Sample Code for an Autoencoder for Dr. Goswami

Shakil Rafi

In this notebook we will do a test-case with synthetic data comparing classic PCA with an autoencoder to see that the autoencoder is able to take high-dimensional data and extract out the relevant features.

Dataset

Our dataset will be a synthetic dataset in which we will use tha sklearn datasets method to create eight "blobs" of data. Each blob has a centroid around which the data is clustered in a Gaussian fashion with a standard deviation of 1, i.e. each element of the cluster c, x_c is such that $x_c \sim \mathcal{N}(\text{centroid}_c, 1)$. We will have nine features to the data, to which we will add an extra feature, consisting of Gaussian noise.

Note on dataset: The author regrets to inform that due to the ongoing shutdowns of internet and communications in Bangladesh that they have had difficulty and delay obtaining actual datasets. This Jupyter notebook therefore serves as a proof of concept that autoencoders can be beneficial and viable in analyzing genomic data.

```
import subprocess
import sys

def install(package):
    subprocess.check_call([sys.executable, "-m", "pip", "install", package])

# List your packages here
packages = ["numpy", "pandas", "seaborn","keras","tensorflow","matplotlib"]

for package in packages:
    install(package)
```

```
Requirement already satisfied: numpy in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3
Requirement already satisfied: pandas in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3
Requirement already satisfied: numpy>=1.26.0 in /Users/shakilrafi/anaconda3/envs/goswami/lib/
Requirement already satisfied: python-dateutil>=2.8.2 in /Users/shakilrafi/anaconda3/envs/gos
Requirement already satisfied: pytz>=2020.1 in /Users/shakilrafi/anaconda3/envs/goswami/lib/p
Requirement already satisfied: tzdata>=2022.7 in /Users/shakilrafi/anaconda3/envs/goswami/lib
Requirement already satisfied: six>=1.5 in /Users/shakilrafi/anaconda3/envs/goswami/lib/pytho
Requirement already satisfied: seaborn in /Users/shakilrafi/anaconda3/envs/goswami/lib/python
Requirement already satisfied: numpy!=1.24.0,>=1.20 in /Users/shakilrafi/anaconda3/envs/goswa
Requirement already satisfied: pandas>=1.2 in /Users/shakilrafi/anaconda3/envs/goswami/lib/py
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in /Users/shakilrafi/anaconda3/envs/go
Requirement already satisfied: contourpy>=1.0.1 in /Users/shakilrafi/anaconda3/envs/goswami/l
Requirement already satisfied: cycler>=0.10 in /Users/shakilrafi/anaconda3/envs/goswami/lib/p
Requirement already satisfied: fonttools>=4.22.0 in /Users/shakilrafi/anaconda3/envs/goswami/
Requirement already satisfied: kiwisolver>=1.3.1 in /Users/shakilrafi/anaconda3/envs/goswami/
Requirement already satisfied: packaging>=20.0 in /Users/shakilrafi/anaconda3/envs/goswami/li
Requirement already satisfied: pillow>=8 in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyth
Requirement already satisfied: pyparsing>=2.3.1 in /Users/shakilrafi/anaconda3/envs/goswami/l
Requirement already satisfied: python-dateutil>=2.7 in /Users/shakilrafi/anaconda3/envs/goswa
Requirement already satisfied: pytz>=2020.1 in /Users/shakilrafi/anaconda3/envs/goswami/lib/p
Requirement already satisfied: tzdata>=2022.7 in /Users/shakilrafi/anaconda3/envs/goswami/lib
Requirement already satisfied: six>=1.5 in /Users/shakilrafi/anaconda3/envs/goswami/lib/pytho
Requirement already satisfied: keras in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3
Requirement already satisfied: absl-py in /Users/shakilrafi/anaconda3/envs/goswami/lib/python
Requirement already satisfied: numpy in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3
Requirement already satisfied: rich in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3.1
Requirement already satisfied: namex in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3
Requirement already satisfied: h5py in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3.1
Requirement already satisfied: optree in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3
Requirement already satisfied: ml-dtypes in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyth
Requirement already satisfied: packaging in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyth
Requirement already satisfied: typing-extensions>=4.5.0 in /Users/shakilrafi/anaconda3/envs/g
Requirement already satisfied: markdown-it-py>=2.2.0 in /Users/shakilrafi/anaconda3/envs/gosw
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /Users/shakilrafi/anaconda3/envs/go
Requirement already satisfied: mdurl~=0.1 in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyt
Requirement already satisfied: tensorflow in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyt
Requirement already satisfied: absl-py>=1.0.0 in /Users/shakilrafi/anaconda3/envs/goswami/lib
Requirement already satisfied: astunparse>=1.6.0 in /Users/shakilrafi/anaconda3/envs/goswami/
Requirement already satisfied: flatbuffers>=24.3.25 in /Users/shakilrafi/anaconda3/envs/goswa
Requirement already satisfied: gast!=0.5.0,!=0.5.1,!=0.5.2,>=0.2.1 in /Users/shakilrafi/anaco
Requirement already satisfied: google-pasta>=0.1.1 in /Users/shakilrafi/anaconda3/envs/goswam
Requirement already satisfied: h5py>=3.10.0 in /Users/shakilrafi/anaconda3/envs/goswami/lib/p
Requirement already satisfied: libclang>=13.0.0 in /Users/shakilrafi/anaconda3/envs/goswami/l
```

```
Requirement already satisfied: ml-dtypes<0.5.0,>=0.3.1 in /Users/shakilrafi/anaconda3/envs/g
Requirement already satisfied: opt-einsum>=2.3.2 in /Users/shakilrafi/anaconda3/envs/goswami/
Requirement already satisfied: packaging in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyt
Requirement already satisfied: protobuf!=4.21.0,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.4,!=4.21.5
Requirement already satisfied: requests<3,>=2.21.0 in /Users/shakilrafi/anaconda3/envs/goswa
Requirement already satisfied: setuptools in /Users/shakilrafi/anaconda3/envs/goswami/lib/py
Requirement already satisfied: six>=1.12.0 in /Users/shakilrafi/anaconda3/envs/goswami/lib/p
Requirement already satisfied: termcolor>=1.1.0 in /Users/shakilrafi/anaconda3/envs/goswami/
Requirement already satisfied: typing-extensions>=3.6.6 in /Users/shakilrafi/anaconda3/envs/
Requirement already satisfied: wrapt>=1.11.0 in /Users/shakilrafi/anaconda3/envs/goswami/lib
Requirement already satisfied: grpcio<2.0,>=1.24.3 in /Users/shakilrafi/anaconda3/envs/goswa
Requirement already satisfied: tensorboard<2.18,>=2.17 in /Users/shakilrafi/anaconda3/envs/g
Requirement already satisfied: keras>=3.2.0 in /Users/shakilrafi/anaconda3/envs/goswami/lib/
Requirement already satisfied: numpy<2.0.0,>=1.26.0 in /Users/shakilrafi/anaconda3/envs/gosw
Requirement already satisfied: wheel<1.0,>=0.23.0 in /Users/shakilrafi/anaconda3/envs/goswam
Requirement already satisfied: rich in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3.
Requirement already satisfied: namex in /Users/shakilrafi/anaconda3/envs/goswami/lib/python3
Requirement already satisfied: optree in /Users/shakilrafi/anaconda3/envs/goswami/lib/python
Requirement already satisfied: charset-normalizer<4,>=2 in /Users/shakilrafi/anaconda3/envs/
Requirement already satisfied: idna<4,>=2.5 in /Users/shakilrafi/anaconda3/envs/goswami/lib/
Requirement already satisfied: urllib3<3,>=1.21.1 in /Users/shakilrafi/anaconda3/envs/goswam
Requirement already satisfied: certifi>=2017.4.17 in /Users/shakilrafi/anaconda3/envs/goswam
Requirement already satisfied: markdown>=2.6.8 in /Users/shakilrafi/anaconda3/envs/goswami/l
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0 in /Users/shakilrafi/an
Requirement already satisfied: werkzeug>=1.0.1 in /Users/shakilrafi/anaconda3/envs/goswami/l
Requirement already satisfied: MarkupSafe>=2.1.1 in /Users/shakilrafi/anaconda3/envs/goswami
Requirement already satisfied: markdown-it-py>=2.2.0 in /Users/shakilrafi/anaconda3/envs/gos
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /Users/shakilrafi/anaconda3/envs/g
Requirement already satisfied: mdurl~=0.1 in /Users/shakilrafi/anaconda3/envs/goswami/lib/py
Requirement already satisfied: matplotlib in /Users/shakilrafi/anaconda3/envs/goswami/lib/py
Requirement already satisfied: contourpy>=1.0.1 in /Users/shakilrafi/anaconda3/envs/goswami/
Requirement already satisfied: cycler>=0.10 in /Users/shakilrafi/anaconda3/envs/goswami/lib/
Requirement already satisfied: fonttools>=4.22.0 in /Users/shakilrafi/anaconda3/envs/goswami
Requirement already satisfied: kiwisolver>=1.3.1 in /Users/shakilrafi/anaconda3/envs/goswami
Requirement already satisfied: numpy>=1.23 in /Users/shakilrafi/anaconda3/envs/goswami/lib/p
Requirement already satisfied: packaging>=20.0 in /Users/shakilrafi/anaconda3/envs/goswami/l
Requirement already satisfied: pillow>=8 in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyt
Requirement already satisfied: pyparsing>=2.3.1 in /Users/shakilrafi/anaconda3/envs/goswami/
Requirement already satisfied: python-dateutil>=2.7 in /Users/shakilrafi/anaconda3/envs/gosw
Requirement already satisfied: six>=1.5 in /Users/shakilrafi/anaconda3/envs/goswami/lib/pyth
```

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

```
from sklearn.datasets import make_blobs
```

```
data = make_blobs(
    n_samples=30000,
    n_features=9,
    centers=8, cluster_std=1, random_state=101
)

X,y = data
```

```
np.random.seed(seed=101)
noise = np.random.normal(size=len(X))
noise = pd.Series(noise)
```

```
feat = pd.DataFrame(X)
feat = pd.concat([feat,noise],axis=1)
```

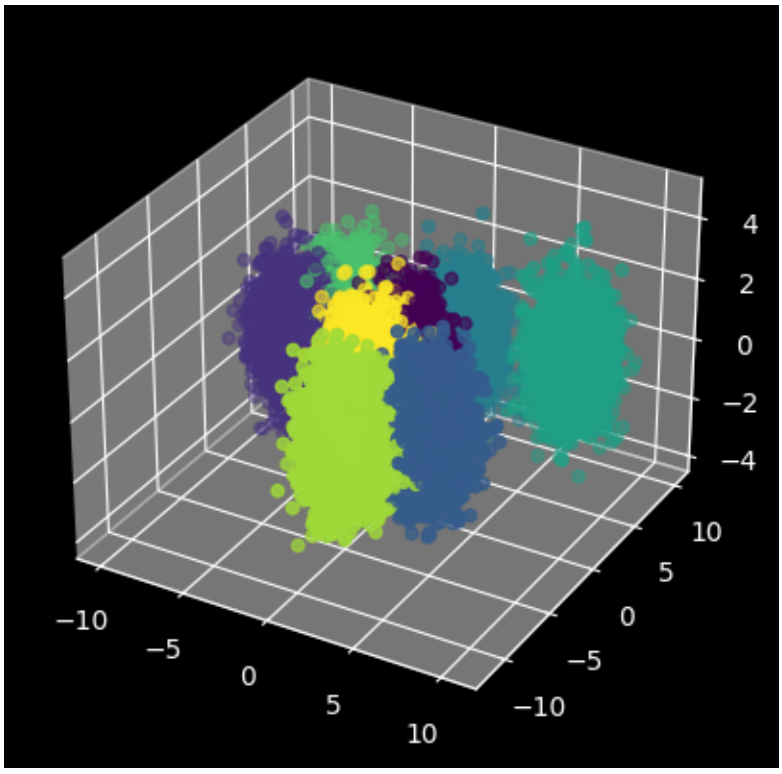
```
feat.columns = [f"X{i+1}" for i in range(len(feat.columns))]
```

```
feat
```

	X1	X2	X3	X4	X5	X6	X7	X8	X9
0	-1.134347	-8.119798	-0.969286	-8.452210	-9.965433	-5.727484	-2.309998	7.648037	9.60767
1	7.194004	7.311195	-1.390716	9.301091	-6.330847	-8.654975	4.595300	8.273353	-4.9361
2	-0.028777	2.151845	-9.137685	-6.243801	2.444628	5.302981	-4.984138	8.185551	4.36234
3	1.076275	4.892708	1.689847	8.103409	6.126228	-2.141447	-0.122088	10.483656	-1.4173
4	6.236355	7.073400	0.687253	7.988951	-7.435968	-8.901483	2.787173	7.181208	-7.0591
...	...	...	...	...	...	...	...	...	...
29995	5.732655	8.271973	0.264257	7.981432	-6.207605	-7.566080	3.201147	8.018709	-7.2761
29996	1.264017	4.182985	1.896195	7.503918	6.432221	-2.004529	-3.277927	10.103653	-4.1108
29997	1.983596	2.385081	-8.652489	-7.857445	4.598875	7.616985	-3.212411	9.231159	4.10015
29998	0.144712	-6.767785	-0.355600	-7.387527	-9.056727	-6.694205	-0.720757	9.221689	8.65236
29999	0.716710	-7.293508	-0.566136	-6.998149	-8.738948	-8.112066	1.402210	7.777800	8.98252

```
feat.to_csv("feat.csv")
```

```
fig = plt.figure()
ax = fig.add_subplot(111, projection = '3d')
ax.scatter(feat['X1'],feat['X2'],feat['X10'], c = y)
```



Preprocessing the data

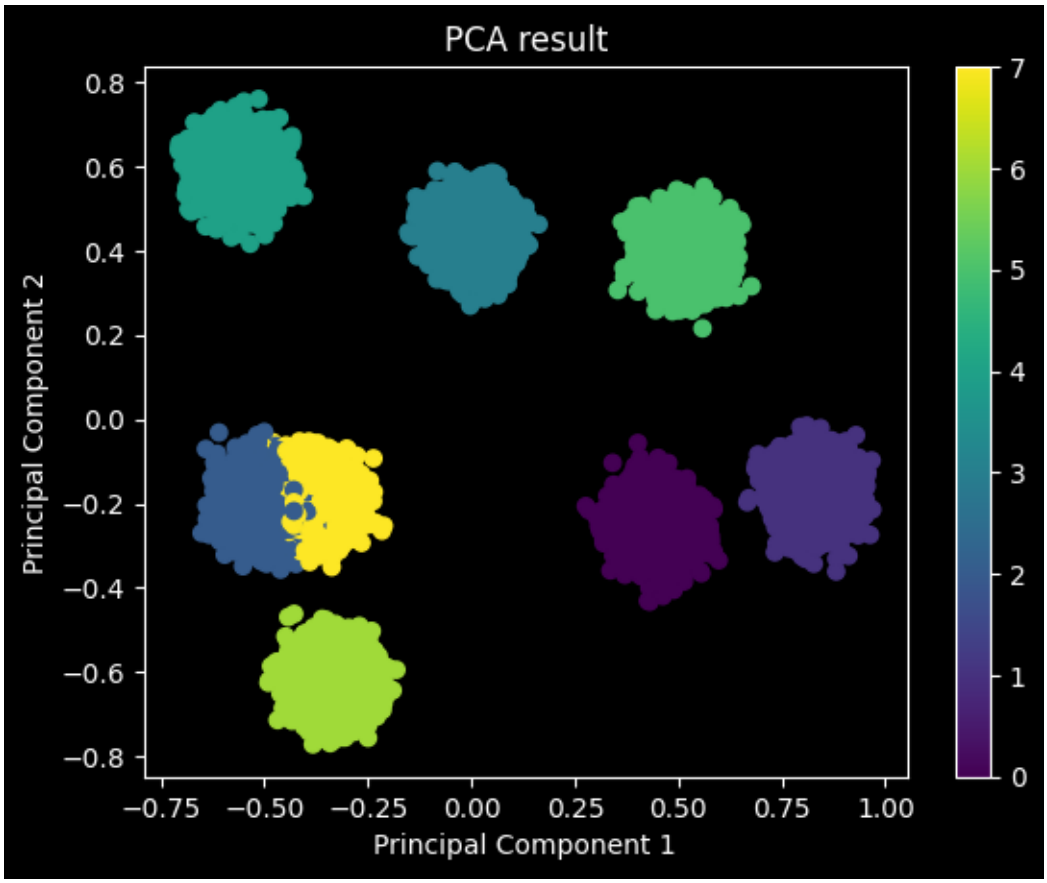
We will use the standard MinMaxScaler from sklearn to scale and preprocess the data

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(feat)
```

The PCA decomposition

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
pca_result = pca.fit_transform(scaled_data)
```

```
plt.scatter(pca_result[:, 0], pca_result[:, 1], c=y)
plt.title('PCA result')
plt.colorbar()
plt.xlabel("Principal Component 1")
plt.ylabel("Principal Component 2")
plt.show()
```



```
from sklearn.metrics import mean_squared_error

X_reconstructed = pca.inverse_transform(pca_result)
print("Reconstruction loss of PCA:",mean_squared_error(X_reconstructed, scaled_data))
```

```
Reconstruction loss of PCA: 0.015083849781626347
```

K-means done on the encoded data