INFR10086 Machine Learning (MLG) Semester 2, 2025/6

Tutorial 5: Autoencoders

In this tutorial, we are going to look at autoencoders and how they relate to other unsupervised learning algorithms, such as PCA and k-means.

1 A simple autoencoder

An autoencoder is defined as an encoder f and a decoder g that both try to minimize

$$\|x - g(f(x))\|^2 \tag{1}$$

for some data point x.

Discussion. Spend some time reading `src/train-autoenc.py`. What are f and g implemented in `src/train-autoenc.py`?

Run the following command.

```
$ mkdir exp1
$ cd exp1
$ ../src/train-autoenc.py init weight-0.ckpt
$ ../src/train-autoenc.py train weight-0.ckpt weight-1.ckpt
```

This trains the autoencoder for 1 epoch, and it took me 2.5 minutes on my desktop. It might take you some time if you run this on your laptop.

Discussion. Run `../src/plot-autoenc.py` in `exp1/`. What are the images saved from the script? Are there any meaningful patterns?

2 Connecting PCA to autoencoding

Given a data set $\{x_1, \ldots, x_n\}$, recall that the variance of all the points along the direction w_i is defined as

$$\sum_{j=1}^{n} (w_i^\top x_j)^2. \tag{2}$$

In other words, principal component analysis (PCA) is trying to maximize

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (w_i^\top x_j)^2 \tag{3}$$

such that $w_i^\top w_j = 0$ if $i \neq j$ and $w_i^\top w_i = 1$ for all $i = 1, \ldots, n$.

Discussion. Show that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (w_i^\top x_j)^2 = \|WX\|_F^2 \tag{4}$$

where $W = \begin{bmatrix} w_1^\top & \cdots & w_n^\top \end{bmatrix}$ and $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$.

Discussion. Show that PCA is optimizing

$$\min_{W} \quad \|X - W^\top WX\|_F^2 \tag{5}$$

$$\text{s.t.} \quad WW^\top = I \tag{6}$$

In other words, PCA can be seen as an autoencoder, where $f(X) = WX$ is the encoder and $g(Z) = W^\top Z$ is the decoder.

Spend some time reading `src/pca.py` and run the following commands.

```
$ mkdir exp2
$ cd exp2
$ ../src/pca.py
```

Discussion. Run `../src/plot-eigen.py` in `exp2/`. What are the images saved from the script? Are there any meaningful patterns?

3 Connecting k-means to autoencoding

Recall that k-means is the problem of grouping a set of n points $\{x_1, \ldots, x_n\}$ into k sets $S_1, \ldots, S_k$, attempting to minimize

$$\sum_{i=1}^{k} \sum_{j \in S_i} \|x_j - \mu_i\|^2, \tag{7}$$

where $\mu_i = \frac{1}{|S_i|} \sum_{j \in S_i} x_j$ for $i = 1, \ldots, k$.

Discussion. Show that the k-means objective can be written as

$$\min_{W,Z} \quad \|X - WZ\|_F^2 \tag{8}$$

$$\text{s.t.} \quad z_{ij} \in \{0,1\} \text{ for } i = 1,\dots,k \text{ and } j = 1,\dots,n \tag{9}$$

$$\sum_{i=1}^{k} z_{ij} = 1 \text{ for } j = 1,\dots,n \tag{10}$$

where $X \in \mathbb{R}^{d\times n}$, d is the dimension of each data point, and n is the number of data points, $W \in \mathbb{R}^{d\times k}$, and $Z \in \mathbb{R}^{k\times n}$.

In class, we presented Lloyd's algorithm for solving k-means. Recall that Lloyd's algorithm iterates between two steps

$$\text{step 1:} \quad z_{ij} = \begin{cases} 1 & \text{if } i = \arg\min_{i=1,\dots,k} \|x_j - \mu_i\| \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$\text{step 2:} \quad \mu_k = \frac{\sum_{j=1}^{n} z_{kj} x_j}{\sum_{j=1}^{n} z_{kj}} \tag{12}$$

Discussion. Show that Lloyd's algorithm is solving

$$\|X - Wq(X)\|_F^2 \tag{13}$$

where $q(X) = Z$ and each column of Z is one-hot. In other words, Lloyd's algorithm can be seen as an autoencoder where q is the encoder and $g(Z) = WZ$ is the decoder.