

QML-Mod2-Classical Machine Learning

Riccardo Marega

March 2025

Indice

1	Introduction to classical machine learning -22/03/2025	3
1.1	Input data	4
1.2	Machine Learning examples	5
1.2.1	Code overview	6
1.3	Cross validation and hyperparameter tuning	6
1.3.1	Evaluation metrics for classification	7
2	Introduction to deep learning -28/03/2025	8
2.1	Training an artificial neural network	8
2.2	Artificial neural networks: tasks	13
2.2.1	Binary classification	13
3	Advanced topics -29/03/2025	16
3.1	Convolutions from scratch	16
3.1.1	Types of convolutions	18
3.2	Convolutional neural networks for image classification	19
3.3	Convolutional neural networks for classification task	20
3.3.1	Image classification	20
3.3.2	Architectures	20
3.4	Training a convolutional neural network	22
3.4.1	Transfer learning and fine tuning	22
3.5	Data preparation	23
3.5.1	Data augmentation	23
3.6	Optimizers	24
3.7	Loss function	25
3.8	Metrics	26
4	AI ETHICS: ethics of artificial intelligence -5/04/2025	27

5 Deep generative models -11/04/2025	31
5.1 Preliminary concepts	31
5.1.1 PDF and joint probability	31
5.1.2 Maximum likelihood estimation (MLE) & Maximum a Posteriori Estimation (MAP)	32
5.2 Variational AutoEncoders (VAE)	33
5.3 Generative Adversarial Networks (GAN)	37
5.4 Diffusion Models (DM)	40
6 Deep Network Training and Testing -18/04/2025	41
6.1 Pytorch practice	41
Exercise 0	42
Exercise 1	48
Exercise 2	57
Exercise 3	66
Exercise 4	77
Exercise 5	82
Exercise 6	87
Exercise 7	92

1 Introduction to classical machine learning -22/03/2025

Despite maybe we were not aware but we've already trained models, in our life, with machine learning algorithms: an example is given by the tools asking the users to spot cars, traffic signals, etc. in some website before entering them. An other example is given by all the filters that one can apply before taking a picture when using social media.

Artificial Intelligence vs Machine Learning vs Deep Learning

- artificial intelligence: any technique that enables computers to mimic human intelligence. It includes machine learning
- machine learning: a subset of AI that includes techniques that enable machines to improve at tasks with experience. It includes deep learning
- deep learning: a subset of machine learning based on neural networks that permit a machine to train itself to perform tasks

Machine learning teaches computers to do what comes naturally to humans and animals: learning from experience. Machine learning algorithms use computational methods to "learn" information directly from data, without relying on a predetermined equation. These algorithms adaptively improve their performance as the number of samples available for learning increases.

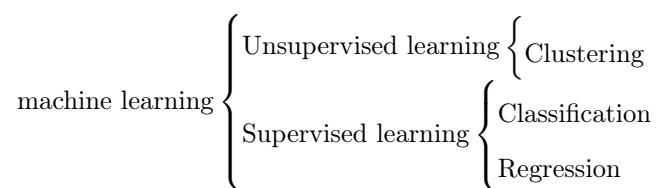
Machine learning includes two kinds of modalities to train the algorithm: supervised and unsupervised learning. In the latter we don't give any label to the algorithm and we let it find itself the correct answer to the given problem.

The choice of using ML algorithms arise when:

- we can't rely on rule-based systems
- the rules depend on too many variables, and many of them overlap or need to be optimized
- scalability becomes an issue

An example of machine learning algorithm is to be found in surgical data science, where AI is at service of surgeons.

The choice of the best algorithm to apply has to be done considering the dimensions and the type of data one has to manage.



Supervised learning: the model is trained under the assumption that for each input the label is known.

Unsupervised learning: the model identifies meaningful patterns within the input data, which does not come with any label.

In supervised learning methods, the output is already known, and the goal of training is to map inputs into the corresponding outputs. To build a model, the machine learning is fed with a large amount of input data along with their corresponding labels. Supervised learning uses classification and regression techniques to develop predictive models. The difference between classification and regression is that the first one does a prediction of discrete outputs while the second one does a prediction of continuous outputs.

For the unsupervised learning case it's not necessary to supervise the model or provide labeled input data. The algorithm begins to learn from the data without guidance. The model uses unlabeled data to identify new information. Since there are no known output values to establish a logical relationship between input and output, specific techniques are used to extract rules, patterns, and grouping of similar types. Machine using unsupervised learning algorithms discover patterns to find meaningful insights. For example, the system can learn to distinguish between dogs and cats by understanding the features and characteristics of each animal.

Every machine learning workflow starts with three key questions:

- What type of data are we working with?
- What do we want to extract or learn from the data?
- What is the context or domain of application?

The main problem with unsupervised learning is that evaluating its performance is not always immediate. We choose unsupervised learning when we aim to understand: the distribution of data, the clustering of data based on similar characteristics and the dimensionality reduction (needed to obtain a more concise and efficient representation).

Note that one can combine supervised and unsupervised learning.

1.1 Input data

Standard algorithms of ML usually take as inputs features and characteristics extracted from the data set. These features are handcrafted or manually designed or selected. Features are relevant pieces of information that help solve the task at hand. **The quality of a machine learning model depends on its ability of extracting features.**

The data is divided in a training set (used exclusively during the training phase) and a validation set (used to monitor the process and optimize model trainings). An algorithm is subject to underfitting of the data set if it has a poor performance on the training set. This happens because the model fails to learn the relationship between input data and their corresponding labels. Otherwise it

is said being subject to overfitting if it works well with the data training but has poor results with data of the test set. This occurs because the model memorize the data it has seen and is unable to generalize to unseen examples.

For a fully functional training of the model we introduce a third data set called validation set. These data will be used exclusively during the training phase to monitor the learning process and optimize model training. This is called validation set because it serves to validate the results obtained in the training set. If the performance is poor, we may need to adjust the model's hyperparameters (in the case of ML) and retrain the model until the validation results are satisfactory. When implementing this third dataset the testing set will be exclusively used after training is complete to evaluate model's performance.

Cross validation strategy: one can repeat the whole training of the model swapping every time the data from one set to another (what was first, for example, in the training set now will go in the validation set).

Generalization In Machine Learning, generalization refers to a model's ability to perform well on unseen data, meaning data that was not used during training. A well-generalized model captures the underlying patterns in the data rather than memorizing specific examples, allowing it to make accurate predictions on new inputs. Poor generalization can lead to overfitting (where the model performs well on training data but poorly on new data) or underfitting (where the model fails to learn meaningful patterns from the training data).

When an algorithm makes consistent mistakes, we say it has a "bias".

A **bias** is a distortion of the training data which is propagated in the algorithm.

The goals related to the supervised trainings are a low error during training, validation and testing phases. The selection of the algorithm is based on:

- velocity of the training
- storage capacity
- accuracy on new data prediction
- transparency and interpretability

1.2 Machine Learning examples

- **Logistic regression:** the model is trained to predict the probability of a binary choice. Due to its simplicity, logistic regression is commonly used as a starting point for binary classification problems. This algorithm is normally used when the data can be clearly separated by

a single decision boundary and in general as baseline to evaluate more complex classification methods.

- **k nearest neighbor (kNN)**: is a pattern recognition algorithm based on distance of the data. If a subject A has, for example, characteristic close to the one of a subject B, then probably they are associated to the same category. This kind of algorithm is normally used: for establishing baseline learning rules, when memory usage is not a major problem and when prediction speed of the trained model is not a major concern.

A **decision tree** is a supervised learning algorithm used for both classification and regression tasks. It has a hierarchical tree structure composed of a root node, branches, internal nodes and leaf nodes. A decision tree starts with the root node, which has no incoming branches. The outgoing branches from the root lead to internal nodes, also known as decisions nodes. Both root and internal nodes perform evaluations to split the data into homogeneous subsets, which are represented by leaf nodes or terminal nodes. Leaf nodes represent all possible outcomes (either continuous or discrete) within the dataset.

ML learning models in training phase learn a set of rules which depends both on the data set and a fixed combination of hyperparameters. The automatic learning of a model is not a single process, indeed it is necessary to experiment different models fixing different values for the hyperparameters.

1.2.1 Code overview

```
1 from sklearn.ensemble import RandomForestClassifier  
2  
3 rf_clf = RandomForestClassifier(n_estimators=?, max_depth=?)
```

Imports the RandomForestClassifier model from the scikit-learn library. This model is a supervised learning algorithm based on decision trees. Creates an instance of the Random Forest classifier with two key parameters:

- `n_estimators=?`: defines the number of trees in the forest (typically between 10 and 1000).
- `max_depth=?`: sets the maximum depth of the trees (can be `None` to allow them to grow until all leaves are pure).

1.3 Cross validation and hyperparameter tuning

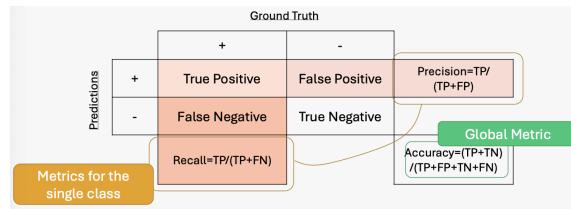
We divide the training and the validation set in multiple sets so that the model can be trained with all the available data. The model's performance for the same model with different hyperparameters combinations strongly depends on the specific data split. Each model is trained and evaluated only once, so its performance is tied to that single evaluation. But this raises an issue: we might get very different results when training and validating on different subsets of the same data. What

if we could split the training and validation sets multiple times, each time using different subsets of data, train and evaluate our models repeatedly, and observe model performance across several rounds of evaluation? → **k-fold Cross Validation**.

Cross validation is statistical method used to estimate the ability of different models of performing automatic learning. This procedure is done by defining a parameter k which represents the number of sets in which every set is subdivided. In this approach the dataset is randomly divided into k groups of roughly equal size. The model is trained on k-1 folds and validated on remaining fold. This process is repeated k times, each time using different fold as the validation set.

We can perform cross validation for hyperparameter tuning in either inner loops or outer loops

1.3.1 Evaluation metrics for classification



$$\text{Accuracy: } \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Recall (Sensitivity/True positive rate): } \frac{TP}{TP+FN}$$

$$\text{Precision: } \frac{TP}{TP+FP}$$

$$\text{Specificity: } \frac{TN}{TN+FP}$$

$$\text{F1 score: } 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC Curve & AUC (Area under the curve)

- **ROC curve:** Plot true positive rate (recall) vs. False Positive rate ($FPR = \frac{FP}{FP+TN}$) at various threshold.
- **AUC (Area Under Curve)** Probability the classifier ranks a randomly chosen positive higher than a randomly chosen negative (AUC = 0.5: random , AUC = 1: perfect).

2 Introduction to deep learning -28/03/2025

2.1 Training an artificial neural network

Modeling problems the first example is defining the position of a car whose position is given exactly by $d(t) = d_0 + vt$. This problem is fully solvable. That is not always the case; indeed, a problem could be characterized by a large number of variables. The problem with working with large number of variables is that not always every variable has the same importance as the others. The problem, can be reduced to: $y = \alpha x_1 + \beta x_2 + \gamma x_3$.

Machine learning allows to solve complex problems in which we can easily tell what variables are involved.

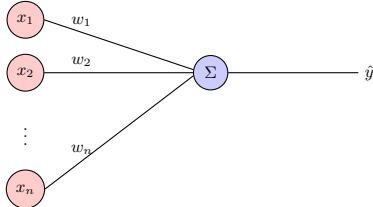
A typical problem that could be modeled is image recognition.

In ML we have a phase that does not appear in deep learning which is feature education: in traditional ML, model require pre-computed features that are manually designed based on domain knowledge. This feature engineering requires human expertise to identify the most relevant attributes for a given task.

Deep learning eliminates the need for manual feature extraction by learning hierarchical representations directly from raw data. Using AI networks, DL can automatically detect patterns at multiple levels of abstraction.

Artificial Neural Networks Biologically, neurons are unique cells that can communicate with one another, thus propagating information.

An artificial neuron receives n inputs (x_1, \dots, x_n) , each scaled by a factor (weights) and sums them all:



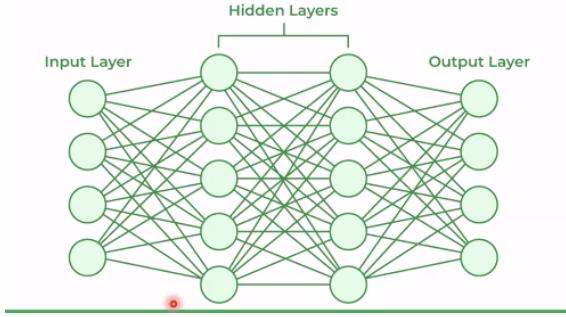
To map all the space is necessary to add a factor b called bias, such that:

$$y = \sum_{i=1}^n w_i \times x_i + b = w_1 \times x_1 + \dots + w_n \times x_n + b.$$

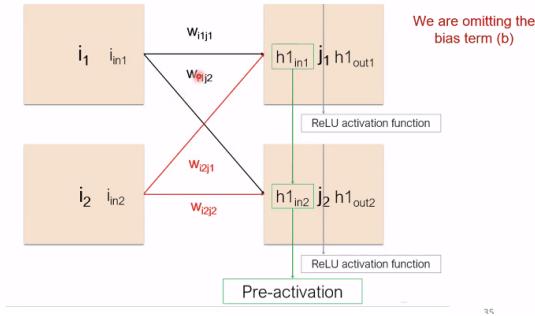
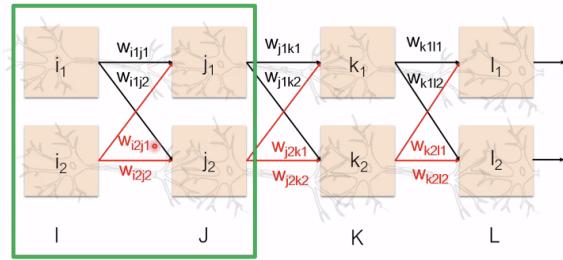
Weights and bias are the parameters of the neural network. Note that very few models are correctly modeled using linear combinations of the variables; indeed, we introduce an activate factor such that:

$$y = \sum_{i=1}^n g(w_i \times x_i + b) = g(w_1 \times x_1 + \dots + w_n \times x_n + b).$$

Typical activation functions are: sigmoid, tanh, ReLU ($\max(0, x)$), Leaky ReLU ($\max(0.1x, x)$). The sigmoid and tanh functions limit the output, indeed, the output is always confined between 0 and 1.



This is what a simple neural network looks like: an input layer, a bunch of hidden layers, and an output layer. Each layer extracts information from the input and transmits it to the next one for further data processing.



$$[i_{in1} i_{in2}] \times \begin{bmatrix} w_{i1j1} w_{i1j2} \\ w_{i2j1} w_{i2j2} \end{bmatrix} = [h_{in1} h_{in2}]$$

$$[h_{out1} h_{out2}] = \text{ReLU}([h_{in1} h_{in2}])$$

What we are building is a neural architecture. All these parameters will have to be adjusted. This process is exactly what is called "train of the neural network". The training process continues by making mistakes.

Let's consider the problem of predicting the cost of a house: y is the real cost and \hat{y} is the predicted one. y is also called ground-truth value and is used to supervise the training. The idea under the training process lays in computing the error, and what we want indeed is to try to minimize this error. Error minimization is an optimization problem: I need a parameter configuration that explains the problem in the best possible way (i.e. I get as close as possible to the ground-truth value)

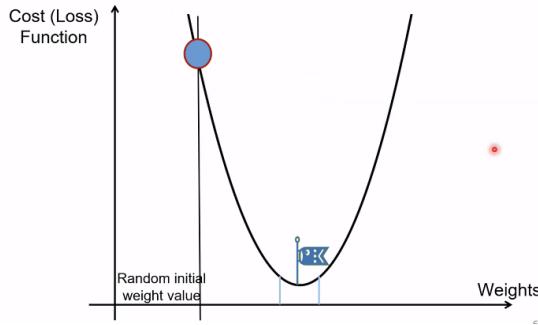
$$\text{error} = |y - \hat{y}|.$$

To optimize the parameters of a neural network we use the **gradient method**. What we always know when computing an algorithm is whether the error is increasing or decreasing.

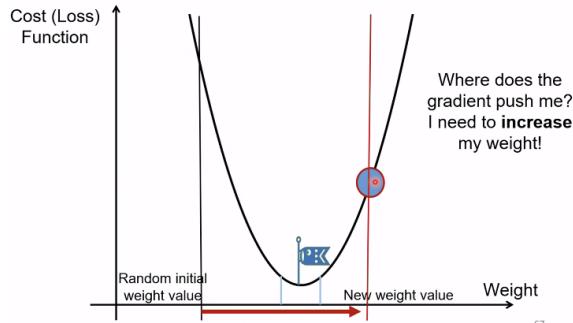
Gradient descent is the most used optimization technique to minimize the error. It means that, for each step, I compute the error and change the parameters in order to minimize it according to its dependence on each parameter.

$$\text{error} = f(w, b)$$

this function is called **cost function**. First order derivative (gradient) is a measure of dependence between each independent variable (w and b) and the dependent one (error).

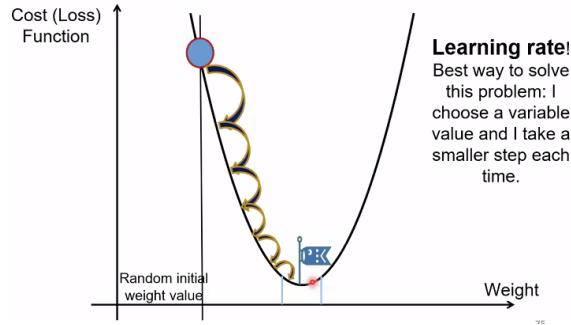


Initializing the parameters randomly gives better results.



and so on till we reach the minimum.

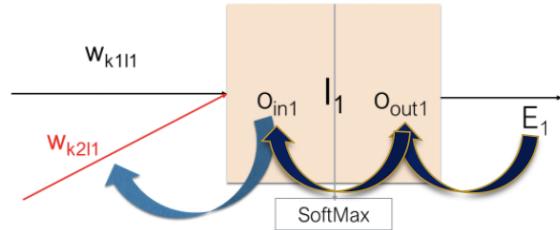
In general the first step to reach the minimum is the bigger among all the steps that will be done. In this case we talk about learning rate. While "learning" we have to decrease the learning rate (it should take smaller steps) in order to reach the target.



Mathematically, training a neural network means to change its parameters N times in order to minimize the cost function (error). Each update can be expressed as

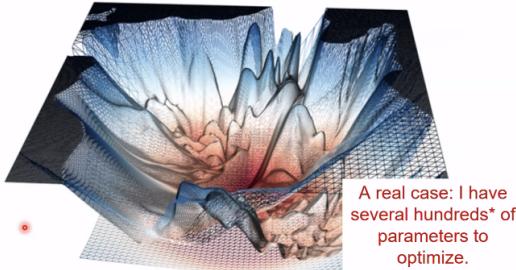
$$w_{t+1} = w_t - \eta \times \nabla \text{error}$$

where η is the learning rate and ∇error is the gradient of the cost function with respect to w . A typical loss function is the MSE. The only constraint that we impose to the loss functions is to be differentiable everywhere. The error can be computed only from the output of the layers:



$$\frac{dE_1}{dw_{k1l1}} = \frac{dE_1}{dO_{out1}} \times \frac{dO_{out1}}{dO_{in1}} \times \frac{dO_{in1}}{dw_{ikl1}}$$

For a forward pass of information we have an error back-propagation. This kind of neural network is called fully connected.

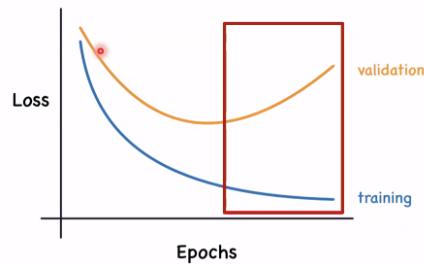


Usually, what we need is a dataset, i.e. a set of (x,y) samples where x is the input and y is the output. The dataset is divided into three sub-sets: a training set, a validation set and a test set. The training process involves parameters optimization on the same training samples several times, or epochs.

AI does not really "learn": it just finds patterns. It's like a student who only studies math by doing every exercise on the book but never reads a page of theory. How can we make sure that the student is learning and not just memorizing? During the test, the student has no way to learn new knowledge because it has no supervision (correct answers). It will be up to the teacher to evaluate the results from the test and assess the student's knowledge. Therefore, our neural network needs to perform well both on the training set and on the two test sets. What happens if it doesn't? The loss function on the training set has a very nice trend: as the network is iteratively optimized on the training sample, its error on these samples decreases. However, the loss function on the validation test doesn't seem as good: while the student improves with those "training" exercises, they make a lots of mistakes during the exam.

Two things are never to be expected:

- no error or 100% accuracy
- better results in validation



However, we should minimize the distance between training and validation performance: **overfitting**.

The easiest task for which ML and DL are used are regression and classification: regression is used to predict a number from a continuous set of numbers (e.g. prediction of the price of a house) while classification is used to predict a number from a discrete set of numbers.

The number of neurons in the final layer will depend on the task: for the regression case we need to predict one number → one neuron is needed, while in case of classification we need to label the inputs as one of the possible N classes → N neurons are needed. The optimal number of neurons in the hidden layers (as well as the number of the hidden layers) cannot be assessed a priori. The more the neurons and the layers, the more abstract information we can extract from the input data.

2.2 Artificial neural networks: tasks

Check out the following site: playground tensorflow

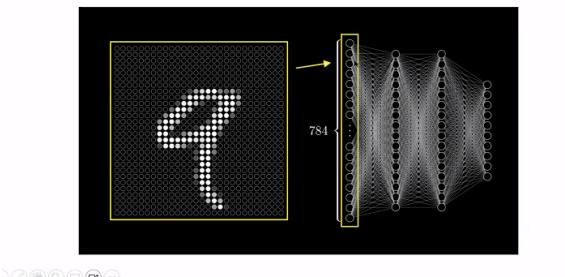
2.2.1 Binary classification

Classification is the most studied problem in DL. To classify means to assign a label to the input data among a closed set of labels.

The **MNIST** dataset is one of the most famous one: our network needs to classify the input image as one of the 10 possible labels (the digits 0-9).

How do we feed these images to the neural network?

Pixel by pixel: MNIST images are $28 \times 28 = 784$ pixels.



Pixels are integer values between 0 (black) and 255 (white) that are processed by neural network as numbers.

Check out: The stilwell brain

In the lecture of today we are going to classify whether a digit (lower than 10) is odd or even. Therefore, ours is a binary classification task. For classification, a perfect loss function is (binary) cross-entropy.

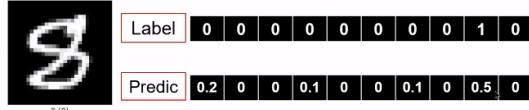


Figura 1: Note: 0.6 instead of 0.5

I want my prediction to be a probability distribution. The sum of all my elements needs to be 1, so that I can "compare" it with the label (also a probability distribution). In case of a odd vs even prediction the binary classification will give more precise results.

Loss function: Binary cross entropy

$$\begin{aligned} & - \sum_{j=1}^M y_j \log(p(y_j)) \\ & - \sum_{i=1}^N y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i)) \end{aligned}$$

Final activation: softmax

In order to obtain a probability distribution, I need to use the softmax activation function in the outer layer.

$$\begin{bmatrix} 1.3 \\ 5.1 \\ 2.2 \\ 0.7 \\ 1.1 \end{bmatrix} = \left[\frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \right] = \begin{bmatrix} 0.02 \\ 0.9 \\ 0.05 \\ 0.01 \\ 0.02 \end{bmatrix}$$

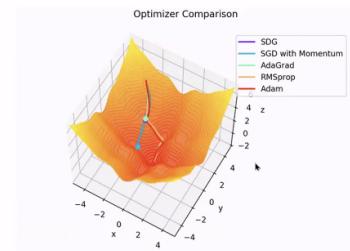
Metric: accuracy

$$\begin{bmatrix} \text{True positive (TP)} & \text{False Positive (FP)} \\ \text{False Negative (FN)} & \text{True Negative (TN)} \end{bmatrix}$$

where

- Recall = $\frac{\sum TP}{\sum TP + FN}$
- Precision = $\frac{\sum TP}{\sum TP + FP}$
- Accuracy = $\frac{\sum TP + TN}{\sum TP + FP + FN + TN}$

Optimizer: Adam (Adaptive moment estimator)



3 Advanced topics -29/03/2025

3.1 Convolutions from scratch

Edge detection is an old but gold problem in computer vision that involves detecting edges in an image to determine object boundaries and thus separate the object of interest.

One of the most popular techniques for edge detection is the Canny Edge Detection algorithm.

An edge is a point of rapid change of intensity of the image function. The gradient points in the direction of the most rapid increase

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}.$$

Using filters (aka matrices) is possible defining the edges in an image. An example is given by:

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

A **convolution** (of images) is simply an elementary multiplication of two matrices followed by a sum:

- take two matrices (which both have the same dimension)
- multiply them, element by element
- add up the elements

Originales	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge-Detect	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
Blur	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Blur	$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	
Blur	$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	

An image is just a multidimensional matrix, but unlike traditional matrices, images (RGB) can also have a depth. The kernel should be thought as a small matrix that is used for blurring, sharpening,

edge detection, and other image processing and functions. It is common to define the kernel by hand to achieve various image processing functions, edge detection: all these operations are hand-defined forms of kernels designed specifically to perform a particular function. A question then arises: is there a way to automatically learn these type of filters? And even use these filters for image classification and object detection? Of course there is: **CNN**.

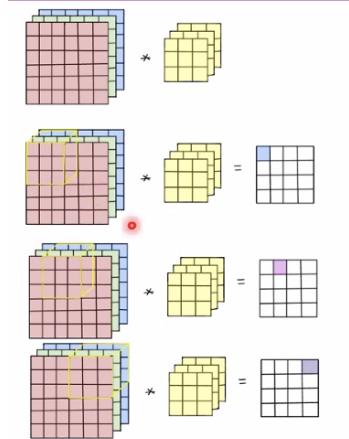
Kernel Most of the kernels we usually see are $N \times N$ square matrices. We use an odd kernel dimension to ensure that there is a valid integer coordinate in the center of the image. In image processing, a convolution requires three components: an input image, a kernel matrix to apply to the input image and an output image to store the results of the input image convolved with the kernel.

The process of "sliding" a convolutional kernel over an image and storing the output decreases the spatial dimensions of the output data. This decrease in spatial dimension is simply a side effect of applying convolutions to images.

Input Matrix	Kernel	Result																														
<table border="1"> <tr><td>45</td><td>12</td><td>5</td><td>17</td></tr> <tr><td>22</td><td>10</td><td>35</td><td>6</td></tr> <tr><td>88</td><td>26</td><td>51</td><td>19</td></tr> <tr><td>9</td><td>77</td><td>42</td><td>3</td></tr> </table>	45	12	5	17	22	10	35	6	88	26	51	19	9	77	42	3	<table border="1"> <tr><td>0</td><td>-1</td><td>0</td></tr> <tr><td>-1</td><td>5</td><td>-1</td></tr> <tr><td>0</td><td>-1</td><td>0</td></tr> </table>	0	-1	0	-1	5	-1	0	-1	0	<table border="1"> <tr><td>-45</td><td></td></tr> <tr><td></td><td></td></tr> </table>	-45				
45	12	5	17																													
22	10	35	6																													
88	26	51	19																													
9	77	42	3																													
0	-1	0																														
-1	5	-1																														
0	-1	0																														
-45																																
<table border="1"> <tr><td>45</td><td>12</td><td>5</td><td>17</td></tr> <tr><td>22</td><td>10</td><td>35</td><td>6</td></tr> <tr><td>88</td><td>26</td><td>51</td><td>19</td></tr> <tr><td>9</td><td>77</td><td>42</td><td>3</td></tr> </table>	45	12	5	17	22	10	35	6	88	26	51	19	9	77	42	3	<table border="1"> <tr><td>0</td><td>-1</td><td>0</td></tr> <tr><td>-1</td><td>5</td><td>-1</td></tr> <tr><td>0</td><td>-1</td><td>0</td></tr> </table>	0	-1	0	-1	5	-1	0	-1	0	<table border="1"> <tr><td>-45</td><td>103</td></tr> <tr><td></td><td></td></tr> </table>	-45	103			$45*0 + 12*(-1) + 5*0 + 22*(-1) + 10*5 + 35*(-1) + 88*0 + 26*(-1) + 51*0$
45	12	5	17																													
22	10	35	6																													
88	26	51	19																													
9	77	42	3																													
0	-1	0																														
-1	5	-1																														
0	-1	0																														
-45	103																															
<table border="1"> <tr><td>45</td><td>12</td><td>5</td><td>17</td></tr> <tr><td>22</td><td>10</td><td>35</td><td>6</td></tr> <tr><td>88</td><td>26</td><td>51</td><td>19</td></tr> <tr><td>9</td><td>77</td><td>42</td><td>3</td></tr> </table>	45	12	5	17	22	10	35	6	88	26	51	19	9	77	42	3	<table border="1"> <tr><td>0</td><td>-1</td><td>0</td></tr> <tr><td>-1</td><td>5</td><td>-1</td></tr> <tr><td>0</td><td>-1</td><td>0</td></tr> </table>	0	-1	0	-1	5	-1	0	-1	0	<table border="1"> <tr><td>-45</td><td>103</td></tr> <tr><td>-96</td><td></td></tr> </table>	-45	103	-96		$12*0 + 5*(-1) + 17*0 + 10*(-1) + 35*5 + 6*(-1) + 26*0 + 51*(-1) + 19*0$
45	12	5	17																													
22	10	35	6																													
88	26	51	19																													
9	77	42	3																													
0	-1	0																														
-1	5	-1																														
0	-1	0																														
-45	103																															
-96																																
<table border="1"> <tr><td>45</td><td>12</td><td>5</td><td>17</td></tr> <tr><td>22</td><td>10</td><td>35</td><td>6</td></tr> <tr><td>88</td><td>26</td><td>51</td><td>19</td></tr> <tr><td>9</td><td>77</td><td>42</td><td>3</td></tr> </table>	45	12	5	17	22	10	35	6	88	26	51	19	9	77	42	3	<table border="1"> <tr><td>0</td><td>-1</td><td>0</td></tr> <tr><td>-1</td><td>5</td><td>-1</td></tr> <tr><td>0</td><td>-1</td><td>0</td></tr> </table>	0	-1	0	-1	5	-1	0	-1	0	<table border="1"> <tr><td>-45</td><td>103</td></tr> <tr><td>-96</td><td>133</td></tr> </table>	-45	103	-96	133	\dots
45	12	5	17																													
22	10	35	6																													
88	26	51	19																													
9	77	42	3																													
0	-1	0																														
-1	5	-1																														
0	-1	0																														
-45	103																															
-96	133																															
<table border="1"> <tr><td>45</td><td>12</td><td>5</td><td>17</td></tr> <tr><td>22</td><td>10</td><td>35</td><td>6</td></tr> <tr><td>88</td><td>26</td><td>51</td><td>19</td></tr> <tr><td>9</td><td>77</td><td>42</td><td>3</td></tr> </table>	45	12	5	17	22	10	35	6	88	26	51	19	9	77	42	3	<table border="1"> <tr><td>0</td><td>-1</td><td>0</td></tr> <tr><td>-1</td><td>5</td><td>-1</td></tr> <tr><td>0</td><td>-1</td><td>0</td></tr> </table>	0	-1	0	-1	5	-1	0	-1	0	<table border="1"> <tr><td>-45</td><td>103</td></tr> <tr><td>-96</td><td>133</td></tr> </table>	-45	103	-96	133	\dots
45	12	5	17																													
22	10	35	6																													
88	26	51	19																													
9	77	42	3																													
0	-1	0																														
-1	5	-1																														
0	-1	0																														
-45	103																															
-96	133																															

However, in most cases, we want the output image to be the same size as the input image. To ensure this, we apply padding. The fact that the output is smaller than the input does not seem to be a big problem: we did not lose much data because most of the important features are located in the central area of the input. The only case when losing this information is a real problem is when much information is concentrated on border of the image. Padding could be done using zero elements or just copies of the border.

For an image we have:



An RGB image is represented as a $6 \times 6 \times 3$ volume, where the 3 correspond to the 3 color channels (RGB). To detect edges or other features in this image, one could convolve the $6 \times 6 \times 3$ with a 3-D filter. Then also the filter itself will have 3 levels corresponding to the red, green and blue channels. So, the filter also has a height, a width and a number of channels. The number of channels in the image must match the number of channels in the filter.

In convolutional neural networks, convolutional layers are not only applied to input data, such as pixel values, but can also be applied to the output layers. The sequence of convolutional layers allows a hierarchical breakdown of the input. Consider that filters operating directly on the raw pixel values will learn to extract low level features from the starting image, such as lines. Filters operating on the output of the first convolutional layer can extract features that are combinations of lower level features, such as features that comprise multiple lines to extract shapes. This process continues until very deep layers extract faces, animals, houses and so on. The abstraction of characteristics to ever higher orders increases with network depth.

Kernel size The kernel size defines the convolution field of view.

Padding The padding defines how the edge of a sample is handled. A convolution with padding will keep the spatial dimensions of the output equal to those of the input, while convolutions without padding will crop some of the edges if kernel is larger than 1.

Stride The stride defines the size of the kernel step when passing through the image. Although the default setting is usually 1, you can use a stride of 2 to downsample an image (this is used to improve the features of the network).

3.1.1 Types of convolutions

Dilated/Atrous convolution The atrous convolutions introduce another parameter called the rate of expansion. This parameter defines the distance between values of a kernel. This way you get

a wider field of view at the same computational cost. That type of convolution is used to understand large partial context.

Spatial separable convolutions A separable spatial convolution simply splits one kernel into smaller kernels. With fewer multiplications, the computational complexity decreases (the network has few parameters to learn).

Depth-wise separable convolutions We have an RGB input image (with 3 channels). After convolutions, a feature map can have more channels (as usually happens). Each channel can be used for a particular interpretation of the image. Similarly to spatial separable convolution, a depthwise separable convolution divides a kernel into two separate kernels that perform two convolutions: the **depthwise convolution** and the **point convolution**.

Width multiplier allows scaling the network's width to control the number of parameters and computation.

Resolution multiplier enables adjusting the input image resolution to further reduce computational requirements.

3.2 Convolutional neural networks for image classification

The visual cortex has hierarchical structure: LGB (lateral geniculate body) -> simple cells -> complex cells -> lower order hypercomplex cells -> higher order hypercomplex cells

We can think that these complex cells are performing an aggregation of activations using functions such as the maximum, the sum. In this way, these cells can recognize edges and orientations.

Deep neural networks are normally organized in alternate repetitions of linear and non-linear operators. The reason for having multiple layers of this type is to build a hierarchical representation of the data.

The local pixels assemble to perform simple patterns like oriented edges. These borders are in turn combined to form patterns that are even more abstract. We can continue to build above these hierarchical representations until we arrive at the objects that we observe in the real world.

This compositional and hierarchical nature that we observe in the natural world is therefore not only the result of our visual perception, but it is something true in the physical world. At the lowest level of description, we have elementary particles, which are composed to form atoms, more atoms form molecules, and we continue to increase this process until we form materials, parts of objects and finally complete objects in the physical world.

3.3 Convolutional neural networks for classification task

3.3.1 Image classification

is the task of assigning to an input image a label belonging to pre-set set of categories.

Pro convolutional neural networks The 3D element that flows along the 3 image channels (RGB) is called convolutional kernel. The convolutional kernel slides over the image (in the case of the input layer) and extract features or features map. Unlike fully connected networks where the input was a carrier, CNNs operate on volumes (multi-channel image).

The pixels of the feature map with the same color are from the same kernel. Both types of networks learn by updating the weights which, in the case of fully connected networks, are the values of the connections whereas, in the case of convolutional neural networks, they are the values of the kernels and connections. In both types of networks, the neuron receives an input which is a combination of weighted inputs. This combination of weighted inputs represents the overall level of neuron excitation and is given as input to an activation function which produces some limited output.

Other layers in CNNs: Pooling Pooling provides a form of translation invariance: small shift in the input can lead to the same output. Statistics over neighboring features to reduce the size of the feature maps:

- separate the image into non-overlapping subimages
- select the maximum/average/... in each layer

3.3.2 Architectures

- LeNet-5 1998

Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	32x32	-	-	-
1	Convolution	6	28x28	5x5	1	tanh
2	Average Pooling	6	14x14	2x2	2	tanh
3	Convolution	16	10x10	5x5	1	tanh
4	Average Pooling	16	5x5	2x2	2	tanh
5	Convolution	120	1x1	5x5	1	tanh
6	FC	-	84	-	-	tanh
Output	FC	-	10	-	-	softmax

- Alexnet 2012

Layer	Feature Map	Size	Kernel Size	Stride	Activation
Input: Image	1	227x227x3	-	-	-
1 Convolution	96	55x55x96	11x11	4	relu
Max Pooling	96	27x27x96	3x3	2	relu
2 Convolution	256	27x27x256	5x5	1	relu
Max Pooling	256	13x13x256	3x3	2	relu
3 Convolution	384	13x13x384	3x3	1	relu
4 Convolution	384	13x13x384	3x3	1	relu
5 Convolution	256	13x13x256	3x3	1	relu
Max Pooling	256	6x6x256	3x3	2	relu
6 FC	-	9216	-	-	relu
7 FC	-	4096	-	-	relu
8 FC	-	4096	-	-	relu
Output FC	-	1000	-	-	Softmax

In deep learning application, the ReLU activation feature is among the most popular. ImageNet is a large dataset of multimedia data annotated manually and divided into 1000 categories.

- **Googlenet inception 2014**

The inception module relies on several convolutions with reduced kernel size to drastically lower the number of parameters.

- **VGG16 2014**

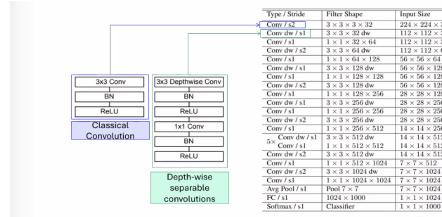
Characterized by a total of 16 layers with weights, that is, of parameters which are implemented.

- **ResNet50 2014**

Particularity: skip connections.

- **MobileNet V1**

MobileNet is designed to be a lightweight architecture optimized for mobile and embedded devices. It uses a new type of convolutional layer, known as Depthwise Separate convolution which comprises a depthwise convolution and a pointwise convolution.



Batch normalization:

$$\hat{x} = \frac{x - \mu}{\sigma},$$

where μ is the mean of x in mini-batch and σ is the std of x in mini-batch.

- **DensNet**

Densnet emphasizes strong feature reuse within the network through dense connections. Each

layer receives inputs from all preceding layers in a dense block. Dense connections include direct connections between layers within dense blocks, reducing the number of parameters and enabling better gradient flow. Growth rate controls the number of output feature maps produced by each layer within a dense block. Compound scaling: scaled the network's depth, width and resolution uniformly to find an optimal balance model size and accuracy.

3.4 Training a convolutional neural network

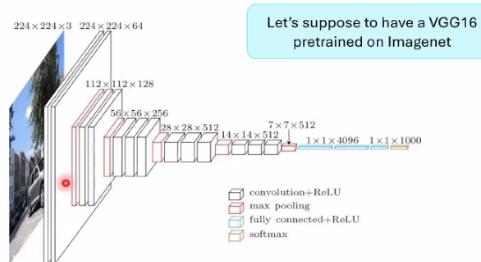
All previous CNNs have been trained on ImageNet to classify 1000 classes. This means that all CNN weights are available online. It's possible to use the pre-trained CNNs.

3.4.1 Transfer learning and fine tuning

It consists in taking the characteristics learned on a problem and exploiting them on a new similar problem. Transfer learning is usually used for tasks where the dataset has too little data to train a complete model from scratch. But how can we implement transfer learning?

What is done is adding new layers (trainable) of top of a set of layers previously frozen (so that the information contained in them is not destroyed during future training cycles) and then pass the new data set into the "new" network and record the output of one (or more) levels from the base model (this operation is called feature extraction). This result is used as input data for a new "smaller" model to be trained. The training of this model is called **fine-tuning**.

How many layers of the original model do I freeze? It depends on the task.



I can try to freeze all layers up to the last convolutional layer and train a mini-network composed of: convolutional block and 3 fully connected blocks

I could also think of not freezing any layer and initializing the net weights with the training weights on Imagenet.

Transfer learning is about "transferring" the representation learned during training of a CNN to another problem. For example, one can use pretrained CNN features to initialize the weights of a new CNN, developed for a different task. Fine tuning is about making fine adjustments. For

example, during transfer learning, you can unfreeze some of (or all) the pre-trained CNN layers and let it adapt more to the task at hand.

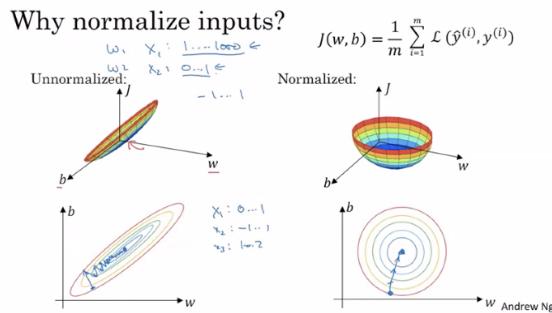
3.5 Data preparation

With the hold-out cross validation, i have:

- 70% training set (development of the model)
- 20% validation set (tuning and selection of the model)
- 10% testing set (reporting of results)

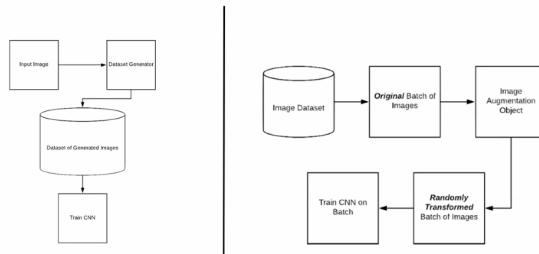
This is often done, unless my data are small size. I can't build statistics on my results, if I have few testing samples.

The data needs to be normalized in order to obtain better results.



3.5.1 Data augmentation

The aim is to increase artificially the dimension of the training set applying a series of transformations.



These transformations could be rotations, flips or a change of the brightness of the image itself. Be careful to consistency in the application of data augmentation techniques. I have to develop an algorithm for monitoring, for example, the movement of preterm infants. Suppose that the position and orientation of the room are fixed in relation to the child. A data augmentation that would be non sense to apply would be a rotation of the images.

Note that some bias of the original adding set persists.

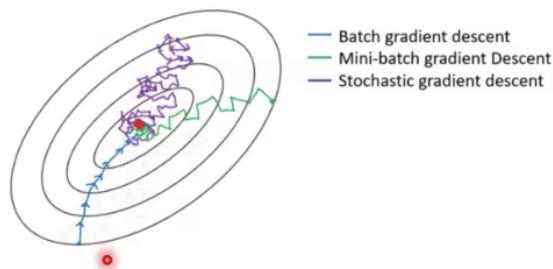
3.6 Optimizers

Tensor A tensor is an N-dimensional array of data (non è vero :_().

SGD (Stochastic gradient descent) is an optimization algorithm used to train machine learning algorithms. The algorithm's task is to find a set of parameters within the model that minimizes the loss or error function. Optimization is a type of research process and you can think of this research as learning. The optimization algorithm is called "gradient descent", where "gradient" refers to the calculation of an error gradient or slope of the error and "descent" refers to the movement along that slope towards a minimum level of error. The algorithm is iterative: **use the error to update the internal parameters of the model (back propagation)**.

A sample is individual data to which, in the case of supervised learning, a label is attached. This is used to compare the forecast and calculate an error. The batch size is a hyperparameter that defines the number of samples to be analyzed before updating the internal parameters of the model. A training dataset can be split into one or more batches. When the batch has sample size, the learning algorithm is called stochastic gradient descent. When the batch size is greater than a sample and less than the size of the training data set, the learning algorithm is called mini-batch gradient descent. The latter is the most common implementation used in the field of deep learning. Mini-batch gradient descent, the most common batch size are 32, 64 and 128.

In the batch gradient descent we use all the training data in a single iteration.



A training epoch means that the learning algorithm has made a pass through the training data

set.

Momentum is a technique used to help the optimizer go faster towards the optimal solution.

3.7 Loss function

For continuous outputs

$$MSE = \frac{1}{n} \sum_{i=1}^N (t_i - p_i)^2$$

while for categorical output we'll implement accuracy (used only for evaluation) and cross entropy (used during training)

$$L_{CE} = - \sum_{i=1}^{c=2} t_i \log p_i = -t_1 \log(p_1) - (1-t_1) \log(p_2).$$

The prediction is a probability vector: it represents the predicted probabilities of all classes with sum equal to 1. In a neural network, this prediction is usually made with the last layer activated by softmax function. We calculate the cross-entropy loss for the image. Loss is a measure of the performance of the model. The lower it is, the better. During learning, the model aims to achieve the lowest possible loss. The target represents the probability for all classes and is one-hot encoded vector, that is has 1 in a single position and 0 in all others. We'll start by calculating the loss for each class separately and then add it up. The loss for each class is calculated as follows:

$$Loss = -p(X) \ln(q(X))$$

where the first term is the probability of class X in target while the second is the probability of class X in predictions. If the probability for the target is 0, the loss is 0.

The loss is 0 if the prediction is 1. The loss tends to infinity if the prediction is 0.

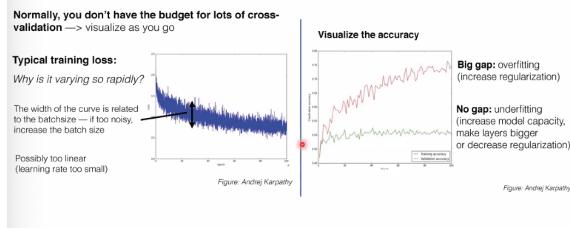
Cross entropy = $\sum_X p(X) \ln q(X)$. This is the cross entropy formula which can be used as a loss function for any two probability vectors. If we want to get the loss for our branch or for the whole set of data? You add up the losses of individual images. If our goal is one-hot encoded vector, we can actually forget the targets and predictions for all other classes and only calculate the loss for the hot class. Cross-entropy loss also works for distributions that are not one-hot vectors. In summary: **General formula used to calculate the loss between two probability vectors.**

The further we get from our target the more the error grows.

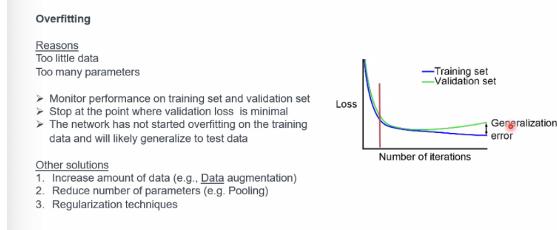
Binary classification = $-[p(X) \ln q(X) + (1-p(X)) \ln(1-q(X))]$ we use binary cross entropy, a specific case of cross entropy where our target is 0 or 1. It can be calculated with the cross entropy formula if you convert the target into a one-hot encoded vector, such as [0,1] or [1,0] and forecasts respectively. We can calculate it with the previous formula.

Multi-label classification = \sum_x Binary cross entropy_x Our target can represent multiple classes (or even zero) at the same time. We calculate the binary cross entropy for each class separately

and the add it up to get the complete loss.



Loss function- learning curves



3.8 Metrics

The **F1 score** is a popular metric for evaluating the performance of a classification model. In the case of multi-class classification, for the calculation, for example, of the F1 score, averaging methods are used, which translate into a series of different averages (macro, weighted, micro) in the rating report.

$$\text{F1 score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

In order to assess the performance of a model in a comprehensive way, we should examine both recall and accuracy. The F1 is a useful metric that considers both.

Instead of having multiple F1 scores per class, it would be better to average for a single number that describes overall performance.

The weighted average F1 score takes into account the support of each class. The weight refers to the proportion of the support for each class in relation to the sum of all the support values.

Micro average F1 score calculates the global mean F1 score by counting the sums of the true positives, false negatives and false positives:

$$\text{Micro-average F1 score} = \frac{TP}{TP + 0.5(FP + FN)}.$$

The micro-average F1 score calculates the percentage of correctly ranked observations on all observations.

4 AI ETHICS: ethics of artificial intelligence -5/04/2025

This philosophical discipline discusses:

- AI impact on society,
- Values and principles of AI,
- Guide lines to ethically design AI systems.

Ethics for AI

- Foundation of ethics of AI
 - Unpacking moral concepts → diverse dimensions = diverse risks
 - Unveiling the values at the core of the ethical principles → **FRIA (EU AI Act)**
- Avoid:
 - Ethical principles as labels → checklist
 - Phenomena:
 - * Ethics washing (and fake ethical self-regulation)
 - * Ethics bashing (trivialization)
 - Criticism: ethics toothless and uselessness
 - Harm and reputation risk
- Ensure:
 - Adequate risk-based assessment (detection and management)
 - Fundamental rights and values impact assessment → **FRIA (EU AI Act)**
 - Develop actionable ethics by design criteria

Benchmarking AI ethics frameworks and principles

+200 ethical frameworks & guidelines:

- EU HLEGAI (2019)
- WHO (2021)
- UNESCO (2021)

Benchmarking AI ethics frameworks and principles

- Benevolence → benefit people/ minimize harm

- Autonomy → AI systems that respect/promote human autonomy
- Justice as fairness → fair AI systems, which do not discriminate
- Explicability → AI systems whose functioning and outputs are explicable
- Accountability & Responsibility → Determine accountability mechanisms and harm redress
- Sustainability → (human and environmental) sustainability

Explicability = understand how AIs achieve a certain goal

- Black Box (Pasquale 2015)
- Transparency(?)
- Explicability ≠ Transparency
- Transparency = Accessibility
- Explicability = Intelligibility
- EXAI

Accountability & Responsibility

- Multi-agent-context = designers, company providers, users - who is responsible if something goes wrong?
- **Accountability mechanisms:**
 - Auditability:
 - * Risk based assessment
 - * Develop ex ante risk scenarios and guidelines for responsibility distribution (shared responsibility)
 - Harm redress:
 - * Develop harm redress mechanisms (people affected negatively should be redressed)

(environmental and human) Sustainability

- Environmental sustainability:
 - Computational power for training models + water for cooling data centers, etc.
 - Carbon footprint → assess the environmental impact
 - Do we adopt technology that enable me to do the same task, but in a more sustainable way?

- Human sustainability
 - Impact of AI on human resources
 - Does this product affect employment? Does it create new or eliminate jobs? of whom? what kind of work? of whom?
 - Sustainability a 360° = impact assessment before implementing AI

Human Autonomy as an ethical value

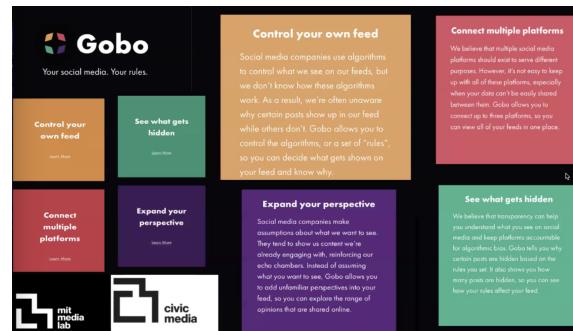
- Respect the individuals as decision-makers
- To choose and act according to their own beliefs, interests, preferences
- Freedom of choice/ human agency
- Necessary conditions:
 - availability of alternative options
 - autonomy as rational deliberation or endorsement of what we choose

Algorithmic governance → pervasive SNS, IoT, etc.

Algorithms = information gatekeepers → choose what information of any kind we can see, what to show us, etc.

Algorithmic choice architecture = they decide what information (contact, opportunity, etc.) we can access to and influence our choices and actions.

Our choice options are always reduced



Autonomy as rational deliberation or endorsement of what we choose.

- Recommender system of highly personalized informational content
- Chosen on the basis of how we respond to the presentation of the sensitive information inferred/discovered on us (profiling: content-based or collaborative based)

- ML can infer choice-driving/eliciting elements or emotionally loaded information
- Use them to nudge our behavior toward pre-defined goals
- Rational deliberation bypassing = our choices/actions/behavior risk to be pre-determined by third parties

ML and DL profiling ad personalization should be ethical:

- We need them to navigate informational overload
- Limited time and cognitive resources
- Based on accurate profile of ourselves, disclosable to the user (do not use people as means: no exploitation of sensitive information).
- **Automation bias**
- Bias by proxy

Problematic: increasing delegation to such systems tasks and decisions → biased/discriminate

Technical solutions:

- Avoid/remove protected attributes
- Cases where protected characteristics are crucial to make fair decisions

Parity models, ensuring the measures of its predictive performance are equal across protected groups = algorithm performance is equal/ treatment is same for diverse groups.

Formal/procedural fairness (ensured by the law): formal equality of opportunity.

Fairness as an ethical value

Substantial fairness: compensate inequalities

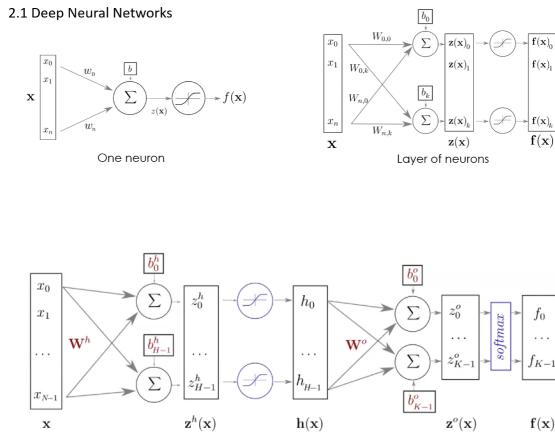
- Using AI to investigate sources of unfair bias in socio-relational conditions
- Design systems capable to compensate them
- Ethics by design (equality of opportunity as a criterion)

5 Deep generative models -11/04/2025

The aim is to understand the structure of DGM, use them, get a general knowledge of model training and as well of fine-tune issues.

In generative models data-driven methods are needed. A first example can be distribution modeling: **It's possible to generate new data given a specific distribution.** The goal is to take as input training samples from some distribution and learn a model that represents that distribution.

5.1 Preliminary concepts



As we saw we implement a loss function as

$$l(f(x^s; \theta), y^s) = nll(x^s, y^s; \theta) = -\log f(x^s; \theta)_{y^s}$$

and the training consists in finding the parameters ($\mathbf{W}^h, \mathbf{b}^h, \mathbf{W}^o, \mathbf{b}^o$) which minimize negative log likelihood. We also saw how we can implement algorithms both supervised and unsupervised.

5.1.1 PDF and joint probability

Given $f(x) = \frac{dF(x)}{dx}$, where $F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx$, we have that $f(x)$ is the PDF and $F(x)$ is the CDF. While the PDF is a function of a single variable, the joint probability is a function of two or more variables. What we have for the joint probability is:

$$P(X_1 \in [a_1, b_1], \dots, X_k \in [a_k, b_k]) = \int_{a_1}^{b_1} \dots \int_{a_k}^{b_k} f_X(x_1, \dots, x_k) dx_k \dots dx_1$$

5.1.2 Maximum likelihood estimation (MLE) & Maximum a Posteriori Estimation (MAP)

The PDF is our model, the question is: how do we estimate the parameters? If the PDF would be Gaussian we would use mean, std, etc. Otherwise we can implement MLE and MAP which are both methods able to estimate the parameters θ of a statistical model based on observed data \mathbf{X} .

MLE Find θ that **maximize** the likelihood $p(X|\theta)$:

$$\hat{\theta}_{MLE} = \arg \max_{\theta} p(X|\theta).$$

Are we making any hypothesis on the prior distribution on θ ? The answer is no. What we have is

$$\theta_{MLE} = \arg \max_{\theta} p(X|\theta) = \arg \max_{\theta} \prod_i p(x_i|\theta).$$

or also, using the logarithm

$$\theta_{MLE} = \arg \max_{\theta} \log P(X|\theta) = \arg \max_{\theta} \log \prod_i P(x_i|\theta) = \arg \max_{\theta} \sum_i \log P(x_i|\theta)$$

Now we maximize w.r.t. θ with our "favorite" optimization algorithm (Gradient descent, PSO, Grid search, etc...).

MAP From a bayesian perspective, includes a prior probability $p(\theta)$ that reflects our belief about parameter values.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \propto P(X|\theta)P(\theta).$$

We compare these two methods:

2.2 MLE and MAP	
Maximum a Posteriori Estimation	
MLE	MAP
$\theta_{MLE} = \arg \max_{\theta} \log P(X \theta)$	$\theta_{MAP} = \arg \max_{\theta} P(X \theta)P(\theta)$
$= \arg \max_{\theta} \log \prod_i P(x_i \theta)$	$= \arg \max_{\theta} \log \prod_i P(x_i \theta) + \log P(\theta)$
$= \arg \max_{\theta} \sum_i \log P(x_i \theta)$	$= \arg \max_{\theta} \sum_i \log P(x_i \theta) + \log P(\theta)$
	$= \arg \max_{\theta} \sum_i \log P(x_i \theta) - \log P(\theta)$
What happens if $\log P(\theta)$ is constant?	

The logarithm $\log P(\theta)$ gives a constant contribution while $P(\theta)$ is constant. That means that MAP coincides with MLE.

In general MLE is simple to implement and is more prone to overfitting if there is limited data

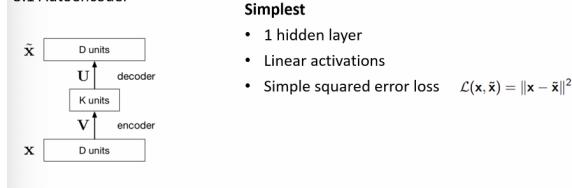
and no regularization/prior is used. However, MAP can help reduce overfitting by imposing a prior (often act as a regularization term) and is useful when we have domain knowledge or want to encode certain parameter restrictions.

To calculate $P(Y|X)$, the first estimate the prior probability $P(Y)$ and the likelihood probability $P(Y|X)$ from the data provided.

$$posterior = \frac{prior \times likelihood}{evidence} \rightarrow P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}.$$

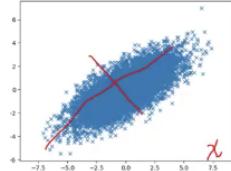
5.2 Variational AutoEncoders (VAE)

An autoencoder is feed-forward neural net whose job is to take an input x and predict \tilde{x} . To make this non-trivial, we need to add a bottleneck layer whose dimension is much smaller than the input.



The simplest autoencoder is composed only of 1 layer, has liner activations and the loss function is simply $L(x, \tilde{x}) = \|x - \tilde{x}\|^2$. This network computer $\tilde{x} = UVx$, which is a linear function. If $k \geq D$, see can choose U and V such that UV is the identity. This isn't very interesting.

So a linear autoencoder with squared error loss is equal to the PCA (principal component analysis). Sample covariance of data \rightarrow minimize covariance \rightarrow diagonal matrix \rightarrow eigenvalues.



The **Principal Component Analysis (PCA)** is a statistical technique used to reduce the dimensionality of a dataset while preserving most of the original variance. PCA finds a linear representation of the data in a lower-dimensional space, where the new variables (the *principal components*) are linear combinations of the original variables. The goal of PCA is to find the principal components that maximize the variance of the data along each axis, in order to obtain a more compact representation of the data.

Mathematically, the PCA process can be formulated as follows:

1. **Centering the data:** Subtract the mean of each variable from the data:

$$X' = X - \mu$$

where $X \in \mathbb{R}^{n \times d}$ is the data matrix (with n observations and d variables), and μ is the mean vector of each column of X .

2. **Calculating the covariance matrix:**

$$S = \frac{1}{n} X'^T X'$$

where $S \in \mathbb{R}^{d \times d}$ is the covariance matrix of the centered data.

3. **Eigenvectors and eigenvalues:** Solve the eigenvalue problem to obtain the **principal components** (eigenvectors) and the variance (eigenvalues):

$$Sw = \lambda w$$

where w is an eigenvector and λ is the associated eigenvalue.

4. **Projection onto the first k principal components:** The original data is projected onto the principal component space, reducing the dimensionality from d to k :

$$Z = X' W_k$$

where $W_k \in \mathbb{R}^{d \times k}$ is the matrix containing the first k eigenvectors (principal components), and $Z \in \mathbb{R}^{n \times k}$ is the matrix of the projected data.

Main limitation: autoencoders and deep autoencoders do not define a distribution → they do not generate anything (they reproduce).

The encoder in generative models is the component that maps the data into a latent space, which becomes the foundation for generating new instances of data through the decoder or the generator. In models like the VAE, the encoder not only reduces the dimensionality but also transforms the input into a probability distribution that allows generating new data consistent with the original distribution.

The loss function is a negative log-likelihood with a regularization term (for each i^{th} sample). In the formula of the log-likelihood, E is the expectation taken w.r.t. to the encoder's distribution over the representations → encourages to learn the data.

- **Encoder:**

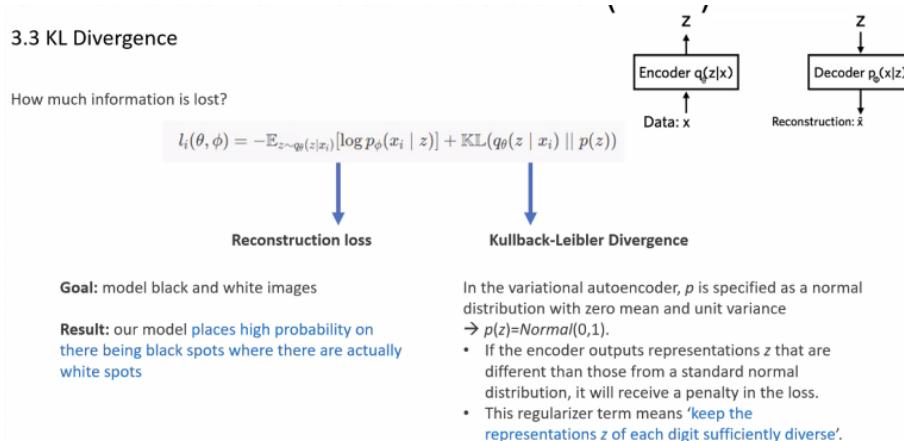
- Input x , hidden output z , reconstruction x
- Encoder is NN with weights/biases θ
- Input 28 x 28 pixel photo → 784- dimensional representation
- Hidden size $\text{len}(z)$ is much less than 784

- Efficient compression with a probability density $q_\theta(z|x)$
- From this distribution, you can sample noisy values of z!

- **Decoder:**

- Input x, hidden output z, reconstruction x
- Decoder is NN with weights/bias Φ
- Its input is z and has decoding function $p_\Phi(x|z)$

In the example we were following before, photo are black and white and represent each pixel as 0 or 1. The probability distribution of a single pixel can be then represented using a PDF (i.e., Bernoulli distribution in this case). The decoder gets as input the latent representation of a digit z and outputs 784 Bernoulli Parameters (one for each of 784 pixels in the image).



The first term represents the **Reconstruction loss**. E is the expectation taken w.r.t. the encoder's distribution over the representations. The second term is the **kullback-Leibler Divergence**. Kullback-Leibler (KL) divergence is a measure of how one probability distribution diverges from a second, expected probability distribution. It is defined as:

$$D_{\text{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

for discrete variables, or

$$D_{\text{KL}}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

for continuous variables. KL divergence is always non-negative and equals zero if and only if the two distributions are the same almost everywhere.

We now move to the **training** part: that is done using backpropagation.

Probabilistic interpretation A variational autoencoder contains a specific probability model of a data x and latent variables z . We can write the joint probability of the model as:

$$P(x, z) = P(x|z)P(z)$$

The generative process:
for each datapoint i

- Draw $z_i \sim p(z) \rightarrow$ draw from a prior
- Draw $x_i \sim p(x|z) \rightarrow$ draw from a probability (Bernoulli distributed for b/w images)

The aim is to infer good values of z given x , or to calculate $P(z|x)$. To calculate the latter we need $P(x)$, but how can I get this probability? That is the prior probability of the data. We can marginalize the latent variables: $P(x) = \int P(x|z)P(z)dz$. However this integral takes time. We know that $q_\lambda(z|x)$ (from the encoder) has a parameter λ which indexes the family distribution. If q were gaussian it would be mean and variance of each data point $\lambda_i = (\mu_{x_i}, \sigma_{x_i})$. We call from now on q : Variational Posterior.

3.5 Probabilistic interpretation

How can we know how well our variational posterior $q_\lambda(z|x)$ approximates the true $p(z|x)$?

Again, KL Divergence

$$\text{KL}(q_\lambda(z|x) || p(z|x)) = E_q[\log q_\lambda(z|x)] - E_q[\log p(x,z)] + \log p(x)$$

But how do we again compute $p(x)????$

Proof is left to exercise (use slide 42)

Define Evidence Lower Bound, ELBO

$$ELBO(\lambda) = E_q[\log p(x,z)] - E_q[\log q_\lambda(z|x)]$$

$$\log P(x) = ELBO(\lambda) = KL(q_\lambda(z|x)||p(z|x)).$$

So if we know that $KL \geq 0$, instead that minimizing KL we can maximize ELBO \rightarrow avoid P(x). To sum up, we recall that

$$ELBO_i(\lambda) = E_{q_\lambda(z|x_i)}[\log p(x_i|z)] - KL(q_\lambda(z|x_i)||p(z)).$$

Let's remember now NN \rightarrow the final step is:

- Parametrize the approximate posterior $q_\theta(z|x, \lambda)$ with an inference network (encoder)

- Parametrize the likelihood $p(x|z)$ with a generative network (decoder) that takes latent variables z and outputs parameters to the data distribution $p_\theta(x|z)$
- $ELBO_i(\theta, \phi) = -l_i(\theta, \phi)$
- The inference and generative networks have model parameters (θ, ϕ) , while λ represents variational parameters

3.7 Other families of variational bounds

Variational Free Energy

$$\mathcal{F}(\mathbf{x}, q) = \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z})] - KL[q(\mathbf{z})||p(\mathbf{z})]$$

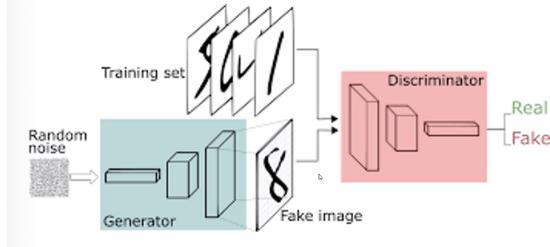
Multi-sample Variational Objective

$$\mathcal{F}(\mathbf{x}, q) = \mathbb{E}_{q(z)} \left[\log \frac{1}{S} \sum_s \frac{p(\mathbf{z})}{q(\mathbf{z})} p(\mathbf{x}|\mathbf{z}) \right]$$

Renyi Variational Objective

$$\mathcal{F}(\mathbf{x}, q) = \frac{1}{1-\alpha} \mathbb{E}_{q(z)} \left[\left(\log \frac{1}{S} \sum_s \frac{p(\mathbf{z})}{q(\mathbf{z})} p(\mathbf{x}|\mathbf{z}) \right)^{1-\alpha} \right]$$

5.3 Generative Adversarial Networks (GAN)

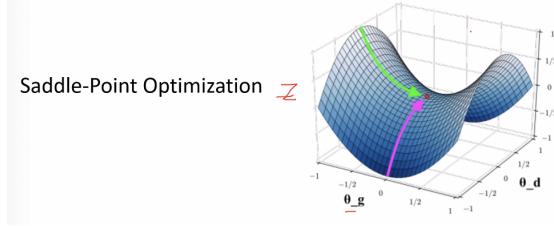


We start by generating a batch of samples (GAN) and then we update the weights based on the results.

Let z and $p(z)$ be a simple base distribution and its probability. The generator $G(z)$ is a deep neural network. D is the manifold and the discriminator $D(x)$ is also a deep neural network. The parameters optimization is given by:

$$\min_{\theta_d} \max_{\theta_g} \underbrace{E_{p_{data}}[\log D_{\theta_d}(x)]}_{\text{likelihood of true data}} + \underbrace{E_{p_z(z)}[\log(1 - D_{\theta_d}(G_{\theta_g}))]}_{\text{likelihood of generated data}}$$

where the first term is the likelihood of true data and the second term is the likelihood of generated data.



An example of training is given:

```

Training
for i = 1 ... N do
    for k = 1 ... K do
        • Sample noise samples {z1, ..., zm} ~ pz(z)
        • Sample examples {x1, ..., xm} from pdata(x).
        • Update the discriminator Dθd:
            
$$\theta_d = \theta_d - \alpha_d \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\log D(x^i) + \log(1 - D(G(z^i)))]$$

    end for
    • Sample noise samples {z1, ..., zm} ~ pz(z).
    • Update the generator Gθg:
        
$$\theta_g = \theta_g - \alpha_g \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z^i)))$$

end for

```

For vanishing gradient, if D is close to perfect, min-max saturates →

- possibilities for non-saturating objectives;
- restrict capability of D (structurally);
- Balance learning D_{θ_d} and G_{θ_g} .

The **mode-collapse** happens when the generator focuses on producing a limited set of data patterns that deceive the discriminator. What happens is that it becomes fixated on a few dominant modes in the training data and fails to capture the full diversity of the data distribution.

Wasserstein GANs → use Wasserstein distance instead than cross-entropy. The Wasserstein-1 distance (or Earth-Mover's distance) is continuous and differentiable almost everywhere with respect to the generator's parameters. Even when generated data has little overlap with the real distribution → W-distance provides non-zero gradients → Always feedback → Steers away from vanishing gradients and thus form → mode collapse.

The Wasserstein-1 distance between two probability distributions P_r (real data) and P_g (generated data) is defined as:

$$W(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|]$$

Here, $\Pi(P_r, P_g)$ represents the set of all joint distributions $\gamma(x, y)$ whose marginals are P_r and P_g , respectively. Intuitively, $W(P_r, P_g)$ measures the minimum "cost" of transporting mass to change one distribution into the other.

Kantorovich-Rubinstein Duality

Directly computing the infimum in the above formulation is intractable in high dimensions. The Kantorovich-Rubinstein duality offers an alternative representation:

$$W(P_r, P_g) = \sup_{\substack{\|f\|_L \leq 1}} (\mathbb{E}_{x \sim P_r} [f(x)] - \mathbb{E}_{x \sim P_g} [f(x)])$$

In this formulation:

- f is any function that is **1-Lipschitz** (i.e., $|f(x_1) - f(x_2)| \leq |x_1 - x_2|$ for all x_1, x_2).
- The supremum is taken over all such functions f .

This dual form transforms the original problem into finding a function f (often referred to as the *critic* in the GAN framework) that maximizes the difference in expectations between the real and generated distributions, under the constraint that f remains 1-Lipschitz.

A critical challenge in implementing WGANs is ensuring that the critic remains 1-Lip. Two common methods are used:

- Weight clipping: Constraining the weights of the critic network to lie within a compact space. This is a crude method that can sometimes limit the capacity of the critic.
- Gradient Penalty (WGAN-GP): A more robust technique where an additional penalty term is added to the critic's loss function:

$$\lambda E_{\tilde{x} \sim P_{\hat{x}}} [(||\nabla_{\hat{x}} f(\hat{x})||_2 - 1)^2]$$

Here, \hat{x} is sampled from a distribution defined along straight lines between real and generated samples, and λ is a hyperparameter controlling the strength of the penalty.

Non-convergence D_{θ_d} and G_{θ_d} are a saddle, that gives a more common non-convergence → saddle is complex to find. Training updates indeed undo each other. To solve this problem one can use:

- Soft and Noisy labels

- DCGAN/Hybrid models
- Stability tricks from RL
- ADAM optimizer

5.4 Diffusion Models (DM)

Based of the diffusion process: two steps (forward & backward). Add gradually to X_0 small amounts of Gaussian noise. Proceed as in Markov Chain ($t+1$ depends only on t).

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{(1-\beta_t)}x_{\{t-1\}}, \beta_t I)$$

with N normal distribution and β the noise schedule (i.e., how much noise is added).

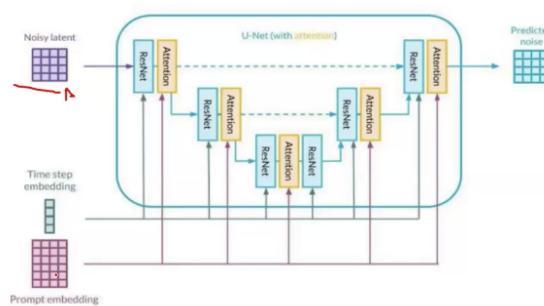
So if N is a normal distribution, our objective is to predict mean and variance to predict the reverse process (invert forward) \rightarrow generate data.

Reverse $\rightarrow p$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

We need to learn this distribution, how?

- Old attempts: with math (Denoising Probabilistic Models)
- New attempts: with Unets (Den. Diffusion Prob. Models)



Training: Minimizing Variational Lower Bound (As VAE...)

$$E[-\log p_\theta(x_\theta)] \leq E_q[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}]$$

The problem is: we have $t/t-1$ process, how much time will it take?

Check out: ControlNet

6 Deep Network Training and Testing -18/04/2025

In 3D image generation the starting point is a multipart view.

6.1 Pytorch practice

Pytorch is an open source machine learning framework based on the Python programming language and the Torch library.

▼ Fibonacci Simple Example Training

Exercise 0

▼ Step0: Import dependencies

```
import torch
import torch.nn as nn
import torch.optim as optim
import torch.nn.functional as F
```

▼ Step1: Get the data

```
train_features, train_targets = [], []
test_features, test_targets = [], []

# generate fibonacci sequence pairs - train
for i in range(7):
    if i == 0:
        train_features.append([i, 1.0])
        train_targets.append([1.0])
        continue

    sum = train_features[-1][1] + train_targets[-1][0]
    train_features.append([train_features[-1][1], train_targets[-1][0]])
    train_targets.append([sum])

# generate fibonacci sequence pairs - test (last index + 1)
sum = train_features[-1][1] + train_targets[-1][0]
test_features.append([train_features[-1][1], train_targets[-1][0]])
test_targets.append([sum])

print(f"Train Features: {train_features}")
print(f"Train Targets: {train_targets}")
print(f"Test Features: {test_features}")
print(f"Test Targets: {test_targets}")

→ Train Features: [[0, 1.0], [1.0, 1.0], [1.0, 2.0], [2.0, 3.0], [3.0, 5.0],
Train Targets: [[1.0], [2.0], [3.0], [5.0], [8.0], [13.0], [21.0]]
Test Features: [[13.0, 21.0]]
Test Targets: [[34.0]]
```

```
# TRAIN: Define your input and target tensors
train_features_tensor = torch.tensor(train_features)
train_target_tensor = torch.tensor(train_targets)

# TESTING: Define your input and target tensors
test_features_tensor = torch.tensor(test_features)
test_target_tensor = torch.tensor(test_targets)

# print shapes
print(f"Train Features Shape: {train_features_tensor.shape}")
print(f"Train Targets Shape: {train_target_tensor.shape}")

print(f"Test Features Shape: {test_features_tensor.shape}")
print(f"Test Targets Shape: {test_target_tensor.shape}")

→ Train Features Shape: torch.Size([7, 2])
    Train Targets Shape: torch.Size([7, 1])
    Test Features Shape: torch.Size([1, 2])
    Test Targets Shape: torch.Size([1, 1])
```

▼ Step3: Define the model

```
class LinearRegression(nn.Module):
    def __init__(self):
        super(LinearRegression, self).__init__()
        self.linear = nn.Linear(2, 1) # Two input feature and single output fe

    def forward(self, x):
        return self.linear(x)

model1 = LinearRegression()
```

▼ Step4: Define training params

```
# Define the loss function
criterion = nn.MSELoss()

# Define the optimizer (SGD with learning rate 0.01 and momentum 0.9)
optimizer = optim.SGD(
    [
        {'params': model1.parameters()},
    ],
    lr=0.01,
    momentum=0.9
)
```

▼ Step5: Train the model

```
# Training loop
for epoch in range(10):
    # Forward pass
    output_tensor1 = model1(train_features_tensor)

    # Compute the loss
    loss1 = criterion(output_tensor1, train_target_tensor)

    # Backward pass and optimization
    optimizer.zero_grad() # Zero the gradients
    loss1.backward() # Compute the gradients for model1
    optimizer.step() # Update the models' parameters

    # Print the losses for monitoring the training progress
    print(f"Epoch {epoch+1}, Loss 1: {loss1.item()}")
```

→ Epoch 1, Loss 1: 92.6393814086914
Epoch 2, Loss 1: 0.710472583770752
Epoch 3, Loss 1: 87.32957458496094
Epoch 4, Loss 1: 46.77772903442383
Epoch 5, Loss 1: 8.138298034667969
Epoch 6, Loss 1: 71.7966537475586
Epoch 7, Loss 1: 18.664520263671875
Epoch 8, Loss 1: 16.93732261657715
Epoch 9, Loss 1: 52.307640075683594
Epoch 10, Loss 1: 4.732697486877441

▼ Step6: Testing Prediction

45

```
prediction = model1(test_features_tensor)
print(f"Prediction: {prediction.item()}")

# calculate error mse
error = (prediction.item() - test_target_tensor.item()) ** 2
print(f"Error: {error}")

→ Prediction: 18.125612258911133
    Error: 251.9961861544325
```

▼ Optional - Step7 - Introspection

```
params_generator = model1.named_parameters()
W_all = next(params_generator)
W_name = W_all[0]
W = W_all[1].detach()[0]
b_all = next(params_generator)
b_name = b_all[0]
b = b_all[1].detach()
print(f"W ({W_name}): {W}")
print(f"b ({b_name}): {b}")

→ W (linear.weight): tensor([0.8981, 0.3096])
    b (linear.bias): tensor([-0.0515])
```

```
pred = test_features_tensor @ W + b
print(f"Prediction: {pred.item()}")
```

```
→ Prediction: 18.125612258911133
```

```
input_tensor = torch.tensor([34.0, 21.0]) # Change the data type to float32

model1(input_tensor)

→ tensor([36.9847], grad_fn=<ViewBackward0>)
```



```
import copy

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import torch
import torch.nn as nn
import torch.optim as optim
import tqdm
from sklearn.model_selection import train_test_split
from sklearn.datasets import fetch_california_housing
```

Exercise 1

```
# Read data
data = fetch_california_housing()
X, y = data.data, data.target

print("Number of elements", len(X))
```

→ Number of elements 20640

```
# visualize X
pd.DataFrame(X).head()
```

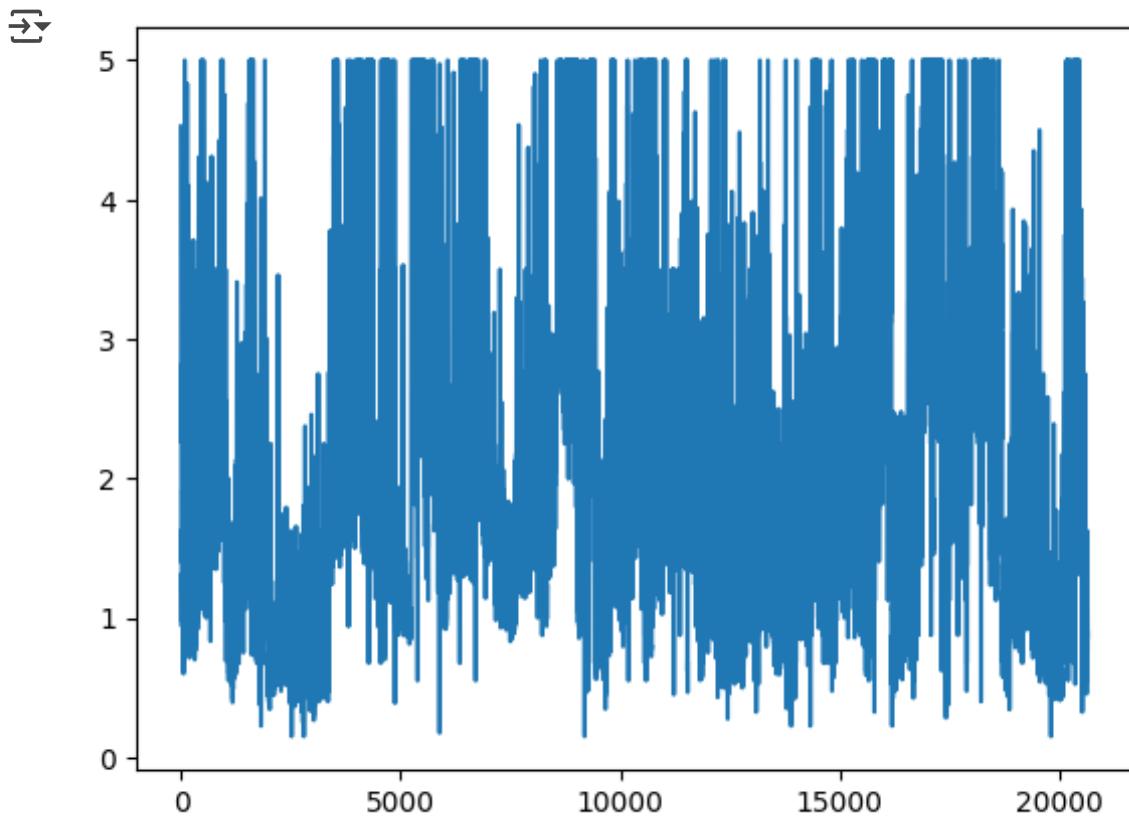
→

	0	1	2	3	4	5	6	7
0	8.3252	41.0	6.984127	1.023810	322.0	2.5555556	37.88	-122.23
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25

```
# visualize y  
pd.DataFrame(y).head()
```

```
0  
0 4.526  
1 3.585  
2 3.521  
3 3.413  
4 3.422
```

```
# plot y  
plt.plot(y)  
plt.show()
```



```
# train-test split for model evaluation
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, shuffle=True)

# Convert to 2D PyTorch tensors
X_train = torch.tensor(X_train, dtype=torch.float32)
y_train = torch.tensor(y_train, dtype=torch.float32).reshape(-1, 1)
X_test = torch.tensor(X_test, dtype=torch.float32)
y_test = torch.tensor(y_test, dtype=torch.float32).reshape(-1, 1)

# check number of elements
print("Number of elements in X_train", len(X_train))
print("Number of elements in y_train", len(y_train))

print("Number of elements in X_test", len(X_test))
print("Number of elements in y_test", len(y_test))

print("len(X) = len(X_train) + len(X_test)")
print(len(X) == len(X_train) + len(X_test))
```

→ Number of elements in X_train 14447
Number of elements in y_train 14447
Number of elements in X_test 6193
Number of elements in y_test 6193
len(X) = len(X_train) + len(X_test)
True

```
# Define the model
# model = nn.Sequential(
#     nn.Linear(8, 24),
#     nn.ReLU(),
#     nn.Linear(24, 12),
#     nn.ReLU(),
#     nn.Linear(12, 6),
#     nn.ReLU(),
#     nn.Linear(6, 1)
# )

model = nn.Sequential(
    nn.Linear(8, 12),
    nn.ReLU(),
    nn.Linear(12, 6),
    nn.ReLU(),
    nn.Linear(6, 1)
)

# loss function and optimizer
loss_fn = nn.MSELoss() # mean square error
optimizer = optim.Adam(model.parameters(), lr=0.0001)
# Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization"

n_epochs = 100 # number of epochs to run
batch_size = 10 # size of each batch
batch_start = torch.arange(0, len(X_train), batch_size)

batch_start
→ tensor([ 0, 10, 20, ..., 14420, 14430, 14440])
```

```
# check elements in batch
print("len(batch_start) = len(X_train) // batch_size")
print(len(batch_start) == len(X_train) // batch_size)

print("---- WHY ? ----")

print(batch_start[-1])
print(batch_start[-2])
print("Length of dataset", len(X_train))

print(len(X_train) == len(batch_start) * batch_size)
print(len(X_train) == len(batch_start) * batch_size - 3)

→ len(batch_start) = len(X_train) // batch_size
False
---- WHY ? ----
tensor(14440)
tensor(14430)
Length of dataset 14447
False
True

# Hold the best model
best_mse = np.inf # init to infinity
best_weights = None
history = []

for epoch in range(n_epochs):
    model.train()
    with tqdm.tqdm(batch_start, unit="batch", mininterval=0, disable=False) as
        bar.set_description(f"Epoch {epoch}")
        for start in bar:
            # take a batch
            X_batch = X_train[start:start+batch_size]
            y_batch = y_train[start:start+batch_size]
            # forward pass
            y_pred = model(X_batch)
            loss = loss_fn(y_pred, y_batch)
            # backward pass
            optimizer.zero_grad()
            loss.backward()
            # update weights
            optimizer.step()
            # print progress
            bar.set_postfix(mse=float(loss))
    # evaluate accuracy at end of each epoch
    model.eval()
```

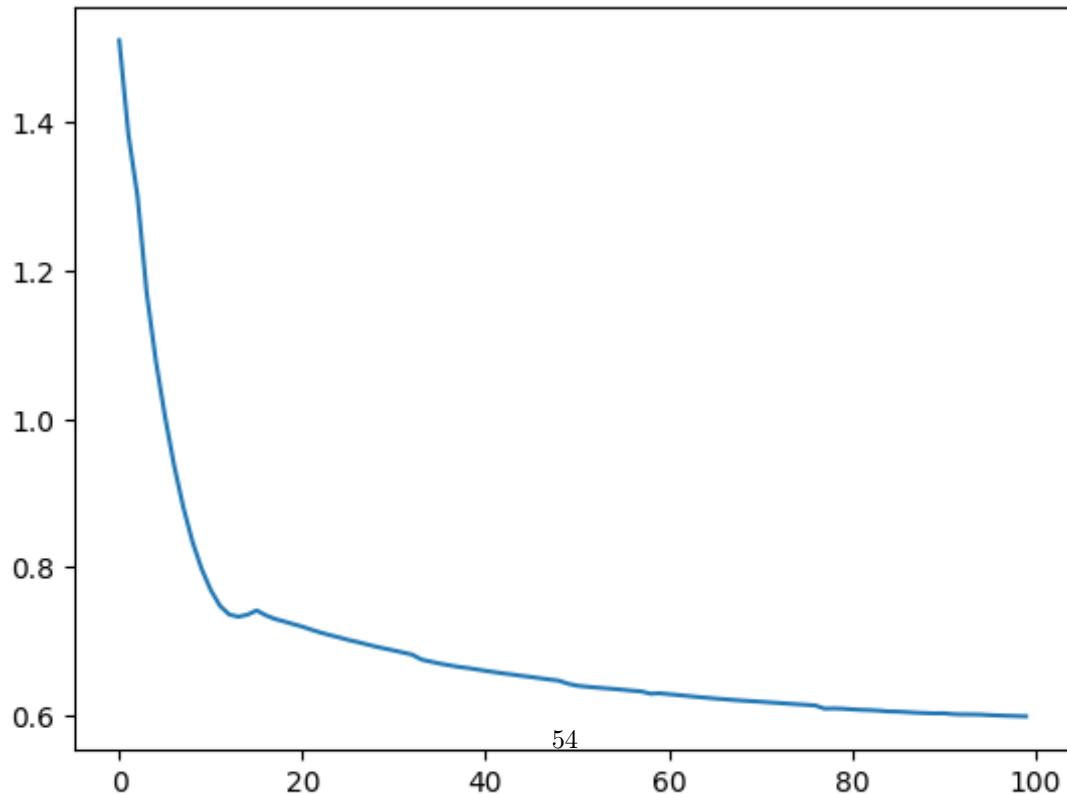
```
y_pred = model(X_test)
mse = loss_fn(y_pred, y_test)
mse = float(mse)
history.append(mse)
if mse < best_mse:
    best_mse = mse
    best_weights = copy.deepcopy(model.state_dict())
```

→ Epoch 0: 100%	██████████	1445/1445 [00:06<00:00, 233.58batch/s, mse=0.877]
Epoch 1: 100%	██████████	1445/1445 [00:06<00:00, 221.91batch/s, mse=0.851]
Epoch 2: 100%	██████████	1445/1445 [00:05<00:00, 244.93batch/s, mse=0.83]
Epoch 3: 100%	██████████	1445/1445 [00:06<00:00, 226.86batch/s, mse=0.728]
Epoch 4: 100%	██████████	1445/1445 [00:06<00:00, 229.80batch/s, mse=0.636]
Epoch 5: 100%	██████████	1445/1445 [00:06<00:00, 216.13batch/s, mse=0.559]
Epoch 6: 100%	██████████	1445/1445 [00:06<00:00, 220.11batch/s, mse=0.479]
Epoch 7: 100%	██████████	1445/1445 [00:06<00:00, 236.69batch/s, mse=0.406]
Epoch 8: 100%	██████████	1445/1445 [00:06<00:00, 233.89batch/s, mse=0.344]
Epoch 9: 100%	██████████	1445/1445 [00:06<00:00, 216.50batch/s, mse=0.295]
Epoch 10: 100%	██████████	1445/1445 [00:06<00:00, 234.38batch/s, mse=0.258]
Epoch 11: 100%	██████████	1445/1445 [00:06<00:00, 229.39batch/s, mse=0.232]
Epoch 12: 100%	██████████	1445/1445 [00:06<00:00, 236.06batch/s, mse=0.215]
Epoch 13: 100%	██████████	1445/1445 [00:06<00:00, 229.49batch/s, mse=0.203]
Epoch 14: 100%	██████████	1445/1445 [00:06<00:00, 230.69batch/s, mse=0.195]
Epoch 15: 100%	██████████	1445/1445 [00:06<00:00, 231.49batch/s, mse=0.19]
Epoch 16: 100%	██████████	1445/1445 [00:06<00:00, 210.46batch/s, mse=0.185]
Epoch 17: 100%	██████████	1445/1445 [00:06<00:00, 231.96batch/s, mse=0.182]
Epoch 18: 100%	██████████	1445/1445 [00:06<00:00, 221.99batch/s, mse=0.179]
Epoch 19: 100%	██████████	1445/1445 [00:06<00:00, 224.34batch/s, mse=0.176]
Epoch 20: 100%	██████████	1445/1445 [00:07<00:00, 205.72batch/s, mse=0.174]
Epoch 21: 100%	██████████	1445/1445 [00:06<00:00, 220.46batch/s, mse=0.171]
Epoch 22: 100%	██████████	1445/1445 [00:06<00:00, 220.36batch/s, mse=0.169]
Epoch 23: 100%	██████████	1445/1445 [00:06<00:00, 219.73batch/s, mse=0.167]
Epoch 24: 100%	██████████	1445/1445 [00:06<00:00, 238.93batch/s, mse=0.165]
Epoch 25: 100%	██████████	1445/1445 [00:06<00:00, 226.14batch/s, mse=0.163]
Epoch 26: 100%	██████████	1445/1445 [00:06<00:00, 217.89batch/s, mse=0.161]
Epoch 27: 100%	██████████	1445/1445 [00:06<00:00, 217.86batch/s, mse=0.16]
Epoch 28: 100%	██████████	1445/1445 [00:06<00:00, 217.70batch/s, mse=0.158]
Epoch 29: 100%	██████████	1445/1445 [00:06<00:00, 222.57batch/s, mse=0.156]
Epoch 30: 100%	██████████	1445/1445 [00:06<00:00, 215.00batch/s, mse=0.155]
Epoch 31: 100%	██████████	1445/1445 [00:06<00:00, 218.28batch/s, mse=0.153]
Epoch 32: 100%	██████████	1445/1445 [00:06<00:00, 208.54batch/s, mse=0.152]
Epoch 33: 100%	██████████	1445/1445 [00:06<00:00, 219.01batch/s, mse=0.15]
Epoch 34: 100%	██████████	1445/1445 [00:06<00:00, 218.01batch/s, mse=0.149]
Epoch 35: 100%	██████████	1445/1445 [00:06<00:00, 217.57batch/s, mse=0.147]
Epoch 36: 100%	██████████	1445/1445 [00:06<00:00, 214.74batch/s, mse=0.146]
Epoch 37: 100%	██████████	1445/1445 [00:06<00:00, 227.47batch/s, mse=0.144]
Epoch 38: 100%	██████████	1445/1445 [00:06<00:00, 230.75batch/s, mse=0.143]
Epoch 39: 100%	██████████	1445/1445 [00:06<00:00, 232.24batch/s, mse=0.142]
Epoch 40: 100%	██████████	1445/1445 [00:06<00:00, 232.92batch/s, mse=0.141]
Epoch 41: 100%	██████████	1445/1445 [00:06<00:00, 219.42batch/s, mse=0.139]

Epoch 42: 100%	[██████████]	1445/1445 [00:06<00:00, 225.29batch/s, mse=0.138
Epoch 43: 100%	[██████████]	1445/1445 [00:06<00:00, 234.91batch/s, mse=0.137
Epoch 44: 100%	[██████████]	1445/1445 [00:06<00:00, 237.03batch/s, mse=0.136
Epoch 45: 100%	[██████████]	1445/1445 [00:06<00:00, 219.72batch/s, mse=0.135
Epoch 46: 100%	[██████████]	1445/1445 [00:06<00:00, 212.72batch/s, mse=0.134
Epoch 47: 100%	[██████████]	1445/1445 [00:06<00:00, 221.94batch/s, mse=0.133
Epoch 48: 100%	[██████████]	1445/1445 [00:06<00:00, 217.84batch/s, mse=0.132
Epoch 49: 100%	[██████████]	1445/1445 [00:06<00:00, 213.98batch/s, mse=0.131
Epoch 50: 100%	[██████████]	1445/1445 [00:06<00:00, 217.73batch/s, mse=0.13]
Epoch 51: 100%	[██████████]	1445/1445 [00:06<00:00, 217.72batch/s, mse=0.129
Epoch 52: 100%	[██████████]	1445/1445 [00:06<00:00, 219.38batch/s, mse=0.128
Epoch 53: 100%	[██████████]	1445/1445 [00:06<00:00, 213.33batch/s, mse=0.127
Epoch 54: 100%	[██████████]	1445/1445 [00:06<00:00, 212.99batch/s, mse=0.127
Epoch 55: 100%	[██████████]	1445/1445 [00:06<00:00, 213.12batch/s, mse=0.126
Epoch 56: 100%	[██████████]	1445/1445 [00:06<00:00, 228.70batch/s, mse=0.125
Epoch 57: 100%	[██████████]	1445/1445 [00:06<00:00, 221.04batch/s, mse=0.124
Epoch 58: 100%	[██████████]	1445/1445 [00:06<00:00, 231.47batch/s, mse=0.123

```
# restore model and return best accuracy
model.load_state_dict(best_weights)
print("MSE: %.2f" % best_mse)
print("RMSE: %.2f" % np.sqrt(best_mse))
plt.plot(history)
plt.show()
```

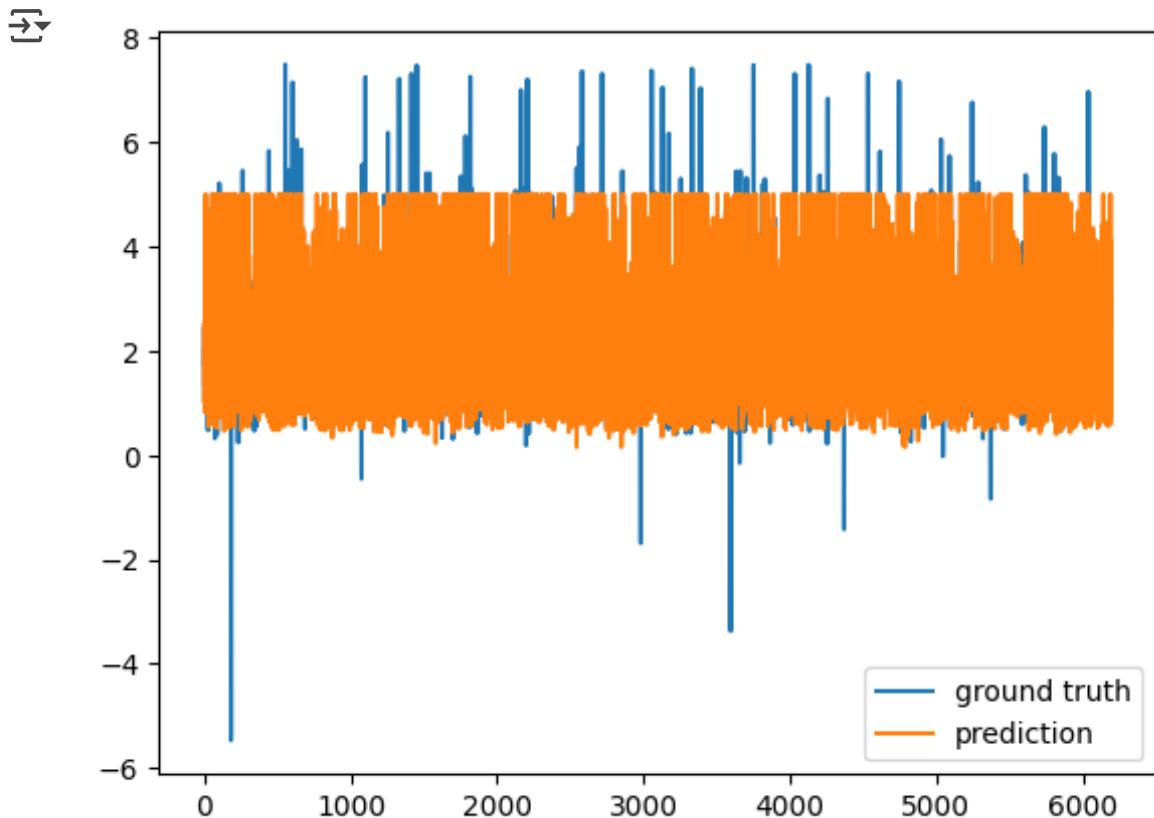
→ MSE: 0.60
RMSE: 0.77



```
# test the model
with torch.no_grad():
    y_pred = model(X_test)
    mse = loss_fn(y_pred, y_test)
    print("MSE: %.2f" % mse)
```

→ MSE: 0.60

```
# plot prediction vs ground truth
plt.plot(range(len(y_pred)), y_pred, label="ground truth")
plt.plot(range(len(y_test)), y_test, label="prediction")
plt.legend()
plt.show()
```



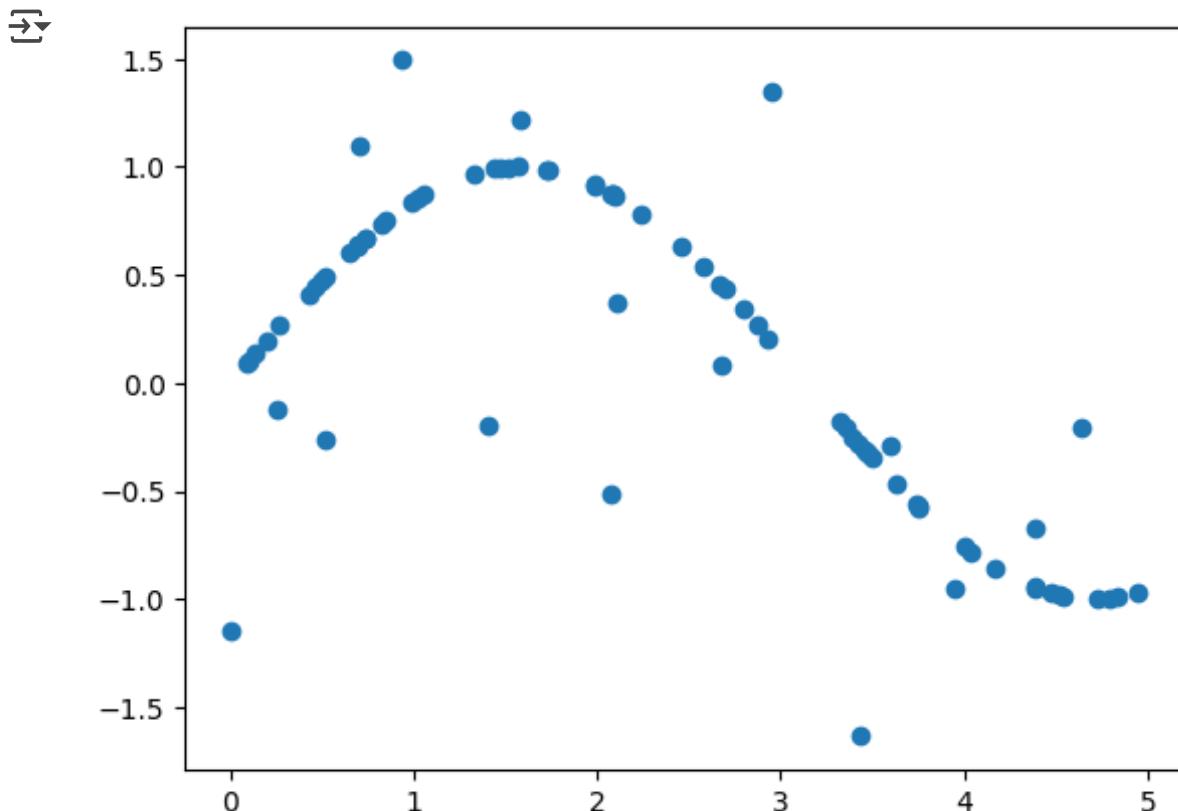

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor
import matplotlib.pyplot as plt
```

Exercise 2

```
rng = np.random.RandomState(1)
X = np.sort(5 * rng.rand(80, 1), axis=0)
y = np.sin(X).ravel()

# introducing outliers
y[::5] += 3 * (0.5 - rng.rand(16))
```

```
# plot X
plt.plot(X, y, 'o')
plt.show()
```



```
regr_1 = DecisionTreeRegressor(max_depth=2)
regr_2 = DecisionTreeRegressor(max_depth=5)
regr_1.fit(X, y)
regr_2.fit(X, y)
```

→ ▾ DecisionTreeRegressor ⓘ ⓘ

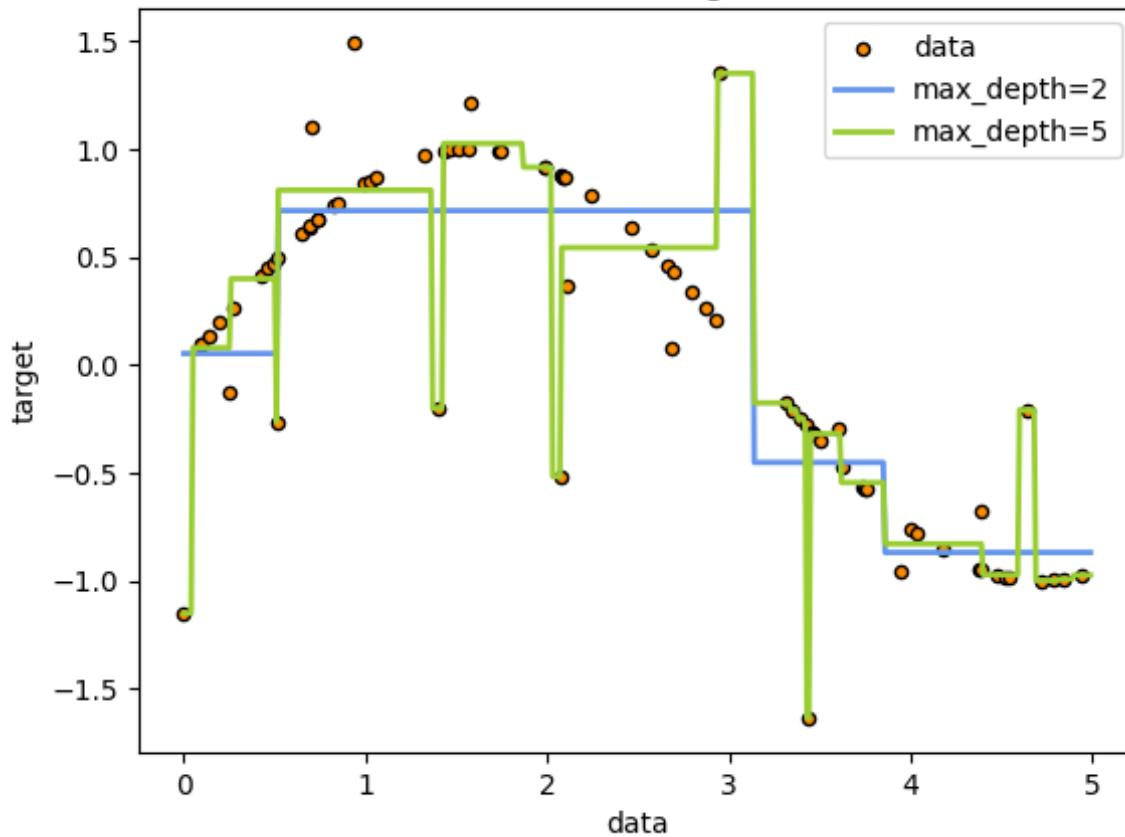
```
DecisionTreeRegressor(max_depth=5)
```

```
X_test = np.arange(0.0, 5.0, 0.01)[:, np.newaxis]
y_1 = regr_1.predict(X_test)
y_2 = regr_2.predict(X_test)
```

```
plt.figure()
plt.scatter(X, y, s=20, edgecolor="black", c="darkorange", label="data")
plt.plot(X_test, y_1, color="cornflowerblue", label="max_depth=2", linewidth=2)
plt.plot(X_test, y_2, color="yellowgreen", label="max_depth=5", linewidth=2)
plt.xlabel("data")
plt.ylabel("target")
plt.title("Decision Tree Regression")
plt.legend()
plt.show()
```



Decision Tree Regression



▼ DNN version

```
import torch
import torch.nn as nn
import torch.optim as optim
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error
```

```
# Generate the same data
rng = np.random.RandomState(1)
X_np = np.sort(5 * rng.rand(80, 1), axis=0).astype(np.float32)
y_np = np.sin(X_np).ravel().astype(np.float32)
y_np[::5] += 3 * (0.5 - rng.rand(16)).astype(np.float32)
X_test_np = np.arange(0.0, 5.0, 0.01)[:, np.newaxis].astype(np.float32)

# Convert NumPy arrays to PyTorch tensors
X = torch.tensor(X_np)
y = torch.tensor(y_np).unsqueeze(1) # Reshape y to [n_samples, 1]
X_test = torch.tensor(X_test_np)

# Define a simple DNN model
class Net(nn.Module):
    def __init__(self, input_size, hidden_size, output_size):
        super(Net, self).__init__()
        self.fc1 = nn.Linear(input_size, hidden_size)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        out = self.fc1(x)
        out = self.relu(out)
        out = self.fc2(out)
        return out

# Instantiate the model
input_size = 1
hidden_size = 10
output_size = 1
model = Net(input_size, hidden_size, output_size)

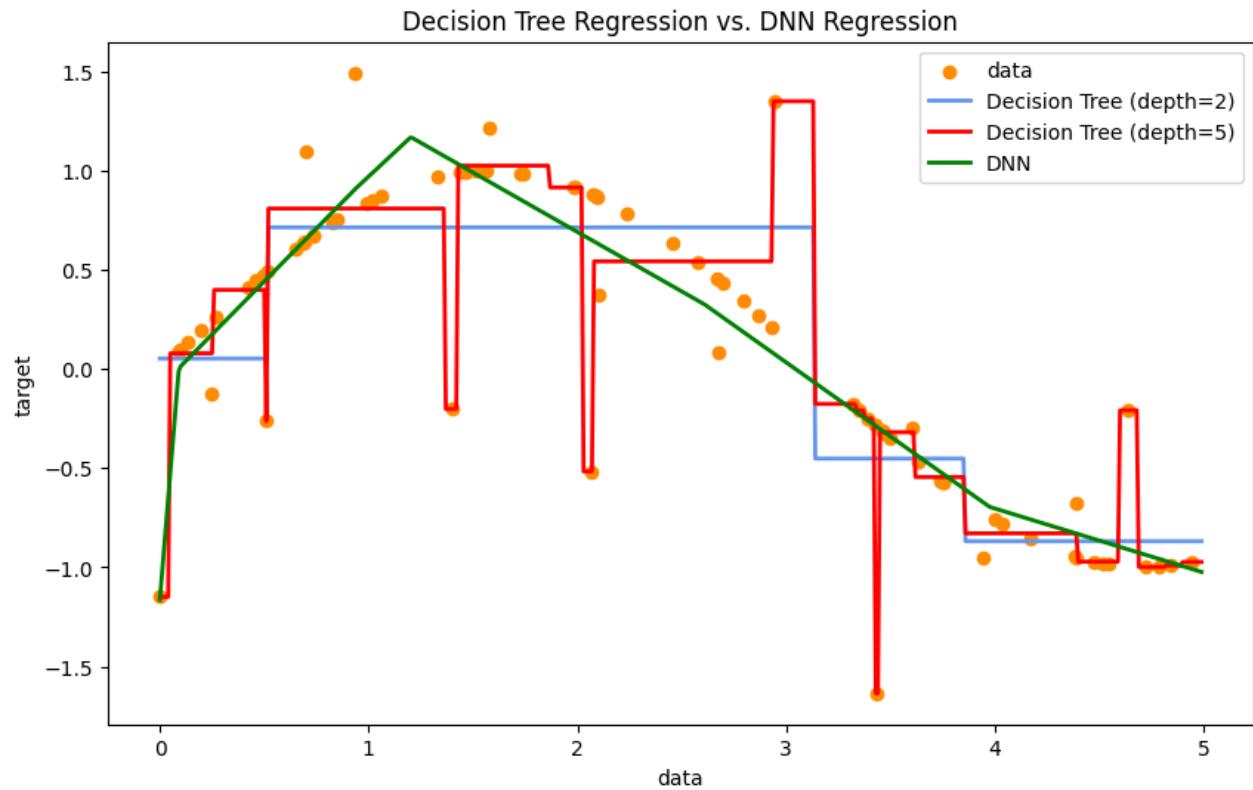
# Define loss function and optimizer
criterion = nn.MSELoss()
optimizer = optim.Adam(model.parameters(), lr=0.01)

# Train the model
num_epochs = 1000
for epoch in range(num_epochs):
    # Forward pass
    outputs = model(X)
    loss = criterion(outputs, y)

    # Backward and optimize
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
```

```
if (epoch + 1) % 100 == 0:  
    print(f'Epoch [{epoch+1}/{num_epochs}], Loss: {loss.item():.4f}')  
  
# Make predictions on the test data  
with torch.no_grad():  
    y_dnn = model(X_test).numpy()  
  
# -----  
# The original RandomForest part (for comparison)  
regr_1 = DecisionTreeRegressor(max_depth=2)  
regr_2 = DecisionTreeRegressor(max_depth=5)  
regr_1.fit(X_np, y_np)  
regr_2.fit(X_np, y_np)  
  
y_1 = regr_1.predict(X_test_np)  
y_2 = regr_2.predict(X_test_np)  
# -----  
  
# Calculate errors  
error_dnn = y_np - model(X).detach().numpy().flatten()  
error_dt1 = y_np - regr_1.predict(X_np)  
error_dt2 = y_np - regr_2.predict(X_np)  
  
# Plot the results (predictions) as well  
plt.figure(figsize=(10, 6))  
plt.scatter(X_np, y_np, color="darkorange", label="data")  
plt.plot(X_test_np, y_1, color="cornflowerblue", label="Decision Tree (depth=2)")  
plt.plot(X_test_np, y_2, color="red", label="Decision Tree (depth=5)", linewidth=2)  
plt.plot(X_test_np, y_dnn, color="green", label="DNN", linewidth=2)  
plt.xlabel("data")  
plt.ylabel("target")  
plt.title("Decision Tree Regression vs. DNN Regression")  
plt.legend()  
plt.show()
```

```
→ Epoch [100/1000], Loss: 0.1593
Epoch [200/1000], Loss: 0.1230
Epoch [300/1000], Loss: 0.1184
Epoch [400/1000], Loss: 0.1171
Epoch [500/1000], Loss: 0.1168
Epoch [600/1000], Loss: 0.1165
Epoch [700/1000], Loss: 0.1161
Epoch [800/1000], Loss: 0.1155
Epoch [900/1000], Loss: 0.1148
Epoch [1000/1000], Loss: 0.1140
```



```
print(y_np.shape)
print(y_dnn.shape)
print(y_1.shape)
print(y_2.shape)

→ (80, )
(500, 1)
(500,)
(500,)

# Calculate Mean Squared Errors (for reference)
mse_dnn = mean_squared_error(y_np, model(X).detach().numpy().flatten())
mse_dt1 = mean_squared_error(y_np, regr_1.predict(X_np))
mse_dt2 = mean_squared_error(y_np, regr_2.predict(X_np))

print(f"Mean Squared Error (DNN): {mse_dnn:.4f}")
print(f"Mean Squared Error (Decision Tree, max_depth=2): {mse_dt1:.4f}")
print(f"Mean Squared Error (Decision Tree, max_depth=5): {mse_dt2:.4f}")

# Plot histograms of the errors
plt.figure(figsize=(12, 6))

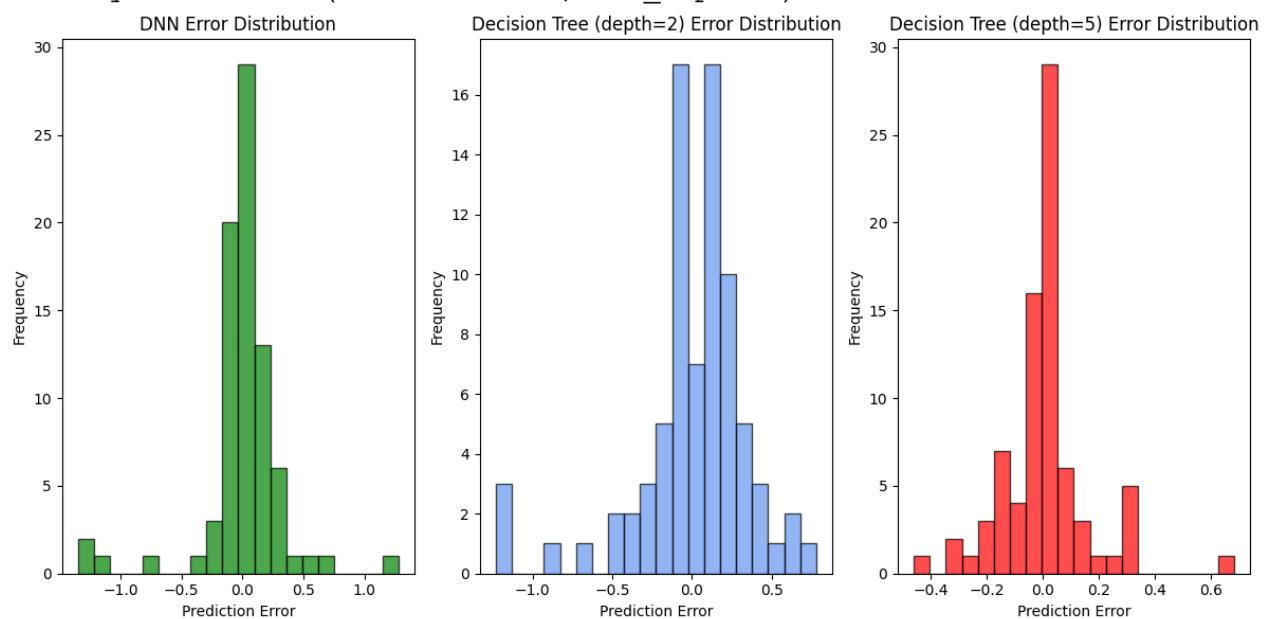
plt.subplot(1, 3, 1)
plt.hist(error_dnn, bins=20, edgecolor='black', alpha=0.7, color='green')
plt.title('DNN Error Distribution')
plt.xlabel('Prediction Error')
plt.ylabel('Frequency')

plt.subplot(1, 3, 2)
plt.hist(error_dt1, bins=20, edgecolor='black', alpha=0.7, color='cornflowerblue')
plt.title('Decision Tree (depth=2) Error Distribution')
plt.xlabel('Prediction Error')
plt.ylabel('Frequency')

plt.subplot(1, 3, 3)
plt.hist(error_dt2, bins=20, edgecolor='black', alpha=0.7, color='red')
plt.title('Decision Tree (depth=5) Error Distribution')
plt.xlabel('Prediction Error')
plt.ylabel('Frequency')

plt.tight_layout()
plt.show()
```

Mean Squared Error (DNN): 0.1140
Mean Squared Error (Decision Tree, max_depth=2): 0.1297
Mean Squared Error (Decision Tree, max_depth=5): 0.0252



▼ PyTorch K-Fold CV for Regression

Exercise 3

```
# -*- coding: utf-8 -*-
"""
PyTorch K-Fold Cross-Validation Example for Regression

This script demonstrates how to implement K-Fold Cross-Validation
with PyTorch for a simple regression problem.
"""

import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import TensorDataset, DataLoader
from sklearn.model_selection import KFold
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_regression # Using make_regression for simplicity
import matplotlib.pyplot as plt

# --- 1. Configuration ---
N_SPLITS = 5          # Number of folds for K-Fold CV
N_EPOCHS = 50          # Number of training epochs per fold
BATCH_SIZE = 16         # Batch size for training
LEARNING_RATE = 0.001 # Learning rate for the optimizer
N_SAMPLES = 500         # Total number of samples in the dataset
N_FEATURES = 10          # Number of features for the synthetic data
RANDOM_SEED = 42         # Seed for reproducibility

# Set random seeds for reproducibility
np.random.seed(RANDOM_SEED)
torch.manual_seed(RANDOM_SEED)
if torch.cuda.is_available():
    torch.cuda.manual_seed_all(RANDOM_SEED)

# --- 2. Generate Synthetic Regression Data ---
print("Generating synthetic data...")
# Using sklearn's make_regression for a controlled example
X, y = make_regression(n_samples=N_SAMPLES, n_features=N_FEATURES, noise=15.0, r

# Reshape y to be a 2D tensor [n_samples, 1] as expected by MSELoss
y = y.reshape(-1, 1)           # 66

# --- 3. Data Preprocessing ---
# It's crucial to scale features, especially for neural networks
```

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X) # Fit and transform on the whole dataset for
                                  # In a real scenario, fit ONLY on training da

# Convert numpy arrays to PyTorch tensors
X_tensor = torch.tensor(X_scaled, dtype=torch.float32)
y_tensor = torch.tensor(y, dtype=torch.float32)

# Create a full dataset
full_dataset = TensorDataset(X_tensor, y_tensor)

print(f"Data generated: {X_tensor.shape[0]} samples, {X_tensor.shape[1]} feature

# --- 4. Define the Neural Network Model ---
class RegressionNet(nn.Module):
    """A simple feed-forward neural network for regression."""
    def __init__(self, input_features):
        super(RegressionNet, self).__init__()
        self.layer_1 = nn.Linear(input_features, 64)
        self.relu1 = nn.ReLU()
        self.layer_2 = nn.Linear(64, 32)
        self.relu2 = nn.ReLU()
        self.output_layer = nn.Linear(32, 1) # Output layer has 1 neuron for reg

    def forward(self, x):
        x = self.layer_1(x)
        x = self.relu1(x)
        x = self.layer_2(x)
        x = self.relu2(x)
        x = self.output_layer(x)
        return x

# --- 5. K-Fold Cross-Validation Setup ---
kf = KFold(n_splits=N_SPLITS, shuffle=True, random_state=RANDOM_SEED)
fold_results = [] # To store validation loss for each fold

print(f"\nStarting {N_SPLITS}-Fold Cross-Validation...")

# --- 6. K-Fold Cross-Validation Loop ---
for fold, (train_idx, val_idx) in enumerate(kf.split(full_dataset)):
    print(f"\n--- Fold {fold+1}/{N_SPLITS} ---")

    # --- Data Splitting for the Current Fold ---
    # Create subsets for training and validation based on indices from KFold
    train_subset = torch.utils.data.Subset(full_dataset, train_idx)
    val_subset = torch.utils.data.Subset(full_dataset, val_idx)

    # Create DataLoaders for the current fold
    train_loader = DataLoader(train_subset, batch_size=BATCH_SIZE, shuffle=True)
    val_loader = DataLoader(val_subset, batch_size=BATCH_SIZE, shuffle=False)
```

```
train_loader = DataLoader(train_subset, batch_size=BATCH_SIZE, shuffle=True)
val_loader = DataLoader(val_subset, batch_size=BATCH_SIZE, shuffle=False) # |

print(f"Train samples: {len(train_subset)}, Validation samples: {len(val_sub}

# --- Model, Loss, Optimizer Initialization (re-initialize for each fold) ---
model = RegressionNet(input_features=N_FEATURES)
criterion = nn.MSELoss() # Mean Squared Error loss for regression
optimizer = optim.Adam(model.parameters(), lr=LEARNING_RATE)

# --- Training Loop for the Current Fold ---
print("Training...")
for epoch in range(N_EPOCHS):
    model.train() # Set model to training mode
    epoch_train_loss = 0.0
    for batch_X, batch_y in train_loader:
        # Zero the parameter gradients
        optimizer.zero_grad()

        # Forward pass
        outputs = model(batch_X)
        loss = criterion(outputs, batch_y)

        # Backward pass and optimize
        loss.backward()
        optimizer.step()

        epoch_train_loss += loss.item() * batch_X.size(0) # Accumulate loss

    avg_epoch_train_loss = epoch_train_loss / len(train_subset)
    # Optional: Print training progress less frequently
    if (epoch + 1) % 10 == 0:
        print(f' Epoch [{epoch+1}/{N_EPOCHS}], Train Loss: {avg_epoch_trai

# --- Validation Loop for the Current Fold ---
print("Validating...")
model.eval() # Set model to evaluation mode
fold_val_loss = 0.0
with torch.no_grad(): # Disable gradient calculation during validation
    for batch_X, batch_y in val_loader:
        outputs = model(batch_X)
        loss = criterion(outputs, batch_y)
        fold_val_loss += loss.item() * batch_X.size(0) # Accumulate loss

avg_fold_val_loss = fold_val_loss / len(val_subset)
fold_results.append(avg_fold_val_loss)
print(f"--- Fold {fold+1} Validation MSE: {avg_fold_val_loss:.4f} ---")

# --- 7. Results ---
```

```
" --- Results
average_val_loss = np.mean(fold_results)
std_dev_val_loss = np.std(fold_results)

print("\n--- Cross-Validation Results ---")
print(f"Validation MSE for each fold: {[f'{loss:.4f}' for loss in fold_results]}")
print(f"Average Validation MSE across {N_SPLITS} folds: {average_val_loss:.4f}")
print(f"Standard Deviation of Validation MSE: {std_dev_val_loss:.4f}")

print("\nK-Fold Cross-Validation finished.")

# --- Optional: Train Final Model on Full Data (if needed) ---
# After finding good hyperparameters using CV, you might train one final
# model on the entire dataset for deployment.
# print("\nTraining final model on full dataset (optional)...")  

# final_model = RegressionNet(input_features=N_FEATURES)
# final_criterion = nn.MSELoss()
# final_optimizer = optim.Adam(final_model.parameters(), lr=LEARNING_RATE)
# full_loader = DataLoader(full_dataset, batch_size=BATCH_SIZE, shuffle=True)
#
# for epoch in range(N_EPOCHS): # Use the same number of epochs or adjust
#     final_model.train()
#     for batch_X, batch_y in full_loader:
#         final_optimizer.zero_grad()
#         outputs = final_model(batch_X)
#         loss = final_criterion(outputs, batch_y)
#         loss.backward()
#         final_optimizer.step()
#     if (epoch + 1) % 10 == 0:
#         print(f' Final Model - Epoch [{epoch+1}/{N_EPOCHS}], Loss: {loss.item():.4f}')
# print("Final model training complete.")
# # You can now save and use 'final_model'
```

→ Generating synthetic data...
Data generated: 500 samples, 10 features.

Starting 5-Fold Cross-Validation...

--- Fold 1/5 ---
Train samples: 400, Validation samples: 100
Training...
Epoch [10/50], Train Loss: 8302.0220
Epoch [20/50], Train Loss: 391.8277
Epoch [30/50], Train Loss: 310.9277
Epoch [40/50], Train Loss: 274.5235
Epoch [50/50], Train Loss: 249.5084
Validating...
--- Fold 1 Validation MSE: 317.9261 ---

```
--- Fold 2/5 ---
Train samples: 400, Validation samples: 100
Training...
Epoch [10/50], Train Loss: 7245.0312
Epoch [20/50], Train Loss: 361.5749
Epoch [30/50], Train Loss: 280.0867
Epoch [40/50], Train Loss: 251.2109
Epoch [50/50], Train Loss: 226.2608
Validating...
--- Fold 2 Validation MSE: 279.0495 ---

--- Fold 3/5 ---
Train samples: 400, Validation samples: 100
Training...
Epoch [10/50], Train Loss: 7414.5510
Epoch [20/50], Train Loss: 354.0877
Epoch [30/50], Train Loss: 293.9479
Epoch [40/50], Train Loss: 261.6438
Epoch [50/50], Train Loss: 235.8877
Validating...
--- Fold 3 Validation MSE: 314.3125 ---

--- Fold 4/5 ---
Train samples: 400, Validation samples: 100
Training...
Epoch [10/50], Train Loss: 8310.9828
Epoch [20/50], Train Loss: 366.8772
Epoch [30/50], Train Loss: 289.3982
Epoch [40/50], Train Loss: 256.3477
Epoch [50/50], Train Loss: 229.4607
Validating...
--- Fold 4 Validation MSE: 325.8239 ---

--- Fold 5/5 ---
Train samples: 400, Validation samples: 100
Training...
Epoch [10/50], Train Loss: 8660.4236
Epoch [20/50], Train Loss: 402.1604
Epoch [30/50], Train Loss: 311.1535
Epoch [40/50], Train Loss: 274.4479
Epoch [50/50], Train Loss: 247.5183
Validating...
--- Fold 5 Validation MSE: 329.3204 ---
```

```
# -*- coding: utf-8 -*-
"""
PyTorch K-Fold Cross-Validation with Hyperparameter Tuning for Regression
```

```
This script demonstrates how to implement K-Fold Cross-Validation
with PyTorch for a simple regression problem and includes a basic
grid search for hyperparameter tuning (learning rate and hidden layer size).
"""
```

```
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import TensorDataset, DataLoader, Subset
from sklearn.model_selection import KFold
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_regression
import itertools # To iterate over hyperparameter combinations
import time

# --- 1. Configuration ---
N_SPLITS = 5          # Number of folds for K-Fold CV
N_EPOCHS = 30          # Number of training epochs per fold (reduced for faster de
BATCH_SIZE = 32         # Batch size for training
N_SAMPLES = 500         # Total number of samples in the dataset
N_FEATURES = 10          # Number of features for the synthetic data
RANDOM_SEED = 42         # Seed for reproducibility

# --- Hyperparameter Search Space ---
# Define the hyperparameters you want to tune
param_grid = {
    'learning_rate': [0.01, 0.001], # Example learning rates
    'hidden_size_1': [32, 64]      # Example sizes for the first hidden layer
    # Add more hyperparameters here if needed (e.g., 'hidden_size_2', 'batch_si
}

# Generate all combinations of hyperparameters
hyperparameter_combinations = [dict(zip(param_grid.keys(), v)) for v in itertools
print(f"Starting Hyperparameter Search with {len(hyperparameter_combinations)}")
print("Hyperparameter combinations to test:")
for combo in hyperparameter_combinations:
    print(f" - {combo}")

# Set random seeds for reproducibility
np.random.seed(RANDOM_SEED)
torch.manual_seed(RANDOM_SEED)
if torch.cuda.is_available():
    torch.cuda.manual_seed_all(RANDOM_SEED)
    torch.backends.cudnn.deterministic = True
    torch.backends.cudnn.benchmark = False
    71
# --- 2. Generate Synthetic Regression Data ---
print("\nGenerating synthetic data...")
X, y = make_regression(n_samples=N_SAMPLES, n_features=N_FEATURES, noise=20.0,
```

```
y = y.reshape(-1, 1) # Reshape y for MSELoss

# --- 3. Data Preprocessing (Outside the loop for this example) ---
# NOTE: Ideally, scaling should be done *inside* the K-Fold loop
#       fitting the scaler ONLY on the training data of that fold.
#       Doing it outside is simpler for demonstration but introduces slight dat
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Convert numpy arrays to PyTorch tensors
X_tensor = torch.tensor(X_scaled, dtype=torch.float32)
y_tensor = torch.tensor(y, dtype=torch.float32)
full_dataset = TensorDataset(X_tensor, y_tensor)

print(f"Data generated: {X_tensor.shape[0]} samples, {X_tensor.shape[1]} featur

# --- 4. Define the Neural Network Model ---
class RegressionNet(nn.Module):
    """A simple feed-forward neural network for regression."""
    def __init__(self, input_features, hidden_size_1=64): # Default value
        super(RegressionNet, self).__init__()
        # Use the passed hidden_size_1 for flexibility
        self.layer_1 = nn.Linear(input_features, hidden_size_1)
        self.relu1 = nn.ReLU()
        # Adjust subsequent layers if needed, here we keep it simple
        self.layer_2 = nn.Linear(hidden_size_1, hidden_size_1 // 2) # Example:
        self.relu2 = nn.ReLU()
        self.output_layer = nn.Linear(hidden_size_1 // 2, 1)

    def forward(self, x):
        x = self.layer_1(x)
        x = self.relu1(x)
        x = self.layer_2(x)
        x = self.relu2(x)
        x = self.output_layer(x)
        return x

# --- 5. Hyperparameter Tuning Loop ---
best_hyperparams = None
best_avg_val_loss = float('inf') # Initialize with infinity
results_log = [] # Store results for each hyperparameter set

start_time_tuning = time.time()

for params in hyperparameter_combinations:
    current_lr = params['learning_rate']
    current_hidden_size = params['hidden_size_1']
    print(f"\n--- Testing Hyperparameters: LR={current_lr}, HiddenSize1={curren
```

```
# --- 6. K-Fold Cross-Validation Setup (Inside Hyperparameter Loop) ---
kf = KFold(n_splits=N_SPLITS, shuffle=True, random_state=RANDOM_SEED)
fold_val_losses = [] # Store validation loss for each fold *for this set of
fold_start_time = time.time()

# --- 7. K-Fold Cross-Validation Loop ---
for fold, (train_idx, val_idx) in enumerate(kf.split(full_dataset)):
    # print(f" Fold {fold+1}/{N_SPLITS}...") # Less verbose output

    # --- Data Splitting for the Current Fold ---
    train_subset = Subset(full_dataset, train_idx)
    val_subset = Subset(full_dataset, val_idx)
    train_loader = DataLoader(train_subset, batch_size=BATCH_SIZE, shuffle=True)
    val_loader = DataLoader(val_subset, batch_size=BATCH_SIZE, shuffle=False)

    # --- Model, Loss, Optimizer Initialization (Using current hyperparameters)
    # Pass the current hidden size to the model
    model = RegressionNet(input_features=N_FEATURES, hidden_size_1=current_hidden_size)
    criterion = nn.MSELoss()
    # Use the current learning rate for the optimizer
    optimizer = optim.Adam(model.parameters(), lr=current_lr)

    # --- Training Loop for the Current Fold ---
    for epoch in range(N_EPOCHS):
        model.train()
        for batch_X, batch_y in train_loader:
            optimizer.zero_grad()
            outputs = model(batch_X)
            loss = criterion(outputs, batch_y)
            loss.backward()
            optimizer.step()
    # No epoch printing during hyperparameter search to reduce clutter

    # --- Validation Loop for the Current Fold ---
    model.eval()
    fold_val_loss = 0.0
    with torch.no_grad():
        for batch_X, batch_y in val_loader:
            outputs = model(batch_X)
            loss = criterion(outputs, batch_y)
            fold_val_loss += loss.item() * batch_X.size(0)

    avg_fold_val_loss = fold_val_loss / len(val_subset)
    fold_val_losses.append(avg_fold_val_loss)
    # print(f" Fold {fold+1} Val MSE: {avg_fold_val_loss:.4f}") # Less verbose output

# --- Calculate Average Performance for this Hyperparameter Set ---
```

```
average_loss_for_params = np.mean(fold_val_losses)
std_dev_for_params = np.std(fold_val_losses)
fold_end_time = time.time()

print(f" Average Validation MSE for {params}: {average_loss_for_params:.4f}")
print(f" Time taken for this set: {fold_end_time - fold_start_time:.2f} seconds")

results_log.append({
    'params': params,
    'avg_val_loss': average_loss_for_params,
    'std_dev_val_loss': std_dev_for_params
})

# --- Update Best Hyperparameters ---
if average_loss_for_params < best_avg_val_loss:
    best_avg_val_loss = average_loss_for_params
    best_hyperparams = params
    print(f" ** New best hyperparameters found! **")

end_time_tuning = time.time()
print(f"\nTotal Hyperparameter Tuning Time: {end_time_tuning - start_time_tuning:.2f} seconds")

# --- 8. Final Results ---
print("\n--- Hyperparameter Tuning Results ---")
if best_hyperparams:
    print(f"Best Hyperparameters Found:")
    for key, value in best_hyperparams.items():
        print(f" - {key}: {value}")
    print(f"Best Average Validation MSE: {best_avg_val_loss:.4f}")
else:
    print("Hyperparameter tuning did not complete successfully or no parameters were found")

# --- Optional: Train Final Model with Best Hyperparameters ---
# if best_hyperparams:
#     print("\nTraining final model on full dataset with best hyperparameters..")
#     final_model = RegressionNet(input_features=N_FEATURES, hidden_size_1=best)
#     final_criterion = nn.MSELoss()
#     final_optimizer = optim.Adam(final_model.parameters(), lr=best_hyperparams['lr'])
#     full_loader = DataLoader(full_dataset, batch_size=BATCH_SIZE, shuffle=True)
#
#     for epoch in range(N_EPOCHS): # Use the same number of epochs or adjust
#         final_model.train()
#         epoch_loss = 0.0
#         for batch_X, batch_y in full_loader:
#             final_optimizer.zero_grad()
#             outputs = final_model(batch_X)
#             loss = final_criterion(outputs, batch_y)
#             loss.backward()
```

```

#             final_optimizer.step()
#             epoch_loss += loss.item() * batch_X.size(0)
#
#             avg_epoch_loss = epoch_loss / len(full_dataset)
#             if (epoch + 1) % 10 == 0:
#                 print(f' Final Model - Epoch [{epoch+1}/{N_EPOCHS}], Train Loss')
#             print("Final model training complete.")
#             # You can now save and use 'final_model'

print("\nScript finished.")

```

→ Starting Hyperparameter Search with 4 combinations.

Hyperparameter combinations to test:

- {'learning_rate': 0.01, 'hidden_size_1': 32}
- {'learning_rate': 0.01, 'hidden_size_1': 64}
- {'learning_rate': 0.001, 'hidden_size_1': 32}
- {'learning_rate': 0.001, 'hidden_size_1': 64}

Generating synthetic data...

Data generated: 500 samples, 10 features.

--- Testing Hyperparameters: LR=0.01, HiddenSize1=32 ---

Average Validation MSE for {'learning_rate': 0.01, 'hidden_size_1': 32}:

Time taken for this set: 3.35 seconds

** New best hyperparameters found! **

--- Testing Hyperparameters: LR=0.01, HiddenSize1=64 ---

Average Validation MSE for {'learning_rate': 0.01, 'hidden_size_1': 64}:

Time taken for this set: 3.96 seconds

--- Testing Hyperparameters: LR=0.001, HiddenSize1=32 ---

Average Validation MSE for {'learning_rate': 0.001, 'hidden_size_1': 32}:

Time taken for this set: 3.27 seconds

--- Testing Hyperparameters: LR=0.001, HiddenSize1=64 ---

Average Validation MSE for {'learning_rate': 0.001, 'hidden_size_1': 64}:

Time taken for this set: 3.05 seconds

Total Hyperparameter Tuning Time: 13.63 seconds

--- Hyperparameter Tuning Results ---

Best Hyperparameters Found:

- learning_rate: 0.01
- hidden_size_1: 32

Best Average Validation MSE: 436.6098

Script finished.


```
# Import necessary libraries
import torch
import torch.nn as nn
import torch.optim as optim
import matplotlib.pyplot as plt
import numpy as np

# --- Configuration ---
INPUT_SIZE = 10      # Dimension of input data
HIDDEN_SIZE = 5       # Dimension of hidden layers
OUTPUT_SIZE = 1        # Dimension of output data
NUM_LAYERS = 10       # Number of hidden layers (make it deep to see the effect)
LEARNING_RATE = 0.01
NUM_EPOCHS = 500
BATCH_SIZE = 16

# --- Create a Deep Neural Network ---
class DeepNet(nn.Module):
    def __init__(self, input_size, hidden_size, output_size, num_layers):
        super(DeepNet, self).__init__()
        layers = []
        # Input layer
        layers.append(nn.Linear(input_size, hidden_size))
        layers.append(nn.Sigmoid()) # Using Sigmoid, prone to vanishing gradient

        # Hidden layers
        for _ in range(num_layers - 1):
            layers.append(nn.Linear(hidden_size, hidden_size))
            layers.append(nn.Sigmoid()) # Using Sigmoid

        # Output layer
        layers.append(nn.Linear(hidden_size, output_size))
        # No activation for regression, or use Sigmoid/Softmax for classification

        # Combine all layers into a sequential model
        self.network = nn.Sequential(*layers)

    def forward(self, x):
        return self.network(x)

# --- Generate Dummy Data ---
# Create random input data (features)
X = torch.randn(BATCH_SIZE * 10, INPUT_SIZE)
# Create corresponding target data (labels) - simple linear relationship for demo
# We add some noise to make it slightly more realistic
true_weights = torch.randn(INPUT_SIZE, OUTPUT_SIZE)
true_bias = torch.randn(OUTPUT_SIZE)
```

Exercise 4

```
y = X @ true_weights + true_bias + torch.randn(BATCH_SIZE * 10, OUTPUT_SIZE) * 0

# --- Initialize Model, Loss, and Optimizer ---
model = DeepNet(INPUT_SIZE, HIDDEN_SIZE, OUTPUT_SIZE, NUM_LAYERS)
criterion = nn.MSELoss() # Mean Squared Error loss for regression
optimizer = optim.SGD(model.parameters(), lr=LEARNING_RATE)

# --- Training Loop ---
losses = []
gradient_norms_first_layer = []
gradient_norms_last_layer = []

print(f"Starting training for {NUM_EPOCHS} epochs...")
for epoch in range(NUM_EPOCHS):
    # Simple batching (using the whole dataset as one batch for simplicity here)
    # In practice, you'd iterate over mini-batches
    optimizer.zero_grad() # Clear previous gradients

    # Forward pass
    outputs = model(X)
    loss = criterion(outputs, y)

    # Backward pass (calculate gradients)
    loss.backward()

    # --- Observe Gradients (Optional but key for vanishing gradient) ---
    # Get the gradient norm for the first linear layer's weights
    first_layer_grad_norm = torch.norm(model.network[0].weight.grad).item()
    gradient_norms_first_layer.append(first_layer_grad_norm)

    # Get the gradient norm for the last linear layer's weights
    # The index depends on the number of layers (Linear + Activation)
    last_linear_layer_index = (NUM_LAYERS * 2) # Index of the last Linear layer
    last_layer_grad_norm = torch.norm(model.network[last_linear_layer_index].wei
    gradient_norms_last_layer.append(last_layer_grad_norm)
    # ----->

    # Update weights
    optimizer.step()

    # Store loss for plotting
    losses.append(loss.item())

    # Print progress
    if (epoch + 1) % 50 == 0: 78
        print(f'Epoch [{epoch+1}/{NUM_EPOCHS}], Loss: {loss.item():.4f}, Grad No

print("Training finished.")
```

```
# --- Plotting ---
plt.figure(figsize=(12, 5))

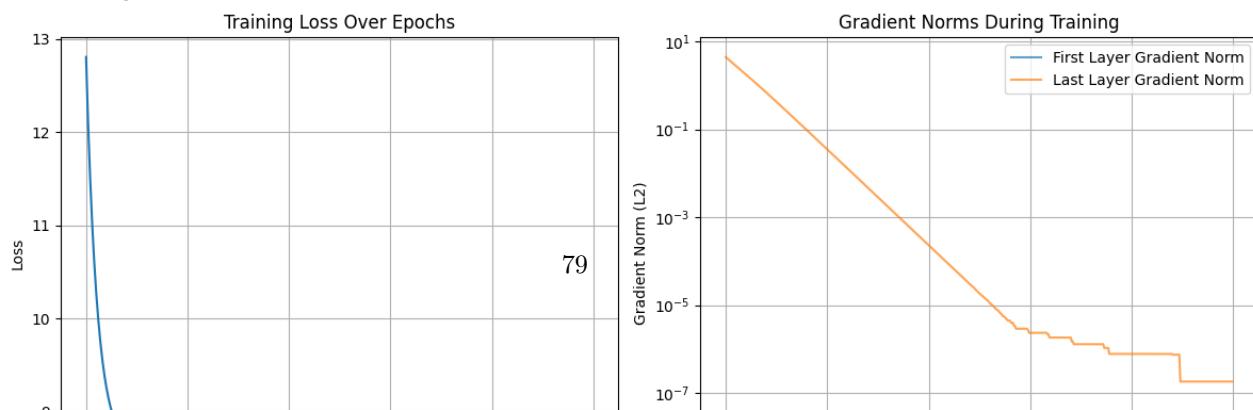
# Plot Loss
plt.subplot(1, 2, 1)
plt.plot(losses)
plt.title('Training Loss Over Epochs')
plt.xlabel('Epoch')
plt.ylabel('Loss')
plt.grid(True)

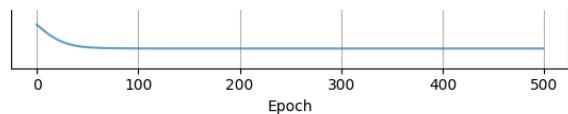
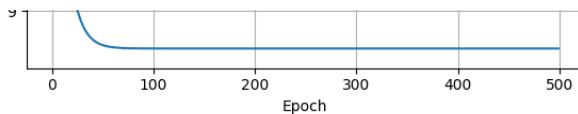
# Plot Gradient Norms
plt.subplot(1, 2, 2)
plt.plot(gradient_norms_first_layer, label='First Layer Gradient Norm', alpha=0)
plt.plot(gradient_norms_last_layer, label='Last Layer Gradient Norm', alpha=0.7)
plt.title('Gradient Norms During Training')
plt.xlabel('Epoch')
plt.ylabel('Gradient Norm (L2)')
plt.yscale('log') # Use log scale to see small values better
plt.legend()
plt.grid(True)

plt.tight_layout()
plt.show()
```

→ Starting training for 500 epochs...

```
Epoch [50/500], Loss: 8.6411, Grad Norm First: 0.000000, Grad Norm Last: 0.
Epoch [100/500], Loss: 8.6016, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [150/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [200/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [250/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [300/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [350/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [400/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [450/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Epoch [500/500], Loss: 8.6014, Grad Norm First: 0.000000, Grad Norm Last: 0
Training finished.
```



**Explanation:**

The left plot shows the training loss. If it decreases very slowly or plateaus quickly, it might indicate issues like vanishing gradients, especially with Sigmoid in deep nets. The right plot shows the L2 norm (magnitude) of the gradients for the weights of the first and last linear layers. In the vanishing gradient problem, the norm of the gradients for the earlier layers (like the first layer) becomes significantly smaller than for later layers, often approaching zero. This means the weights of the earlier layer are updated very slowly or not at all, hindering learning.

Note: Using ReLU activation instead of Sigmoid usually mitigates this problem.

Explanation: The left plot shows the training loss. If it decreases very slowly or plateaus quickly, it might indicate issues like vanishing gradients, especially with Sigmoid in deep nets. The right plot shows the L2 norm (magnitude) of the gradients for the weights of the first and last linear layers. In the vanishing gradient problem, the norm of the gradients for the earlier layers (like the first layer) becomes significantly smaller than for later layers, often approaching zero. This means the weights of the earlier layers are updated very slowly or not at all, hindering learning.

Note: Using ReLU activation instead of Sigmoid usually mitigates this problem.

Inizia a programmare o genera codice con l'IA.


```
# Import necessary libraries
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import TensorDataset, DataLoader
import numpy as np
import matplotlib.pyplot as plt # For plotting results

# --- Configuration ---
NUM_SAMPLES = 1000
NUM_FEATURES = 2
NUM_CLASSES = 2
NUM_EPOCHS = 50
LEARNING_RATE = 0.01
BATCH_SIZE = 64
SEED = 42 # for reproducibility

# Set random seeds for reproducibility
torch.manual_seed(SEED)
np.random.seed(SEED)

# --- 1. Generate Synthetic Data ---
# Create data where features have very different scales
# Feature 1: Small scale (e.g., between 0 and 1)
# Feature 2: Large scale (e.g., between 0 and 1000)
X = np.zeros((NUM_SAMPLES, NUM_FEATURES), dtype=np.float32)
X[:, 0] = np.random.rand(NUM_SAMPLES) * 1.0 # Scale: 0 to 1
X[:, 1] = np.random.rand(NUM_SAMPLES) * 1000.0 # Scale: 0 to 1000

# Create simple linear labels based on a weighted sum
#  $y = 1 \text{ if } 10*X1 + 0.01*X2 > 5 \text{ else } 0$  (example decision boundary)
# We scale weights inversely to feature scale to make it learnable
weights = np.array([10.0, 0.01], dtype=np.float32)
bias = -5.0
logits = X @ weights + bias
y = (logits > 0).astype(np.int64) # Binary classification labels (0 or 1)

# Convert numpy arrays to PyTorch tensors
X_tensor = torch.from_numpy(X)
y_tensor = torch.from_numpy(y)

# --- 2. Define a Simple Neural Network Model ---
class SimpleMLP(nn.Module):
    def __init__(self, input_size, hidden_size, output_size):82
        super(SimpleMLP, self).__init__()
        # A simple network with one hidden layer
        self.layer_1 = nn.Linear(input_size, hidden_size)
```

Exercise 5

```
        self.relu = nn.ReLU()
        self.layer_2 = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        x = self.layer_1(x)
        x = self.relu(x)
        x = self.layer_2(x)
        return x

# --- 3. Define Training Loop ---
def train_model(model, dataloader, criterion, optimizer, num_epochs):
    """Trains the model and returns loss history."""
    loss_history = []
    model.train() # Set model to training mode
    print(f"\n--- Starting Training ({'Normalized' if model.normalized_input else 'Unnormalized'}) ---")
    for epoch in range(num_epochs):
        epoch_loss = 0.0
        for batch_X, batch_y in dataloader:
            # Zero the gradients
            optimizer.zero_grad()

            # Forward pass
            outputs = model(batch_X)

            # Calculate loss
            loss = criterion(outputs, batch_y)

            # Backward pass and optimization
            loss.backward()
            optimizer.step()

            epoch_loss += loss.item() * batch_X.size(0) # Accumulate loss

        # Calculate average loss for the epoch
        avg_epoch_loss = epoch_loss / len(dataloader.dataset)
        loss_history.append(avg_epoch_loss)

        if (epoch + 1) % 10 == 0 or epoch == 0: # Print every 10 epochs
            print(f'Epoch [{epoch+1}/{num_epochs}], Loss: {avg_epoch_loss:.6f}')

    print("---- Training Finished ----")
    return loss_history

# --- 4. Scenario 1: Train without Normalization ---
# Create dataset and dataloader          83
dataset_unnormalized = TensorDataset(X_tensor, y_tensor)
dataloader_unnormalized = DataLoader(dataset_unnormalized, batch_size=BATCH_SIZE)
```

```
# Initialize model, loss, and optimizer
model_unnormalized = SimpleMLP(input_size=NUM_FEATURES, hidden_size=10, output_size=1)
model_unnormalized.normalized_input = False # Add flag for printing
criterion = nn.CrossEntropyLoss() # Suitable for classification
optimizer_unnormalized = optim.SGD(model_unnormalized.parameters(), lr=LEARNING_RATE)

# Train the model
loss_unnormalized = train_model(model_unnormalized, dataloader_unnormalized, criterion)

# --- 5. Scenario 2: Train with Normalization ---
# Calculate mean and std deviation for normalization (only from training data!)
mean = X_tensor.mean(dim=0, keepdim=True)
std = X_tensor.std(dim=0, keepdim=True)

# Handle potential zero standard deviation (though unlikely with random data)
std[std == 0] = 1.0

# Apply normalization: (X - mean) / std
X_normalized_tensor = (X_tensor - mean) / std

# Create dataset and dataloader for normalized data
dataset_normalized = TensorDataset(X_normalized_tensor, y_tensor)
dataloader_normalized = DataLoader(dataset_normalized, batch_size=BATCH_SIZE, shuffle=True)

# Initialize a *new* instance of the same model
model_normalized = SimpleMLP(input_size=NUM_FEATURES, hidden_size=10, output_size=1)
model_normalized.normalized_input = True # Add flag for printing
# Use the same loss function
optimizer_normalized = optim.SGD(model_normalized.parameters(), lr=LEARNING_RATE)

# Train the model on normalized data
loss_normalized = train_model(model_normalized, dataloader_normalized, criterion)

# --- 6. Compare Results ---
print("\n--- Comparison ---")
print(f"Final Loss (Unnormalized): {loss_unnormalized[-1]:.6f}")
print(f"Final Loss (Normalized): {loss_normalized[-1]:.6f}")

# Plotting the loss curves
plt.figure(figsize=(10, 5))
plt.plot(range(1, NUM_EPOCHS + 1), loss_unnormalized, label='Unnormalized Loss', color='blue')
plt.plot(range(1, NUM_EPOCHS + 1), loss_normalized, label='Normalized Loss', color='orange')
plt.xlabel('Epoch')
plt.ylabel('Average Loss')
plt.title('Training Loss: Normalized vs Unnormalized Data')
plt.legend()
plt.grid(True)
plt.show()
```

```
print("\nObserve the plot: Normalization typically leads to faster convergence (")
print("Without normalization, the large scale difference between features can ma")
```

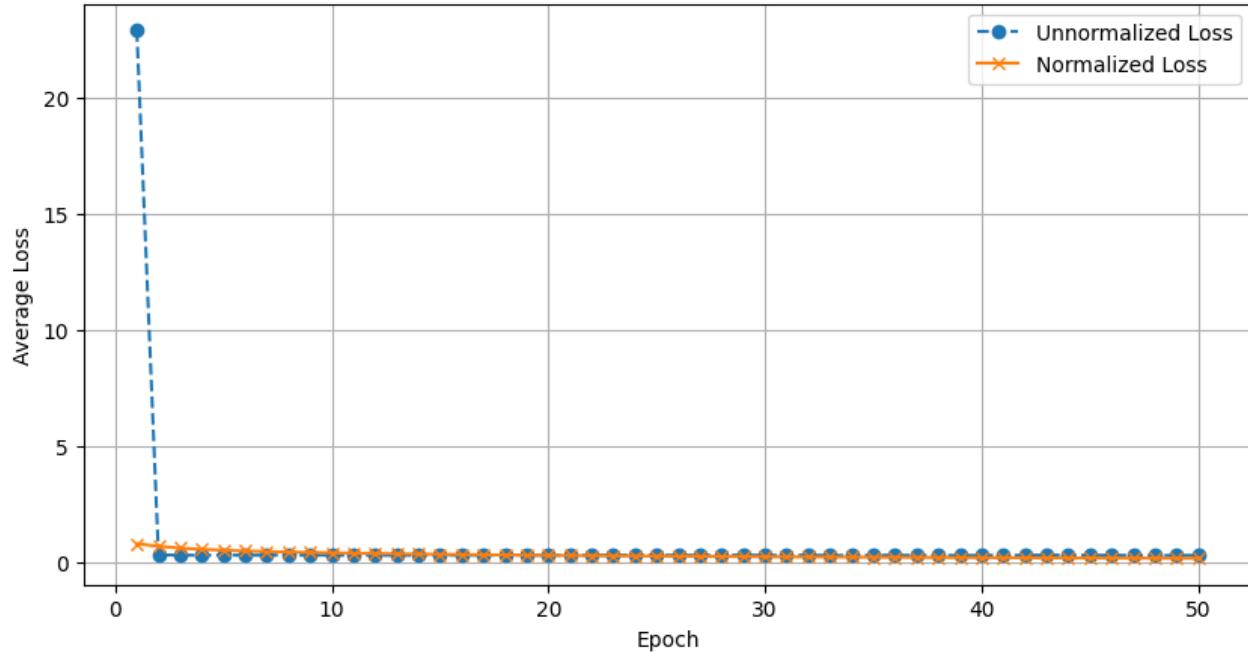


```
--- Starting Training (Unnormalized) ---
Epoch [1/50], Loss: 22.923474
Epoch [10/50], Loss: 0.278562
Epoch [20/50], Loss: 0.275233
Epoch [30/50], Loss: 0.279166
Epoch [40/50], Loss: 0.271082
Epoch [50/50], Loss: 0.276131
--- Training Finished ---

--- Starting Training (Normalized) ---
Epoch [1/50], Loss: 0.777922
Epoch [10/50], Loss: 0.381256
Epoch [20/50], Loss: 0.275293
Epoch [30/50], Loss: 0.214452
Epoch [40/50], Loss: 0.175974
Epoch [50/50], Loss: 0.150031
--- Training Finished ---

--- Comparison ---
Final Loss (Unnormalized): 0.276131
Final Loss (Normalized): 0.150031
```

Training Loss: Normalized vs. Unnormalized Data



Observe the plot: Normalization typically leads to faster convergence (loss Without normalization, the large scale difference between features can make


```

# Import necessary libraries
import torch
import torch.nn as nn
import torch.optim as optim
import matplotlib.pyplot as plt
import numpy as np

# --- 1. Data Generation ---

# Define the true relationship (y = 2x + 1)
true_weight = 2.0
true_bias = 1.0
num_samples = 100

# Generate 'Good' Data (Clean)
X_good = torch.randn(num_samples, 1) * 10 # Input features
y_good = true_weight * X_good + true_bias # Corresponding labels based on true r
print(f"Sample of Good Data (X, y):\n{torch.cat((X_good, y_good), 1)[:5]}\n")

# Generate 'Garbage' Data (Noisy)
# Add significant noise to the labels
noise_level = 15.0
noise = torch.randn(num_samples, 1) * noise_level
X_garbage = X_good # Use the same input features for direct comparison
y_garbage = true_weight * X_garbage + true_bias + noise # Add noise to labels
print(f"Sample of Garbage Data (X, y):\n{torch.cat((X_garbage, y_garbage), 1)[:5]}\n")

# --- 2. Model Definition ---

# Simple Linear Regression Model
class LinearRegression(nn.Module):
    def __init__(self):
        super(LinearRegression, self).__init__()
        # Define a single linear layer
        self.linear = nn.Linear(1, 1) # One input feature, one output feature

    def forward(self, x):
        # Pass input through the linear layer
        return self.linear(x)

# --- 3. Training Function ---

def train_model(model, criterion, optimizer, X_train, y_train, epochs=100):
    """Trains the linear regression model.87"""
    losses = []
    for epoch in range(epochs):

```

Exercise 6

```
# Forward pass: Compute predicted y by passing X to the model
outputs = model(X_train)
loss = criterion(outputs, y_train)

# Backward pass and optimization
optimizer.zero_grad() # Clear previous gradients
loss.backward()         # Compute gradients
optimizer.step()        # Update weights

losses.append(loss.item())
if (epoch + 1) % 10 == 0:
    print(f'Epoch [{epoch+1}/{epochs}], Loss: {loss.item():.4f}')
return losses

# --- 4. Training on 'Good' Data ---

print("\n--- Training on Good Data ---")
model_good = LinearRegression()
criterion_good = nn.MSELoss() # Mean Squared Error loss
optimizer_good = optim.SGD(model_good.parameters(), lr=0.001) # Stochastic Gradient Descent

losses_good = train_model(model_good, criterion_good, optimizer_good, X_good, y_good)

# Get learned parameters
learned_weight_good, learned_bias_good = model_good.parameters()
print(f"\nLearned Parameters (Good Data):")
print(f"  Weight: {learned_weight_good.item():.4f} (True: {true_weight})")
print(f"  Bias:   {learned_bias_good.item():.4f} (True: {true_bias})")
print(f"Final Loss (Good Data): {losses_good[-1]:.4f}")

# --- 5. Training on 'Garbage' Data ---

print("\n--- Training on Garbage Data ---")
model_garbage = LinearRegression() # Create a new instance
criterion_garbage = nn.MSELoss()
optimizer_garbage = optim.SGD(model_garbage.parameters(), lr=0.001) # Use same optimizer

losses_garbage = train_model(model_garbage, criterion_garbage, optimizer_garbage)

# Get learned parameters
learned_weight_garbage, learned_bias_garbage = model_garbage.parameters()
print(f"\nLearned Parameters (Garbage Data):")
print(f"  Weight: {learned_weight_garbage.item():.4f} (True: {true_weight})")
print(f"  Bias:   {learned_bias_garbage.item():.4f} (True: {true_bias})")
print(f"Final Loss (Garbage Data): {losses_garbage[-1]:.4f}")
```

```
# --- 6. Visualization (Optional - Requires Matplotlib) ---  
  
# Function to detach tensors for plotting  
def detach_numpy(tensor):  
    return tensor.detach().cpu().numpy()  
  
plt.figure(figsize=(12, 5))  
  
# Plot for Good Data  
plt.subplot(1, 2, 1)  
plt.scatter(detach_numpy(X_good), detach_numpy(y_good), label='Good Data Points'  
plt.plot(detach_numpy(X_good), detach_numpy(model_good(X_good)), color='red', la  
plt.plot(detach_numpy(X_good), detach_numpy(true_weight * X_good + true_bias), c  
plt.title('Training on Good Data')  
plt.xlabel('X')  
plt.ylabel('y')  
plt.legend()  
plt.grid(True)  
  
# Plot for Garbage Data  
plt.subplot(1, 2, 2)  
plt.scatter(detach_numpy(X_garbage), detach_numpy(y_garbage), label='Garbage Dat  
plt.plot(detach_numpy(X_garbage), detach_numpy(model_garbage(X_garbage)), color=  
plt.plot(detach_numpy(X_garbage), detach_numpy(true_weight * X_garbage + true_bi  
plt.title('Training on Garbage Data')  
plt.xlabel('X')  
plt.ylabel('y')  
plt.legend()  
plt.grid(True)  
  
plt.tight_layout()  
# To display the plot in environments that support it:  
# plt.show()  
# In other environments, save the plot to a file:  
plt.savefig('gigo_comparison.png')  
print("\nPlot saved as gigo_comparison.png")  
  
# --- 7. Comparison of Losses ---  
plt.figure(figsize=(8, 5))  
plt.plot(losses_good, label='Loss (Good Data)')  
plt.plot(losses_garbage, label='Loss (Garbage Data)')  
plt.title('Training Loss Comparison')  
plt.xlabel('Epoch')  
plt.ylabel('MSE Loss')  
plt.legend()  
plt.grid(True)  
plt.yscale('log') # Use log scale as garbage loss might be much higher  
plt.savefig('gigo_loss_comparison.png')
```

```
print("Loss comparison plot saved as gigo_loss_comparison.png")
```

→ Sample of Good Data (X, y):
tensor([[10.1579, 21.3159],
 [-13.7729, -26.5459],
 [8.5747, 18.1494],
 [-3.5829, -6.1658],
 [12.6501, 26.3001]]))

Sample of Garbage Data (X, y):
tensor([[10.1579, 16.3769],
 [-13.7729, -16.3491],
 [8.5747, -7.6126],
 [-3.5829, -11.8858],
 [12.6501, 17.5613]]))

--- Training on Good Data ---
Epoch [10/100], Loss: 5.2646
Epoch [20/100], Loss: 2.9034
Epoch [30/100], Loss: 2.7653
Epoch [40/100], Loss: 2.6588
Epoch [50/100], Loss: 2.5567
Epoch [60/100], Loss: 2.4585
Epoch [70/100], Loss: 2.3641
Epoch [80/100], Loss: 2.2733
Epoch [90/100], Loss: 2.1859
Epoch [100/100], Loss: 2.1020

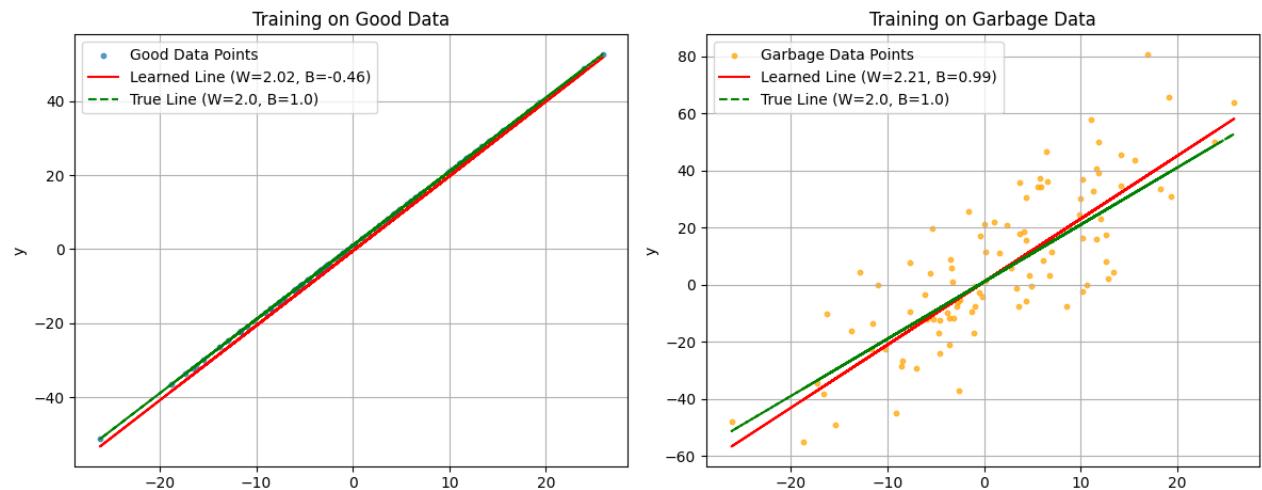
Learned Parameters (Good Data):
Weight: 2.0219 (True: 2.0)
Bias: -0.4629 (True: 1.0)
Final Loss (Good Data): 2.1020

--- Training on Garbage Data ---
Epoch [10/100], Loss: 264.5396
Epoch [20/100], Loss: 251.8058
Epoch [30/100], Loss: 251.3786
Epoch [40/100], Loss: 251.1133
Epoch [50/100], Loss: 250.8600
Epoch [60/100], Loss: 250.6164
Epoch [70/100], Loss: 250.3822
Epoch [80/100], Loss: 250.1570
Epoch [90/100], Loss: 249.9404
Epoch [100/100], Loss: 249.7322

Learned Parameters (Garbage Data):
Weight: 2.2055 (True: 2.0) 90
Bias: 0.9894 (True: 1.0)
Final Loss (Garbage Data): 249.7322

Plot saved as gigo_comparison.png

Loss comparison plot saved as gigo_loss_comparison.png



Training Loss Comparison



▼ Air passenger Regression

Exercise 7

Dataset: <https://www.kaggle.com/datasets/rakannimer/air-passengers>

```
import os
import urllib.request
from pandas import read_csv
from matplotlib import pyplot

import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import Dataset, DataLoader
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
import numpy as np
import matplotlib.pyplot as plt

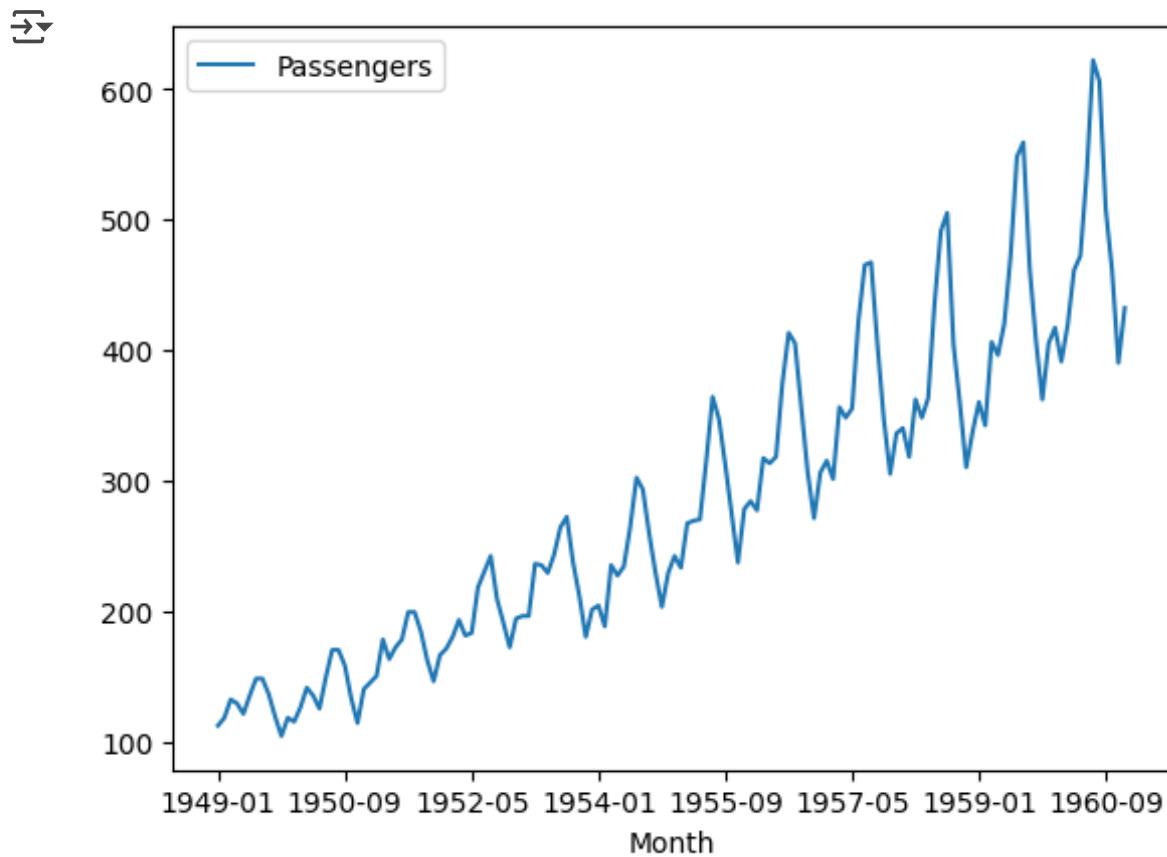
urllib.request.urlretrieve('https://raw.githubusercontent.com/jbrownlee/Datasets/1f9b79c/airline-passengers.csv')

# check file exist
os.path.exists('airline-passengers.csv')
```

→ True

▼ Visualize the data

```
series = read_csv('airline-passengers.csv', header=0, index_col=0)
series.plot()
pyplot.show()
```



```
# Load the dataset
series = pd.read_csv('airline-passengers.csv', header=0, index_col=0)
values = series['Passengers'].values.astype('float32').reshape(-1, 1)

# Scale the data
scaler = MinMaxScaler(feature_range=(-1, 1))
scaled_values = scaler.fit_transform(values)

# Define a function to create sequences for autoregression
def create_sequences(data, seq_length):
    xs = []
    ys = []
    for i in range(len(data) - seq_length):
        x = data[i:(i + seq_length)]
        y = data[i + seq_length]
        xs.append(x)
        ys.append(y)
    return np.array(xs), np.array(ys)

# Define the sequence length (how many past values to use for prediction)
seq_length = 12 # You can experiment with this value

# Create sequences
X, y = create_sequences(scaled_values, seq_length)

# Split the data into training and testing sets
train_size = int(len(X) * 0.8)
X_train, X_test = X[:train_size], X[train_size:]
y_train, y_test = y[:train_size], y[train_size:]

# Convert to PyTorch tensors
X_train_tensor = torch.tensor(X_train, dtype=torch.float32)
X_test_tensor = torch.tensor(X_test, dtype=torch.float32)
y_train_tensor = torch.tensor(y_train, dtype=torch.float32)
y_test_tensor = torch.tensor(y_test, dtype=torch.float32)

print("X_train_tensor shape:", X_train_tensor.shape)
print("X_test_tensor shape:", X_test_tensor.shape)
print("y_train_tensor shape:", y_train_tensor.shape)
print("y_test_tensor shape:", y_test_tensor.shape)

→ X_train_tensor shape: torch.Size([105, 12, 1])
   X_test_tensor shape: torch.Size([27, 12, 1])
   y_train_tensor shape: torch.Size([105, 1])
   y_test_tensor shape: torch.Size([27, 1])
```

```
# Create Datasets and DataLoaders
class SequenceDataset(Dataset):
    def __init__(self, x, y):
        self.x = x
        self.y = y

    def __len__(self):
        return len(self.x)

    def __getitem__(self, idx):
        return self.x[idx], self.y[idx]

batch_size = 32
train_dataset = SequenceDataset(X_train_tensor, y_train_tensor)
test_dataset = SequenceDataset(X_test_tensor, y_test_tensor)
train_loader = DataLoader(train_dataset, batch_size=batch_size, shuffle=True)
test_loader = DataLoader(test_dataset, batch_size=batch_size, shuffle=False)

# Define the Autoregressive DNN model
class AutoregressiveDNN(nn.Module):
    def __init__(self, input_size, hidden_size1, hidden_size2, output_size):
        super(AutoregressiveDNN, self).__init__()
        self.fc1 = nn.Linear(input_size, hidden_size1)
        self.relu1 = nn.ReLU()
        self.fc2 = nn.Linear(hidden_size1, hidden_size2)
        self.relu2 = nn.ReLU()
        self.fc3 = nn.Linear(hidden_size2, output_size)

    def forward(self, x):
        # Flatten the input sequence
        x = x.view(x.size(0), -1)
        x = self.fc1(x)
        x = self.relu1(x)
        x = self.fc2(x)
        x = self.relu2(x)
        x = self.fc3(x)
        return x
```

```
# Instantiate the model
input_size = seq_length # Input is a sequence of past 'seq_length' values
hidden_size1 = 64
hidden_size2 = 32
output_size = 1 # Output is the prediction for the next single value
model = AutoregressiveDNN(input_size, hidden_size1, hidden_size2, output_size)

# Define loss function and optimizer
criterion = nn.MSELoss()
optimizer = optim.Adam(model.parameters(), lr=0.001)

# Training loop
num_epochs = 100
for epoch in range(num_epochs):
    model.train()
    for sequences, targets in train_loader:
        # Forward pass
        outputs = model(sequences)
        loss = criterion(outputs, targets)

        # Backward and optimize
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

    if (epoch + 1) % 10 == 0:
        print(f'Epoch [{epoch+1}/{num_epochs}], Loss: {loss.item():.4f}')
```

→ Epoch [10/100], Loss: 0.0083
Epoch [20/100], Loss: 0.0149
Epoch [30/100], Loss: 0.0025
Epoch [40/100], Loss: 0.0049
Epoch [50/100], Loss: 0.0012
Epoch [60/100], Loss: 0.0040
Epoch [70/100], Loss: 0.0017
Epoch [80/100], Loss: 0.0014
Epoch [90/100], Loss: 0.0018
Epoch [100/100], Loss: 0.0014

```
# Evaluation
model.eval()
with torch.no_grad():
    test_predictions = model(X_test_tensor)
    test_loss = criterion(test_predictions, y_test_tensor)
    print(f'Test Loss: {test_loss.item():.4f}')

# Inverse transform the predictions and actual values
predicted_values = scaler.inverse_transform(test_predictions.numpy())
actual_values = scaler.inverse_transform(y_test_tensor.numpy())

# Prepare for plotting
train_actual = scaler.inverse_transform(y_train_tensor.numpy())
train_predictions = model(X_train_tensor).detach().numpy()
train_predictions = scaler.inverse_transform(train_predictions)

# Create a sequence of the original dates (aligning with the test set)
original_dates = pd.to_datetime(series.index)
test_dates = original_dates[-len(actual_values):]
train_dates_plot = original_dates[seq_length:train_size + seq_length]
```

→ Test Loss: 0.0581

```
# Plot the results
plt.figure(figsize=(12, 6))
plt.plot(original_dates, scaler.inverse_transform(scaled_values), label='Original Data')
# plt.plot(train_dates_plot, train_actual, label='Actual Train Data (Aligned)')
plt.plot(train_dates_plot, train_predictions, label='Train Predictions')
# plt.plot(test_dates, actual_values, label='Actual Test Data')
plt.plot(test_dates, predicted_values, label='DNN Autoregression Predictions')
plt.xlabel('Date')
plt.ylabel('Number of Passengers')
plt.title('Airline Passengers Autoregression using DNN')
plt.legend()
plt.grid(True)
plt.show()
```

