Sieci neuronowe

Wstęp

Celem laboratorium jest zapoznanie się z podstawami sieci neuronowych oraz uczeniem głębokim (*deep learning*). Zapoznasz się na nim z następującymi tematami:

- treningiem prostych sieci neuronowych, w szczególności z:
  - regresją liniową w sieciach neuronowych
  - optymalizacją funkcji kosztu
  - algorytmem spadku wzdłuż gradientu
  - siecią typu Multilayer Perceptron (MLP)
- frameworkiem PyTorch, w szczególności z:
  - ładowaniem danych
  - preprocessingiem danych
  - pisaniem pętli treningowej i walidacyjnej
  - walidacją modeli
- architekturą i hiperaprametrami sieci MLP, w szczególności z:
  - warstwami gęstymi (w pełni połączonymi)
  - funkcjami aktywacji
  - regularyzacją: L2, dropout

Wykorzystywane biblioteki

Zaczniemy od pisania ręcznie prostych sieci w bibliotece Numpy, służącej do obliczeń numerycznych na CPU. Później przejdziemy do wykorzystywania frameworka PyTorch, służącego do obliczeń numerycznych na CPU, GPU oraz automatycznego różniczkowania, wykorzystywanego głównie do treningu sieci neuronowych.

Wykorzystamy PyTorcha ze względu na popularność, łatwość instalacji i użycia, oraz dużą kontrolę nad niskopoziomowymi aspektami budowy i treningu sieci neuronowych. Framework ten został stworzony do zastosowań badawczych i naukowych, ale ze względu na wygodę użycia stał się bardzo popularny także w przemyśle. W szczególności całkowicie zdominował przetwarzanie języka naturalnego (NLP) oraz uczenie na grafach.

Pierwszy duży framework do deep learningu, oraz obecnie najpopularniejszy, to TensorFlow, wraz z wysokopoziomową nakładką Keras. Są jednak szanse, że Google (autorzy) będzie go powoli porzucać na rzecz ich nowego frameworka JAX (dyskusja, artykuł Business Insidera), który jest bardzo świeżym, ale ciekawym narzędziem.

Trzecia, ale znacznie mniej popularna od powyższych opcja to Apache MXNet.

Konfiguracja własnego komputera

Jeżeli korzystasz z własnego komputera, to musisz zainstalować trochę więcej bibliotek (Google Colab ma je już zainstalowane).

Jeżeli nie masz GPU lub nie chcesz z niego korzystać, to wystarczy znaleźć odpowiednią komendę CPU na stronie PyTorcha. Dla Anacondy odpowiednia komenda została podana poniżej, dla pip'a znajdź ją na stronie.

Jeżeli chcesz korzystać ze wsparcia GPU (na tym laboratorium nie będzie potrzebne, na kolejnych może przyspieszyć nieco obliczenia), to musi być to odpowiednio nowa karta NVidii, mająca CUDA compatibility (lista). Poza PyTorchem będzie potrzebne narzędzie NVidia CUDA w wersji 11.6 lub 11.7. Instalacja na Windowsie jest bardzo prosta (wystarczy ściągnąć plik EXE i zainstalować jak każdy inny program). Instalacja na Linuxie jest trudna i można względnie łatwo zepsuć sobie system, ale jeżeli chcesz spróbować, to ten tutorial jest bardzo dobry.

In [1]:
```
# for conda users
!conda install -y matplotlib pandas pytorch torchvision -c pytorch -c conda-forge
```

```
Retrieving notices: ...working... done
Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

## Package Plan ##

  environment location: G:\anaconda3

  added / updated specs:
    - matplotlib
    - pandas
    - pytorch
    - torchvision


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    libjpeg-turbo-2.0.0        |       h196d8e1_0         618 KB
    libuv-1.44.2               |       h8ffe710_0         362 KB  conda-forge
    mpmath-1.3.0               |     pyhd8ed1ab_0         428 KB  conda-forge
    networkx-3.2.1             |     pyhd8ed1ab_0         1.1 MB  conda-forge
    pytorch-2.1.0              |     py3.11_cpu_0       166.3 MB  pytorch
    pytorch-mutex-1.0          |             cpu           3 KB  pytorch
    sympy-1.12                 |     pyh04b8f61_3         4.0 MB  conda-forge
    torchvision-0.16.0         |        py311_cpu         6.9 MB  pytorch
    ------------------------------------------------------------
                                           Total:       179.7 MB

The following NEW packages will be INSTALLED:

  libjpeg-turbo      pkgs/main/win-64::libjpeg-turbo-2.0.0-h196d8e1_0
  libuv              conda-forge/win-64::libuv-1.44.2-h8ffe710_0
  matplotlib         pkgs/main/win-64::matplotlib-3.8.0-py311haa95532_0
  mpmath             conda-forge/noarch::mpmath-1.3.0-pyhd8ed1ab_0
  networkx           conda-forge/noarch::networkx-3.2.1-pyhd8ed1ab_0
  pytorch            pytorch/win-64::pytorch-2.1.0-py3.11_cpu_0
  pytorch-mutex      pytorch/noarch::pytorch-mutex-1.0-cpu
  sympy              conda-forge/noarch::sympy-1.12-pyh04b8f61_3
  torchvision        pytorch/win-64::torchvision-0.16.0-py311_cpu

The following packages will be SUPERSEDED by a higher-priority channel:
```

```
  ca-certificates    pkgs/main::ca-certificates-2023.08.22~ --> conda-forge::ca-certific
ates-2023.7.22-h56e8100_0
  certifi            pkgs/main/win-64::certifi-2023.7.22-p~ --> conda-forge/noarch::cert
ifi-2023.7.22-pyhd8ed1ab_0



Downloading and Extracting Packages: ...working... done
Preparing transaction: ...working... done
Verifying transaction: ...working... done
Executing transaction: ...working... done

==> WARNING: A newer version of conda exists. <==
  current version: 23.9.0
  latest version: 23.10.0

Please update conda by running

    $ conda update -n base -c defaults conda

Or to minimize the number of packages updated during conda update use

     conda install conda=23.10.0
```

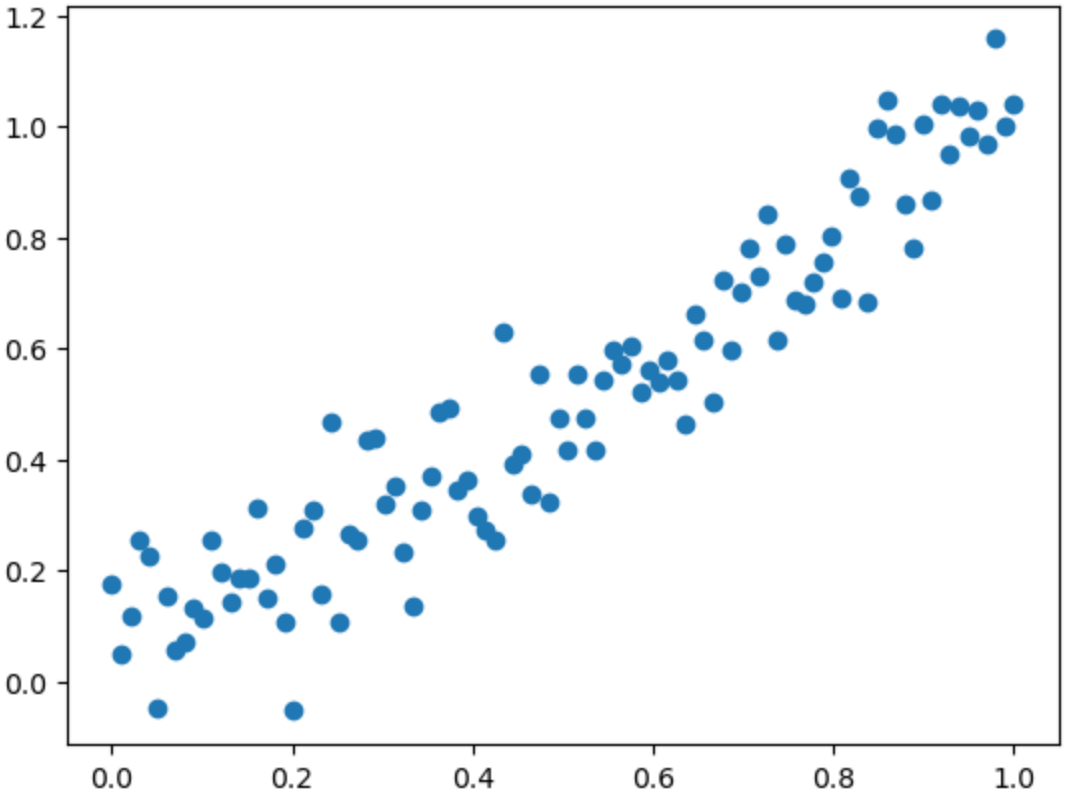
Wprowadzenie

Zanim zaczniemy naszą przygodę z sieciami neuronowymi, przyjrzyjmy się prostemu przykładowi regresji liniowej na syntetycznych danych:

In [2]:
```
from typing import Tuple, Dict

import numpy as np
import matplotlib.pyplot as plt
```

In [3]:
```
np.random.seed(0)

x = np.linspace(0, 1, 100)
y = x + np.random.normal(scale=0.1, size=x.shape)

plt.scatter(x, y)
```

Out[3]:
```
<matplotlib.collections.PathCollection at 0x13f98209110>
```



W przeciwieństwie do laboratorium 1, tym razem będziemy chcieli rozwiązać ten problem własnoręcznie, bez użycia wysokopoziomowego interfejsu Scikit-learn'a. W tym celu musimy sobie przypomnieć sformułowanie naszego **problemu optymalizacyjnego (optimization problem)**.

W przypadku prostej regresji liniowej (1 zmienna) mamy model postaci $\hat{y} = \alpha x + \beta$, z dwoma parametrami, których będziemy się uczyć. Miarą niedopasowania modelu o danych parametrach jest **funkcja kosztu (cost function)**, nazywana też funkcją celu. Najczęściej używa się **błędu średniokwadratowego (mean squared error, MSE)**:

$$MSE = \frac{1}{N}\sum_{i}^{N}(y - \hat{y})^2$$

Od jakich α i β zacząć? W najprostszym wypadku wystarczy po prostu je wylosować jako niewielkie liczby zmiennoprzecinkowe.

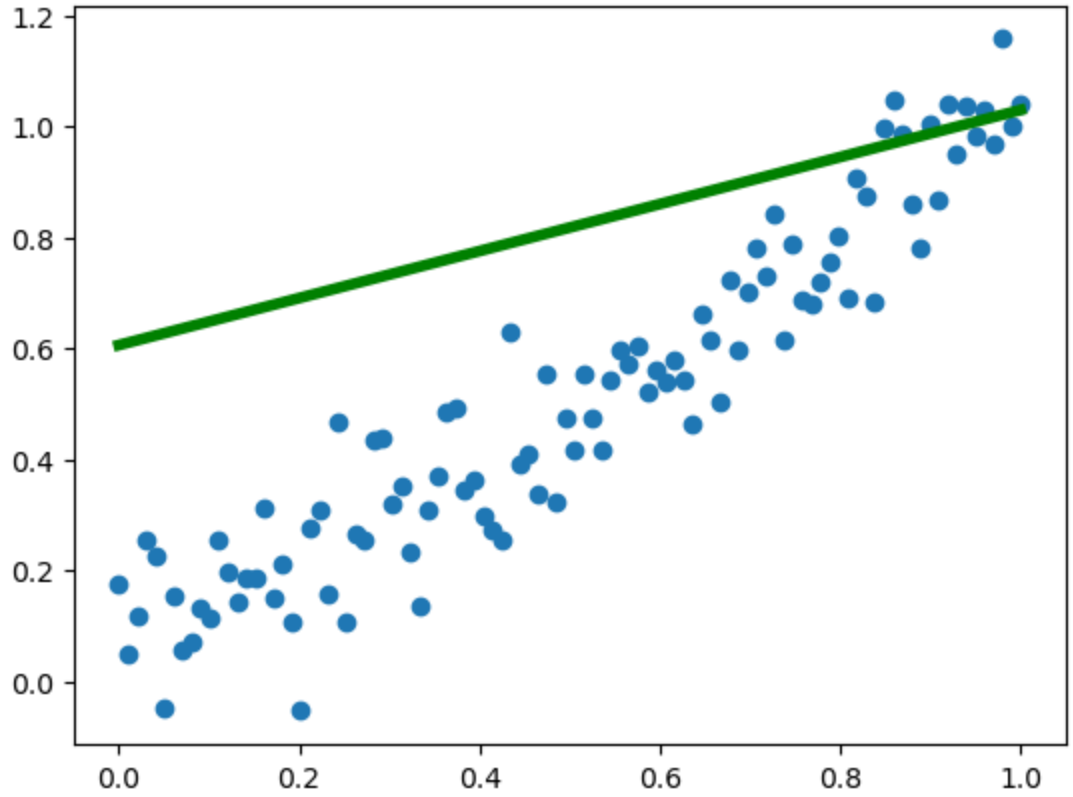
Zadanie 1 (0.5 punkt)

Uzupełnij kod funkcji `mse`, obliczającej błąd średniokwadratowy. Wykorzystaj Numpy'a w celu wektoryzacji obliczeń dla wydajności.

In [4]:
```
def mse(y: np.ndarray, y_hat: np.ndarray) -> float:
    return (np.square(y - y_hat)).sum() / len(y)
```

In [5]:
```
a = np.random.rand()
b = np.random.rand()
print(f"MSE: {mse(y, a * x + b):.3f}")

plt.scatter(x, y)
plt.plot(x, a * x + b, color="g", linewidth=4)
```

Out[5]:
```
MSE: 0.133
[<matplotlib.lines.Line2D at 0x13f9838bd10>]
```



Losowe parametry radzą sobie nie najlepiej. Jak lepiej dopasować naszą prostą do danych? Zawsze możemy starać się wyprowadzić rozwiązanie analitycznie, i w tym wypadku nawet nam się uda. Jest to jednak szczególny i dość rzadki przypadek, a w szczególności nie będzie to możliwe w większych sieciach neuronowych.

Potrzebna nam będzie **metoda optymalizacji (optimization method)**, dającą wartości parametrów minimalizujące dowolną różniczkowalną funkcję kosztu. Zdecydowanie najpopularniejszy jest tutaj **spadek wzdłuż gradientu (gradient descent)**.

Metoda ta wywodzi się z prostych obserwacji, które tutaj przedstawimy. Bardziej szczegółowe rozwinięcie dla zainteresowanych: sekcja 4.3 "Deep Learning Book", ten praktyczny kurs, analiza oryginalnej publikacji Cauchy'ego (oryginał w języku francuskim).

Pochodna jest dokładnie równa granicy funkcji. Dla małego ϵ można ją przybliżyć jako:

$$\frac{f(x)}{dx} \approx \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

Przyglądając się temu równaniu widzimy, że:

- dla funkcji rosnącej ($f(x+\epsilon) > f(x)$) wyrażenie $\frac{f(x)}{dx}$ będzie miało znak dodatni
- dla funkcji malejącej ($f(x+\epsilon) < f(x)$) wyrażenie $\frac{f(x)}{dx}$ będzie miało znak ujemny

Widzimy więc, że potrafimy wskazać kierunek zmniejszenia wartości funkcji, patrząc na znak pochodnej. Zaobserwowano także, że amplituda wartości w $\frac{f(x)}{dx}$ jest tym większa, im dalej jesteśmy od minimum (maximum). Pochodna wyznacza więc, w jakim kierunku funkcja najszybciej rośnie, zaś przeciwny zwrot to ten, w którym funkcja najszybciej spada.

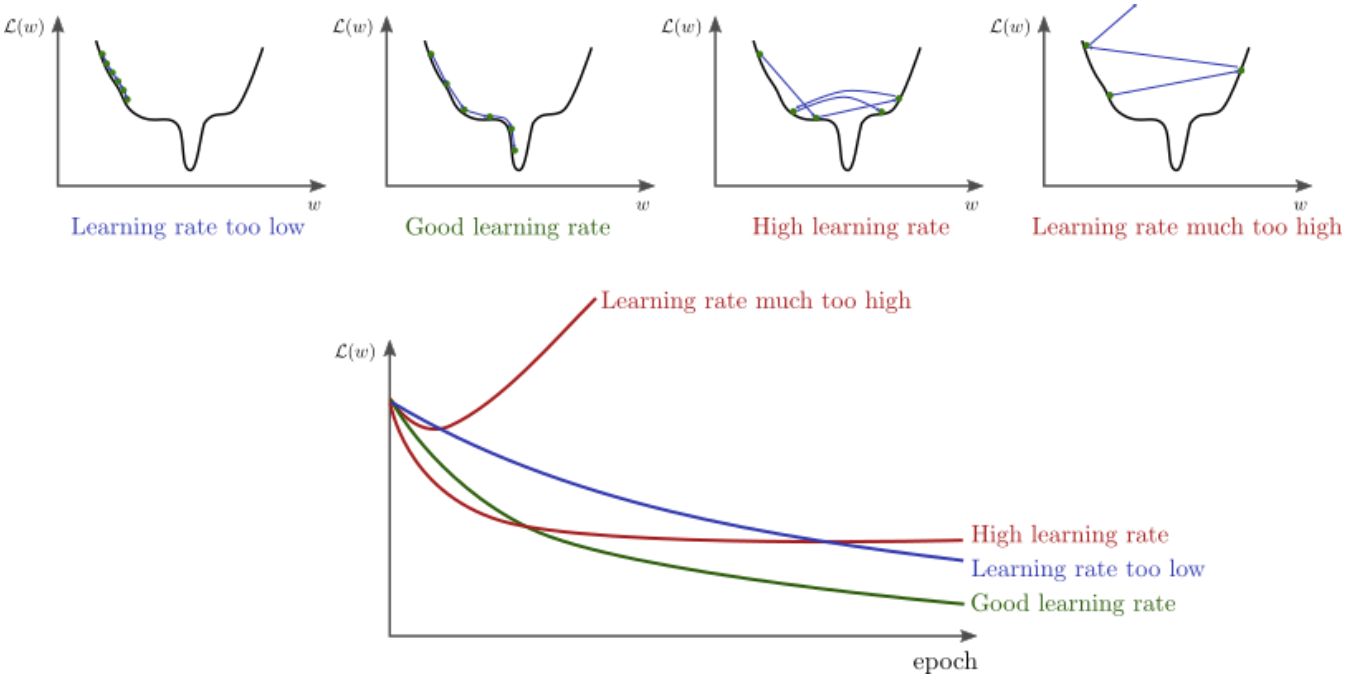
Stosując powyższe do optymalizacji, mamy:

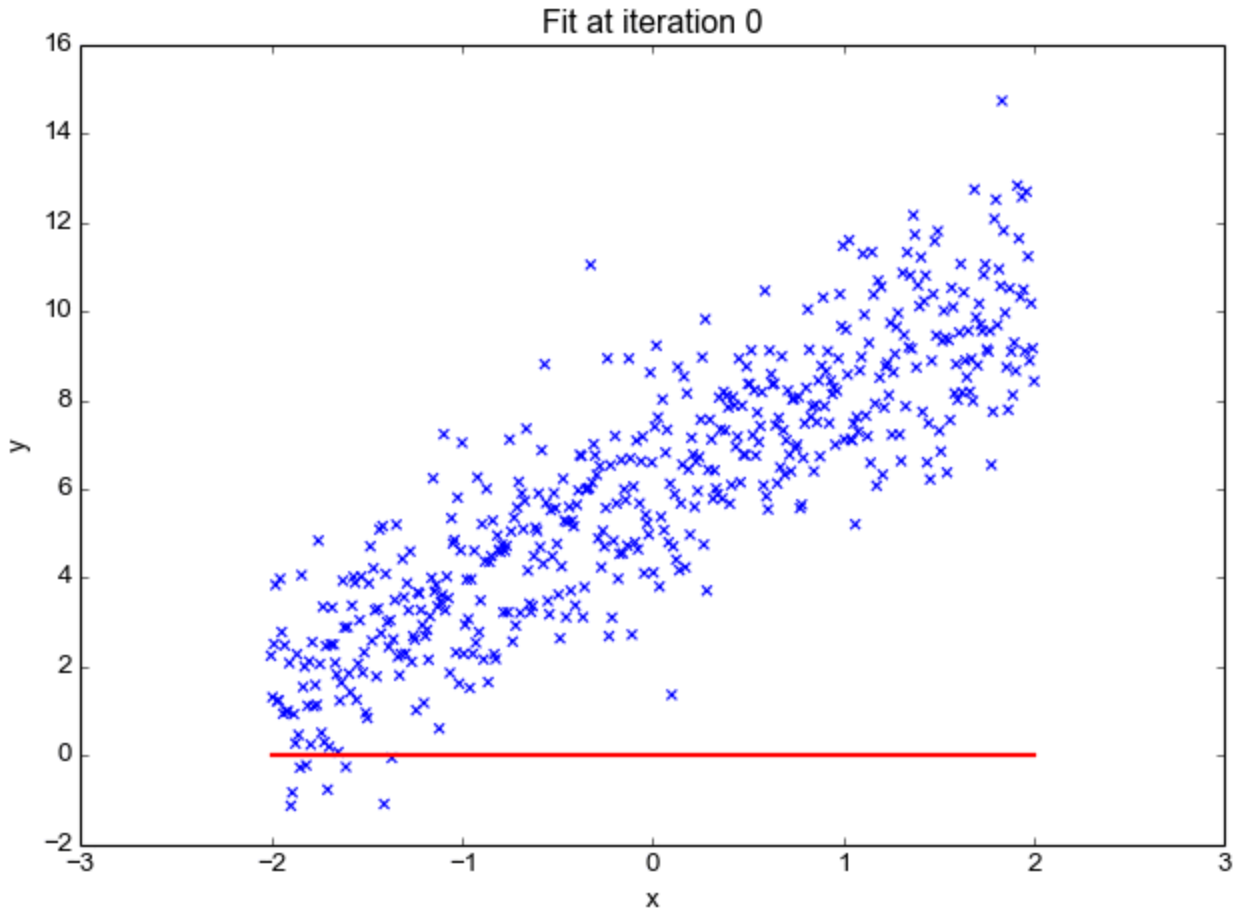
$$x_{t+1} = x_t - \alpha * \frac{f(x)}{dx}$$

α to niewielka wartość (rzędu zwykle 10^{-5} - 10^{-2}), wprowadzona, aby trzymać się założenia o małej zmianie parametrów (ϵ). Nazywa się ją **stałą uczącą (learning rate)** i jest zwykle najważniejszym hiperparametrem podczas nauki sieci.

Metoda ta zakłada, że używamy całego zbioru danych do aktualizacji parametrów w każdym kroku, co nazywa się po prostu GD (od *gradient descent*) albo *full batch GD*. Wtedy każdy krok optymalizacji nazywa się **epoką (epoch)**.

Im większa stała ucząca, tym większe nasze kroki podczas minimalizacji. Możemy więc uczyć szybciej, ale istnieje ryzyko, że będziemy "przeskakiwać" minima. Mniejsza stała ucząca to wolniejszy, ale dokładniejszy trening. Jednak nie zawsze ona pozwala osiągnąć lepsze wyniki, bo może okazać się, że utkniemy w minimum lokalnym. Można także zmieniać stałą uczącą podczas treningu, co nazywa się **learning rate scheduling (LR scheduling)**. Obrazowo:





Policzmy więc pochodną dla naszej funkcji kosztu MSE. Pochodną liczymy po parametrach naszego modelu, bo to właśnie ich chcemy dopasować tak, żeby koszt był jak najmniejszy:

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2$$

W powyższym wzorze tylko y_i jest zależny od a oraz b. Możemy wykorzystać tu regułę łańcuchową (*chain rule*) i policzyć pochodne po naszych parametrach w sposób następujący:

$$\frac{\mathrm{d}MSE}{\mathrm{d}a} = \frac{1}{N}\sum_{i}^{N}\frac{\mathrm{d}(y_i - \hat{y}_i)^2}{\mathrm{d}\hat{y}_i}\frac{\mathrm{d}\hat{y}_i}{\mathrm{d}a}$$

$$\frac{\mathrm{d}MSE}{\mathrm{d}b} = \frac{1}{N}\sum_{i}^{N}\frac{\mathrm{d}(y_i - \hat{y}_i)^2}{\mathrm{d}\hat{y}_i}\frac{\mathrm{d}\hat{y}_i}{\mathrm{d}b}$$

Policzmy te pochodne po kolei:

$$\frac{\mathrm{d}(y_i - \hat{y}_i)^2}{\mathrm{d}\hat{y}_i} = -2 \cdot (y_i - \hat{y}_i)$$

$$\frac{\mathrm{d}\hat{y}_i}{\mathrm{d}a} = x_i$$

$$\frac{\mathrm{d}\hat{y}_i}{\mathrm{d}b} = 1$$

Łącząc powyższe wyniki dostaniemy:

$$\frac{\mathrm{d}MSE}{\mathrm{d}a} = \frac{-2}{N}\sum_{i}^{N}(y_i - \hat{y}_i)\cdot x_i$$

$$\frac{\mathrm{d}MSE}{\mathrm{d}b} = \frac{-2}{N}\sum_{i}^{N}(y_i - \hat{y}_i)$$

Aktualizacja parametrów wygląda tak:

$$a' = a - \alpha * \left(\frac{-2}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)\cdot x_i\right)$$

$$b' = b - \alpha * \left(\frac{-2}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)\right)$$

Liczymy więc pochodną funkcji kosztu, a potem za pomocą reguły łańcuchowej "cofamy się", dochodząc do tego, jak każdy z parametrów wpływa na błąd i w jaki sposób powinniśmy go zmienić. Nazywa się to **propagacją wsteczną (backpropagation)** i jest podstawowym mechanizmem umożliwiającym naukę sieci neuronowych za pomocą spadku wzdłuż gradientu. Więcej możesz o tym przeczytać tutaj.

Zadanie 2 (1.0 punkt)

Zaimplementuj funkcję realizującą jedną epokę treningową. Zauważ, że `x` oraz `y` są wektorami. Oblicz predykcję przy aktualnych parametrach oraz zaktualizuj je zgodnie z powyższymi wzorami.

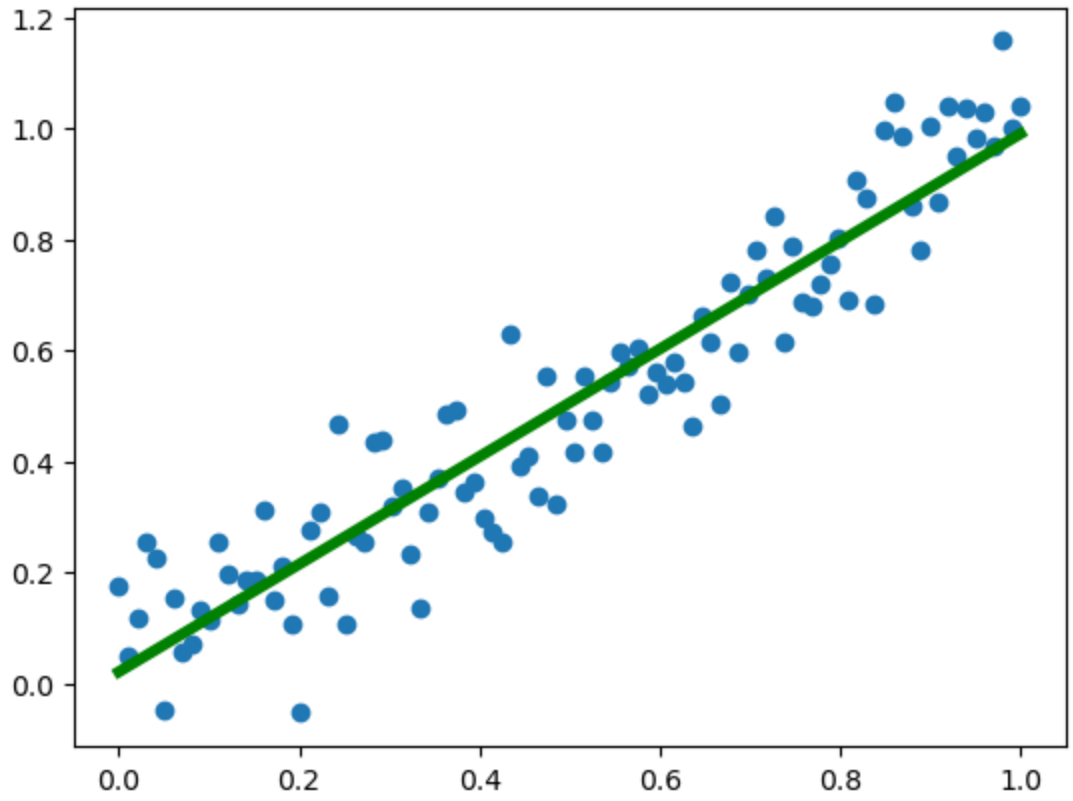
In [6]:
```
def optimize(
    x: np.ndarray, y: np.ndarray, a: float, b: float, learning_rate: float = 0.1
):
    y_hat = a * x + b
    errors = y - y_hat
    n = len(x)
    new_a = a - learning_rate * ((-2/n) * sum(x * errors))
    new_b = b - learning_rate * ((-2/n) * sum(errors))
    return new_a, new_b
```

In [7]:
```
for i in range(1000):
    loss = mse(y, a * x + b)
    a, b = optimize(x, y, a, b)
    if i % 100 == 0:
        print(f"step {i} loss: ", loss)

print("final loss:", loss)
```

```
step 0 loss:  0.1330225119404028
step 100 loss:  0.012673197778527 68
step 200 loss:  0.010257153540857817
step 300 loss:  0.0100948037549359
step 400 loss:  0.010083894412889118
step 500 loss:  0.010083161342973332
step 600 loss:  0.010083112083219709
step 700 loss:  0.010083108773135263
step 800 loss:  0.010083108550709076
step 900 loss:  0.01008310853576281
final loss: 0.010083108534760455
```

In [8]:
```
plt.scatter(x, y)
plt.plot(x, a * x + b, color="g", linewidth=4)
```

Out[8]:
```
[<matplotlib.lines.Line2D at 0x13f98c49610>]
```



Udało ci się wytrenować swoją pierwszą sieć neuronową. Czemu? Otóż neuron to po prostu wektor parametrów, a zwykle robimy iloczyn skalarny tych parametrów z wejściem. Dodatkowo na wyjście nakłada się **funkcję aktywacji (activation function)**, która przekształca wyjście. Tutaj takiej nie było, a właściwie była to po prostu funkcja identyczności.

Oczywiście w praktyce korzystamy z odpowiedniego frameworka, który w szczególności:

- ułatwia budowanie sieci, np. ma gotowe klasy dla warstw neuronów
- ma zaimplementowane funkcje kosztu oraz ich pochodne
- sam różniczkuje ze względu na odpowiednie parametry i aktualizuje je odpowiednio podczas treningu

Wprowadzenie do PyTorcha

PyTorch to w gruncie rzeczy narzędzie do algebry liniowej z automatycznym rożniczkowaniem, z możliwością przyspieszenia obliczeń z pomocą GPU. Na tych fundamentach zbudowany jest pełny framework do uczenia głębokiego. Można spotkać się ze stwierdzenie, że PyTorch to NumPy + GPU + opcjonalne różniczkowanie, co jest całkiem celne. Plus można łatwo debugować printem :)

PyTorch używa dynamicznego grafu obliczeń, który sami definiujemy w kodzie. Takie podejście jest bardzo wygodne, elastyczne i pozwala na łatwe eksperymentowanie. Odbywa się to potencjalnie kosztem wydajności, ponieważ pozostawia kwestię optymalizacji programiście. Więcej na ten temat dla zainteresowanych na końcu laboratorium.

Samo API PyTorcha bardzo przypomina Numpy'a, a podstawowym obiektem jest `Tensor`, klasa reprezentująca tensory dowolnego wymiaru. Dodatkowo niektóre tensory będą miały automatycznie

obliczony gradient. Co ważne, tensor jest na pewnym urządzeniu, CPU lub GPU, a przenosić między nimi trzeba explicite.

Najważniejsze moduły:

- `torch` - podstawowe klasy oraz funkcje, np. `Tensor`, `from_numpy()`
- `torch.nn` - klasy związane z sieciami neuronowymi, np. `Linear`, `Sigmoid`
- `torch.optim` - wszystko związane z optymalizacją, głównie spadkiem wzdłuż gradientu

In [9]:
```
import torch
import torch.nn as nn
import torch.optim as optim
```

In [10]:
```
ones = torch.ones(10)
noise = torch.ones(10) * torch.rand(10)

# elementwise sum
print(ones + noise)

# elementwise multiplication
print(ones * noise)

# dot product
print(ones @ noise)
```

```
tensor([1.7475, 1.8848, 1.5492, 1.2163, 1.2911, 1.5440, 1.7616, 1.7570, 1.0739,
        1.9237])
tensor([0.7475, 0.8848, 0.5492, 0.2163, 0.2911, 0.5440, 0.7616, 0.7570, 0.0739,
        0.9237])
tensor(5.7492)
```

In [11]:
```
# beware - shares memory with original Numpy array!
# very fast, but modifications are visible to original variable
x = torch.from_numpy(x)
y = torch.from_numpy(y)
```

Jeżeli dla stworzonych przez nas tensorów chcemy śledzić operacje i obliczać gradient, to musimy oznaczyć `requires_grad=True`.

In [12]:
```
a = torch.rand(1, requires_grad=True)
b = torch.rand(1, requires_grad=True)
a, b
```

Out[12]:
```
(tensor([0.2601], requires_grad=True), tensor([0.3181], requires_grad=True))
```

PyTorch zawiera większość powszechnie używanych funkcji kosztu, np. MSE. Mogą być one używane na 2 sposoby, z czego pierwszy jest popularniejszy:

- jako klasy wywoływalne z modułu `torch.nn`
- jako funkcje z modułu `torch.nn.functional`

Po wykonaniu poniższego kodu widzimy, że zwraca on nam tensor z dodatkowymi atrybutami. Co ważne, jest to skalar (0-wymiarowy tensor), bo potrzebujemy zwyczajnej liczby do obliczania propagacji wstecznych (pochodnych cząstkowych).

In [13]:
```
mse = nn.MSELoss()
mse(y, a * x + b)
```

Out[13]:
```
tensor(0.0563, dtype=torch.float64, grad_fn=<MseLossBackward0>)
```

Atrybutu `grad_fn` nie używamy wprost, bo korzysta z niego w środku PyTorch, ale widać, że tensor jest "świadomy", że liczy się na nim pochodną. Możemy natomiast skorzystać z atrybutu `grad`, który zawiera faktyczny gradient. Zanim go jednak dostaniemy, to trzeba powiedzieć PyTorchowi, żeby policzył gradient. Służy do tego metoda `.backward()`, wywoływana na obiekcie zwracanym przez funkcję kosztu.

In [14]:
```
loss = mse(y, a * x + b)
loss.backward()
```

In [15]:
```
print(a.grad)
```

```
tensor([-0.1786])
```

Ważne jest, że PyTorch nie liczy za każdym razem nowego gradientu, tylko dodaje go do istniejącego, czyli go akumuluje. Jest to przydatne w niektórych sieciach neuronowych, ale zazwyczaj trzeba go zerować. Jeżeli tego nie zrobimy, to dostaniemy coraz większe gradienty.

Do zerowania służy metoda `.zero_()`. W PyTorchu wszystkie metody modyfikujące tensor w miejscu mają `_` na końcu nazwy. Jest to dość niskopoziomowa operacja dla pojedynczych tensorów - zobaczymy za chwilę, jak to robić łatwiej dla całej sieci.

In [16]:
```
loss = mse(y, a * x + b)
loss.backward()
a.grad
```

Out[16]:
```
tensor([-0.3571])
```

Zobaczmy, jak wyglądałaby regresja liniowa, ale napisana w PyTorchu. Jest to oczywiście bardzo niskopoziomowa implementacja - za chwilę zobaczymy, jak to wygląda w praktyce.

In [17]:
```
learning_rate = 0.1
for i in range(1000):
    loss = mse(y, a * x + b)

    # compute gradients
    loss.backward()

    # update parameters
    a.data -= learning_rate * a.grad
    b.data -= learning_rate * b.grad

    # zero gradients
    a.grad.data.zero_()
    b.grad.data.zero_()

    if i % 100 == 0:
        print(f"step {i} loss: ", loss)

print("final loss:", loss)
```

```
step 0 loss:  tensor(0.0563, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 100 loss:  tensor(0.0126, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 200 loss:  tensor(0.0103, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 300 loss:  tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 400 loss:  tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 500 loss:  tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 600 loss:  tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 700 loss:  tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
step 800 loss:  tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
```

```
step 900 loss:  tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
final loss: tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
```

Trening modeli w PyTorchu jest dosyć schematyczny i najczęściej rozdziela się go na kilka bloków, dających razem **pętlę uczącą (training loop)**, powtarzaną w każdej epoce:

1. Forward pass - obliczenie predykcji sieci
2. Loss calculation
3. Backpropagation - obliczenie pochodnych oraz zerowanie gradientów
4. Optimalization - aktualizacja wag
5. Other - ewaluacja na zbiorze walidacyjnym, logging etc.

In [18]:
```python
# initialization
learning_rate = 0.1
a = torch.rand(1, requires_grad=True)
b = torch.rand(1, requires_grad=True)
optimizer = torch.optim.SGD([a, b], lr=learning_rate)
best_loss = float("inf")

# training loop in each epoch
for i in range(1000):
    # forward pass
    y_hat = a * x + b

    # loss calculation
    loss = mse(y, y_hat)

    # backpropagation
    loss.backward()

    # optimization
    optimizer.step()
    optimizer.zero_grad()  # zeroes all gradients - very convenient!

    if i % 100 == 0:
        if loss < best_loss:
            best_model = (a.clone(), b.clone())
            best_loss = loss
        print(f"step {i} loss: {loss.item():.4f}")

print("final loss:", loss)
```

```
step 0 loss: 0.2709
step 100 loss: 0.0149
step 200 loss: 0.0104
step 300 loss: 0.0101
step 400 loss: 0.0101
step 500 loss: 0.0101
step 600 loss: 0.0101
step 700 loss: 0.0101
step 800 loss: 0.0101
step 900 loss: 0.0101
final loss: tensor(0.0101, dtype=torch.float64, grad_fn=<MseLossBackward0>)
```

Przejdziemy teraz do budowy sieci neuronowej do klasyfikacji. Typowo implementuje się ją po prostu jako sieć dla regresji, ale zwracającą tyle wyników, ile mamy klas, a potem aplikuje się na tym funkcję sigmoidalną (2 klasy) lub softmax (>2 klasy). W przypadku klasyfikacji binarnej zwraca się czasem tylko 1 wartość, przepuszczaną przez sigmoidę - wtedy wyjście z sieci to prawdopodobieństwo klasy pozytywnej.

Funkcją kosztu zwykle jest **entropia krzyżowa (cross-entropy)**, stosowana też w klasycznej regresji logistycznej. Co ważne, sieci neuronowe, nawet tak proste, uczą się szybciej i stabilniej, gdy dane na wejściu

(a przynajmniej zmienne numeryczne) są **ustandaryzowane (standardized)**. Operacja ta polega na odjęciu średniej i podzieleniu przez odchylenie standardowe (tzw. *Z-score transformation*).

Uwaga - PyTorch wymaga tensora klas będącego liczbami zmiennoprzecinkowymi!

Zbiór danych

Na tym laboratorium wykorzystamy zbiór Adult Census. Dotyczy on przewidywania na podstawie danych demograficznych, czy dany człowiek zarabia powyżej 50 tysięcy dolarów rocznie, czy też mniej. Jest to cenna informacja np. przy planowaniu kampanii marketingowych. Jak możesz się domyślić, zbiór pochodzi z czasów, kiedy inflacja była dużo niższa :)

Poniżej znajduje się kod do ściągnięcia i preprocessingu zbioru. Nie musisz go dokładnie analizować.

In [19]:
```
!wget https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
```

In [19]:
```
import pandas as pd


columns = [
    "age",
    "workclass",
    "fnlwgt",
    "education",
    "education-num",
    "marital-status",
    "occupation",
    "relationship",
    "race",
    "sex",
    "capital-gain",
    "capital-loss",
    "hours-per-week",
    "native-country",
    "wage"
]

"""
age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, W
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-spec
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying
"""

df = pd.read_csv("adult.data", header=None, names=columns)
df.wage.unique()
```

Out[19]:
```
array([' <=50K', ' >50K'], dtype=object)
```

In [21]:
```
# attribution: https://www.kaggle.com/code/royshih23/topic7-classification-in-python
```

```
df['education'].replace('Preschool', 'dropout',inplace=True)
df['education'].replace('10th', 'dropout',inplace=True)
df['education'].replace('11th', 'dropout',inplace=True)
df['education'].replace('12th', 'dropout',inplace=True)
df['education'].replace('1st-4th', 'dropout',inplace=True)
df['education'].replace('5th-6th', 'dropout',inplace=True)
df['education'].replace('7th-8th', 'dropout',inplace=True)
df['education'].replace('9th', 'dropout',inplace=True)
df['education'].replace('HS-Grad', 'HighGrad',inplace=True)
df['education'].replace('HS-grad', 'HighGrad',inplace=True)
df['education'].replace('Some-college', 'CommunityCollege',inplace=True)
df['education'].replace('Assoc-acdm', 'CommunityCollege',inplace=True)
df['education'].replace('Assoc-voc', 'CommunityCollege',inplace=True)
df['education'].replace('Bachelors', 'Bachelors',inplace=True)
df['education'].replace('Masters', 'Masters',inplace=True)
df['education'].replace('Prof-school', 'Masters',inplace=True)
df['education'].replace('Doctorate', 'Doctorate',inplace=True)

df['marital-status'].replace('Never-married', 'NotMarried',inplace=True)
df['marital-status'].replace(['Married-AF-spouse'], 'Married',inplace=True)
df['marital-status'].replace(['Married-civ-spouse'], 'Married',inplace=True)
df['marital-status'].replace(['Married-spouse-absent'], 'NotMarried',inplace=True)
df['marital-status'].replace(['Separated'], 'Separated',inplace=True)
df['marital-status'].replace(['Divorced'], 'Separated',inplace=True)
df['marital-status'].replace(['Widowed'], 'Widowed',inplace=True)
```

In [46]:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, OneHotEncoder, StandardScaler


X = df.copy()
y = (X.pop("wage") == ' >50K').astype(int).values

train_valid_size = 0.2

X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=train_valid_size,
    random_state=0,
    shuffle=True,
    stratify=y
)
X_train, X_valid, y_train, y_valid = train_test_split(
    X_train, y_train,
    test_size=train_valid_size,
    random_state=0,
    shuffle=True,
    stratify=y_train
)

continuous_cols = ['age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss', 'ho
continuous_X_train = X_train[continuous_cols]
categorical_X_train = X_train.loc[:, ~X_train.columns.isin(continuous_cols)]

continuous_X_valid = X_valid[continuous_cols]
categorical_X_valid = X_valid.loc[:, ~X_valid.columns.isin(continuous_cols)]

continuous_X_test = X_test[continuous_cols]
categorical_X_test = X_test.loc[:, ~X_test.columns.isin(continuous_cols)]

categorical_encoder = OneHotEncoder(sparse=False, handle_unknown='ignore')
continuous_scaler = StandardScaler() #MinMaxScaler(feature_range=(-1, 1))

categorical_encoder.fit(categorical_X_train)
continuous_scaler.fit(continuous_X_train)
```

```
continuous_X_train = continuous_scaler.transform(continuous_X_train)
continuous_X_valid = continuous_scaler.transform(continuous_X_valid)
continuous_X_test = continuous_scaler.transform(continuous_X_test)

categorical_X_train = categorical_encoder.transform(categorical_X_train)
categorical_X_valid = categorical_encoder.transform(categorical_X_valid)
categorical_X_test = categorical_encoder.transform(categorical_X_test)

X_train = np.concatenate([continuous_X_train, categorical_X_train], axis=1)
X_valid = np.concatenate([continuous_X_valid, categorical_X_valid], axis=1)
X_test = np.concatenate([continuous_X_test, categorical_X_test], axis=1)

X_train.shape, y_train.shape
```

Out[46]:
```
((20838, 108), (20838,))
```

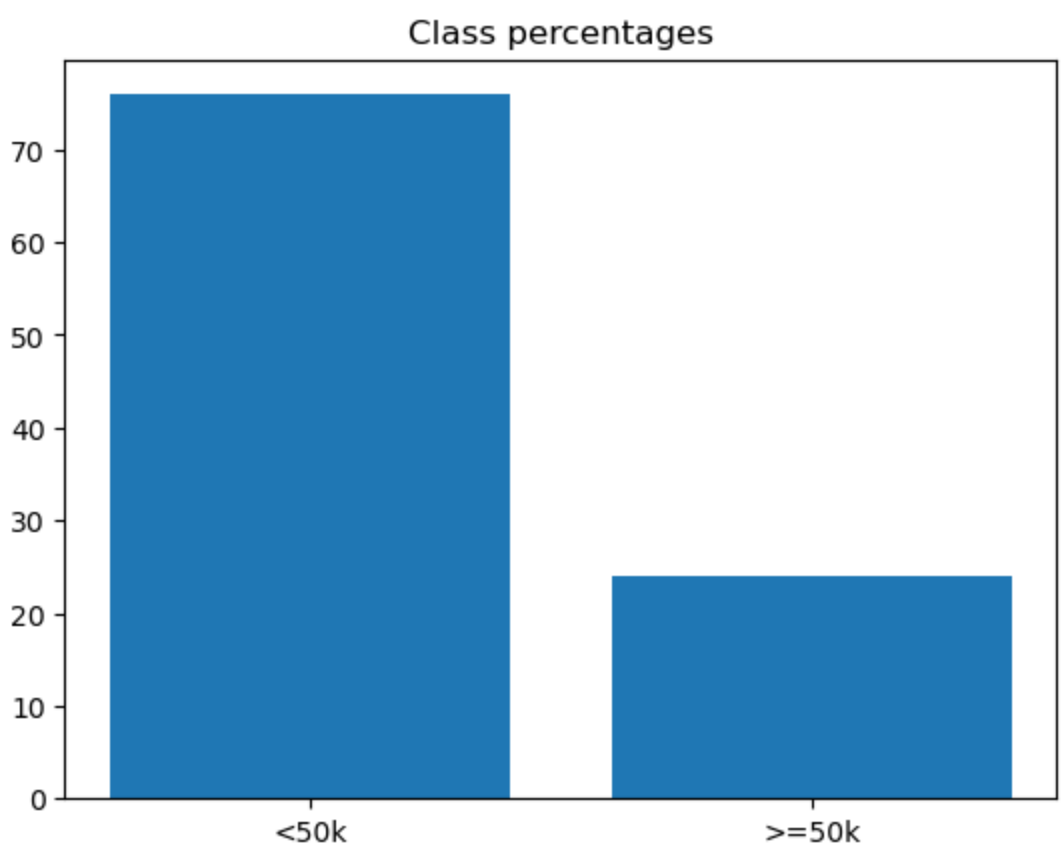
Uwaga co do typów - PyTorchu wszystko w sieci neuronowej musi być typu `float32`. W szczególności trzeba uważać na konwersje z Numpy'a, który używa domyślnie typu `float64`. Może ci się przydać metoda `.float()`.

Uwaga co do kształtów wyjścia - wejścia do `nn.BCELoss` muszą być tego samego kształtu. Może ci się przydać metoda `.squeeze()` lub `.unsqueeze()`.

In [21]:
```
X_train = torch.from_numpy(X_train).float()
y_train = torch.from_numpy(y_train).float().unsqueeze(-1)

X_valid = torch.from_numpy(X_valid).float()
y_valid = torch.from_numpy(y_valid).float().unsqueeze(-1)

X_test = torch.from_numpy(X_test).float()
y_test = torch.from_numpy(y_test).float().unsqueeze(-1)
```

Podobnie jak w laboratorium 2, mamy tu do czynienia z klasyfikacją niezbalansowaną:

In [22]:
```
import matplotlib.pyplot as plt

y_pos_perc = 100 * y_train.sum().item() / len(y_train)
y_neg_perc = 100 - y_pos_perc

plt.title("Class percentages")
plt.bar(["<50k", ">=50k"], [y_neg_perc, y_pos_perc])
plt.show()
```



W związku z powyższym będziemy używać odpowiednich metryk, czyli AUROC, precyzji i czułości.

Zadanie 3 (1.0 punkt)

Zaimplementuj regresję logistyczną dla tego zbioru danych, używając PyTorcha. Dane wejściowe zostały dla ciebie przygotowane w komórkach poniżej.

Sama sieć składa się z 2 elementów:

- warstwa liniowa `nn.Linear`, przekształcająca wektor wejściowy na 1 wyjście - logit
- aktywacja sigmoidalna `nn.Sigmoid`, przekształcająca logit na prawdopodobieństwo klasy pozytywnej

Użyj binarnej entropii krzyżowej `nn.BCELoss` jako funkcji kosztu. Użyj optymalizatora SGD ze stałą uczącą `1e-3`. Trenuj przez 3000 epok. Pamiętaj, aby przekazać do optymalizatora `torch.optim.SGD` parametry sieci (metoda `.parameters()`). Dopisz logowanie kosztu raz na 100 epok.

In [23]:
```
learning_rate = 1e-3

model = nn.Linear(X_train.shape[1], 1)
activation = nn.Sigmoid()
optimizer = torch.optim.SGD(model.parameters(), lr=learning_rate)
loss_fn = nn.BCELoss()

for t in range(3000):
    y_pred = activation(model(X_train))

    loss = loss_fn(y_pred, y_train)
    if t % 100 == 99: print(t, loss.item())
    loss.backward()

    optimizer.step()
    optimizer.zero_grad()

print(f"final loss: {loss.item():.4f}")
```

```
99 0.656265914440155
199 0.6300392746925354
299 0.6080109477043152
399 0.5893544554710388
499 0.5734135508537292
599 0.5596693754196167
699 0.5477116703987122
799 0.5372163653373718
899 0.527925968170166
999 0.5196357369422913
1099 0.5121819376945496
1199 0.5054330229759216
1299 0.4992825388908386
1399 0.4936443269252777
1499 0.4884475767612457
1599 0.48363426327705383
1699 0.47915640473365784
1799 0.4749738872051239
1899 0.4710533916950226
1999 0.46736642718315125
2099 0.4638892114162445
2199 0.46060124039649963
2299 0.4574849009513855
2399 0.45452508330345154
2499 0.4517085552215576
2599 0.4490237534046173
2699 0.4464605748653412
2799 0.44401004910469055
2899 0.44166409969329834
2999 0.43941569328308105
final loss: 0.4394
```

Teraz trzeba sprawdzić, jak poszło naszej sieci. W PyTorchu sieć pracuje zawsze w jednym z dwóch trybów: treningowym lub ewaluacyjnym (predykcyjnym). Ten drugi wyłącza niektóre mechanizmy, które są używane tylko podczas treningu, w szczególności regularyzację dropout. Do przełączania służą metody modelu `.train()` i `.eval()`.

Dodatkowo podczas liczenia predykcji dobrze jest wyłączyć liczenie gradientów, bo nie będą potrzebne, a oszczędza to czas i pamięć. Używa się do tego menadżera kontekstu `with torch.no_grad():`.

In [24]:
```python
from sklearn.metrics import precision_recall_curve, precision_recall_fscore_support, roc


model.eval()
with torch.no_grad():
    y_score = activation(model(X_test))

auroc = roc_auc_score(y_test, y_score)
print(f"AUROC: {100 * auroc:.2f}%")
```

```
AUROC: 84.78%
```

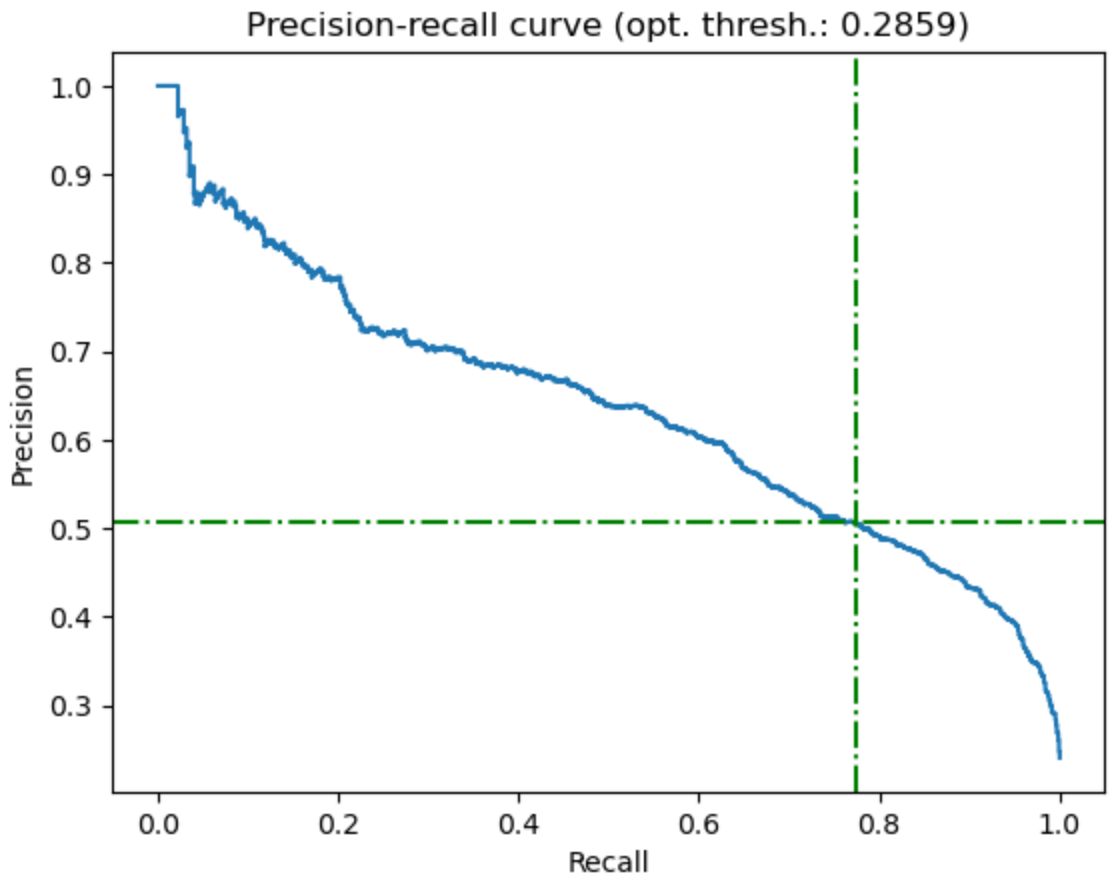
Jest to całkiem dobry wynik, a może być jeszcze lepszy. Sprawdźmy dla pewności jeszcze inne metryki: precyzję, recall oraz F1-score. Dodatkowo narysujemy krzywą precision-recall, czyli jak zmieniają się te metryki w zależności od przyjętego progu (threshold) prawdopodobieństwa, powyżej którego przyjmujemy klasę pozytywną. Taką krzywą należy rysować na zbiorze walidacyjnym, bo później chcemy wykorzystać tę informację do doboru progu, a nie chcemy mieć wycieku danych testowych (data leakage).

Poniżej zaimplementowano także funkcję `get_optimal_threshold()`, która sprawdza, dla którego progu uzyskujemy maksymalny F1-score, i zwraca indeks oraz wartość optymalnego progu. Przyda ci się ona w

dalszej części laboratorium.

In [25]:
```python
from sklearn.metrics import PrecisionRecallDisplay


def get_optimal_threshold(
    precisions: np.array,
    recalls: np.array,
    thresholds: np.array
) -> Tuple[int, float]:

    numerator = 2 * precisions * recalls
    denominator = precisions + recalls
    f1_scores = np.divide(numerator, denominator, out=np.zeros_like(numerator), where=de

    optimal_idx = np.argmax(f1_scores)
    optimal_threshold = thresholds[optimal_idx]

    return optimal_idx, optimal_threshold


def plot_precision_recall_curve(y_true, y_pred_score) -> None:
    precisions, recalls, thresholds = precision_recall_curve(y_true, y_pred_score)
    optimal_idx, optimal_threshold = get_optimal_threshold(precisions, recalls, threshol

    disp = PrecisionRecallDisplay(precisions, recalls)
    disp.plot()
    plt.title(f"Precision-recall curve (opt. thresh.: {optimal_threshold:.4f})")
    plt.axvline(recalls[optimal_idx], color="green", linestyle="-.")
    plt.axhline(precisions[optimal_idx], color="green", linestyle="-.")
    plt.show()
```

In [26]:
```python
model.eval()
with torch.no_grad():
    y_pred_valid_score = activation(model(X_valid))

plot_precision_recall_curve(y_valid, y_pred_valid_score)
```



Jak widać, chociaż AUROC jest wysokie, to dla optymalnego F1-score recall nie jest zbyt wysoki, a precyzja jest już dość niska. Być może wynik uda się poprawić, używając modelu o większej pojemności - pełnej, głębokiej sieci neuronowej.

Sieci neuronowe

Wszystko zaczęło się od inspirowanych biologią sztucznych neuronów, których próbowano użyć do symulacji mózgu. Naukowcy szybko odeszli od tego podejścia (sam problem modelowania okazał się też znacznie trudniejszy, niż sądzono), zamiast tego używając neuronów jako jednostek reprezentującą dowolną funkcję parametryczną $f(x, \Theta)$. Każdy neuron jest zatem bardzo elastyczny, bo jedyne wymagania to funkcja różniczkowalna, a mamy do tego wektor parametrów Θ.

W praktyce najczęściej można spotkać się z kilkoma rodzinami sieci neuronowych:

1. Perceptrony wielowarstwowe (*MultiLayer Perceptron*, MLP) - najbardziej podobne do powyższego opisu, niezbędne do klasyfikacji i regresji
2. Konwolucyjne (*Convolutional Neural Networks*, CNNs) - do przetwarzania danych z zależnościami przestrzennymi, np. obrazów czy dźwięku
3. Rekurencyjne (*Recurrent Neural Networks*, RNNs) - do przetwarzania danych z zależnościami sekwencyjnymi, np. szeregi czasowe, oraz kiedyś do języka naturalnego
4. Transformacyjne (*Transformers*), oparte o mechanizm atencji (*attention*) - do przetwarzania języka naturalnego (NLP), z którego wyparły RNNs, a coraz częściej także do wszelkich innych danych, np. obrazów, dźwięku
5. Grafowe (*Graph Neural Networks*, GNNS) - do przetwarzania grafów

Na tym laboratorium skupimy się na najprostszej architekturze, czyli MLP. Jest ona powszechnie łączona z wszelkimi innymi architekturami, bo pozwala dokonywać klasyfikacji i regresji. Przykładowo, klasyfikacja

obrazów to zwykle CNN + MLP, klasyfikacja tekstów to transformer + MLP, a regresja na grafach to GNN + MLP.

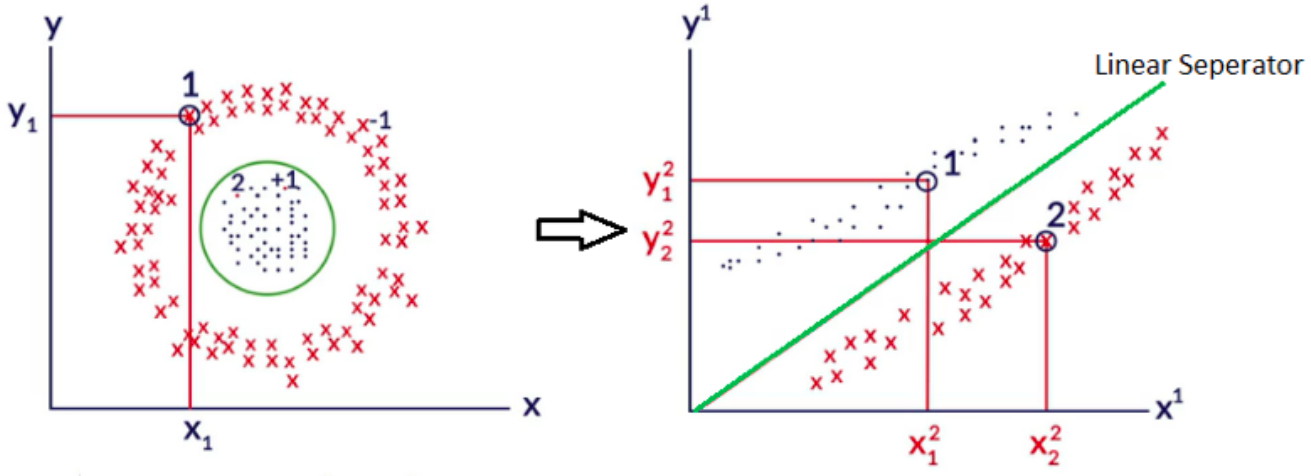
Dodatkowo, pomimo prostoty MLP są bardzo potężne - udowodniono, że perceptrony (ich powszechna nazwa) są uniwersalnym aproksymatorem, będącym w stanie przybliżyć dowolną funkcję z odpowiednio małym błędem, zakładając wystarczającą wielkość warstw sieci. Szczególne ich wersje potrafią nawet reprezentować drzewa decyzyjne.

Dla zainteresowanych polecamy doskonałą książkę "Dive into Deep Learning", z implementacjami w PyTorchu, klasyczną książkę "Deep Learning Book", oraz ten filmik, jeśli zastanawiałeś/-aś się, czemu używamy deep learning, a nie naprzykład (wide?) learning. (aka. czemu staramy się budować głębokie sieci, a nie płytkie za to szerokie)

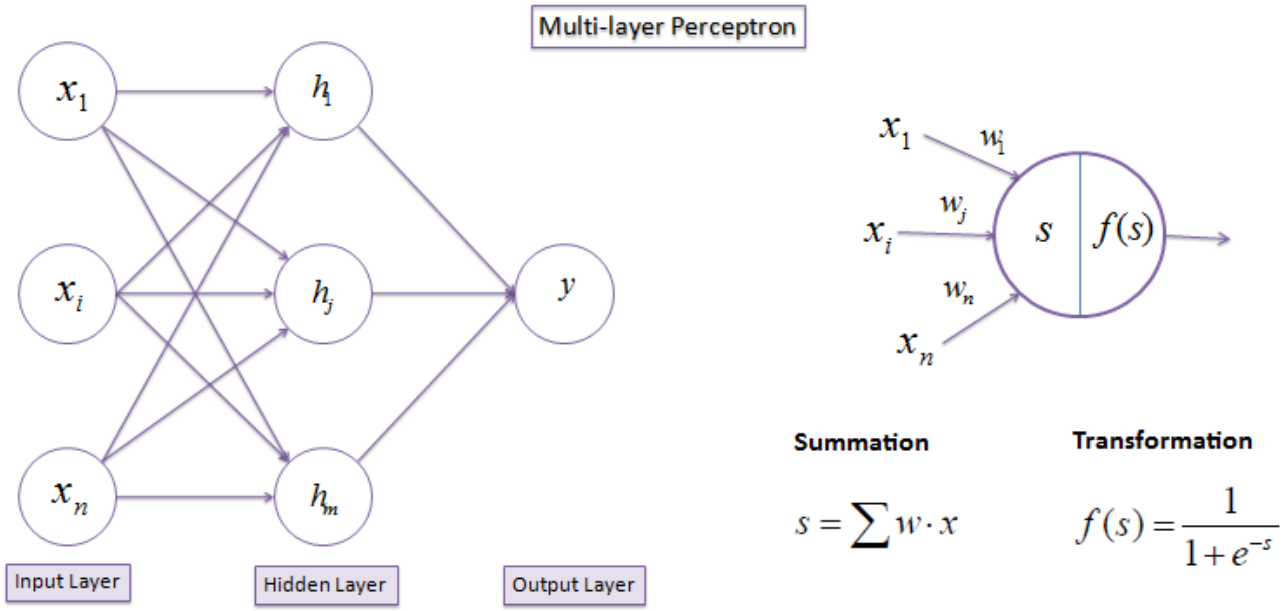
Sieci MLP

Dla przypomnienia, na wejściu mamy punkty ze zbioru treningowego, czyli d-wymiarowe wektory. W klasyfikacji chcemy znaleźć granicę decyzyjną, czyli krzywą, która oddzieli od siebie klasy. W wejściowej przestrzeni może być to trudne, bo chmury punktów z poszczególnych klas mogą być ze sobą dość pomieszane. Pamiętajmy też, że regresja logistyczna jest klasyfikatorem liniowym, czyli w danej przestrzeni potrafi oddzielić punkty tylko linią prostą.

Sieć MLP składa się z warstw. Każda z nich dokonuje nieliniowego przekształcenia przestrzeni (można o tym myśleć jak o składaniu przestrzeni jakąś prostą/łamaną), tak, aby w finalnej przestrzeni nasze punkty były możliwie liniowo separowalne. Wtedy ostatnia warstwa z sigmoidą będzie potrafiła je rozdzielić od siebie.



Poszczególne neurony składają się z iloczynu skalarnego wejść z wagami neuronu, oraz nieliniowej funkcji aktywacji. W PyTorchu są to osobne obiekty - `nn.Linear` oraz np. `nn.Sigmoid`. Funkcja aktywacji przyjmuje wynik iloczynu skalarnego i przekształca go, aby sprawdzić, jak mocno reaguje neuron na dane wejście. Musi być nieliniowa z dwóch powodów. Po pierwsze, tylko nieliniowe przekształcenia są na tyle potężne, żeby umożliwić liniową separację danych w ostatniej warstwie. Po drugie, liniowe przekształcenia zwyczajnie nie działają. Aby zrozumieć czemu, trzeba zobaczyć, co matematycznie oznacza sieć MLP.



Zapisane matematycznie MLP to:

$$
\begin{aligned}
h_1 &= f_1(x) \\
h_2 &= f_2(h_1) \\
h_3 &= f_3(h_2) \\
&\dots \\
h_n &= f_n(h_{n-1})
\end{aligned}
$$

gdzie x to wejście f_i to funkcja aktywacji i-tej warstwy, a h_i to wyjście i-tej warstwy, nazywane **ukrytą reprezentacją (hidden representation)**, lub *latent representation*. Nazwa bierze się z tego, że w środku sieci wyciągamy cechy i wzorce w danych, które nie są widoczne na pierwszy rzut oka na wejściu.

Załóżmy, że uczymy się na danych x o jednym wymiarze (dla uproszczenia wzorów) oraz nie mamy funkcji aktywacji, czyli wykorzystujemy tak naprawdę aktywację liniową $f(x) = x$. Zobaczmy jak będą wyglądać dane przechodząc przez kolejne warstwy:

$$
\begin{aligned}
h_1 &= f_1(xw_1) = xw_1 \\
h_2 &= f_2(h_1w_2) = xw_1w_2 \\
&\dots \\
h_n &= f_n(h_{n-1}w_n) = xw_1w_2 \dots w_n
\end{aligned}
$$

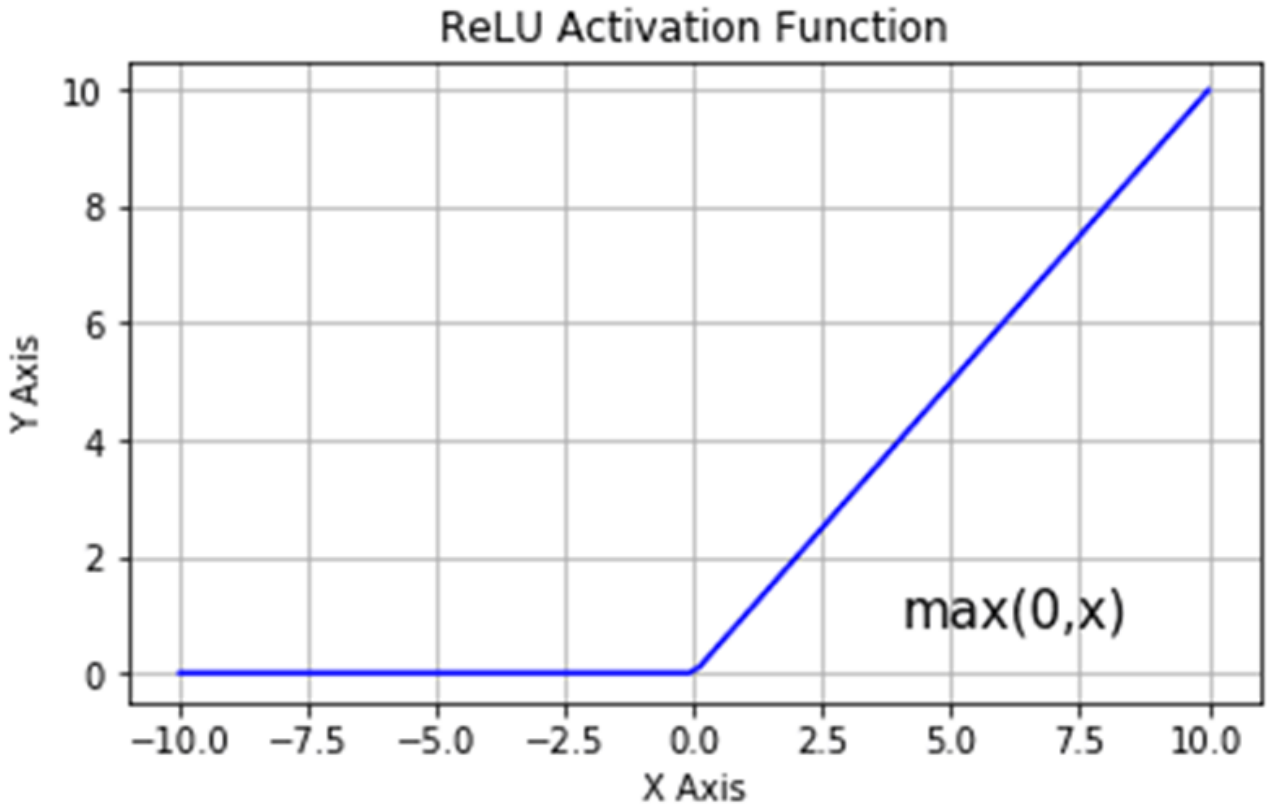
gdzie w_i to jest parametr i-tej warstwy sieci, x to są dane (w naszym przypadku jedna liczba) wejściowa, a h_i to wyjście i-tej warstwy.

Jak widać, taka sieć o n warstwach jest równoważna sieci o jednej warstwie z parametrem $w = w_1w_2 \dots w_n$. Wynika to z tego, że złożenie funkcji liniowych jest także funkcją liniową - patrz notatki z algebry :)

Jeżeli natomiast użyjemy nieliniowej funkcji aktywacji, często oznaczanej jako σ, to wszystko będzie działać. Co ważne, ostatnia warstwa, dająca wyjście sieci, ma zwykle inną aktywację od warstw wewnątrz sieci, bo też ma inne zadanie - zwrócić wartość dla klasyfikacji lub regresji. Na wyjściu korzysta się z funkcji liniowej (regresja), sigmoidalnej (klasyfikacja binarna) lub softmax (klasyfikacja wieloklasowa).

Wewnątrz sieci używano kiedyś sigmoidy oraz tangensa hiperbolicznego `tanh`, ale okazało się to nieefektywne przy uczeniu głębokich sieci o wielu warstwach. Nowoczesne sieci korzystają zwykle z funkcji ReLU (*rectified linear unit*), która jest zaskakująco prosta: $ReLU(x) = \max(0, x)$. Okazało się, że bardzo

dobrze nadaje się do treningu nawet bardzo głębokich sieci neuronowych. Nowsze funkcje aktywacji są głównie modyfikacjami ReLU.



MLP w PyTorchu

Warstwę neuronów w MLP nazywa się warstwą gęstą (*dense layer*) lub warstwą w pełni połączoną (*fully-connected layer*), i taki opis oznacza zwykle same neurony oraz funkcję aktywacji. PyTorch, jak już widzieliśmy, definiuje osobno transformację liniową oraz aktywację, a więc jedna warstwa składa się de facto z 2 obiektów, wywoływanych jeden po drugim. Inne frameworki, szczególnie wysokopoziomowe (np. Keras) łączą to często w jeden obiekt.

MLP składa się zatem z sekwencji obiektów, które potem wywołuje się jeden po drugim, gdzie wyjście poprzedniego to wejście kolejnego. Ale nie można tutaj używać Pythonowych list! Z perspektywy PyTorcha to wtedy niezależne obiekty i nie zostanie wtedy przekazany między nimi gradient. Trzeba tutaj skorzystać z `nn.Sequential`, aby tworzyć taki pipeline.

Rozmiary wejścia i wyjścia dla każdej warstwy trzeba w PyTorchu podawać explicite. Jest to po pierwsze edukacyjne, a po drugie często ułatwia wnioskowanie o działaniu sieci oraz jej debugowanie - mamy jasno podane, czego oczekujemy. Niektóre frameworki (np. Keras) obliczają to automatycznie.

Co ważne, ostatnia warstwa zwykle nie ma funkcji aktywacji. Wynika to z tego, że obliczanie wielu funkcji kosztu (np. entropii krzyżowej) na aktywacjach jest często niestabilne numerycznie. Z tego powodu PyTorch oferuje funkcje kosztu zawierające w środku aktywację dla ostatniej warstwy, a ich implementacje są stabilne numerycznie. Przykładowo, `nn.BCELoss` przyjmuje wejście z zaaplikowanymi już aktywacjami, ale może skutkować under/overflow, natomiast `nn.BCEWithLogitsLoss` przyjmuje wejście bez aktywacji, a w środku ma specjalną implementację łączącą binarną entropię krzyżową z aktywacją sigmoidalną. Oczywiście w związku z tym aby dokonać potem predykcji w praktyce, trzeba pamiętać o użyciu funkcji aktywacji.

Często korzysta się przy tym z funkcji z modułu `torch.nn.functional`, które są w tym wypadku nieco wygodniejsze od klas wywoływalnych z `torch.nn`.

Całe sieci w PyTorchu tworzy się jako klasy dziedziczące po `nn.Module`. Co ważne, obiekty, z których tworzymy sieć, np. `nn.Linear`, także dziedziczą po tej klasie. Pozwala to na bardzo modułową budowę kodu, zgodną z zasadami OOP. W konstruktorze najpierw trzeba zawsze wywołać konstruktor rodzica - `super().__init__()`, a później tworzy się potrzebne obiekty i zapisuje jako atrybuty. Każdy atrybut dziedziczący po `nn.Module` lub `nn.Parameter` jest uważany za taki, który zawiera parametry sieci, a więc przy wywołaniu metody `parameters()` - parametry z tych atrybutów pojawią się w liście wszystkich parametrów. Musimy też zdefiniować metodę `forward()`, która przyjmuje tensor `x` i zwraca wynik. Typowo ta metoda po prostu używa obiektów zdefiniowanych w konstruktorze.

UWAGA: nigdy w normalnych warunkach się nie woła metody `forward` ręcznie

Zadanie 4 (0.5 punktu)

Uzupełnij implementację 3-warstwowej sieci MLP. Użyj rozmiarów:

- pierwsza warstwa: input_size x 256
- druga warstwa: 256 x 128
- trzecia warstwa: 128 x 1

Użyj funkcji aktywacji ReLU.

Przydatne klasy:

- `nn.Sequential`
- `nn.Linear`
- `nn.ReLU`

In [27]:
```
from torch import sigmoid


class MLP(nn.Module):
    def __init__(self, input_size: int):
        super().__init__()

        self.mlp = nn.Sequential(
            nn.Linear(input_size, 256),
            nn.ReLU(),
            nn.Linear(256, 128),
            nn.ReLU(),
            nn.Linear(128, 1)
        )

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

    def predict_proba(self, x):
        return sigmoid(self(x))

    def predict(self, x, threshold: float = 0.5):
        y_pred_score = self.predict_proba(x)
        return (y_pred_score > threshold).to(torch.int32)
```

In [28]:
```
learning_rate = 1e-3
model = MLP(input_size=X_train.shape[1])
```

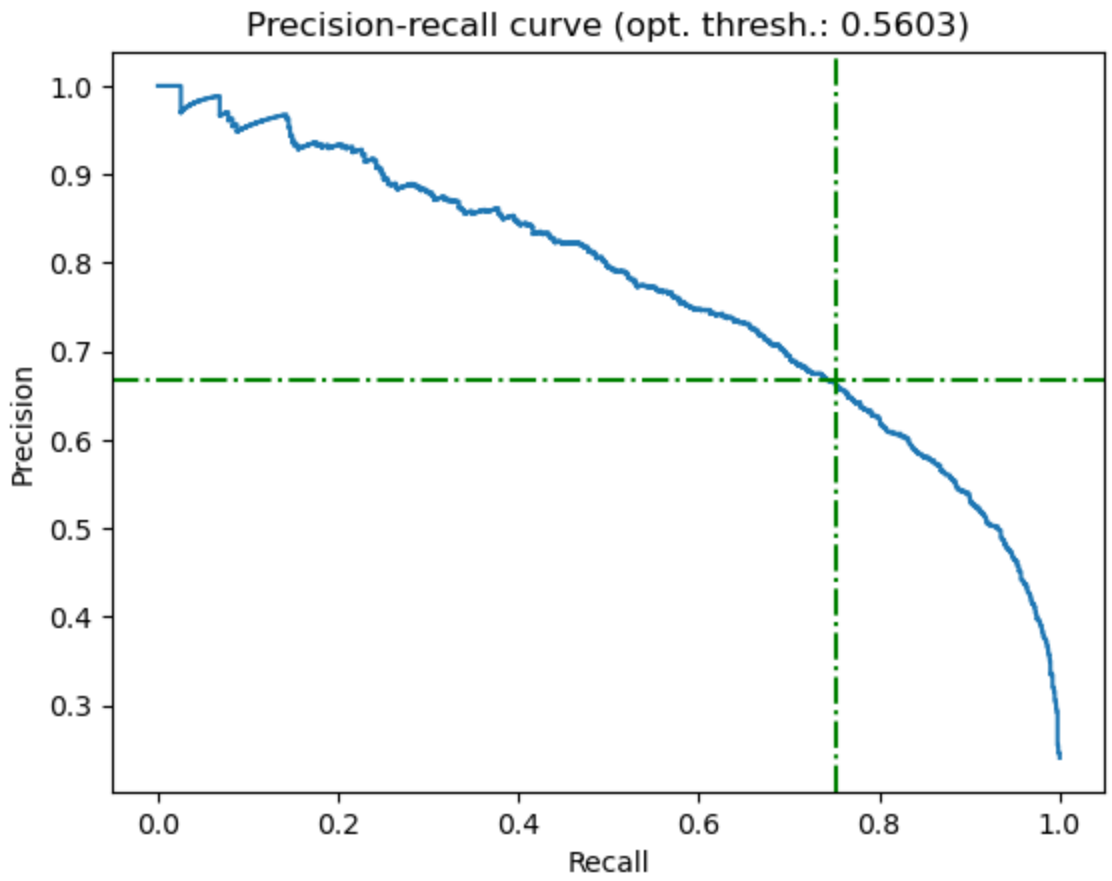
```
optimizer = torch.optim.SGD(model.parameters(), lr=learning_rate)

# note that we are using loss function with sigmoid built in
loss_fn = torch.nn.BCEWithLogitsLoss()
num_epochs = 2000
evaluation_steps = 200

for i in range(num_epochs):
    y_pred = model(X_train)
    loss = loss_fn(y_pred, y_train)
    loss.backward()

    optimizer.step()
    optimizer.zero_grad()

    if i % evaluation_steps == 0:
        print(f"Epoch {i} train loss: {loss.item():.4f}")

print(f"final loss: {loss.item():.4f}")
```

```
Epoch 0 train loss: 0.6781
Epoch 200 train loss: 0.6583
Epoch 400 train loss: 0.6413
Epoch 600 train loss: 0.6265
Epoch 800 train loss: 0.6134
Epoch 1000 train loss: 0.6018
Epoch 1200 train loss: 0.5913
Epoch 1400 train loss: 0.5820
Epoch 1600 train loss: 0.5736
Epoch 1800 train loss: 0.5660
final loss: 0.5592
```

In [44]:

```
model.eval()
with torch.no_grad():
    # positive class probabilities
    y_pred_valid_score = model.predict_proba(X_valid)
    y_pred_test_score = model.predict_proba(X_test)

auroc = roc_auc_score(y_test, y_pred_test_score)
print(f"AUROC: {100 * auroc:.2f}%")

plot_precision_recall_curve(y_valid, y_pred_valid_score)
```

```
AUROC: 90.72%
```



AUROC jest podobne, a precision i recall spadły - wypadamy wręcz gorzej od regresji liniowej! Skoro dodaliśmy więcej warstw, to może pojemność modelu jest teraz za duża i trzeba by go zregularyzować?

Sieci neuronowe bardzo łatwo przeuczają, bo są bardzo elastycznymi i pojemnymi modelami. Dlatego mają wiele różnych rodzajów regularyzacji, których używa się razem. Co ciekawe, udowodniono eksperymentalnie, że zbyt duże sieci z mocną regularyzacją działają lepiej niż mniejsze sieci, odpowiedniego rozmiaru, za to ze słabszą regularyzacją.

Pierwszy rodzaj regularyzacji to znana nam już **regularyzacja L2**, czyli penalizacja zbyt dużych wag. W kontekście sieci neuronowych nazywa się też ją czasem *weight decay*. W PyTorchu dodaje się ją jako argument do optymalizatora.

Regularyzacja specyficzna dla sieci neuronowych to **dropout**. Polega on na losowym wyłączaniu zadanego procenta neuronów podczas treningu. Pomimo prostoty okazała się niesamowicie skuteczna, szczególnie w treningu bardzo głębokich sieci. Co ważne, jest to mechanizm używany tylko podczas treningu - w trakcie predykcji za pomocą sieci wyłącza się ten mechanizm i dokonuje normalnie predykcji całą siecią. Podejście to można potraktować jak ensemble learning, podobny do lasów losowych - wyłączając losowe części sieci, w każdej iteracji trenujemy nieco inną sieć, co odpowiada uśrednianiu predykcji różnych algorytmów. Typowo stosuje się dość mocny dropout, rzędu 25-50%. W PyTorchu implementuje go warstwa `nn.Dropout`, aplikowana zazwyczaj po funkcji aktywacji.

Ostatni, a być może najważniejszy rodzaj regularyzacji to **wczesny stop (early stopping)**. W każdym kroku mocniej dostosowujemy terenową sieć do zbioru treningowego, a więc zbyt długi trening będzie skutkował przeuczeniem. W metodzie wczesnego stopu używamy wydzielonego zbioru walidacyjnego (pojedynczego, metoda holdout), sprawdzając co określoną liczbę epok wynik na tym zbiorze. Jeżeli nie uzyskamy wyniku lepszego od najlepszego dotychczas uzyskanego przez określoną liczbę epok, to przerywamy trening. Okres, przez który czekamy na uzyskanie lepszego wyniku, to cierpliwość (*patience*). Im mniejsze, tym mocniejszy jest ten rodzaj regularyzacji, ale trzeba z tym uważać, bo łatwo jest przesadzić i zbyt szybko przerywać

trening. Niektóre implementacje uwzględniają tzw. *grace period*, czyli gwarantowaną minimalną liczbę epok, przez którą będziemy trenować sieć, niezależnie od wybranej cierpliwości.

Dodatkowo ryzyko przeuczenia można zmniejszyć, używając mniejszej stałej uczącej.

Zadanie 5 (1.5 punktu)

Zaimplementuj funkcję `evaluate_model()`, obliczającą metryki na zbiorze testowym:

- wartość funkcji kosztu (loss)
- AUROC
- optymalny próg
- F1-score przy optymalnym progu
- precyzję oraz recall dla optymalnego progu

Jeżeli podana jest wartość argumentu `threshold`, to użyj jej do zamiany prawdopodobieństw na twarde predykcje. W przeciwnym razie użyj funkcji `get_optimal_threshold` i oblicz optymalną wartość progu.

Pamiętaj o przełączeniu modelu w tryb ewaluacji oraz o wyłączeniu obliczania gradientów.

In [38]:
```python
from typing import Optional

from sklearn.metrics import precision_score, recall_score, f1_score
from torch import sigmoid


def evaluate_model(
    model: nn.Module,
    X: torch.Tensor,
    y: torch.Tensor,
    loss_fn: nn.Module,
    threshold: Optional[float]= None
) -> Dict[str, float]:
    model.eval()
    with torch.no_grad():
        y_pred_score = model.predict_proba(X)

    auroc = roc_auc_score(y, y_pred_score)
    loss = loss_fn(y_pred_score, y)
    print(y_pred_score)

    if threshold is None:
        precisions, recalls, thresholds = precision_recall_curve(y, y_pred_score)
        _, threshold = get_optimal_threshold(precisions, recalls, thresholds)

    y_pred = (y_pred_score>threshold).float()

    precision = precision_score(y, y_pred)
    recall = recall_score(y, y_pred)
    f1 = f1_score(y, y_pred)

    results = {
        "loss": loss,
        "AUROC": auroc,
        "optimal_threshold": threshold,
        "precision": precision,
        "recall": recall,
        "F1-score": f1,
    }
    return results
```

Zadanie 6 (0.5 punktu)

Zaimplementuj 3-warstwową sieć MLP z dropout (50%). Rozmiary warstw ukrytych mają wynosić 256 i 128.

In [35]:
```
class RegularizedMLP(nn.Module):
    def __init__(self, input_size: int, dropout_p: float = 0.5):
        super().__init__()

        self.mlp = nn.Sequential(
            nn.Linear(input_size, 256),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(256, 128),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(128, 1)
        )

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

    def predict_proba(self, x):
        return sigmoid(self(x))

    def predict(self, x, threshold: float = 0.5):
        y_pred_score = self.predict_proba(x)
        return (y_pred_score > threshold).to(torch.int32)
```

Opisaliśmy wcześniej podstawowy optymalizator w sieciach neuronowych - spadek wzdłuż gradientu. Jednak wymaga on użycia całego zbioru danych, aby obliczyć gradient, co jest często niewykonalne przez rozmiar zbioru. Dlatego wymyślono **stochastyczny spadek wzdłuż gradientu (stochastic gradient descent, SGD)**, w którym używamy 1 przykładu naraz, liczymy gradient tylko po nim i aktualizujemy parametry. Jest to oczywiście dość grube przybliżenie gradientu, ale pozwala robić szybko dużo małych kroków. Kompromisem, którego używa się w praktyce, jest **minibatch gradient descent**, czyli używanie batchy np. 32, 64 czy 128 przykładów.

Rzadko wspominanym, a ważnym faktem jest także to, że stochastyczność metody optymalizacji jest sama w sobie też metodą regularyzacji, a więc `batch_size` to także hiperparametr.

Obecnie najpopularniejszą odmianą SGD jest Adam, gdyż uczy on szybko sieć oraz daje bardzo dobre wyniki nawet przy niekoniecznie idealnie dobranych hiperparametrach. W PyTorchu najlepiej korzystać z jego implementacji `AdamW`, która jest nieco lepsza niż implementacja `Adam`. Jest to zasadniczo zawsze wybór domyślny przy treningu współczesnych sieci neuronowych.

Na razie użyjemy jednak minibatch SGD.

Poniżej znajduje się implementacja prostej klasy dziedziczącej po `Dataset` - tak w PyTorchu implementuje się własne zbiory danych. Użycie takich klas umożliwia użycie klas ładujących dane (`DataLoader`), które z kolei pozwalają łatwo ładować batche danych. Trzeba w takiej klasie zaimplementować metody:

- `__len__` - zwraca ilość punktów w zbiorze
- `__getitem__` - zwraca przykład ze zbioru pod danym indeksem oraz jego klasę

In [32]:
```
from torch.utils.data import Dataset


class MyDataset(Dataset):
    def __init__(self, data, y):
```

```
        super().__init__()

        self.data = data
        self.y = y

    def __len__(self):
        return self.data.shape[0]

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

Zadanie 7 (1.5 punktu)

Zaimplementuj pętlę treningowo-walidacyjną dla sieci neuronowej. Wykorzystaj podane wartości hiperparametrów do treningu (stała ucząca, prawdopodobieństwo dropoutu, regularyzacja L2, rozmiar batcha, maksymalna liczba epok). Użyj optymalizatora SGD.

Dodatkowo zaimplementuj regularyzację przez early stopping. Sprawdzaj co epokę wynik na zbiorze walidacyjnym. Użyj podanej wartości patience, a jako metryki po prostu wartości funkcji kosztu. Może się tutaj przydać zaimplementowana funkcja `evaluate_model()`.

Pamiętaj o tym, aby przechowywać najlepszy dotychczasowy wynik walidacyjny oraz najlepszy dotychczasowy model. Zapamiętaj też optymalny próg do klasyfikacji dla najlepszego modelu.

In [33]:
```
from copy import deepcopy

from torch.utils.data import DataLoader


learning_rate = 1e-3
dropout_p = 0.5
l2_reg = 1e-4
batch_size = 128
max_epochs = 300

early_stopping_patience = 4
```

In [36]:
```
model = RegularizedMLP(
    input_size=X_train.shape[1],
    dropout_p=dropout_p
)
optimizer = torch.optim.SGD(
    model.parameters(),
    lr=learning_rate,
    weight_decay=l2_reg
)
loss_fn = torch.nn.BCEWithLogitsLoss()

train_dataset = MyDataset(X_train, y_train)
train_dataloader = DataLoader(train_dataset, batch_size=batch_size)

steps_without_improvement = 0

best_val_loss = np.inf
best_model = None
best_threshold = None

for epoch_num in range(max_epochs):
    model.train()

    # note that we are using DataLoader to get batches
```

```
    for X_batch, y_batch in train_dataloader:
        y_batch_pred = model(X_batch)
        loss = loss_fn(y_batch_pred, y_batch)
        loss.backward()

        optimizer.step()
        optimizer.zero_grad()

    # model evaluation, early stopping
    model.eval()
    valid_metrics = evaluate_model(model, X_valid, y_valid, loss_fn)
    if valid_metrics["loss"] < best_val_loss:
        best_val_loss = valid_metrics["loss"]
        steps_without_improvement = 0
        best_model = deepcopy(model)
        best_threshold = valid_metrics["optimal_threshold"]
    else:
        steps_without_improvement += 1
        if steps_without_improvement == early_stopping_patience: break

    print(f"Epoch {epoch_num} train loss: {loss.item():.4f}, eval loss {valid_metrics['l
```

```
tensor([[0.4674],
        [0.4846],
        [0.4702],
        ...,
        [0.4786],
        [0.4734],
        [0.4719]])
Epoch 0 train loss: 0.6671, eval loss 0.8437523245811462
tensor([[0.4503],
        [0.4694],
        [0.4539],
        ...,
        [0.4611],
        [0.4577],
        [0.4545]])
Epoch 1 train loss: 0.6464, eval loss 0.837480902671814
tensor([[0.4342],
        [0.4551],
        [0.4386],
        ...,
        [0.4446],
        [0.4425],
        [0.4380]])
Epoch 2 train loss: 0.6327, eval loss 0.8315537571907043
tensor([[0.4187],
        [0.4414],
        [0.4241],
        ...,
        [0.4287],
        [0.4279],
        [0.4221]])
Epoch 3 train loss: 0.6197, eval loss 0.8258987665176392
tensor([[0.4035],
        [0.4286],
        [0.4103],
        ...,
        [0.4133],
        [0.4137],
        [0.4070]])
Epoch 4 train loss: 0.6063, eval loss 0.8205088376998901
tensor([[0.3890],
        [0.4166],
        [0.3971],
        ...,
```

```
        [0.3986],
        [0.4001],
        [0.3925]])
Epoch 5 train loss: 0.5991, eval loss 0.8153571486473083
tensor([[0.3749],
        [0.4056],
        [0.3846],
        ...,
        [0.3843],
        [0.3870],
        [0.3786]])
Epoch 6 train loss: 0.5951, eval loss 0.8104532361030579
tensor([[0.3615],
        [0.3954],
        [0.3729],
        ...,
        [0.3702],
        [0.3744],
        [0.3654]])
Epoch 7 train loss: 0.5832, eval loss 0.8057981729507446
tensor([[0.3487],
        [0.3859],
        [0.3620],
        ...,
        [0.3567],
        [0.3622],
        [0.3529]])
Epoch 8 train loss: 0.5712, eval loss 0.8013750314712524
tensor([[0.3368],
        [0.3772],
        [0.3519],
        ...,
        [0.3440],
        [0.3506],
        [0.3410]])
Epoch 9 train loss: 0.5646, eval loss 0.7972067594528198
tensor([[0.3255],
        [0.3694],
        [0.3426],
        ...,
        [0.3320],
        [0.3395],
        [0.3298]])
Epoch 10 train loss: 0.5597, eval loss 0.7932718396186829
tensor([[0.3150],
        [0.3624],
        [0.3341],
        ...,
        [0.3208],
        [0.3290],
        [0.3192]])
Epoch 11 train loss: 0.5530, eval loss 0.7895668745040894
tensor([[0.3051],
        [0.3562],
        [0.3263],
        ...,
        [0.3101],
        [0.3189],
        [0.3091]])
Epoch 12 train loss: 0.5541, eval loss 0.7860631942749023
tensor([[0.2960],
        [0.3509],
        [0.3193],
        ...,
        [0.3002],
        [0.3094],
```

```
        [0.2998]])
Epoch 13 train loss: 0.5402, eval loss 0.7828163504600525
tensor([[0.2875],
        [0.3465],
        [0.3130],
        ...,
        [0.2910],
        [0.3004],
        [0.2910]])
Epoch 14 train loss: 0.5412, eval loss 0.7797814607620239
tensor([[0.2796],
        [0.3427],
        [0.3075],
        ...,
        [0.2825],
        [0.2919],
        [0.2828]])
Epoch 15 train loss: 0.5383, eval loss 0.7769457101821899
tensor([[0.2723],
        [0.3396],
        [0.3025],
        ...,
        [0.2745],
        [0.2837],
        [0.2751]])
Epoch 16 train loss: 0.5302, eval loss 0.7742713689804077
tensor([[0.2656],
        [0.3373],
        [0.2983],
        ...,
        [0.2671],
        [0.2760],
        [0.2680]])
Epoch 17 train loss: 0.5216, eval loss 0.7717950940132141
tensor([[0.2594],
        [0.3355],
        [0.2946],
        ...,
        [0.2602],
        [0.2687],
        [0.2613]])
Epoch 18 train loss: 0.5304, eval loss 0.7694524526596069
tensor([[0.2537],
        [0.3344],
        [0.2915],
        ...,
        [0.2538],
        [0.2617],
        [0.2551]])
Epoch 19 train loss: 0.5241, eval loss 0.7672763466835022
tensor([[0.2484],
        [0.3340],
        [0.2890],
        ...,
        [0.2479],
        [0.2552],
        [0.2493]])
Epoch 20 train loss: 0.5236, eval loss 0.7652437090873718
tensor([[0.2435],
        [0.3340],
        [0.2867],
        ...,
        [0.2422],
        [0.2488],
        [0.2438]])
Epoch 21 train loss: 0.5170, eval loss 0.7632789611816406
```

```
tensor([[0.2390],
        [0.3346],
        [0.2851],
        ...,
        [0.2370],
        [0.2428],
        [0.2389]])
Epoch 22 train loss: 0.5098, eval loss 0.761475145816803
tensor([[0.2348],
        [0.3357],
        [0.2838],
        ...,
        [0.2321],
        [0.2369],
        [0.2341]])
Epoch 23 train loss: 0.5167, eval loss 0.759737491607666
tensor([[0.2310],
        [0.3374],
        [0.2831],
        ...,
        [0.2275],
        [0.2313],
        [0.2297]])
Epoch 24 train loss: 0.4979, eval loss 0.7581110000610352
tensor([[0.2272],
        [0.3393],
        [0.2825],
        ...,
        [0.2230],
        [0.2257],
        [0.2254]])
Epoch 25 train loss: 0.5059, eval loss 0.756522536277771
tensor([[0.2237],
        [0.3416],
        [0.2823],
        ...,
        [0.2189],
        [0.2203],
        [0.2213]])
Epoch 26 train loss: 0.4845, eval loss 0.754999577999115
tensor([[0.2207],
        [0.3444],
        [0.2827],
        ...,
        [0.2151],
        [0.2151],
        [0.2177]])
Epoch 27 train loss: 0.4935, eval loss 0.7535885572433472
tensor([[0.2177],
        [0.3475],
        [0.2832],
        ...,
        [0.2113],
        [0.2101],
        [0.2142]])
Epoch 28 train loss: 0.5082, eval loss 0.7522171139717102
tensor([[0.2149],
        [0.3510],
        [0.2841],
        ...,
        [0.2078],
        [0.2051],
        [0.2107]])
Epoch 29 train loss: 0.4940, eval loss 0.7508841156959534
tensor([[0.2122],
        [0.3549],
```

```
        [0.2852],
        ...,
        [0.2044],
        [0.2002],
        [0.2074]])
Epoch 30 train loss: 0.4876, eval loss 0.7495970726013184
tensor([[0.2096],
        [0.3589],
        [0.2865],
        ...,
        [0.2010],
        [0.1953],
        [0.2042]])
Epoch 31 train loss: 0.4875, eval loss 0.7483287453651428
tensor([[0.2074],
        [0.3633],
        [0.2884],
        ...,
        [0.1979],
        [0.1906],
        [0.2013]])
Epoch 32 train loss: 0.4897, eval loss 0.7471542954444885
tensor([[0.2051],
        [0.3676],
        [0.2903],
        ...,
        [0.1947],
        [0.1859],
        [0.1983]])
Epoch 33 train loss: 0.4890, eval loss 0.7459630370140076
tensor([[0.2027],
        [0.3718],
        [0.2922],
        ...,
        [0.1916],
        [0.1811],
        [0.1953]])
Epoch 34 train loss: 0.4780, eval loss 0.7447528839111328
tensor([[0.2006],
        [0.3764],
        [0.2945],
        ...,
        [0.1885],
        [0.1764],
        [0.1926]])
Epoch 35 train loss: 0.4716, eval loss 0.7436038255691528
tensor([[0.1984],
        [0.3811],
        [0.2970],
        ...,
        [0.1855],
        [0.1717],
        [0.1898]])
Epoch 36 train loss: 0.4695, eval loss 0.7424644231796265
tensor([[0.1964],
        [0.3860],
        [0.2997],
        ...,
        [0.1825],
        [0.1672],
        [0.1872]])
Epoch 37 train loss: 0.4690, eval loss 0.7413651943206787
tensor([[0.1945],
        [0.3909],
        [0.3028],
        ...,
```

```
        [0.1796],
        [0.1626],
        [0.1846]])
Epoch 38 train loss: 0.4731, eval loss 0.7402745485305786
tensor([[0.1929],
        [0.3961],
        [0.3063],
        ...,
        [0.1770],
        [0.1583],
        [0.1824]])
Epoch 39 train loss: 0.4666, eval loss 0.7392732501029968
tensor([[0.1913],
        [0.4011],
        [0.3100],
        ...,
        [0.1743],
        [0.1540],
        [0.1801]])
Epoch 40 train loss: 0.4687, eval loss 0.7382758855819702
tensor([[0.1895],
        [0.4060],
        [0.3136],
        ...,
        [0.1715],
        [0.1496],
        [0.1777]])
Epoch 41 train loss: 0.4607, eval loss 0.7372434139251709
tensor([[0.1878],
        [0.4109],
        [0.3174],
        ...,
        [0.1687],
        [0.1452],
        [0.1754]])
Epoch 42 train loss: 0.4643, eval loss 0.736240804195404
tensor([[0.1862],
        [0.4158],
        [0.3216],
        ...,
        [0.1662],
        [0.1410],
        [0.1733]])
Epoch 43 train loss: 0.4562, eval loss 0.735284149646759
tensor([[0.1847],
        [0.4207],
        [0.3258],
        ...,
        [0.1636],
        [0.1368],
        [0.1712]])
Epoch 44 train loss: 0.4610, eval loss 0.7343284487724304
tensor([[0.1834],
        [0.4257],
        [0.3305],
        ...,
        [0.1612],
        [0.1328],
        [0.1692]])
Epoch 45 train loss: 0.4466, eval loss 0.7334338426589966
tensor([[0.1820],
        [0.4308],
        [0.3353],
        ...,
        [0.1587],
        [0.1288],
```

```
        [0.1672]])
Epoch 46 train loss: 0.4471, eval loss 0.7325394749641418
tensor([[0.1807],
        [0.4357],
        [0.3402],
        ...,
        [0.1564],
        [0.1248],
        [0.1653]])
Epoch 47 train loss: 0.4501, eval loss 0.7316687703132629
tensor([[0.1794],
        [0.4405],
        [0.3452],
        ...,
        [0.1540],
        [0.1210],
        [0.1635]])
Epoch 48 train loss: 0.4443, eval loss 0.730816662311554
tensor([[0.1781],
        [0.4451],
        [0.3503],
        ...,
        [0.1516],
        [0.1171],
        [0.1615]])
Epoch 49 train loss: 0.4404, eval loss 0.7299630641937256
tensor([[0.1770],
        [0.4497],
        [0.3559],
        ...,
        [0.1494],
        [0.1135],
        [0.1599]])
Epoch 50 train loss: 0.4492, eval loss 0.7291789650917053
tensor([[0.1758],
        [0.4544],
        [0.3612],
        ...,
        [0.1471],
        [0.1098],
        [0.1580]])
Epoch 51 train loss: 0.4455, eval loss 0.7283633947372437
tensor([[0.1746],
        [0.4592],
        [0.3668],
        ...,
        [0.1448],
        [0.1062],
        [0.1562]])
Epoch 52 train loss: 0.4352, eval loss 0.7275556325912476
tensor([[0.1734],
        [0.4640],
        [0.3724],
        ...,
        [0.1424],
        [0.1026],
        [0.1544]])
Epoch 53 train loss: 0.4423, eval loss 0.7267788052558899
tensor([[0.1724],
        [0.4690],
        [0.3783],
        ...,
        [0.1402],
        [0.0992],
        [0.1528]])
Epoch 54 train loss: 0.4348, eval loss 0.7260212898254395
```

```
tensor([[0.1716],
        [0.4741],
        [0.3846],
        ...,
        [0.1381],
        [0.0959],
        [0.1513]])
Epoch 55 train loss: 0.4389, eval loss 0.7253273725509644
tensor([[0.1706],
        [0.4789],
        [0.3905],
        ...,
        [0.1360],
        [0.0926],
        [0.1496]])
Epoch 56 train loss: 0.4290, eval loss 0.7245904207229614
tensor([[0.1697],
        [0.4840],
        [0.3968],
        ...,
        [0.1339],
        [0.0895],
        [0.1481]])
Epoch 57 train loss: 0.4212, eval loss 0.7239089608192444
tensor([[0.1690],
        [0.4889],
        [0.4031],
        ...,
        [0.1319],
        [0.0865],
        [0.1466]])
Epoch 58 train loss: 0.4410, eval loss 0.7232421636581421
tensor([[0.1681],
        [0.4936],
        [0.4093],
        ...,
        [0.1298],
        [0.0835],
        [0.1451]])
Epoch 59 train loss: 0.4343, eval loss 0.7225755453109741
tensor([[0.1675],
        [0.4983],
        [0.4154],
        ...,
        [0.1278],
        [0.0806],
        [0.1437]])
Epoch 60 train loss: 0.4149, eval loss 0.721943199634552
tensor([[0.1667],
        [0.5032],
        [0.4215],
        ...,
        [0.1256],
        [0.0777],
        [0.1422]])
Epoch 61 train loss: 0.4362, eval loss 0.7212890982627869
tensor([[0.1660],
        [0.5081],
        [0.4276],
        ...,
        [0.1236],
        [0.0750],
        [0.1407]])
Epoch 62 train loss: 0.4149, eval loss 0.7206574082374573
tensor([[0.1654],
        [0.5131],
```

```
        [0.4340],
        ...,
        [0.1215],
        [0.0724],
        [0.1392]])
Epoch 63 train loss: 0.4213, eval loss 0.7200643420219421
tensor([[0.1645],
        [0.5181],
        [0.4400],
        ...,
        [0.1193],
        [0.0697],
        [0.1377]])
Epoch 64 train loss: 0.4055, eval loss 0.7194331288337708
tensor([[0.1640],
        [0.5230],
        [0.4462],
        ...,
        [0.1175],
        [0.0673],
        [0.1365]])
Epoch 65 train loss: 0.4066, eval loss 0.7188980579376221
tensor([[0.1634],
        [0.5279],
        [0.4523],
        ...,
        [0.1155],
        [0.0650],
        [0.1352]])
Epoch 66 train loss: 0.4182, eval loss 0.7183459997177124
tensor([[0.1627],
        [0.5327],
        [0.4581],
        ...,
        [0.1134],
        [0.0627],
        [0.1338]])
Epoch 67 train loss: 0.4031, eval loss 0.7177897691726685
tensor([[0.1620],
        [0.5376],
        [0.4636],
        ...,
        [0.1114],
        [0.0605],
        [0.1323]])
Epoch 68 train loss: 0.3989, eval loss 0.7172337770462036
tensor([[0.1614],
        [0.5424],
        [0.4694],
        ...,
        [0.1095],
        [0.0584],
        [0.1312]])
Epoch 69 train loss: 0.4303, eval loss 0.7167323231697083
tensor([[0.1612],
        [0.5472],
        [0.4752],
        ...,
        [0.1080],
        [0.0565],
        [0.1303]])
Epoch 70 train loss: 0.4084, eval loss 0.7163107395172119
tensor([[0.1606],
        [0.5520],
        [0.4806],
        ...,
```

```
        [0.1062],
        [0.0545],
        [0.1290]])
Epoch 71 train loss: 0.4030, eval loss 0.7158381938934326
tensor([[0.1605],
        [0.5571],
        [0.4864],
        ...,
        [0.1047],
        [0.0527],
        [0.1282]])
Epoch 72 train loss: 0.4069, eval loss 0.7154344320297241
tensor([[0.1597],
        [0.5617],
        [0.4915],
        ...,
        [0.1027],
        [0.0509],
        [0.1267]])
Epoch 73 train loss: 0.3973, eval loss 0.7149480581283569
tensor([[0.1591],
        [0.5664],
        [0.4968],
        ...,
        [0.1010],
        [0.0492],
        [0.1256]])
Epoch 74 train loss: 0.4154, eval loss 0.7145039439201355
tensor([[0.1586],
        [0.5709],
        [0.5019],
        ...,
        [0.0992],
        [0.0475],
        [0.1244]])
Epoch 75 train loss: 0.4057, eval loss 0.7140629887580872
tensor([[0.1583],
        [0.5758],
        [0.5075],
        ...,
        [0.0977],
        [0.0459],
        [0.1236]])
Epoch 76 train loss: 0.4257, eval loss 0.7136954069137573
tensor([[0.1576],
        [0.5803],
        [0.5123],
        ...,
        [0.0958],
        [0.0443],
        [0.1223]])
Epoch 77 train loss: 0.4040, eval loss 0.7132340669631958
tensor([[0.1571],
        [0.5848],
        [0.5172],
        ...,
        [0.0943],
        [0.0429],
        [0.1213]])
Epoch 78 train loss: 0.4014, eval loss 0.7128432989120483
tensor([[0.1566],
        [0.5892],
        [0.5220],
        ...,
        [0.0927],
        [0.0415],
```

```
        [0.1203]])
Epoch 79 train loss: 0.4169, eval loss 0.7124595642089844
tensor([[0.1558],
        [0.5936],
        [0.5265],
        ...,
        [0.0909],
        [0.0401],
        [0.1190]])
Epoch 80 train loss: 0.4084, eval loss 0.7120310068130493
tensor([[0.1554],
        [0.5980],
        [0.5312],
        ...,
        [0.0895],
        [0.0389],
        [0.1181]])
Epoch 81 train loss: 0.4030, eval loss 0.7116968631744385
tensor([[0.1548],
        [0.6020],
        [0.5355],
        ...,
        [0.0879],
        [0.0377],
        [0.1171]])
Epoch 82 train loss: 0.3756, eval loss 0.7113170027732849
tensor([[0.1546],
        [0.6064],
        [0.5406],
        ...,
        [0.0865],
        [0.0366],
        [0.1163]])
Epoch 83 train loss: 0.4000, eval loss 0.7110137343406677
tensor([[0.1541],
        [0.6105],
        [0.5451],
        ...,
        [0.0852],
        [0.0355],
        [0.1154]])
Epoch 84 train loss: 0.3671, eval loss 0.7106923460960388
tensor([[0.1533],
        [0.6143],
        [0.5491],
        ...,
        [0.0835],
        [0.0344],
        [0.1141]])
Epoch 85 train loss: 0.3975, eval loss 0.7103023529052734
tensor([[0.1529],
        [0.6179],
        [0.5531],
        ...,
        [0.0822],
        [0.0335],
        [0.1133]])
Epoch 86 train loss: 0.4247, eval loss 0.7100095152854919
tensor([[0.1528],
        [0.6218],
        [0.5574],
        ...,
        [0.0810],
        [0.0326],
        [0.1127]])
Epoch 87 train loss: 0.3973, eval loss 0.7097485661506653
```

```
tensor([[0.1525],
        [0.6255],
        [0.5615],
        ...,
        [0.0797],
        [0.0317],
        [0.1119]])
Epoch 88 train loss: 0.4150, eval loss 0.7094582319259644
tensor([[0.1518],
        [0.6289],
        [0.5652],
        ...,
        [0.0783],
        [0.0308],
        [0.1108]])
Epoch 89 train loss: 0.3975, eval loss 0.7091214060783386
tensor([[0.1513],
        [0.6325],
        [0.5690],
        ...,
        [0.0769],
        [0.0299],
        [0.1099]])
Epoch 90 train loss: 0.4042, eval loss 0.7088332176208496
tensor([[0.1509],
        [0.6360],
        [0.5729],
        ...,
        [0.0757],
        [0.0292],
        [0.1091]])
Epoch 91 train loss: 0.3896, eval loss 0.7085840702056885
tensor([[0.1507],
        [0.6393],
        [0.5764],
        ...,
        [0.0745],
        [0.0284],
        [0.1085]])
Epoch 92 train loss: 0.3886, eval loss 0.7083477973937988
tensor([[0.1502],
        [0.6426],
        [0.5797],
        ...,
        [0.0732],
        [0.0277],
        [0.1077]])
Epoch 93 train loss: 0.4078, eval loss 0.708077073097229
tensor([[0.1497],
        [0.6459],
        [0.5829],
        ...,
        [0.0721],
        [0.0270],
        [0.1069]])
Epoch 94 train loss: 0.3756, eval loss 0.7078351974487305
tensor([[0.1489],
        [0.6488],
        [0.5856],
        ...,
        [0.0707],
        [0.0263],
        [0.1058]])
Epoch 95 train loss: 0.3918, eval loss 0.7075007557868958
tensor([[0.1486],
        [0.6516],
```

```
        [0.5886],
        ...,
        [0.0697],
        [0.0257],
        [0.1051]])
Epoch 96 train loss: 0.3818, eval loss 0.7072798013687134
tensor([[0.1487],
        [0.6545],
        [0.5916],
        ...,
        [0.0688],
        [0.0252],
        [0.1048]])
Epoch 97 train loss: 0.3852, eval loss 0.7070964574813843
tensor([[0.1481],
        [0.6572],
        [0.5942],
        ...,
        [0.0676],
        [0.0246],
        [0.1039]])
Epoch 98 train loss: 0.4010, eval loss 0.7068347930908203
tensor([[0.1476],
        [0.6598],
        [0.5968],
        ...,
        [0.0666],
        [0.0240],
        [0.1032]])
Epoch 99 train loss: 0.3974, eval loss 0.7066034078598022
tensor([[0.1472],
        [0.6621],
        [0.5991],
        ...,
        [0.0654],
        [0.0235],
        [0.1025]])
Epoch 100 train loss: 0.3942, eval loss 0.7063566446304321
tensor([[0.1469],
        [0.6645],
        [0.6016],
        ...,
        [0.0645],
        [0.0230],
        [0.1020]])
Epoch 101 train loss: 0.3981, eval loss 0.7061589360237122
tensor([[0.1464],
        [0.6669],
        [0.6041],
        ...,
        [0.0636],
        [0.0225],
        [0.1012]])
Epoch 102 train loss: 0.3893, eval loss 0.7059480547904968
tensor([[0.1462],
        [0.6694],
        [0.6067],
        ...,
        [0.0627],
        [0.0221],
        [0.1008]])
Epoch 103 train loss: 0.3802, eval loss 0.7057801485061646
tensor([[0.1457],
        [0.6718],
        [0.6091],
        ...,
```

```
        [0.0618],
        [0.0217],
        [0.1001]])
Epoch 104 train loss: 0.4136, eval loss 0.7055513262748718
tensor([[0.1454],
        [0.6740],
        [0.6115],
        ...,
        [0.0609],
        [0.0212],
        [0.0996]])
Epoch 105 train loss: 0.4110, eval loss 0.7053661942481995
tensor([[0.1449],
        [0.6760],
        [0.6136],
        ...,
        [0.0600],
        [0.0209],
        [0.0989]])
Epoch 106 train loss: 0.4034, eval loss 0.7051578760147095
tensor([[0.1446],
        [0.6783],
        [0.6160],
        ...,
        [0.0591],
        [0.0205],
        [0.0984]])
Epoch 107 train loss: 0.3661, eval loss 0.7049967050552368
tensor([[0.1445],
        [0.6805],
        [0.6185],
        ...,
        [0.0584],
        [0.0201],
        [0.0980]])
Epoch 108 train loss: 0.3828, eval loss 0.7048390507698059
tensor([[0.1441],
        [0.6827],
        [0.6207],
        ...,
        [0.0576],
        [0.0198],
        [0.0975]])
Epoch 109 train loss: 0.3846, eval loss 0.7046543955802917
tensor([[0.1435],
        [0.6845],
        [0.6226],
        ...,
        [0.0568],
        [0.0194],
        [0.0969]])
Epoch 110 train loss: 0.3707, eval loss 0.7044621706008911
tensor([[0.1434],
        [0.6863],
        [0.6246],
        ...,
        [0.0561],
        [0.0191],
        [0.0966]])
Epoch 111 train loss: 0.4017, eval loss 0.7043234705924988
tensor([[0.1431],
        [0.6880],
        [0.6264],
        ...,
        [0.0556],
        [0.0189],
```

```
        [0.0963]])
Epoch 112 train loss: 0.3954, eval loss 0.7041837573051453
tensor([[0.1428],
        [0.6895],
        [0.6281],
        ...,
        [0.0549],
        [0.0186],
        [0.0959]])
Epoch 113 train loss: 0.3754, eval loss 0.7040092945098877
tensor([[0.1426],
        [0.6911],
        [0.6299],
        ...,
        [0.0544],
        [0.0183],
        [0.0956]])
Epoch 114 train loss: 0.3733, eval loss 0.7038753032684326
tensor([[0.1427],
        [0.6927],
        [0.6317],
        ...,
        [0.0539],
        [0.0181],
        [0.0955]])
Epoch 115 train loss: 0.3942, eval loss 0.7037782669067383
tensor([[0.1428],
        [0.6941],
        [0.6333],
        ...,
        [0.0535],
        [0.0179],
        [0.0954]])
Epoch 116 train loss: 0.3799, eval loss 0.703717052936554
tensor([[0.1422],
        [0.6955],
        [0.6348],
        ...,
        [0.0529],
        [0.0176],
        [0.0949]])
Epoch 117 train loss: 0.4017, eval loss 0.7035383582115173
tensor([[0.1416],
        [0.6968],
        [0.6362],
        ...,
        [0.0522],
        [0.0174],
        [0.0943]])
Epoch 118 train loss: 0.4019, eval loss 0.7033478021621704
tensor([[0.1412],
        [0.6982],
        [0.6377],
        ...,
        [0.0518],
        [0.0171],
        [0.0940]])
Epoch 119 train loss: 0.4146, eval loss 0.7032114863395691
tensor([[0.1411],
        [0.6994],
        [0.6392],
        ...,
        [0.0513],
        [0.0170],
        [0.0937]])
Epoch 120 train loss: 0.3975, eval loss 0.7031040787696838
```

```
tensor([[0.1407],
        [0.7007],
        [0.6406],
        ...,
        [0.0508],
        [0.0168],
        [0.0934]])
Epoch 121 train loss: 0.4005, eval loss 0.7029848694801331
tensor([[0.1399],
        [0.7018],
        [0.6419],
        ...,
        [0.0501],
        [0.0165],
        [0.0926]])
Epoch 122 train loss: 0.3939, eval loss 0.7027607560157776
tensor([[0.1395],
        [0.7031],
        [0.6434],
        ...,
        [0.0496],
        [0.0163],
        [0.0923]])
Epoch 123 train loss: 0.3882, eval loss 0.7026083469390869
tensor([[0.1389],
        [0.7039],
        [0.6444],
        ...,
        [0.0490],
        [0.0161],
        [0.0917]])
Epoch 124 train loss: 0.3880, eval loss 0.7024297118186951
tensor([[0.1384],
        [0.7048],
        [0.6455],
        ...,
        [0.0485],
        [0.0159],
        [0.0913]])
Epoch 125 train loss: 0.3651, eval loss 0.7022769451141357
tensor([[0.1383],
        [0.7059],
        [0.6468],
        ...,
        [0.0481],
        [0.0158],
        [0.0912]])
Epoch 126 train loss: 0.3928, eval loss 0.702196478843689
tensor([[0.1373],
        [0.7071],
        [0.6480],
        ...,
        [0.0475],
        [0.0156],
        [0.0905]])
Epoch 127 train loss: 0.3907, eval loss 0.7019831538200378
tensor([[0.1370],
        [0.7085],
        [0.6496],
        ...,
        [0.0470],
        [0.0154],
        [0.0901]])
Epoch 128 train loss: 0.3986, eval loss 0.701878547668457
tensor([[0.1365],
        [0.7093],
```

```
        [0.6506],
        ...,
        [0.0466],
        [0.0153],
        [0.0897]])
Epoch 129 train loss: 0.4065, eval loss 0.7017270922660828
tensor([[0.1370],
        [0.7108],
        [0.6524],
        ...,
        [0.0463],
        [0.0152],
        [0.0900]])
Epoch 130 train loss: 0.3987, eval loss 0.7017460465431213
tensor([[0.1372],
        [0.7122],
        [0.6541],
        ...,
        [0.0462],
        [0.0151],
        [0.0902]])
Epoch 131 train loss: 0.3748, eval loss 0.7017617225646973
tensor([[0.1369],
        [0.7130],
        [0.6552],
        ...,
        [0.0458],
        [0.0149],
        [0.0901]])
Epoch 132 train loss: 0.3822, eval loss 0.7016583681106567
tensor([[0.1362],
        [0.7139],
        [0.6562],
        ...,
        [0.0454],
        [0.0148],
        [0.0896]])
Epoch 133 train loss: 0.3661, eval loss 0.7014980912208557
tensor([[0.1358],
        [0.7148],
        [0.6571],
        ...,
        [0.0449],
        [0.0146],
        [0.0892]])
Epoch 134 train loss: 0.3976, eval loss 0.7013714909553528
tensor([[0.1357],
        [0.7157],
        [0.6583],
        ...,
        [0.0446],
        [0.0145],
        [0.0891]])
Epoch 135 train loss: 0.3980, eval loss 0.7013109922409058
tensor([[0.1352],
        [0.7163],
        [0.6592],
        ...,
        [0.0441],
        [0.0144],
        [0.0887]])
Epoch 136 train loss: 0.3908, eval loss 0.7011807560920715
tensor([[0.1354],
        [0.7173],
        [0.6605],
        ...,
```

```
        [0.0439],
        [0.0143],
        [0.0888]])
Epoch 137 train loss: 0.3726, eval loss 0.70115727186203
tensor([[0.1347],
        [0.7182],
        [0.6615],
        ...,
        [0.0435],
        [0.0142],
        [0.0884]])
Epoch 138 train loss: 0.3864, eval loss 0.7010329365730286
tensor([[0.1347],
        [0.7191],
        [0.6626],
        ...,
        [0.0432],
        [0.0141],
        [0.0884]])
Epoch 139 train loss: 0.3572, eval loss 0.7009764313697815
tensor([[0.1341],
        [0.7197],
        [0.6633],
        ...,
        [0.0429],
        [0.0140],
        [0.0881]])
Epoch 140 train loss: 0.3868, eval loss 0.7008467316627502
tensor([[0.1338],
        [0.7207],
        [0.6645],
        ...,
        [0.0425],
        [0.0138],
        [0.0879]])
Epoch 141 train loss: 0.3921, eval loss 0.7007395625114441
tensor([[0.1336],
        [0.7217],
        [0.6656],
        ...,
        [0.0422],
        [0.0137],
        [0.0876]])
Epoch 142 train loss: 0.4068, eval loss 0.7006398439407349
tensor([[0.1331],
        [0.7224],
        [0.6664],
        ...,
        [0.0419],
        [0.0136],
        [0.0873]])
Epoch 143 train loss: 0.3653, eval loss 0.700533390045166
tensor([[0.1329],
        [0.7231],
        [0.6674],
        ...,
        [0.0416],
        [0.0135],
        [0.0872]])
Epoch 144 train loss: 0.3755, eval loss 0.7004663944244385
tensor([[0.1325],
        [0.7239],
        [0.6683],
        ...,
        [0.0414],
        [0.0135],
```

```
        [0.0870]])
Epoch 145 train loss: 0.3637, eval loss 0.7004027366638184
tensor([[0.1323],
        [0.7247],
        [0.6693],
        ...,
        [0.0410],
        [0.0133],
        [0.0869]])
Epoch 146 train loss: 0.3880, eval loss 0.7003120183944702
tensor([[0.1319],
        [0.7255],
        [0.6702],
        ...,
        [0.0407],
        [0.0133],
        [0.0867]])
Epoch 147 train loss: 0.3769, eval loss 0.7002341151237488
tensor([[0.1315],
        [0.7262],
        [0.6711],
        ...,
        [0.0404],
        [0.0132],
        [0.0864]])
Epoch 148 train loss: 0.4116, eval loss 0.7001532912254333
tensor([[0.1308],
        [0.7269],
        [0.6718],
        ...,
        [0.0401],
        [0.0131],
        [0.0861]])
Epoch 149 train loss: 0.3598, eval loss 0.7000499963760376
tensor([[0.1303],
        [0.7273],
        [0.6724],
        ...,
        [0.0398],
        [0.0130],
        [0.0859]])
Epoch 150 train loss: 0.3649, eval loss 0.6999582052230835
tensor([[0.1298],
        [0.7278],
        [0.6729],
        ...,
        [0.0395],
        [0.0129],
        [0.0857]])
Epoch 151 train loss: 0.3771, eval loss 0.6998475790023804
tensor([[0.1297],
        [0.7285],
        [0.6738],
        ...,
        [0.0394],
        [0.0129],
        [0.0857]])
Epoch 152 train loss: 0.3928, eval loss 0.6998289823532104
tensor([[0.1293],
        [0.7291],
        [0.6745],
        ...,
        [0.0391],
        [0.0128],
        [0.0855]])
Epoch 153 train loss: 0.3835, eval loss 0.699730396270752
```

```
tensor([[0.1291],
        [0.7298],
        [0.6754],
        ...,
        [0.0389],
        [0.0127],
        [0.0855]])
Epoch 154 train loss: 0.3770, eval loss 0.6996609568595886
tensor([[0.1285],
        [0.7302],
        [0.6759],
        ...,
        [0.0386],
        [0.0126],
        [0.0852]])
Epoch 155 train loss: 0.3974, eval loss 0.6995344758033752
tensor([[0.1283],
        [0.7307],
        [0.6765],
        ...,
        [0.0384],
        [0.0126],
        [0.0852]])
Epoch 156 train loss: 0.3882, eval loss 0.6994751691818237
tensor([[0.1278],
        [0.7313],
        [0.6772],
        ...,
        [0.0382],
        [0.0125],
        [0.0850]])
Epoch 157 train loss: 0.3776, eval loss 0.6994084119796753
tensor([[0.1273],
        [0.7318],
        [0.6777],
        ...,
        [0.0380],
        [0.0124],
        [0.0848]])
Epoch 158 train loss: 0.4001, eval loss 0.6993196606636047
tensor([[0.1267],
        [0.7321],
        [0.6781],
        ...,
        [0.0377],
        [0.0123],
        [0.0845]])
Epoch 159 train loss: 0.3816, eval loss 0.6992015242576599
tensor([[0.1264],
        [0.7325],
        [0.6786],
        ...,
        [0.0376],
        [0.0123],
        [0.0846]])
Epoch 160 train loss: 0.3661, eval loss 0.699142336845398
tensor([[0.1260],
        [0.7327],
        [0.6790],
        ...,
        [0.0375],
        [0.0122],
        [0.0844]])
Epoch 161 train loss: 0.4215, eval loss 0.6990565061569214
tensor([[0.1256],
        [0.7333],
```

```
        [0.6797],
        ...,
        [0.0372],
        [0.0122],
        [0.0842]])
Epoch 162 train loss: 0.3888, eval loss 0.6989603638648987
tensor([[0.1252],
        [0.7337],
        [0.6803],
        ...,
        [0.0370],
        [0.0121],
        [0.0841]])
Epoch 163 train loss: 0.3765, eval loss 0.6989077925682068
tensor([[0.1249],
        [0.7343],
        [0.6809],
        ...,
        [0.0368],
        [0.0120],
        [0.0840]])
Epoch 164 train loss: 0.3715, eval loss 0.6988497972488403
tensor([[0.1242],
        [0.7349],
        [0.6816],
        ...,
        [0.0364],
        [0.0119],
        [0.0836]])
Epoch 165 train loss: 0.3778, eval loss 0.6987264156341553
tensor([[0.1235],
        [0.7352],
        [0.6820],
        ...,
        [0.0361],
        [0.0118],
        [0.0832]])
Epoch 166 train loss: 0.3750, eval loss 0.6985686421394348
tensor([[0.1228],
        [0.7357],
        [0.6826],
        ...,
        [0.0358],
        [0.0117],
        [0.0828]])
Epoch 167 train loss: 0.3541, eval loss 0.6984437108039856
tensor([[0.1230],
        [0.7365],
        [0.6836],
        ...,
        [0.0357],
        [0.0117],
        [0.0831]])
Epoch 168 train loss: 0.3963, eval loss 0.698489785194397
tensor([[0.1229],
        [0.7371],
        [0.6844],
        ...,
        [0.0357],
        [0.0117],
        [0.0833]])
Epoch 169 train loss: 0.3742, eval loss 0.6984850168228149
tensor([[0.1222],
        [0.7375],
        [0.6848],
        ...,
```

```
        [0.0354],
        [0.0116],
        [0.0830]])
Epoch 170 train loss: 0.3437, eval loss 0.6983683109283447
tensor([[0.1219],
        [0.7382],
        [0.6855],
        ...,
        [0.0353],
        [0.0116],
        [0.0829]])
Epoch 171 train loss: 0.3932, eval loss 0.6983386278152466
tensor([[0.1219],
        [0.7389],
        [0.6863],
        ...,
        [0.0352],
        [0.0115],
        [0.0830]])
Epoch 172 train loss: 0.3558, eval loss 0.6983299851417542
tensor([[0.1217],
        [0.7396],
        [0.6872],
        ...,
        [0.0351],
        [0.0115],
        [0.0831]])
Epoch 173 train loss: 0.3849, eval loss 0.6982997059822083
tensor([[0.1210],
        [0.7400],
        [0.6876],
        ...,
        [0.0348],
        [0.0114],
        [0.0827]])
Epoch 174 train loss: 0.3934, eval loss 0.6981901526451111
tensor([[0.1211],
        [0.7406],
        [0.6883],
        ...,
        [0.0347],
        [0.0114],
        [0.0829]])
Epoch 175 train loss: 0.4002, eval loss 0.6981723308563232
tensor([[0.1206],
        [0.7411],
        [0.6888],
        ...,
        [0.0345],
        [0.0113],
        [0.0827]])
Epoch 176 train loss: 0.3823, eval loss 0.6981101036071777
tensor([[0.1203],
        [0.7417],
        [0.6895],
        ...,
        [0.0344],
        [0.0113],
        [0.0827]])
Epoch 177 train loss: 0.3760, eval loss 0.6980744004249573
tensor([[0.1195],
        [0.7416],
        [0.6895],
        ...,
        [0.0341],
        [0.0112],
```

```
        [0.0823]])
Epoch 178 train loss: 0.3856, eval loss 0.6979422569274902
tensor([[0.1189],
        [0.7421],
        [0.6900],
        ...,
        [0.0340],
        [0.0112],
        [0.0821]])
Epoch 179 train loss: 0.3677, eval loss 0.6978561282157898
tensor([[0.1184],
        [0.7426],
        [0.6905],
        ...,
        [0.0339],
        [0.0111],
        [0.0819]])
Epoch 180 train loss: 0.3811, eval loss 0.6977962851524353
tensor([[0.1184],
        [0.7433],
        [0.6914],
        ...,
        [0.0338],
        [0.0111],
        [0.0821]])
Epoch 181 train loss: 0.3643, eval loss 0.6977999210357666
tensor([[0.1181],
        [0.7438],
        [0.6918],
        ...,
        [0.0337],
        [0.0110],
        [0.0820]])
Epoch 182 train loss: 0.3840, eval loss 0.6977445483207703
tensor([[0.1175],
        [0.7441],
        [0.6921],
        ...,
        [0.0335],
        [0.0110],
        [0.0818]])
Epoch 183 train loss: 0.3574, eval loss 0.6976673603057861
tensor([[0.1174],
        [0.7446],
        [0.6927],
        ...,
        [0.0335],
        [0.0110],
        [0.0820]])
Epoch 184 train loss: 0.3664, eval loss 0.6976714134216309
tensor([[0.1171],
        [0.7449],
        [0.6931],
        ...,
        [0.0334],
        [0.0109],
        [0.0820]])
Epoch 185 train loss: 0.4004, eval loss 0.6976303458213806
tensor([[0.1164],
        [0.7451],
        [0.6933],
        ...,
        [0.0332],
        [0.0109],
        [0.0817]])
Epoch 186 train loss: 0.4101, eval loss 0.6975318193435669
```

```
tensor([[0.1160],
        [0.7455],
        [0.6937],
        ...,
        [0.0331],
        [0.0108],
        [0.0817]])
Epoch 187 train loss: 0.3677, eval loss 0.6975052952766418
tensor([[0.1159],
        [0.7461],
        [0.6945],
        ...,
        [0.0331],
        [0.0108],
        [0.0819]])
Epoch 188 train loss: 0.3503, eval loss 0.6974973678588867
tensor([[0.1156],
        [0.7464],
        [0.6948],
        ...,
        [0.0330],
        [0.0108],
        [0.0820]])
Epoch 189 train loss: 0.3962, eval loss 0.6974585056304932
tensor([[0.1154],
        [0.7470],
        [0.6955],
        ...,
        [0.0329],
        [0.0107],
        [0.0820]])
Epoch 190 train loss: 0.3664, eval loss 0.6974256038665771
tensor([[0.1153],
        [0.7473],
        [0.6960],
        ...,
        [0.0328],
        [0.0107],
        [0.0821]])
Epoch 191 train loss: 0.3866, eval loss 0.6973950266838074
tensor([[0.1142],
        [0.7475],
        [0.6960],
        ...,
        [0.0325],
        [0.0106],
        [0.0816]])
Epoch 192 train loss: 0.3903, eval loss 0.6972233653068542
tensor([[0.1135],
        [0.7478],
        [0.6964],
        ...,
        [0.0323],
        [0.0105],
        [0.0814]])
Epoch 193 train loss: 0.3850, eval loss 0.697119414806366
tensor([[0.1127],
        [0.7480],
        [0.6966],
        ...,
        [0.0322],
        [0.0105],
        [0.0811]])
Epoch 194 train loss: 0.3947, eval loss 0.6970065832138062
tensor([[0.1125],
        [0.7486],
```

```
        [0.6972],
        ...,
        [0.0321],
        [0.0104],
        [0.0812]])
Epoch 195 train loss: 0.4148, eval loss 0.6969975233078003
tensor([[0.1125],
        [0.7492],
        [0.6979],
        ...,
        [0.0320],
        [0.0104],
        [0.0814]])
Epoch 196 train loss: 0.4023, eval loss 0.6970105171203613
tensor([[0.1127],
        [0.7497],
        [0.6986],
        ...,
        [0.0320],
        [0.0104],
        [0.0818]])
Epoch 197 train loss: 0.3837, eval loss 0.6970665454864502
tensor([[0.1117],
        [0.7501],
        [0.6988],
        ...,
        [0.0318],
        [0.0103],
        [0.0814]])
Epoch 198 train loss: 0.3603, eval loss 0.6969547867774963
tensor([[0.1115],
        [0.7505],
        [0.6993],
        ...,
        [0.0317],
        [0.0103],
        [0.0815]])
Epoch 199 train loss: 0.3781, eval loss 0.6969383358955383
tensor([[0.1106],
        [0.7511],
        [0.6997],
        ...,
        [0.0315],
        [0.0102],
        [0.0810]])
Epoch 200 train loss: 0.3506, eval loss 0.6968243718147278
tensor([[0.1103],
        [0.7516],
        [0.7004],
        ...,
        [0.0313],
        [0.0102],
        [0.0810]])
Epoch 201 train loss: 0.3718, eval loss 0.6967495083808899
tensor([[0.1103],
        [0.7522],
        [0.7011],
        ...,
        [0.0313],
        [0.0102],
        [0.0811]])
Epoch 202 train loss: 0.3844, eval loss 0.6967585682868958
tensor([[0.1098],
        [0.7523],
        [0.7012],
        ...,
```

```
        [0.0312],
        [0.0101],
        [0.0810]])
Epoch 203 train loss: 0.3450, eval loss 0.6966646909713745
tensor([[0.1095],
        [0.7525],
        [0.7015],
        ...,
        [0.0310],
        [0.0101],
        [0.0810]])
Epoch 204 train loss: 0.4018, eval loss 0.6966320276260376
tensor([[0.1094],
        [0.7529],
        [0.7020],
        ...,
        [0.0309],
        [0.0101],
        [0.0812]])
Epoch 205 train loss: 0.4051, eval loss 0.6966316103935242
tensor([[0.1090],
        [0.7531],
        [0.7023],
        ...,
        [0.0308],
        [0.0100],
        [0.0812]])
Epoch 206 train loss: 0.3766, eval loss 0.6965577006340027
tensor([[0.1082],
        [0.7534],
        [0.7025],
        ...,
        [0.0306],
        [0.0099],
        [0.0808]])
Epoch 207 train loss: 0.3823, eval loss 0.6964257955551147
tensor([[0.1079],
        [0.7538],
        [0.7028],
        ...,
        [0.0306],
        [0.0099],
        [0.0809]])
Epoch 208 train loss: 0.3605, eval loss 0.6963928937911987
tensor([[0.1080],
        [0.7543],
        [0.7034],
        ...,
        [0.0307],
        [0.0099],
        [0.0813]])
Epoch 209 train loss: 0.3937, eval loss 0.6964395046234131
tensor([[0.1077],
        [0.7547],
        [0.7038],
        ...,
        [0.0306],
        [0.0099],
        [0.0813]])
Epoch 210 train loss: 0.3598, eval loss 0.6963790059089661
tensor([[0.1074],
        [0.7550],
        [0.7042],
        ...,
        [0.0306],
        [0.0099],
```

```
        [0.0815]])
Epoch 211 train loss: 0.3757, eval loss 0.6963608264923096
tensor([[0.1070],
        [0.7551],
        [0.7043],
        ...,
        [0.0304],
        [0.0098],
        [0.0812]])
Epoch 212 train loss: 0.3679, eval loss 0.6962800621986389
tensor([[0.1068],
        [0.7556],
        [0.7048],
        ...,
        [0.0304],
        [0.0098],
        [0.0815]])
Epoch 213 train loss: 0.3495, eval loss 0.6962953805923462
tensor([[0.1058],
        [0.7557],
        [0.7047],
        ...,
        [0.0301],
        [0.0097],
        [0.0809]])
Epoch 214 train loss: 0.3986, eval loss 0.6961293816566467
tensor([[0.1056],
        [0.7564],
        [0.7055],
        ...,
        [0.0300],
        [0.0097],
        [0.0810]])
Epoch 215 train loss: 0.3620, eval loss 0.6961299777030945
tensor([[0.1051],
        [0.7567],
        [0.7058],
        ...,
        [0.0300],
        [0.0097],
        [0.0809]])
Epoch 216 train loss: 0.3684, eval loss 0.6960800290107727
tensor([[0.1051],
        [0.7575],
        [0.7067],
        ...,
        [0.0300],
        [0.0096],
        [0.0812]])
Epoch 217 train loss: 0.3500, eval loss 0.6961080431938171
tensor([[0.1049],
        [0.7580],
        [0.7072],
        ...,
        [0.0299],
        [0.0096],
        [0.0813]])
Epoch 218 train loss: 0.3674, eval loss 0.6960762143135071
tensor([[0.1045],
        [0.7585],
        [0.7077],
        ...,
        [0.0298],
        [0.0096],
        [0.0812]])
Epoch 219 train loss: 0.3861, eval loss 0.6960244178771973
```

```
tensor([[0.1039],
        [0.7586],
        [0.7077],
        ...,
        [0.0296],
        [0.0095],
        [0.0812]])
Epoch 220 train loss: 0.3619, eval loss 0.69594806432724
tensor([[0.1036],
        [0.7590],
        [0.7081],
        ...,
        [0.0295],
        [0.0095],
        [0.0811]])
Epoch 221 train loss: 0.3663, eval loss 0.6959074139595032
tensor([[0.1031],
        [0.7594],
        [0.7085],
        ...,
        [0.0294],
        [0.0094],
        [0.0810]])
Epoch 222 train loss: 0.3841, eval loss 0.695855438709259
tensor([[0.1026],
        [0.7596],
        [0.7086],
        ...,
        [0.0294],
        [0.0094],
        [0.0810]])
Epoch 223 train loss: 0.3523, eval loss 0.6957951784133911
tensor([[0.1022],
        [0.7601],
        [0.7090],
        ...,
        [0.0293],
        [0.0094],
        [0.0809]])
Epoch 224 train loss: 0.3687, eval loss 0.6957592964172363
tensor([[0.1019],
        [0.7608],
        [0.7096],
        ...,
        [0.0292],
        [0.0093],
        [0.0811]])
Epoch 225 train loss: 0.3563, eval loss 0.6957342028617859
tensor([[0.1013],
        [0.7610],
        [0.7099],
        ...,
        [0.0291],
        [0.0093],
        [0.0809]])
Epoch 226 train loss: 0.3612, eval loss 0.6956626176834106
tensor([[0.1009],
        [0.7614],
        [0.7103],
        ...,
        [0.0290],
        [0.0092],
        [0.0809]])
Epoch 227 train loss: 0.3462, eval loss 0.6956233978271484
tensor([[0.1007],
        [0.7617],
```

```
        [0.7107],
        ...,
        [0.0289],
        [0.0092],
        [0.0809]])
Epoch 228 train loss: 0.3771, eval loss 0.6955837607383728
tensor([[0.1003],
        [0.7622],
        [0.7112],
        ...,
        [0.0288],
        [0.0091],
        [0.0808]])
Epoch 229 train loss: 0.3728, eval loss 0.6955260038375854
tensor([[0.0994],
        [0.7623],
        [0.7112],
        ...,
        [0.0286],
        [0.0091],
        [0.0805]])
Epoch 230 train loss: 0.3526, eval loss 0.6953924298286438
tensor([[0.0994],
        [0.7627],
        [0.7116],
        ...,
        [0.0286],
        [0.0091],
        [0.0808]])
Epoch 231 train loss: 0.3736, eval loss 0.6954019665718079
tensor([[0.0990],
        [0.7629],
        [0.7119],
        ...,
        [0.0285],
        [0.0090],
        [0.0808]])
Epoch 232 train loss: 0.3492, eval loss 0.695358395576477
tensor([[0.0990],
        [0.7635],
        [0.7126],
        ...,
        [0.0285],
        [0.0090],
        [0.0810]])
Epoch 233 train loss: 0.3644, eval loss 0.6953827738761902
tensor([[0.0987],
        [0.7639],
        [0.7130],
        ...,
        [0.0284],
        [0.0090],
        [0.0810]])
Epoch 234 train loss: 0.3760, eval loss 0.6953495144844055
tensor([[0.0984],
        [0.7641],
        [0.7132],
        ...,
        [0.0282],
        [0.0089],
        [0.0810]])
Epoch 235 train loss: 0.3636, eval loss 0.6952821016311646
tensor([[0.0980],
        [0.7645],
        [0.7137],
        ...,
```

```
        [0.0282],
        [0.0089],
        [0.0810]])
Epoch 236 train loss: 0.3599, eval loss 0.6952510476112366
tensor([[0.0978],
        [0.7647],
        [0.7138],
        ...,
        [0.0281],
        [0.0089],
        [0.0811]])
Epoch 237 train loss: 0.3794, eval loss 0.6952342987060547
tensor([[0.0974],
        [0.7648],
        [0.7140],
        ...,
        [0.0280],
        [0.0088],
        [0.0809]])
Epoch 238 train loss: 0.3693, eval loss 0.6951684355735779
tensor([[0.0970],
        [0.7649],
        [0.7141],
        ...,
        [0.0279],
        [0.0088],
        [0.0808]])
Epoch 239 train loss: 0.3769, eval loss 0.6950883865356445
tensor([[0.0967],
        [0.7653],
        [0.7145],
        ...,
        [0.0278],
        [0.0088],
        [0.0808]])
Epoch 240 train loss: 0.3760, eval loss 0.6950602531433105
tensor([[0.0964],
        [0.7653],
        [0.7144],
        ...,
        [0.0277],
        [0.0087],
        [0.0808]])
Epoch 241 train loss: 0.3565, eval loss 0.6949852108955383
tensor([[0.0963],
        [0.7655],
        [0.7147],
        ...,
        [0.0277],
        [0.0087],
        [0.0809]])
Epoch 242 train loss: 0.3605, eval loss 0.694980800151825
tensor([[0.0961],
        [0.7660],
        [0.7152],
        ...,
        [0.0277],
        [0.0087],
        [0.0810]])
Epoch 243 train loss: 0.3706, eval loss 0.6949658989906311
tensor([[0.0957],
        [0.7661],
        [0.7153],
        ...,
        [0.0276],
        [0.0087],
```

```
        [0.0810]])
Epoch 244 train loss: 0.3734, eval loss 0.69492506980896
tensor([[0.0953],
        [0.7663],
        [0.7154],
        ...,
        [0.0276],
        [0.0086],
        [0.0811]])
Epoch 245 train loss: 0.3540, eval loss 0.6948869824409485
tensor([[0.0957],
        [0.7672],
        [0.7165],
        ...,
        [0.0277],
        [0.0086],
        [0.0815]])
Epoch 246 train loss: 0.3409, eval loss 0.6949678063392639
tensor([[0.0953],
        [0.7676],
        [0.7169],
        ...,
        [0.0276],
        [0.0086],
        [0.0815]])
Epoch 247 train loss: 0.4071, eval loss 0.6949251890182495
tensor([[0.0947],
        [0.7677],
        [0.7169],
        ...,
        [0.0274],
        [0.0085],
        [0.0811]])
Epoch 248 train loss: 0.3553, eval loss 0.694819986820221
tensor([[0.0946],
        [0.7682],
        [0.7173],
        ...,
        [0.0275],
        [0.0085],
        [0.0814]])
Epoch 249 train loss: 0.3744, eval loss 0.6948317289352417
tensor([[0.0940],
        [0.7682],
        [0.7172],
        ...,
        [0.0273],
        [0.0085],
        [0.0811]])
Epoch 250 train loss: 0.3607, eval loss 0.6947374939918518
tensor([[0.0937],
        [0.7682],
        [0.7173],
        ...,
        [0.0272],
        [0.0084],
        [0.0810]])
Epoch 251 train loss: 0.3875, eval loss 0.6946656703948975
tensor([[0.0934],
        [0.7686],
        [0.7175],
        ...,
        [0.0271],
        [0.0084],
        [0.0810]])
Epoch 252 train loss: 0.3466, eval loss 0.6946229934692383
```

```
tensor([[0.0929],
        [0.7689],
        [0.7177],
        ...,
        [0.0271],
        [0.0083],
        [0.0808]])
Epoch 253 train loss: 0.3471, eval loss 0.6945545077323914
tensor([[0.0926],
        [0.7693],
        [0.7182],
        ...,
        [0.0269],
        [0.0083],
        [0.0806]])
Epoch 254 train loss: 0.3955, eval loss 0.6945109963417053
tensor([[0.0927],
        [0.7696],
        [0.7186],
        ...,
        [0.0269],
        [0.0083],
        [0.0809]])
Epoch 255 train loss: 0.3687, eval loss 0.6945532560348511
tensor([[0.0922],
        [0.7698],
        [0.7187],
        ...,
        [0.0268],
        [0.0083],
        [0.0809]])
Epoch 256 train loss: 0.3495, eval loss 0.6945010423660278
tensor([[0.0916],
        [0.7701],
        [0.7189],
        ...,
        [0.0267],
        [0.0082],
        [0.0807]])
Epoch 257 train loss: 0.3705, eval loss 0.6943989992141724
tensor([[0.0915],
        [0.7701],
        [0.7190],
        ...,
        [0.0267],
        [0.0082],
        [0.0808]])
Epoch 258 train loss: 0.4148, eval loss 0.6943970918655396
tensor([[0.0912],
        [0.7704],
        [0.7193],
        ...,
        [0.0267],
        [0.0082],
        [0.0809]])
Epoch 259 train loss: 0.3626, eval loss 0.6943908929824829
tensor([[0.0908],
        [0.7708],
        [0.7196],
        ...,
        [0.0266],
        [0.0081],
        [0.0808]])
Epoch 260 train loss: 0.3526, eval loss 0.6943176388874054
tensor([[0.0907],
        [0.7712],
```

```
        [0.7200],
        ...,
        [0.0267],
        [0.0081],
        [0.0811]])
Epoch 261 train loss: 0.3760, eval loss 0.6943569183349609
tensor([[0.0904],
        [0.7716],
        [0.7203],
        ...,
        [0.0266],
        [0.0081],
        [0.0810]])
Epoch 262 train loss: 0.3774, eval loss 0.6943163275718689
tensor([[0.0899],
        [0.7719],
        [0.7205],
        ...,
        [0.0264],
        [0.0081],
        [0.0808]])
Epoch 263 train loss: 0.3610, eval loss 0.6942456364631653
tensor([[0.0898],
        [0.7724],
        [0.7210],
        ...,
        [0.0264],
        [0.0080],
        [0.0809]])
Epoch 264 train loss: 0.3336, eval loss 0.6942481994628906
tensor([[0.0895],
        [0.7725],
        [0.7211],
        ...,
        [0.0264],
        [0.0080],
        [0.0810]])
Epoch 265 train loss: 0.3514, eval loss 0.69419926404953
tensor([[0.0892],
        [0.7728],
        [0.7212],
        ...,
        [0.0264],
        [0.0080],
        [0.0810]])
Epoch 266 train loss: 0.3638, eval loss 0.694172203540802
tensor([[0.0889],
        [0.7731],
        [0.7215],
        ...,
        [0.0264],
        [0.0080],
        [0.0810]])
Epoch 267 train loss: 0.3560, eval loss 0.6941599249839783
tensor([[0.0891],
        [0.7739],
        [0.7225],
        ...,
        [0.0264],
        [0.0080],
        [0.0814]])
Epoch 268 train loss: 0.4142, eval loss 0.6942111849784851
tensor([[0.0887],
        [0.7742],
        [0.7226],
        ...,
```

```
        [0.0264],
        [0.0079],
        [0.0812]])
Epoch 269 train loss: 0.3691, eval loss 0.6941578388214111
tensor([[0.0881],
        [0.7745],
        [0.7227],
        ...,
        [0.0262],
        [0.0079],
        [0.0809]])
Epoch 270 train loss: 0.3693, eval loss 0.6940632462501526
tensor([[0.0881],
        [0.7749],
        [0.7231],
        ...,
        [0.0263],
        [0.0079],
        [0.0812]])
Epoch 271 train loss: 0.3902, eval loss 0.6940929889678955
tensor([[0.0876],
        [0.7749],
        [0.7231],
        ...,
        [0.0261],
        [0.0078],
        [0.0810]])
Epoch 272 train loss: 0.3519, eval loss 0.6940025091171265
tensor([[0.0870],
        [0.7751],
        [0.7232],
        ...,
        [0.0260],
        [0.0078],
        [0.0807]])
Epoch 273 train loss: 0.3533, eval loss 0.6939234733581543
tensor([[0.0867],
        [0.7754],
        [0.7236],
        ...,
        [0.0260],
        [0.0077],
        [0.0807]])
Epoch 274 train loss: 0.3711, eval loss 0.6938757300376892
tensor([[0.0865],
        [0.7758],
        [0.7242],
        ...,
        [0.0259],
        [0.0077],
        [0.0808]])
Epoch 275 train loss: 0.3743, eval loss 0.693890392780304
tensor([[0.0860],
        [0.7761],
        [0.7243],
        ...,
        [0.0258],
        [0.0077],
        [0.0805]])
Epoch 276 train loss: 0.3534, eval loss 0.6937983632087708
tensor([[0.0860],
        [0.7764],
        [0.7246],
        ...,
        [0.0259],
        [0.0077],
```

```
        [0.0809]])
Epoch 277 train loss: 0.3402, eval loss 0.6938261985778809
tensor([[0.0857],
        [0.7764],
        [0.7244],
        ...,
        [0.0259],
        [0.0077],
        [0.0809]])
Epoch 278 train loss: 0.3858, eval loss 0.6937993168830872
tensor([[0.0856],
        [0.7771],
        [0.7251],
        ...,
        [0.0258],
        [0.0076],
        [0.0810]])
Epoch 279 train loss: 0.3764, eval loss 0.6938161849975586
tensor([[0.0849],
        [0.7773],
        [0.7251],
        ...,
        [0.0257],
        [0.0076],
        [0.0806]])
Epoch 280 train loss: 0.3631, eval loss 0.6937028765678406
tensor([[0.0844],
        [0.7775],
        [0.7252],
        ...,
        [0.0257],
        [0.0076],
        [0.0805]])
Epoch 281 train loss: 0.3679, eval loss 0.6936661005020142
tensor([[0.0837],
        [0.7778],
        [0.7252],
        ...,
        [0.0255],
        [0.0075],
        [0.0801]])
Epoch 282 train loss: 0.3623, eval loss 0.6935320496559143
tensor([[0.0832],
        [0.7780],
        [0.7253],
        ...,
        [0.0254],
        [0.0075],
        [0.0799]])
Epoch 283 train loss: 0.3489, eval loss 0.6934666633605957
tensor([[0.0832],
        [0.7783],
        [0.7255],
        ...,
        [0.0254],
        [0.0074],
        [0.0800]])
Epoch 284 train loss: 0.3732, eval loss 0.6934781074523926
tensor([[0.0830],
        [0.7788],
        [0.7261],
        ...,
        [0.0253],
        [0.0074],
        [0.0801]])
Epoch 285 train loss: 0.3840, eval loss 0.6934815645217896
```

```
tensor([[0.0826],
        [0.7785],
        [0.7258],
        ...,
        [0.0253],
        [0.0074],
        [0.0802]])
Epoch 286 train loss: 0.3880, eval loss 0.6934249997138977
tensor([[0.0824],
        [0.7786],
        [0.7258],
        ...,
        [0.0252],
        [0.0074],
        [0.0802]])
Epoch 287 train loss: 0.3586, eval loss 0.6934030652046204
tensor([[0.0821],
        [0.7792],
        [0.7262],
        ...,
        [0.0252],
        [0.0074],
        [0.0803]])
Epoch 288 train loss: 0.3707, eval loss 0.6934023499488831
tensor([[0.0817],
        [0.7792],
        [0.7261],
        ...,
        [0.0251],
        [0.0073],
        [0.0802]])
Epoch 289 train loss: 0.3754, eval loss 0.6933225393295288
tensor([[0.0817],
        [0.7796],
        [0.7266],
        ...,
        [0.0251],
        [0.0073],
        [0.0804]])
Epoch 290 train loss: 0.3613, eval loss 0.6933565139770508
tensor([[0.0816],
        [0.7800],
        [0.7270],
        ...,
        [0.0252],
        [0.0073],
        [0.0806]])
Epoch 291 train loss: 0.3342, eval loss 0.6933692097663879
tensor([[0.0816],
        [0.7802],
        [0.7272],
        ...,
        [0.0252],
        [0.0073],
        [0.0808]])
Epoch 292 train loss: 0.3524, eval loss 0.6933732032775879
tensor([[0.0812],
        [0.7803],
        [0.7273],
        ...,
        [0.0252],
        [0.0073],
        [0.0808]])
```

In [39]:
```
test_metrics = evaluate_model(best_model, X_test, y_test, loss_fn, best_threshold)
```

```
print(f"AUROC: {100 * test_metrics['AUROC']:.2f}%")
print(f"F1: {100 * test_metrics['F1-score']:.2f}%")
print(f"Precision: {100 * test_metrics['precision']:.2f}%")
print(f"Recall: {100 * test_metrics['recall']:.2f}%")
```

```
tensor([[0.4111],
        [0.0155],
        [0.0967],
        ...,
        [0.0044],
        [0.4716],
        [0.8707]])
AUROC: 90.08%
F1: 68.28%
Precision: 64.06%
Recall: 73.09%
```

Wyniki wyglądają już dużo lepiej.

Na koniec laboratorium dołożymy do naszego modelu jeszcze 3 powrzechnie używane techniki, które są bardzo proste, a pozwalają często ulepszyć wynik modelu.

Pierwszą z nich są **warstwy normalizacji (normalization layers)**. Powstały one początkowo z założeniem, że przez przekształcenia przestrzeni dokonywane przez sieć zmienia się rozkład prawdopodobieństw pomiędzy warstwami, czyli tzw. *internal covariate shift*. Później okazało się, że zastosowanie takiej normalizacji wygładza powierzchnię funkcji kosztu, co ułatwia i przyspiesza optymalizację. Najpowszechniej używaną normalizacją jest **batch normalization (batch norm)**.

Drugim ulepszeniem jest dodanie **wag klas (class weights)**. Mamy do czynienia z problemem klasyfikacji niezbalansowanej, więc klasa mniejszościowa, ważniejsza dla nas, powinna dostać większą wagę. Implementuje się to trywialnie prosto - po prostu mnożymy wartość funkcji kosztu dla danego przykładu przez wagę dla prawdziwej klasy tego przykładu. Praktycznie każdy klasyfikator operujący na jakiejś ważonej funkcji może działać w ten sposób, nie tylko sieci neuronowe.

Ostatnim ulepszeniem jest zamiana SGD na optymalizator Adam, a konkretnie na optymalizator `AdamW`. Jest to przykład **optymalizatora adaptacyjnego (adaptive optimizer)**, który potrafi zaadaptować stałą uczącą dla każdego parametru z osobna w trakcie treningu. Wykorzystuje do tego gradienty - w uproszczeniu, im większa wariancja gradientu, tym mniejsze kroki w tym kierunku robimy.

Zadanie 8 (0.5 punktu)

Zaimplementuj model `NormalizingMLP`, o takiej samej strukturze jak `RegularizedMLP`, ale dodatkowo z warstwami `BatchNorm1d` pomiędzy warstwami `Linear` oraz `ReLU`.

Za pomocą funkcji `compute_class_weight()` oblicz wagi dla poszczególnych klas. Użyj opcji `"balanced"`. Przekaż do funkcji kosztu wagę klasy pozytywnej (pamiętaj, aby zamienić ją na tensor).

Zamień używany optymalizator na `AdamW`.

Na koniec skopiuj resztę kodu do treningu z poprzedniego zadania, wytrenuj sieć i oblicz wyniki na zbiorze testowym.

In [40]:
```
class NormalizingMLP(nn.Module):
    def __init__(self, input_size: int, dropout_p: float = 0.5):
        super().__init__()
```

```
        self.mlp = nn.Sequential(
            nn.Linear(input_size, 256),
            nn.BatchNorm1d(256),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(256, 128),
            nn.BatchNorm1d(128),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(128, 1)
        )

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

    def predict_proba(self, x):
        return sigmoid(self(x))

    def predict(self, x, threshold: float = 0.5):
        y_pred_score = self.predict_proba(x)
        return (y_pred_score > threshold).to(torch.int32)
```

In [41]:
```
from sklearn.utils.class_weight import compute_class_weight

weights = compute_class_weight(
    class_weight = "balanced",
    classes = np.unique(y),
    y = y
)

learning_rate = 1e-3
dropout_p = 0.5
l2_reg = 1e-4
batch_size = 128
max_epochs = 300
early_stopping_rate = 4
```

In [42]:
```
model = NormalizingMLP(
    input_size=X_train.shape[1],
    dropout_p=dropout_p
)
optimizer = torch.optim.AdamW(
    model.parameters(),
    lr=learning_rate,
    weight_decay=l2_reg
)
loss_fn = torch.nn.BCEWithLogitsLoss(pos_weight=torch.from_numpy(weights)[1])

train_dataset = MyDataset(X_train, y_train)
train_dataloader = DataLoader(train_dataset, batch_size=batch_size)

steps_without_improvement = 0

best_val_loss = np.inf
best_model = None
best_threshold = None

for epoch_num in range(max_epochs):
    model.train()

    # note that we are using DataLoader to get batches
    for X_batch, y_batch in train_dataloader:
        y_batch_pred = model(X_batch)
        loss = loss_fn(y_batch_pred, y_batch)
        loss.backward();
```

```
        optimizer.step()
        optimizer.zero_grad()

    # model evaluation, early stopping
    model.eval()
    valid_metrics = evaluate_model(model, X_valid, y_valid, loss_fn)
    if valid_metrics["loss"] < best_val_loss:
        best_val_loss = valid_metrics["loss"]
        steps_without_improvement = 0
        best_model = deepcopy(model)
        best_threshold = valid_metrics["optimal_threshold"]
    else:
        steps_without_improvement += 1
        if steps_without_improvement == early_stopping_patience: break

    print(f"Epoch {epoch_num} train loss: {loss.item():.4f}, eval loss {valid_metrics['l
```

```
tensor([[0.1321],
        [0.8681],
        [0.8495],
        ...,
        [0.0871],
        [0.0059],
        [0.2024]])
Epoch 0 train loss: 0.6065, eval loss 0.8210509419441223
tensor([[0.0893],
        [0.8803],
        [0.8407],
        ...,
        [0.0584],
        [0.0053],
        [0.2178]])
Epoch 1 train loss: 0.5615, eval loss 0.8184170126914978
tensor([[0.0623],
        [0.8707],
        [0.8313],
        ...,
        [0.0789],
        [0.0028],
        [0.2261]])
Epoch 2 train loss: 0.5796, eval loss 0.8153061866760254
tensor([[0.0624],
        [0.8801],
        [0.8223],
        ...,
        [0.0467],
        [0.0029],
        [0.1883]])
Epoch 3 train loss: 0.5104, eval loss 0.8132386207580566
tensor([[0.0725],
        [0.8858],
        [0.8205],
        ...,
        [0.0303],
        [0.0020],
        [0.1927]])
Epoch 4 train loss: 0.5415, eval loss 0.8125532269477844
tensor([[0.0526],
        [0.8840],
        [0.8230],
        ...,
        [0.0362],
        [0.0017],
        [0.2309]])
Epoch 5 train loss: 0.5356, eval loss 0.812799334526062
```

```
tensor([[0.0391],
        [0.8843],
        [0.8313],
        ...,
        [0.0246],
        [0.0012],
        [0.1901]])
Epoch 6 train loss: 0.5747, eval loss 0.8107834458351135
tensor([[0.0366],
        [0.8873],
        [0.8246],
        ...,
        [0.0281],
        [0.0021],
        [0.2230]])
Epoch 7 train loss: 0.5564, eval loss 0.8116219639778137
tensor([[0.0410],
        [0.8843],
        [0.8110],
        ...,
        [0.0356],
        [0.0021],
        [0.2066]])
Epoch 8 train loss: 0.5029, eval loss 0.8105090260505676
tensor([[3.4218e-02],
        [8.9003e-01],
        [8.3122e-01],
        ...,
        [2.4498e-02],
        [8.5456e-04],
        [1.9794e-01]])
Epoch 9 train loss: 0.5496, eval loss 0.8108126521110535
tensor([[0.0436],
        [0.8745],
        [0.8171],
        ...,
        [0.0219],
        [0.0013],
        [0.2420]])
Epoch 10 train loss: 0.4780, eval loss 0.8115437030792236
tensor([[0.0266],
        [0.8941],
        [0.8122],
        ...,
        [0.0137],
        [0.0010],
        [0.2061]])
Epoch 11 train loss: 0.5903, eval loss 0.8090687394142151
tensor([[3.3586e-02],
        [8.8962e-01],
        [8.1483e-01],
        ...,
        [2.2318e-02],
        [7.4556e-04],
        [2.3403e-01]])
Epoch 12 train loss: 0.5941, eval loss 0.8097663521766663
tensor([[0.0449],
        [0.9006],
        [0.8067],
        ...,
        [0.0196],
        [0.0012],
        [0.2438]])
Epoch 13 train loss: 0.4679, eval loss 0.8101270794868469
tensor([[4.0408e-02],
        [8.9688e-01],
```

```
        [8.2190e-01],
        ...,
        [1.7118e-02],
        [7.2984e-04],
        [2.3558e-01]])
Epoch 14 train loss: 0.5078, eval loss 0.8099202513694763
tensor([[3.6673e-02],
        [9.0071e-01],
        [8.2511e-01],
        ...,
        [2.0775e-02],
        [8.8225e-04],
        [2.2009e-01]])
```

In [43]:
```
test_metrics = evaluate_model(best_model, X_test, y_test, loss_fn, best_threshold)

print(f"AUROC: {100 * test_metrics['AUROC']:.2f}%")
print(f"F1: {100 * test_metrics['F1-score']:.2f}%")
print(f"Precision: {100 * test_metrics['precision']:.2f}%")
print(f"Recall: {100 * test_metrics['recall']:.2f}%")
```

```
tensor([[0.5707],
        [0.0027],
        [0.1831],
        ...,
        [0.0055],
        [0.7412],
        [0.9679]])
AUROC: 90.76%
F1: 69.35%
Precision: 65.51%
Recall: 73.66%
```

Akceleracja sprzętowa (dla zainteresowanych)

Jak wcześniej wspominaliśmy, użycie akceleracji sprzętowej, czyli po prostu GPU do obliczeń, jest bardzo efektywne w przypadku sieci neuronowych. Karty graficzne bardzo efektywnie mnożą macierze, a sieci neuronowe to, jak można było się przekonać, dużo mnożenia macierzy.

W PyTorchu jest to dosyć łatwe, ale trzeba robić to explicite. Służy do tego metoda `.to()`, która przenosi tensory między CPU i GPU. Poniżej przykład, jak to się robi (oczywiście trzeba mieć skonfigurowane GPU, żeby działało):

In [51]:
```
import time


class CudaMLP(nn.Module):
    def __init__(self, input_size: int, dropout_p: float = 0.5):
        super().__init__()

        self.mlp = nn.Sequential(
            nn.Linear(input_size, 512),
            nn.BatchNorm1d(512),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(512, 256),
            nn.BatchNorm1d(256),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(256, 256),
            nn.BatchNorm1d(256),
            nn.ReLU(),
```

```
            nn.Dropout(dropout_p),
            nn.Linear(256, 128),
            nn.BatchNorm1d(128),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(128, 1),
        )

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

    def predict_proba(self, x):
        return sigmoid(self(x))

    def predict(self, x, threshold: float = 0.5):
        y_pred_score = self.predict_proba(x)
        return (y_pred_score > threshold).to(torch.int32)


model = CudaMLP(X_train.shape[1]).to('cuda')

optimizer = torch.optim.AdamW(model.parameters(), lr=learning_rate, weight_decay=1e-4)

# note that we are using loss function with sigmoid built in
loss_fn = torch.nn.BCEWithLogitsLoss(pos_weight=torch.from_numpy(weights)[1].to('cuda'))

step_counter = 0
time_from_eval = time.time()
for epoch_id in range(30):
    for batch_x, batch_y in train_dataloader:
        batch_x = batch_x.to('cuda')
        batch_y = batch_y.to('cuda')

        loss = loss_fn(model(batch_x), batch_y)
        loss.backward()

        optimizer.step()
        optimizer.zero_grad()

        if step_counter % evaluation_steps == 0:
            print(f"Epoch {epoch_id} train loss: {loss.item():.4f}, time: {time.time() -
            time_from_eval = time.time()

        step_counter += 1

test_res = evaluate_model(model.to('cpu'), X_test, y_test, loss_fn.to('cpu'), threshold=

print(f"AUROC: {100 * test_res['AUROC']:.2f}%")
print(f"F1: {100 * test_res['F1-score']:.2f}%")
print(test_res)
```

```
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
Cell In[51], line 39
     35         y_pred_score = self.predict_proba(x)
     36         return (y_pred_score > threshold).to(torch.int32)
---> 39 model = CudaMLP(X_train.shape[1]).to('cuda')
     41 optimizer = torch.optim.AdamW(model.parameters(), lr=learning_rate, weight_decay
=1e-4)
     43 # note that we are using loss function with sigmoid built in

File G:\anaconda3\Lib\site-packages\torch\nn\modules\module.py:1160, in Module.to(self,
 *args, **kwargs)
   1156         return t.to(device, dtype if t.is_floating_point() or t.is_complex() els
e None,
```

```
   1157                     non_blocking, memory_format=convert_to_format)
   1158     return t.to(device, dtype if t.is_floating_point() or t.is_complex() else No
ne, non_blocking)
-> 1160 return self._apply(convert)

File G:\anaconda3\Lib\site-packages\torch\nn\modules\module.py:810, in Module._apply(sel
f, fn, recurse)
    808 if recurse:
    809     for module in self.children():
--> 810         module._apply(fn)
    812 def compute_should_use_set_data(tensor, tensor_applied):
    813     if torch._has_compatible_shallow_copy_type(tensor, tensor_applied):
    814         # If the new tensor has compatible tensor type as the existing tensor,
    815         # the current behavior is to change the tensor in-place using `.data =`,
   (...)
    820         # global flag to let the user control whether they want the future
    821         # behavior of overwriting the existing tensor or not.

File G:\anaconda3\Lib\site-packages\torch\nn\modules\module.py:810, in Module._apply(sel
f, fn, recurse)
    808 if recurse:
    809     for module in self.children():
--> 810         module._apply(fn)
    812 def compute_should_use_set_data(tensor, tensor_applied):
    813     if torch._has_compatible_shallow_copy_type(tensor, tensor_applied):
    814         # If the new tensor has compatible tensor type as the existing tensor,
    815         # the current behavior is to change the tensor in-place using `.data =`,
   (...)
    820         # global flag to let the user control whether they want the future
    821         # behavior of overwriting the existing tensor or not.

File G:\anaconda3\Lib\site-packages\torch\nn\modules\module.py:833, in Module._apply(sel
f, fn, recurse)
    829 # Tensors stored in modules are graph leaves, and we don't want to
    830 # track autograd history of `param_applied`, so we have to use
    831 # `with torch.no_grad():`
    832 with torch.no_grad():
--> 833     param_applied = fn(param)
    834 should_use_set_data = compute_should_use_set_data(param, param_applied)
    835 if should_use_set_data:

File G:\anaconda3\Lib\site-packages\torch\nn\modules\module.py:1158, in Module.to.<local
s>.convert(t)
   1155 if convert_to_format is not None and t.dim() in (4, 5):
   1156     return t.to(device, dtype if t.is_floating_point() or t.is_complex() else No
ne,
   1157                 non_blocking, memory_format=convert_to_format)
-> 1158 return t.to(device, dtype if t.is_floating_point() or t.is_complex() else None,
 non_blocking)

File G:\anaconda3\Lib\site-packages\torch\cuda\__init__.py:289, in _lazy_init()
    284     raise RuntimeError(
    285         "Cannot re-initialize CUDA in forked subprocess. To use CUDA with "
    286         "multiprocessing, you must use the 'spawn' start method"
    287     )
    288 if not hasattr(torch._C, "_cuda_getDeviceCount"):
--> 289     raise AssertionError("Torch not compiled with CUDA enabled")
    290 if _cudart is None:
    291     raise AssertionError(
    292         "libcudart functions unavailable. It looks like you have a broken build?
"
    293     )

AssertionError: Torch not compiled with CUDA enabled
```

Co prawda ten model nie będzie tak dobry jak ten z laboratorium, ale zwróć uwagę, o ile jest większy, a przy tym szybszy.

Dla zainteresowanych polecamy tę serie artykułów

Zadanie dla chętnych

Jak widzieliśmy, sieci neuronowe mają bardzo dużo hiperparametrów. Przeszukiwanie ich grid search'em jest więc niewykonalne, a chociaż random search by działał, to potrzebowałby wielu iteracji, co też jest kosztowne obliczeniowo.

Zaimplementuj inteligentne przeszukiwanie przestrzeni hiperparametrów za pomocą biblioteki Optuna. Implementuje ona między innymi algorytm Tree Parzen Estimator (TPE), należący do grupy algorytmów typu Bayesian search. Typowo osiągają one bardzo dobre wyniki, a właściwie zawsze lepsze od przeszukiwania losowego. Do tego wystarcza im często niewielka liczba kroków.

Zaimplementuj 3-warstwową sieć MLP, gdzie pierwsza warstwa ma rozmiar ukryty N, a druga N // 2. Ucz ją optymalizatorem Adam przez maksymalnie 300 epok z cierpliwością 10.

Przeszukaj wybrane zakresy dla hiperparametrów:

- rozmiar warstw ukrytych (N)
- stała ucząca
- batch size
- siła regularyzacji L2
- prawdopodobieństwo dropoutu

Wykorzystaj przynajmniej 30 iteracji. Następnie przełącz algorytm na losowy (Optuna także jego implementuje), wykonaj 30 iteracji i porównaj jakość wyników.

Przydatne materiały:

- Optuna code examples - PyTorch
- Auto-Tuning Hyperparameters with Optuna and PyTorch
- Hyperparameter Tuning of Neural Networks with Optuna and PyTorch
- Using Optuna to Optimize PyTorch Hyperparameters

In [ ]: