Elements of Machine Learning

Assignment 4

Xingrong Zong, tnd179

March 27th, 2022

Introduction

Generative modelling in Machine Learning is formulated as the task of estimating the underlying data-generating distribution, $P_x(x)$, where X is the random variable of interest. X can represent any kind of signal ranging from scalar measurements to high dimensional data such as text, time series data or even videos.

Except in toy examples we never have access to the distribution $P_x(x)$ in analytical form or either via an efficient sampler. We can only observe this distribution from a subset of data samples. In this assignment we look at some interesting Machine Learning tasks based on modelling the data generating distribution in the context of Unsupervised learning.

Monte Carlo Estimation

Variance of Sample Mean Estimator

$$E[X] \approx \hat{\mu} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

Starting with the definition of the sample mean,

$$Var(X) = Var(\frac{x[0] + x[1] + x[2] + \cdots + x[n]}{N})$$
$$Var(X) = Var(\frac{1}{N}x[0] + \frac{1}{N}x[1] + \cdots + \frac{1}{N}x[n]$$
$$Var(X) = \frac{1}{N^2}Var(x[0]) + \frac{1}{N^2}Var(x[1]) + \cdots \frac{1}{N^2}Var(x[n])$$
$$Var(X) = \frac{1}{N^2}[\sigma^2 + \sigma^2 + \cdots + sigma^2]$$
$$Var(X) = \frac{sigma^2}{N}$$

Representation Learning Generative Modelling

Getting to know the data

Download and load MNIST data

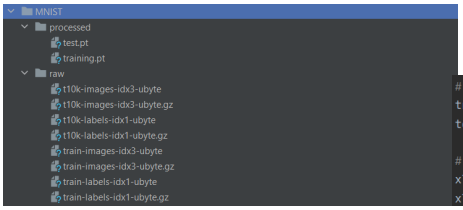


Figure 1: Download the data

```
# Download MNIST data
train_set = MNIST('.', download=True)
test_set = MNIST('.', train=False)

# Load the training and test data [x], along with labels [y]
xTrain, yTrain = torch.load('MNIST/processed/training.pt')
xTest, yTest = torch.load('MNIST/processed/test.pt')
```

Figure 2: Load the data

Data points

Training set has 60000 data points.

Test set has 10000 data points.

```
Dataset MNIST
    Number of datapoints: 60000
    Root location: .
    Split: Train
```

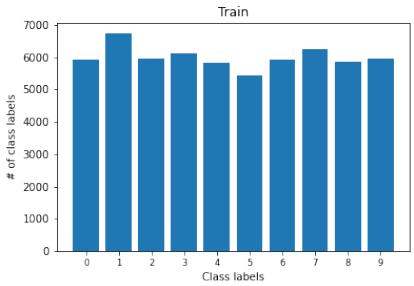
Figure 3: Training Set

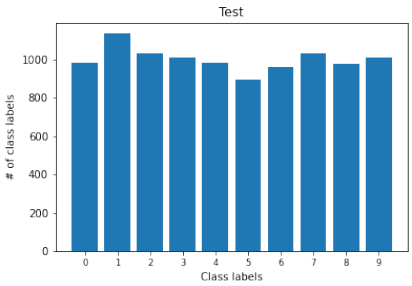
```
Dataset MNIST
    Number of datapoints: 10000
    Root location: .
    Split: Test
```

Figure 4: Test Set

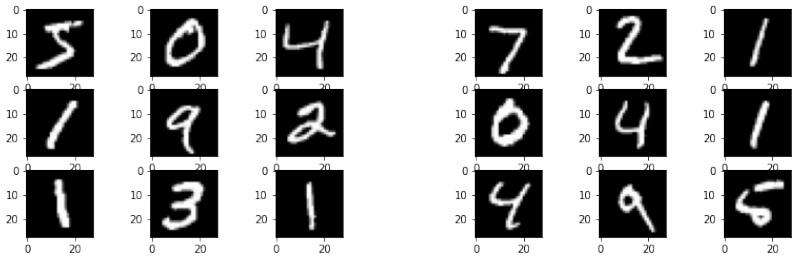
Histogram of different classes

Through the histograms we can see the data distribution in both training and test sets are quite balanced.





Visualize samples



Reshape the images

Principal component analysis (PCA) on MNIST

Smaller dataset

The smaller dataset:

xDataset shape: (30596, 784)

yDataset shape: (30596, 1)

```
### Reshape the images from Nx28x28 to Nx784
xTrain = xTrain.reshape(xTrain.shape[0], 784)
print(xTrain.shape)
xTest = xTest.reshape(xTest.shape[0], 784)
print(xTest.shape)
```

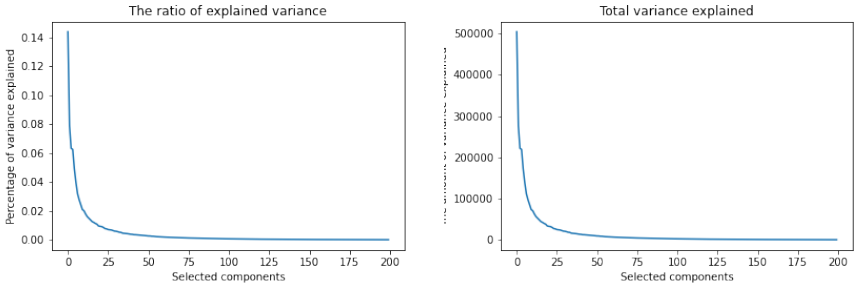
```
torch.Size([60000, 784])
torch.Size([10000, 784])
```

```
count = 0
xDataset = np.zeros(shape=(60000,784))
yDataset = []
for i in range(60000):
    if yTrainNp[i] == 0 or yTrainNp[i] == 1 or yTrainNp[i] == 2 or yTrainNp[i] == 3 or yTrainNp[i] == 4:
        count+=1
        xDataset[i] = xTrainNp[i]
        yDataset.append([yTrainNp[i]])
print(count)
m, n = xDataset.shape
rows = [row for row in range(m) if not all(xDataset[row] == 0)]
xDataset = xDataset[rows]
print(xDataset.shape)
yDataset = np.array(yDataset)
print(yDataset.shape)
```

Import and Perform PCA with D=200 components

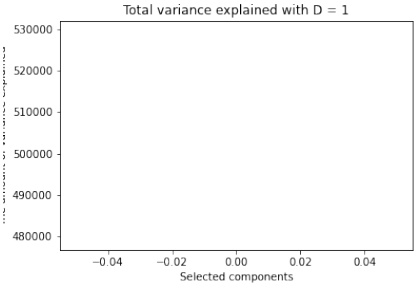
```
from sklearn.decomposition import PCA
X = xDataset
pca = PCA(n_components=200)
pca.fit(X)
```

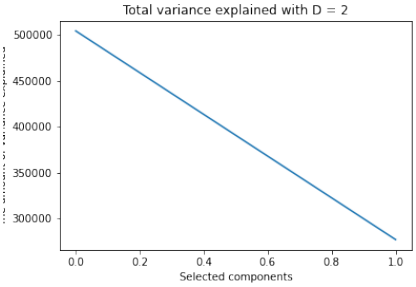
Plot Eigen spectrum

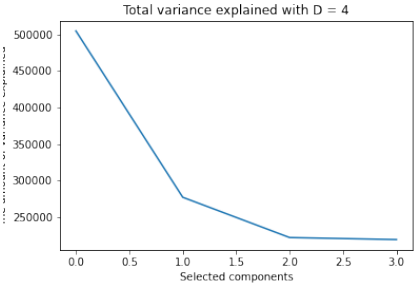


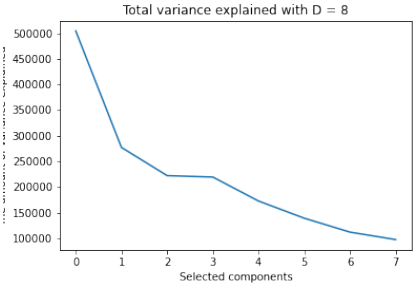
Various number of principal components

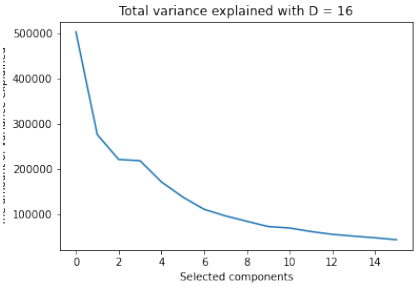
While D increasing, the graph gets more accurate and precise.

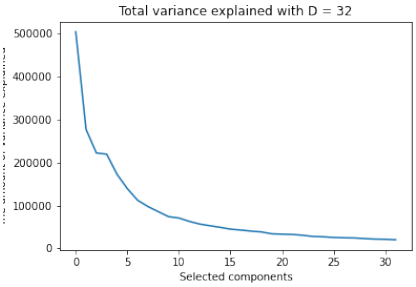

Total variance explained with D = 1
Selected components

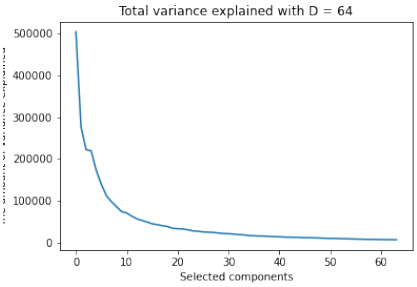

Total variance explained with D = 2
Selected components

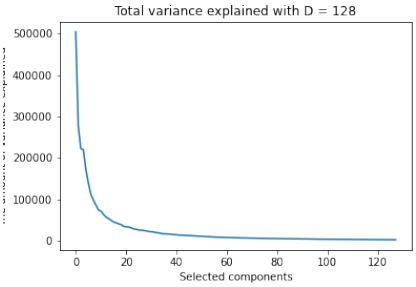

Total variance explained with D = 4
Selected components

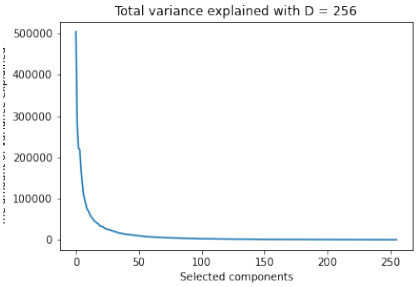

Total variance explained with D = 8
Selected components

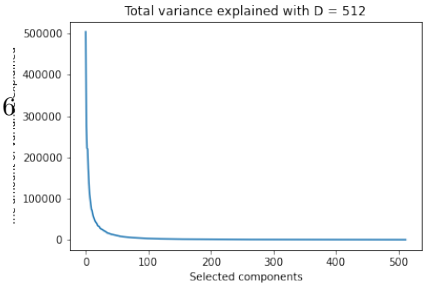

Total variance explained with D = 16
Selected components

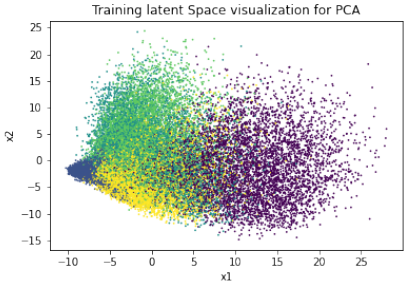

Total variance explained with D = 32
Selected components


Total variance explained with D = 64
Selected components

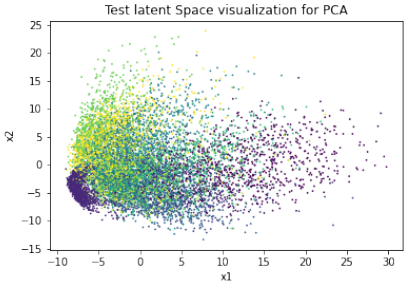

Total variance explained with D = 128
Selected components


Total variance explained with D = 256
Selected components


Total variance explained with D = 512
Selected components



2D scatter plot



Test data

Antoencoders on MNIST