

In autoencoders, the activation function of the output layer plays a particularly important role. The most used possibilities are two: ReLU and sigmoid. Let’s look at both and give some tips on when to use which and why you should choose one instead of the other.

### ReLU

The ReLU activation function can assume all values in the range . Note that it cannot assume negative values. If your input also has negative values, the ReLU is, of course, a terrible choice, and the identity function would be a much better choice.

**Note** The ReLU activation function for the output layer is well suited for cases when your input observations assume positive real values.

### Sigmoid

The sigmoid function can assume all values in the range . This activation function can only be used if your input observations are all in the range or if you have normalized them to be in that range. Consider as an example the MNIST dataset. Each value of an image is the gray values of a pixel that can assume any value from to. Normalizing the data by dividing the pixel values by would make each observation (each image) have only pixel values between and . In this case, the sigmoid would be a good choice for the output layer's activation function.

**Note** The sigmoid activation function for the output layer is a good choice in all cases where the input observations assume only values between and or if you have normalized them to assume values in the range .

## Loss Function

As with any neural network, we need a loss function to minimize. This loss functions should measure how big is the difference between the input and output . Remember that an autoencoder is trying to learn the identity function; therefore, you want to find the weights in the network that gives you the smallest difference according to some metric between and . This metric will be the loss function that we will use to train our autoencoder with backpropagation. Two loss functions are widely used for autoencoders: the Mean Squared Error (MSE) and the Binary Cross Entropy (BCE). Let’s have a more in-depth look at both since they can only be used when specific requirements are met.

### Mean Square Error

Since an autoencoder is trying to solve a regression problem, the most common choice as a loss function is the Mean Square Error (MSE):

Where the symbol indicates the norm of a vector[[4]](#footnote-4). This loss function is particularly well suited when the inputs are real values that can assume any value. It can be used in almost all cases, independently of how you choose your output layer activation function or of how you normalize your input data.

It is easy to show that the minimum of is found for . To prove it, let’s calculate the derivative of with respect to a specific observation Remember that the minimum is found when the condition

Is met for all . To simplify the calculations, let’s assume that the inputs are one dimensional[[5]](#footnote-5) and let’s indicate them with . We can write

Equation (25.5) is satisfied when as can be easily seen from Equation (25.6), as we wanted to prove. To be really precise, we need to show that

As you should know from calculus. This is easily proved as we have

That is greater than zero, therefore confirming our assumption that for we indeed have a minimum.

### Binary Cross Entropy

If the activation function of the output layer of the FFA is a sigmoid function, thus limiting neuron outputs to be between 0 and 1, and the input features are normalized to be between 0 and 1 we can use as loss function the binary cross entropy, indicated here with . The formula for it is

Where the sum is over the entire set of observations. Can we prove that minimizing this loss function is equivalent to reconstructing the input as well as possible? Let’s calculate where has a minimum with respect to . In other words, we need to find out what values should assume to minimize . As we have done for the MSE, to make the calculations easier, let’s consider the simplified case where and are one-dimensional and let’s indicate them with and .

To find the minimum of a function, as you should know from calculus, we need the first derivative of . In particular we need to solve the set of equations

In this case it is easy to show that the binary cross entropy is minimized when for .

To find when the is minimized we can derive with respect to a specific input

Now remember that we need to satisfy the condition

That can clearly happen only if as can be seen from Equation (25.11). To make sure that this is a minimum we need to evaluate the second derivative, since as you should know from calculus the point for which the first derivative is zero is a minimum only if

We can calculate the second derivative at the minimum point easily

Now remember that . We can immediately see that the denominator of the previous formula is greater than zero. The nominator is also clearly greater than zero since . Dividing two positive numbers gives a positive number, thus we have just proved that

The minimum of the cost function is reached when the output is exactly equal to the inputs, as we wanted to prove.

**Note** an important prerequisite for using the binary cross entropy loss function is that the inputs **must** be normalized between 0 and 1 and that the activation function for the last layer must be a *sigmoid* or *softmax* function.

## Reconstruction Error

The reconstruction error (RE) is a metric that gives you an indication of how good (or bad) the autoencoder was able to reconstruct the input observation . The most typical RE used is the MSE

That can be easily calculated. The RE is used often when doing anomaly detection with autoencoders, as we will explain later. There is an easy intuitive explanation of the reconstruction error. When the RE is big the autoencoder was not able to reconstruct the input well, while when it is small, the reconstruction was successful. In Figure (25.2) you can see an example of big and small reconstruction error when an autoencoder tries to reconstruct an image.

Diagram

Description automatically generated

Figure 25.2: An example of big and small reconstruction error when an autoencoder tries to reconstruct an image.

## Example: reconstruction of hand-written digits

Let’s now see how an autoencoder perform with a real example, using the MNIST dataset. This dataset[[6]](#footnote-6) contains hand-written digits from 0 to 9. Each image is pixels with only gray values, that means that we have 784 features (the pixel gray values) as inputs. Let’s start with an autoencoder with 3 layers with the numbers of neurons in each layer equal to . Note that the first and last layers **must** have a dimension equal to the input dimensions. For this example, we used the Adam optimizer[[7]](#footnote-7), as loss function the cross-entropy and we trained the model for epochs with a batch size of . In Figure (25.3) you can see two lines of images of digits. The line at the top contains ten random images from the original dataset, while the one at the bottom are the reconstructed images with the autoencoder we just described.

A picture containing sign

Description automatically generated

Figure 25.3: In the top line you can see the original digits from the MNIST dataset. While the line below are the digits reconstructed by the autoencoder with number of neurons equal to (784, 16, 784).

It is impressive that to reconstruct an image with 784 pixels only 16 features are needed to have a result that, although not perfect, allows us to understand almost perfectly what digits was used as input. Increasing the size of the middle layer to (and leaving all other parameters the same) gets a much better result as you can see in Figure (25.4).

A picture containing text

Description automatically generated

Figure 25.4: In the top line you can see the original digits from the MNIST dataset. While the line below are the digits reconstructed by the autoencoder with number of neuraons equal to (784, 64, 784).

This tells us that the information of the images is really contained in a much lower number of features than 784.

**Note** An autoencoder with a middle layer smaller than the input dimensions can be used to extract the important features of an input dataset creating a learned representation of the inputs given by the function . Effectively an FFA can be used to perform dimensionality reduction.

The FFA will not be able to recreate the input digits well, if the number of neurons in the middle layer is reduced too much. In Figure (25.5) you can see the reconstruction of the same digits with an autoencoder with only 8 neurons in the middle layer. With only 8 neurons in the middle layer you can see that some reconstructed digits are wrong. As you can see in Figure (25.5) the 4 is reconstructed as a 9 and a 2 is reconstructed to something that resembles a 3.

A close up of a sign

Description automatically generated

Figure 25.5: In the top line you can see the original digits from the MNIST dataset. While the line below are the digits reconstructed by the autoencoder with number of neuraons equal to (784, 8, 784).

In Figure (25.6) you can see a comparison of the reconstructed digits by all the FFAs we have discussed.



Figure 25.6: In the top line you can see the original digits from the MNIST dataset. The second line of digits are the digits reconsructed by the FFA (784,8,784), the third by the FFA (784,16,784) and the last one by the FFA (784,64,784).

From Figure (25.6) you can clearly see how, increasing the size of the middle layer, the reconstruction gets better and better, as we expected.

For these examples we have used the binary cross entropy as loss function, but the MSE would have worked exactly as well as can be seen in Figure (25.7).

Calendar

Description automatically generated

Figure 25.7: In the top line you can see ten random original digits from the MNIST dataset. The second line of digits are the digits reconstructed with an FFA with 16 neurons in the middle layer and the binary cross entropy as loss function. The last line are images reconstructed with the MSE as loss function.

# Sparse Autoencoders

asd

# Autoencoders Applications

## Dimensionality Reduction

As we mentioned in this chapter, the latent features have a dimension that is smaller than the dimensions of the input observations . The *encoder* part (once trained) does naturally (by design) dimension reduction producing real numbers. One can use the latent features for various tasks, as classification for example (as we will see in the next section). I would like to point out some of the advantages of doing dimensionality reduction with an autoencoder compared to a more classical approach as PCA. The autoencoder has one main advantage from a computational point of view: it can deal with very big amount of data efficiently, since its training can be done with mini batches, while PCA needs to do its calculations using the entire dataset. This may seem trivial but in many practical applications the amount of data and the number of features is so big that PCA is not really a practical solution to do dimensionality reduction.

**Note** The use of an autoencoder for dimensionality reduction has one main advantage from a computational point of view: it can deal with very big amount of data efficiently, since its training can be done with mini batches.

## Classification

### Classification with Latent Features

Let’s now suppose that we want to classify our input images. Of course, we can simply use all the features, in our case the pixel values of the images. We can simply use an algorithm as kNN, for illustrative purposes. Note all the results you see in this section have been run on Google Colab. Doing it with nearest neighbors on the training MNIST dataset (with 60000 images) will take around 16.6 minutes (ca.1000 sec) and gets you an accuracy on the test dataset of 10000 images of . But what happens if we use this algorithm not with the original dataset, but with the latent features ? For example, if we consider an FFA with neurons in the middle layer and we again train a kNN algorithm on the latent features we get an accuracy of 89% in 1.1 sec. We get a gain of a factor of 1000 in running time, for a loss of 7.4% in accuracy[[8]](#footnote-8).

|  |  |  |
| --- | --- | --- |
| Input Data | Accuracy | Running Time |
| Original Data | 96.4% | sec. min. |
| Latent Features | 89% | 1.1 sec. |

Table 25.1: the different in accuracy and running time when applying the kNN algorithm to the original 784 features or the 8 latent features for the MNIST dataset.

Keep in mind that using features allow us to get a very high accuracy in one second.

We can do the same analysis with the Fashion MNIST dataset. In Table 25.2 you can see the summary of the results.

|  |  |  |
| --- | --- | --- |
| Input Data | Accuracy | Running Time |
| Original Data | 85.4% | sec. min. |
| Latent Features | 79.9% | 1.2 sec. |
| Latent Features | 83.6% | 3.0 sec. |

Table 25.2: the different in accuracy and running time when applying the kNN algorithm to the original 784 features with a FFA with 8 neurons and with a FFA with 16 neurons for the Fashion MNIST dataset.

It is very interesting to note that with an FFA with 16 neurons in the middle layer we reach an accuracy of 83.6% in just 3 sec. When applying a kNN algorithm to the original features (784) we reach an accuracy just 1.8% higher but with a running time ca. 330 times longer.

**Note** Using autoencoders and doing classification with the latent features is a very viable techniques to reduce the training time by several order of magnitude while incurring in a minor drop in accuracy.

### Curse of dimensionality – a small detour

Is there any other reason why we want to do dimensionality reduction before doing any classification? This becomes important with datasets that have a very high number of features (or dimensions). To understand why we need to make a quick detour in the problem of high dimensionality classification. Let’s consider the unit cube with an integer. Let’s suppose we consider points in the unit cube. How big would be the length of the smallest hyper-cube that contains point? We can easily calculate it as

We can easily calculate this value for various . Let’s suppose that we consider .

|  |  |
| --- | --- |
| d | l |
| 2 | 0.003 |
| 10 | 0.50 |
| 100 | 0.93 |
| 1000 | 0.99 |

And as you can see the data becomes so sparse in high dimensions that you need to consider the entire hyper cube to be able be sure to capture one single observation. When the data becomes so sparse the number of observations that you will need to train an algorithm properly becomes much bigger than the size of existing datasets.

We could look at this in a different way. Let’s consider now a small hypercube of side . How many observations we will find on average in this small portion of the hypercube? This is easy to calculate and is given by

You can clearly see that this number is very small for high values of . For example, if we consider is easy to see that we would need more observations than atoms in the universe[[9]](#footnote-9) to find at least one observation in that small portion of the hypercube.

This is one reason why doing dimensionality reduction may be very important in high dimensionality datasets.

## Anomaly Detection

Autoencoders are often used to perform anomaly detection on the most different datasets. The best way to understand how anomaly detection works with an autoencoder is to do a simple example. Let’s consider an autoencoder with only three layers with 784 neurons in the first, 64 in the latent feature generation layer, and again 784 neurons in the output layers. We will train it with the MNIST dataset and in particular with the 60000 training portion of it as we have done in the previous sections of the chapter. Now let’s consider the Fashion MNIST dataset[[10]](#footnote-10). This dataset is very similar to the MNIST one, with the main difference being that the 10 classes of images are of clothing items instead of hand-written digits. The images are also only gray levels and have a resolution of 28x28 as the MNIST ones. Let’s choose an image of a shoe (see Figure 25.8) from this dataset

A picture containing drawing

Description automatically generated

Figure 25.8: one random image from the Zalando MNIST dataset.

and add it to the testing portion of the MNIST dataset. The original testing portion of MNIST has 10000 images. With the shoe we will have a 10001 images dataset. How can we use an autoencoder to find the shoe automatically in those 10001 images? Note that the shoe is an “outsider” since is a completely different class of images than hand-written digits. To do that we will take the autoencoder we trained with the 60000 MNIST images and with it we will calculate the reconstruction error for the 10001 test images.

The main idea is that since the autoencoder has only seen hand-written digits images, it will not be able to reconstruct the shoe image, and therefore we expect this image to have the biggest reconstruction error. We can check if that is the case by taking the top 3 reconstruction errors we just calculated. For the image we will the MSE for the reconstruction error. You can actually check the code of this example at <https://adl.toelt.ai> yourself. The shoe has the highest reconstruction error: 0.062. It is clear that the autoencoder is absolutely not able to reconstruct the image as it can be seen from Figure 25.9.

Graphical user interface

Description automatically generated

Figure 25.9: the shoe and the reconstruction of the autoencoder trained on the 60000 hand-written images of the MNIST dataset. This image has the biggest RE in the entire 10001 test dataset we built with a value of 0.062.

The second biggest RE is one third of that of the shoe: 0.022, indicating that the autoencoder is doing quite a good job in understand how to reconstruct hand-written digits. You can see the image with the second biggest RE in Figure 25.10. Sincerely also this image could be classified as an outlier, as is not completely clear if is a 4 or an incomplete 9 for example.

A picture containing drawing

Description automatically generatedFigure 25.10: the image with the second biggest RE in the 10001 test dataset: 0.022.

The readers with most experience have noted that we trained our autoencoders on a dataset without any outliers and applied it to a second dataset with outliers. This is not always possible as very often the outliers are not known and are lost in a big dataset. In general, one wants to find outliers in a single big dataset without any information on how many there are or how they look like. Generally speaking, anomaly detection can be done following the main steps below.

1. One train an autoencoder on the entire dataset (or if possible, on a portion of the dataset that is known **not** to have any outlier).
2. For each observation (or input) of the portion of the dataset known to have the wanted outliers one calculates the RE.
3. One sorts the observations by the RE.
4. One classifies the observations with the highest RE as outliers. Note that how many observations are really outliers will depend on the problem at hand and will require an analysis of the results and usually lot of knowledge of the data and the problem.

Note that if one train the autoencoder on the entire dataset at disposal, there is an essential assumption: the outliers are a negligible part of the dataset and their presence will not influence (or will influence in a negligible way) how the autoencoder learns to reconstruct the observations.

A classic example of anomaly detection is trying to find the fraudulent credit card transactions (the outliers). This case usually presents ca. 0.1% fraudulent transactions and therefore this would be a case that would allow us to train the autoencoder on the entire dataset.

**Note** If one train the autoencoder on the entire dataset at disposal, there is an essential assumption: the outliers are a negligible part of the dataset and their presence will not influence (or will influence in a negligible way) how the autoencoder learns to reconstruct the observations.

### Model Stability

Note that doing anomaly detection as described in the previous section seems easy, but those methods are prone to overfitting and give often unstable results. This means that training an autoencoder with a different architecture may well give different REs and therefore different outliers. One the simplest way of dealing with instability of results is to train different models and then take the average of the RE. Another technique involves taking the maximum of the REs evaluated from several models.

**Note** Anomaly detection done with autoencoders is prone to problems as overfitting and unstable results. It is important to be aware of these problems and to check the results coming from different models to be able to interpret the results correctly.

Note that this section serves the purpose of giving you some pointers and is not meant to be an exhaustive overview on how to solve this problem.

More advanced techniques, like autoencoders ensembles[[11]](#footnote-11), are also used to deal with problems of instable results coming from small datasets for example.

# Exercises

Exercise 1

List the most useful tasks you can use an autoencoder for. Can you think of an application in your field of work?

Exercise 2

Can you explain briefly what is a sparse autoencoder? Why is different from a classical autoencoder?

Exercise 3

How do you measure (with which metric) the performance of an autoencoder? List the most commonly used metrics that you can use. Can you think of any additional metric, in addition to those discussed in this chapter, that could be used?

Exercise 4

Describe how anomaly detection work with autoencoders.

# Further Readings

Deep Learning Tutorial from Stanford University

<http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/>

Building autoencoders in Keras

<https://blog.keras.io/building-autoencoders-in-keras.html>

Introduction to autoencoders in TensorFlow

<https://www.tensorflow.org/tutorials/generative/autoencoder>

Bank, D., Koenigstein, N., and Giryes, R., “Autoencoders”, arXiv e-prints, 2020,

<https://arxiv.org/abs/2003.05991>

R. Grosse, University of Toronto, Lecture on autoencoders

<http://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/slides/lec20.pdf>

1. In this chapter we will discuss at length what we mean with error here. [↑](#footnote-ref-1)
2. Bank, D., Koenigstein, N., and Giryes, R., Autoencoders, <https://arxiv.org/abs/2003.05991> [↑](#footnote-ref-2)
3. A dense layer is simply a set of neurons that gets their inputs from the previous layer. Each neuron in a dense layer gets as input the output of all neurons in the previous layer. [↑](#footnote-ref-3)
4. The norm of a vector is simply the square root of the sum of the square of the components. [↑](#footnote-ref-4)
5. If we did not have this assumptions, one would have to calculate the gradient of the loss function instead of the simple derivative. [↑](#footnote-ref-5)
6. More information on the dataset can be found here: <http://yann.lecun.com/exdb/mnist/>. [↑](#footnote-ref-6)
7. You can find the entire code at the address <https://adl.toelt.ai>. [↑](#footnote-ref-7)
8. You can run those tests yourself going to Chapter 25 at <https://adl.toelt.ai>. [↑](#footnote-ref-8)
9. <https://www.universetoday.com/36302/atoms-in-the-universe/> [↑](#footnote-ref-9)
10. <https://research.zalando.com/welcome/mission/research-projects/fashion-mnist/> [↑](#footnote-ref-10)
11. See for example <https://saketsathe.net/downloads/autoencode.pdf>. [↑](#footnote-ref-11)