

09_additional_pytorch_model

May 26, 2024

1 Pytorch Model (additional, not necessary for thesis)

1.1 Content

1. Import Files
2. ????

```
[1]: # imports
import numpy as np
import pandas as pd
import time
import torch
from torch.utils.data import Dataset
import matplotlib.pyplot as plt
import torch.nn as nn
import torch.optim as optim
from sklearn.model_selection import train_test_split
from torch import nn
from sklearn.preprocessing import StandardScaler
import torch.onnx
```

import

```
[2]: # import files
probabilities = pd.read_csv('files/switzerland_probabilities.csv').
    ↳drop(columns=["Name_de", "Berufshauptgruppe"])
df = pd.read_csv('files/switzerland_occupations.csv')
not_automatable = [item[0] for item in pd.read_csv("files/not_automatable.csv").
    ↳values.tolist()]
```

```
[3]: # remove cols
keep_cols = ['isco08', 'Name_de', 'probability', "Berufshauptgruppe"]
keep_cols.extend(not_automatable)

df = df.merge(probabilities, on='isco08')
df = df[keep_cols]
df.head(5)
```

```
[3]: isco08                                     Name_de probability \
0    2655                                     Schauspieler      0.4618
1    2612                                     Richter           0.2638
2    3115  Maschinenbautechniker, Techniker im Bereich Sy...  0.3666
3    2120                                     Mathematiker, Aktuare und Statistiker  0.4381
4    1222  Führungskräfte in Werbung und Öffentlichkeitsa...  0.2940
```


	Berufshauptgruppe	a12	s4	s27	s15	s8	s31	s26	s24
0	2	0.47	0.69	0.00	0.38	0.47	0.0	0.00	0.06
1	2	0.78	0.72	0.00	0.56	0.72	0.0	0.00	0.19
2	3	0.72	0.47	0.28	0.38	0.63	0.5	0.41	0.28
3	2	0.75	0.50	0.00	0.44	0.75	0.0	0.00	0.22
4	1	0.72	0.75	0.00	0.53	0.63	0.0	0.03	0.19

```
[4]: # calculate the amount of inputs
a_cols = df.columns[df.columns.str.startswith('a')].shape[0]
s_cols = df.columns[df.columns.str.startswith('s')].shape[0]

total_inputs = a_cols + s_cols
print(f"Total inputs: {total_inputs}")
```

Total inputs: 8

```
[5]: ## Global settings
# Train/validation/test split: 80% train, 10% validation, 10% testsplit
split_size = (0.8, 0.1, 0.1)

# Dropout rate: 5%
dropout_rate = 0.05

# Batch import size: 10 jobs in a minibatch
batch_size = 20

# Number of epochs
num_epochs = 100

# Define L2 regularization strength (penalty parameter)
l2_reg = 0.01

# Early stopping parameters
early_stopping = False

# Initial best loss set to infinity
best_val_loss = float('inf')

# Number of epochs with no improvement to wait before stopping
patience = 10
```

```
# Counter for epochs without improvement
no_improve_epochs = 0
```

```
[6]: class TabularDataset(Dataset):
    def __init__(self, df, target_col, ignore_cols):
        self.data = df.drop(columns=ignore_cols).copy()
        self.target_col = target_col

        # Standardize the features
        scaler = StandardScaler()
        self.data.loc[:, self.data.columns != target_col] = scaler.
        fit_transform(self.data.loc[:, self.data.columns != target_col])

        # Create targets tensor
        self.targets = torch.tensor(self.data.loc[:, target_col].values,
        dtype=torch.float)

        # Convert data to tensor
        self.data = torch.tensor(self.data.drop(columns=target_col).values,
        dtype=torch.float)

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

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

# Specify columns to ignore
ignore_cols = ["Name_de", "isco08", "Berufshauptgruppe"]

# Load data
dataset = TabularDataset(df, 'probability', ignore_cols)

# Visualize one sample
print(dataset[0])
```

```
(tensor([-1.1356,  1.3300, -0.4782, -0.1548, -0.4538, -1.4646, -1.0840,
-1.1903]), tensor(0.4618))
```

```
[7]: # Prepare the data
feature_cols = [col for col in df.columns if col.startswith(('s', 'a'))]
X = df[feature_cols]
y = df['probability']

# Define split sizes
split_size = (0.8, 0.1, 0.1)
train_size = int(split_size[0] * len(dataset))
```

```

val_size = int(split_size[1] * len(dataset))
test_size = len(dataset) - train_size - val_size

# Randomly split dataset into train, validation and test sets
train_dataset, val_dataset, test_dataset = torch.utils.data.
    random_split(dataset, [train_size, val_size, test_size])

print(f'Training: {len(train_dataset)}, validation: {len(val_dataset)}, test:
    {len(test_dataset)}')

# Define batch size
batch_size = 10

# Create data loaders
train_loader = torch.utils.data.DataLoader(dataset=train_dataset,
                                             batch_size=batch_size,
                                             shuffle=True)

val_loader = torch.utils.data.DataLoader(dataset=val_dataset,
                                          batch_size=batch_size,
                                          shuffle=False)

test_loader = torch.utils.data.DataLoader(dataset=test_dataset,
                                           batch_size=batch_size,
                                           shuffle=False)

```

Training: 276, validation: 34, test: 36

```

[8]: class pytorch_probabilities(nn.Module):
    def __init__(self):
        super(pytorch_probabilities, self).__init__()
        # The architecture of the neural network is based on the following
        ↪ paper:
            # Yar Muhammad, Mohammad Dahman Alshehri, Wael Mohammed Alenazy, Truong
            ↪ Vinh Hoang, Ryan Alturki
            # "Identification of Pneumonia Disease Applying an Intelligent
            ↪ Computational Framework Based on Deep Learning and Machine Learning
            ↪ Techniques", Mobile Information Systems, vol. 2021, Article ID 9989237, 20
            ↪ pages, 2021. https://doi.org/10.1155/2021/9989237

            # The architecture consists of a convolutional layer, a pooling layer,
            ↪ and a fully connected layer.
            # The convolutional layer consists of two convolutional layers.
            # That layer is then ran through a max-pooling layer.
            # Finally, classification is done by a fully connected layer consisting
            ↪ of two linear layers.

```

```

    # Dropout is performed both in the convolutional layer and the fully
    ↪connected layer in order to increase the generalization of the model.

    # Convolutional layer: 2 layers with 32 filters each
    self.fc1 = nn.Linear(8, 64) # 87 input features, 50 output features
    self.fc2 = nn.Linear(64, 2) # 250 input features, 125 output features
    self.softmax = nn.Softmax(dim=1)

    # Activation functions
    self.relu = nn.ReLU()
    self.sigmoid = nn.Sigmoid()
    self.batch_norm1 = nn.BatchNorm1d(64)

    # Validation criterion
    self.criterion = nn.CrossEntropyLoss()

    # Forward pass through the network
    def forward(self, x):
        # Pass input through first fully connected layer
        x = self.fc1(x)

        # Apply batch normalization
        x = self.batch_norm1(x)

        # Pass through second fully connected layer
        x = self.fc2(x)

        # Apply softmax activation function
        x = self.softmax(x)

        return x

    # Validate model on dataset
    def validate(self, loader):
        # Set to evaluation mode
        self.eval()

        with torch.no_grad():
            # Track loss values and true/predicted values for MAPE
            losses = []
            all_targets = []
            all_predicted = []

            for data, targets in loader:
                # Generate predictions
                outputs = self(data)

```

```

        # Predict validation data
        _, predicted = torch.max(outputs.detach(), dim=1)

        # Store targets and predictions for MAPE calculation
        all_targets.append(targets)
        all_predicted.append(predicted)

        # Calculate and track loss for predictions outside the range
        out_of_range = (predicted - targets).abs() > 0.05
        if out_of_range.any():
            losses.append(self.criterion(outputs[out_of_range],
→targets[out_of_range].long()).item())

        # Calculate loss and MAPE
        loss = np.mean(losses) if losses else 0
        all_targets = torch.cat(all_targets)
        all_predicted = torch.cat(all_predicted)
        mape = self.mape(all_targets, all_predicted)

        return loss, mape

    def mape(self, y_true, y_pred):
        return torch.mean(torch.abs((y_true - y_pred) / y_true)) * 100

```

```

[9]: # Instantiate the model
model = pytorch_probabilities()

# Define loss function and optimizer
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(), lr=0.005, momentum=0.9)

# Keep track of losses per epoch
train_losses = []
validation_losses = []

# Keep track of epoch start/end times
start_time, end_time = None, None

# Train model
for epoch in range(num_epochs):
    # Start timer
    start_time = time.perf_counter()

    # Set model to training mode
    model.train()

    # Track loss values (inside epoch)

```

```

losses = []

# Train model on training data
for data, targets in train_loader:
    # Reset gradients
    optimizer.zero_grad()

    # Generate predictions
    outputs = model(data)

    # Calculate and track loss
    loss = criterion(outputs, targets.long())
    losses.append(loss.item())

    # Regularization using L2 norm
    l2_loss = 0
    for param in model.parameters():
        l2_loss += torch.norm(param, p=2)**2

    loss += l2_reg * l2_loss

    losses.append(loss.item())

    # Back-propagate loss
    loss.backward()

    # Update weights
    optimizer.step()

# End timer
end_time = time.perf_counter()

train_losses.append(np.mean(losses))

# Evaluate model on validation data
val_loss, val_accuracy = model.validate(val_loader)

validation_losses.append(val_loss)

# Print summary each epoch
print(f'Epoch: {epoch + 1}/{num_epochs} (train_loss: {np.mean(losses):.2f},  

↪ val_loss: {val_loss:.2f}, accuracy: {val_accuracy:.2f}%, time: {end_time -  

↪ start_time:.2f}s)')

# Stop early if validation loss does not improve anymore
if early_stopping:
    current_val_loss = val_loss

```

```

    if current_val_loss < best_val_loss:
        best_val_loss = current_val_loss

        # Save best model
        torch.save(model.state_dict(), f'pytorch_model/
↪calculate_probability.pth')

        # Reset counter
        no_improve_epochs = 0
    else:
        # Increase counter
        no_improve_epochs += 1

    # Stop training if patience is exceeded
    if no_improve_epochs >= patience:
        print("Early stopping due to no improvement.")
        break
    else:
        # Save model weights (each epoch)
        torch.save(model.state_dict(), f'pytorch_model/calculate_probability.
↪pth')

```

```

Epoch: 1/100 (train_loss: 1.12, val_loss: 0.60, accuracy: 100.00%, time: 0.07s))
Epoch: 2/100 (train_loss: 0.94, val_loss: 0.47, accuracy: 100.00%, time: 0.02s))
Epoch: 3/100 (train_loss: 0.83, val_loss: 0.41, accuracy: 100.00%, time: 0.02s))
Epoch: 4/100 (train_loss: 0.77, val_loss: 0.38, accuracy: 100.00%, time: 0.02s))
Epoch: 5/100 (train_loss: 0.73, val_loss: 0.36, accuracy: 100.00%, time: 0.02s))
Epoch: 6/100 (train_loss: 0.70, val_loss: 0.35, accuracy: 100.00%, time: 0.01s))
Epoch: 7/100 (train_loss: 0.67, val_loss: 0.35, accuracy: 100.00%, time: 0.02s))
Epoch: 8/100 (train_loss: 0.65, val_loss: 0.34, accuracy: 100.00%, time: 0.02s))
Epoch: 9/100 (train_loss: 0.63, val_loss: 0.34, accuracy: 100.00%, time: 0.02s))
Epoch: 10/100 (train_loss: 0.62, val_loss: 0.34, accuracy: 100.00%, time:
0.02s))
Epoch: 11/100 (train_loss: 0.60, val_loss: 0.34, accuracy: 100.00%, time:
0.01s))
Epoch: 12/100 (train_loss: 0.59, val_loss: 0.34, accuracy: 100.00%, time:
0.01s))
Epoch: 13/100 (train_loss: 0.57, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 14/100 (train_loss: 0.56, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 15/100 (train_loss: 0.55, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 16/100 (train_loss: 0.54, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 17/100 (train_loss: 0.53, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))

```


Epoch: 18/100 (train_loss: 0.52, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 19/100 (train_loss: 0.51, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 20/100 (train_loss: 0.50, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 21/100 (train_loss: 0.49, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 22/100 (train_loss: 0.48, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 23/100 (train_loss: 0.47, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 24/100 (train_loss: 0.47, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 25/100 (train_loss: 0.46, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 26/100 (train_loss: 0.45, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 27/100 (train_loss: 0.45, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 28/100 (train_loss: 0.44, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 29/100 (train_loss: 0.44, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 30/100 (train_loss: 0.43, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 31/100 (train_loss: 0.43, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 32/100 (train_loss: 0.42, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 33/100 (train_loss: 0.42, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 34/100 (train_loss: 0.42, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 35/100 (train_loss: 0.41, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 36/100 (train_loss: 0.41, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 37/100 (train_loss: 0.41, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 38/100 (train_loss: 0.40, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 39/100 (train_loss: 0.40, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 40/100 (train_loss: 0.40, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 41/100 (train_loss: 0.39, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))

Epoch: 42/100 (train_loss: 0.39, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 43/100 (train_loss: 0.39, val_loss: 0.33, accuracy: 100.00%, time: 0.01s))
Epoch: 44/100 (train_loss: 0.39, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 45/100 (train_loss: 0.38, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 46/100 (train_loss: 0.38, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 47/100 (train_loss: 0.38, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 48/100 (train_loss: 0.38, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 49/100 (train_loss: 0.38, val_loss: 0.33, accuracy: 100.00%, time: 0.01s))
Epoch: 50/100 (train_loss: 0.38, val_loss: 0.33, accuracy: 100.00%, time: 0.01s))
Epoch: 51/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 52/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 53/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.01s))
Epoch: 54/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 55/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 56/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 57/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 58/100 (train_loss: 0.37, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 59/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 60/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 61/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 62/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.01s))
Epoch: 63/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 64/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 65/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))

Epoch: 66/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 67/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 68/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 69/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.01s))
Epoch: 70/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 71/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 72/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 73/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 74/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 75/100 (train_loss: 0.36, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 76/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 77/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 78/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 79/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 80/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 81/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 82/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 83/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 84/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 85/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 86/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 87/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 88/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))
Epoch: 89/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time: 0.02s))

```

Epoch: 90/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 91/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 92/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 93/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.01s))
Epoch: 94/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 95/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 96/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.01s))
Epoch: 97/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 98/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))
Epoch: 99/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.01s))
Epoch: 100/100 (train_loss: 0.35, val_loss: 0.33, accuracy: 100.00%, time:
0.02s))

```

```

[10]: ## Plot the training and validation losses (Dries)
plt.figure(figsize=(8, 6))

plt.plot(range(len(train_losses)), train_losses, label="Training loss")
plt.plot(range(len(validation_losses)), validation_losses, label="Validation_
↪loss")

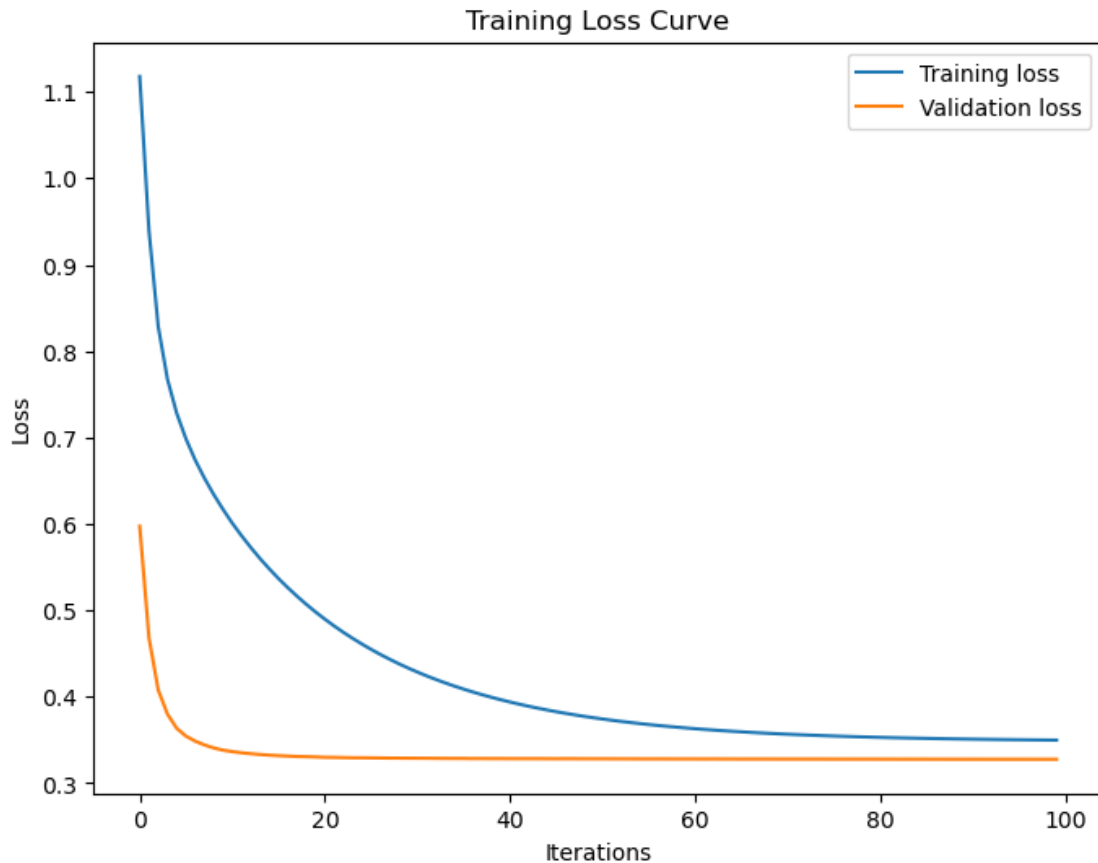
plt.xlabel('Iterations')
plt.ylabel('Loss')

plt.title('Training Loss Curve')

plt.legend()

plt.show()

```



```
[11]: # Load the saved model
model = pytorch_probabilities()
model.load_state_dict(torch.load('pytorch_model/calculate_probability.pth'))

# Set the model to evaluation mode
model.eval()

# An example input you would normally provide to your model's forward() method.
# This will be used to run a forward pass in your model, so it should be
# representative of your model's input.
# Replace it with the correct input for your model.
example_input = torch.randn(1, 8)

# Export the model
torch.onnx.export(model,                # model being run
                  example_input,        # model input (or a tuple for multiple
                  # inputs)
                  "pytorch_model/model.onnx", # where to save the model (can
                  # be a file or file-like object)
```

```

        export_params=True, # store the trained parameter weights
↪inside the model file
        opset_version=7,    # the ONNX version to export the model to
        do_constant_folding=True, # whether to execute constant
↪folding for optimization
        input_names = ['input'], # the model's input names
        output_names = ['output'], # the model's output names
        dynamic_axes={'input' : {0 : 'batch_size'}, # variable
↪length axes
                        'output' : {0 : 'batch_size'}})

# Evaluate the model on the test data
test_loss, test_accuracy = model.validate(test_loader)

# Print summary
print(f'Test loss: {test_loss:.2f}, test accuracy: {test_accuracy:.2f}%')

```

Test loss: 0.33, test accuracy: 100.00%