PCA + AutoEncoders

CMSC 422

SOHEIL FEIZI

sfeizi@cs.umd.edu

Slides adapted from MARINE CARPUAT and GUY GOLAN

Today’s topics

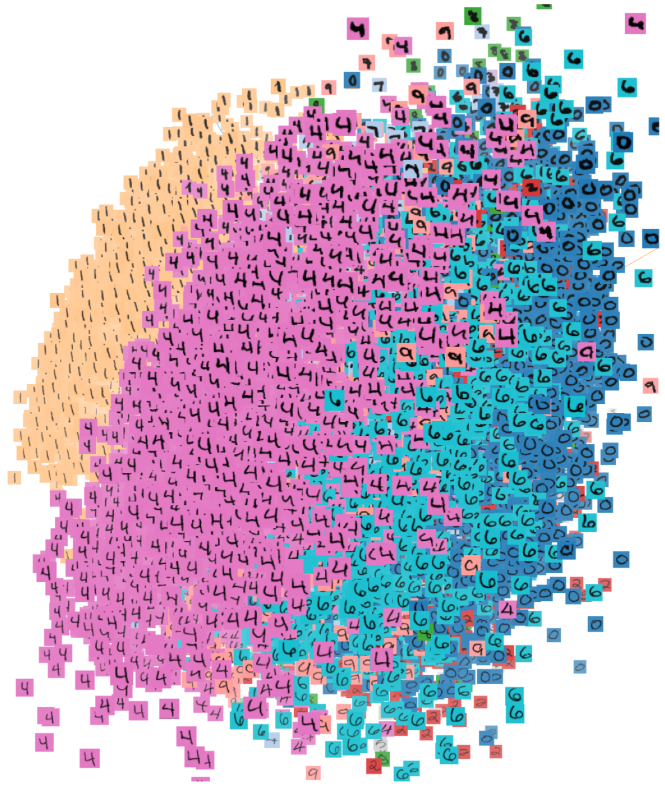
- PCA
- Nonlinear dimensionality reduction

Unsupervised Learning

Unsupervised Learning

Data: X (no labels!)

Goal: Learn the hidden structure of the data



Dimensionality Reduction

- Goal: extract hidden lower-dimensional structure from high dimensional datasets

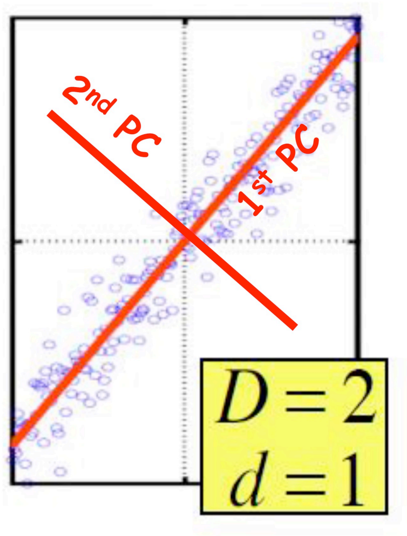
- Why?
  - To visualize data more easily
  - To remove noise in data
  - To lower resource requirements for storing/processing data
  - To improve classification/clustering

Principal Component Analysis

- Goal: Find a **projection** of the data onto directions that **maximize variance** of the original data set
  - Intuition: those are directions in which most information is encoded

- Definition: **Principal Components** are orthogonal directions that capture most of the variance in the data

PCA: finding principal components



- 1st PC
  - Projection of data points along 1st PC discriminates data most along any one direction
- 2nd PC
  - next orthogonal direction of greatest variability
- And so on…

PCA: notation

- Data points
  - Represented by matrix X of size NxD
  - Let’s assume data is centered
- Principal components are d vectors: $v_1, v_2, \dots v_d$

  $v_i . v_j = 0, i \neq j$ and $v_i . v_i = 1$
- The sample variance data projected on vector v is

$$\sum_{i=1}^{n} ({x_i}^T v)^2 = (Xv)^T (Xv)$$

PCA formally

- Finding vector that maximizes sample variance of projected data:

$$argmax_v \, v^T X^T \, Xv \text{ such that } v^T v = 1$$

- A constrained optimization problem
  - Lagrangian folds constraint into objective: $argmax_v \, v^T X^T \, Xv \; - \; \lambda(v^T v \; - 1)$
  - Solutions are vectors v such that $X^T \, Xv = \; \lambda v$
    - i.e. eigenvectors of $X^T \, X$(sample covariance matrix)

PCA formally

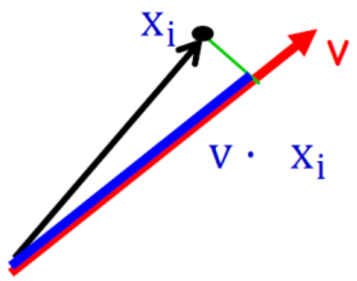
- The eigenvalue λ denotes the amount of variability captured along dimension v
  - Sample variance of projection $v^T X^T X v = \lambda$

- If we rank eigenvalues from large to small
  - The 1st PC is the eigenvector of $X^T X$ associated with largest eigenvalue
  - The 2nd PC is the eigenvector of $X^T X$ associated with 2nd largest eigenvalue
  - …

Alternative interpretation of PCA

- PCA finds vectors v such that projection on to these vectors minimizes reconstruction error

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$



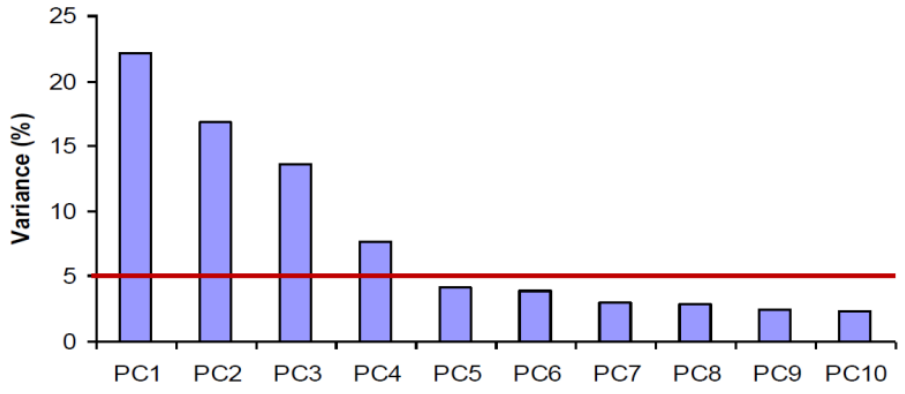
Resulting PCA algorithm

Algorithm 36 PCA($\mathbf{D}$, K)

```
1: $\boldsymbol{\mu} \leftarrow$ MEAN($\mathbf{X}$)                                   // compute data mean for centering
2: $\mathbf{D} \leftarrow \left(\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^\top\right)^\top \left(\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^\top\right)$    // compute covariance, $\mathbf{1}$ is a vector of ones
3: $\{\lambda_k, \boldsymbol{u}_k\} \leftarrow$ top $K$ eigenvalues/eigenvectors of $\mathbf{D}$
4: return $(\mathbf{X} - \boldsymbol{\mu}\mathbf{1})\,\mathbf{U}$                          // project data using $\mathbf{U}$
```

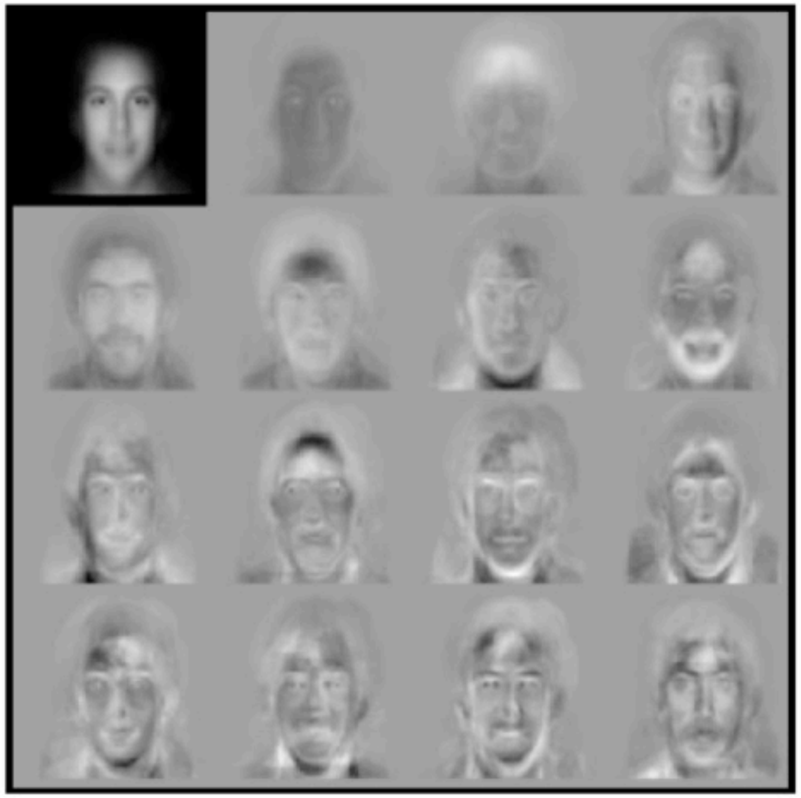
How to choose the hyperparameter K?

- i.e. the number of dimensions



- We can ignore the components of smaller significance

An example: Eigenfaces



Eigenfaces from 7562 images:

top left image is linear combination of rest.

Sirovich & Kirby (1987)
Turk & Pentland (1991)

PCA pros and cons

- Pros
  - Eigenvector method
  - No tuning of the parameters
  - No local optima

- Cons
  - Only based on covariance (2nd order statistics)
  - Limited to linear projections

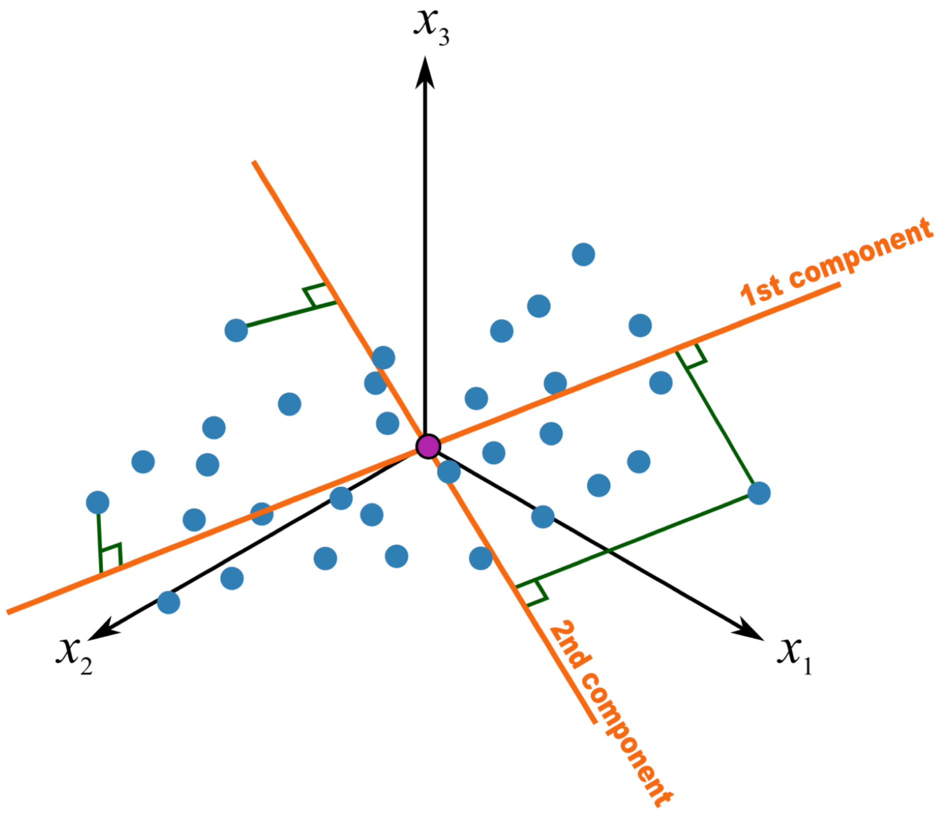
What you should know

- Principal Components Analysis
  - Goal: Find a **projection** of the data onto directions that **maximize variance** of the original data set
  - PCA **optimization objectives** and resulting **algorithm**
  - Why this is useful!

PCA – Principal Component analysis

- Statistical approach for data compression and visualization
- Invented by Karl Pearson in 1901

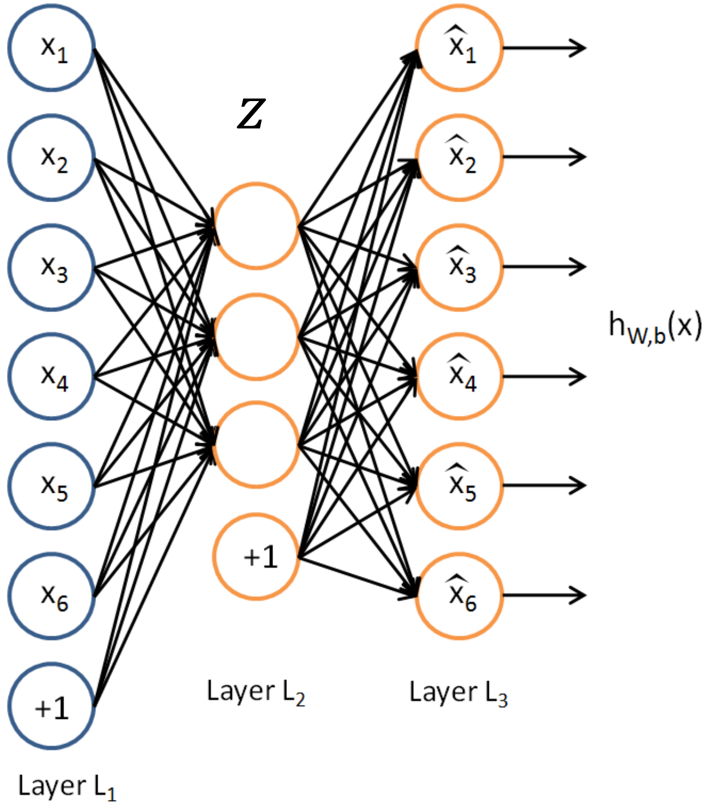
- Weakness: linear components only.



Autoencoder

- Unlike the **PCA** now we can use activation functions to achieve non-linearity.

- It has been shown that an AE without activation functions achieves the **PCA** capacity.



Uses

- The autoencoder idea was a part of NN history for decades (LeCun et al, 1987).

- Traditionally an autoencoder is used for dimensionality reduction and feature learning.

- Recently, the connection between autoencoders and latent space modeling has brought autoencoders to the front of generative modeling, as we will see in the next lecture.

Simple Idea

- Given data x (no labels) we would like to learn the functions f (encoder) and g (decoder) where:

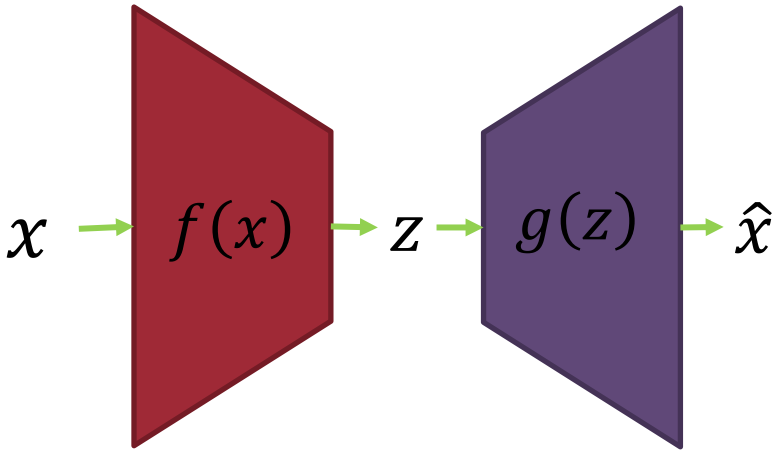
$$f(x) = s(wx + b) = z$$

and

$$g(z) = s(w'z + b') = \hat{x}$$

s.t $h(x) = g\big(f(x)\big) = \hat{x}$

where h is an **approximation** of the identity function.



(z is some **latent** representation or **code** and s is a non-linearity such as the sigmoid)

($\hat{x}$ is x's reconstruction)

Simple Idea

Learning the identity function seems trivial, but with added constraints on the network (such as limiting the number of hidden neurons or regularization) we can learn information about the structure of the data.

Trying to capture the distribution of the data (data specific!)

Training the AE

Using **Gradient Descent** we can simply train the model as any other FC NN with:

- Traditionally with *squared error* loss function

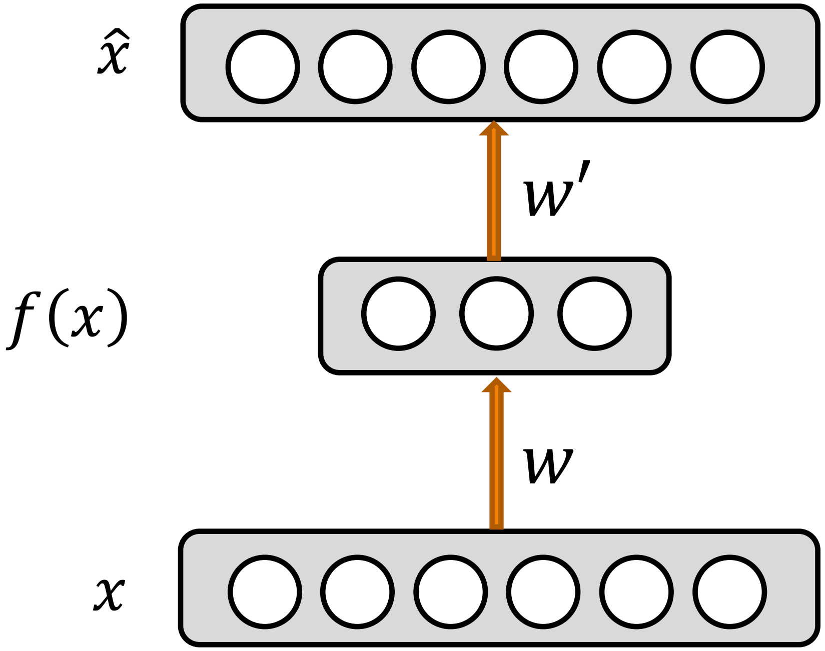
$$L(x, \hat{x}) = \|x - \hat{x}\|^2$$

- Why?

AE Architecture

- Hidden layer is **Undercomplete** if smaller than the input layer
  - ❑Compresses the input
  - ❑Compresses well only for the training dist.

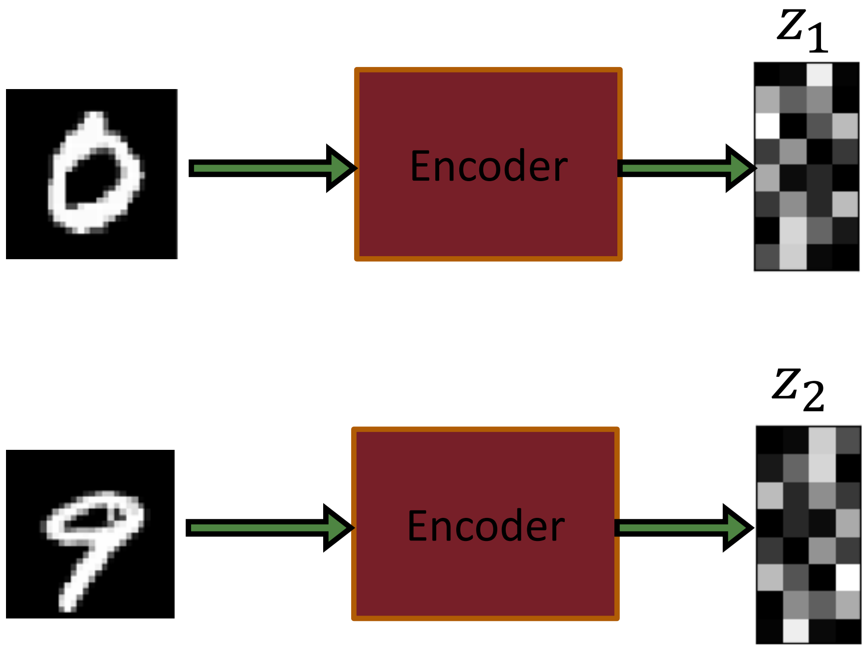
- Hidden nodes will be
  - ❑Good features for the training distribution.
  - ❑Bad for other types on input



Deep Autoencoder Example

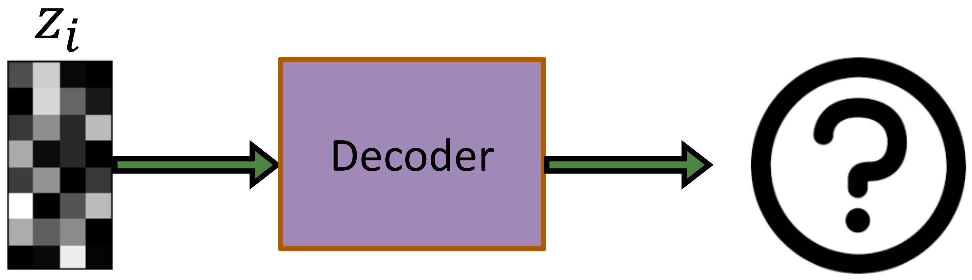
- https://cs.stanford.edu/people/karpathy/convnetjs/demo/autoencoder.html - By Andrej Karpathy

Simple latent space interpolation



Simple latent space interpolation

$$z_i = \alpha \overset{z_1}{\boxed{}} + (1-\alpha) \overset{z_2}{\boxed{}}$$



Simple latent space interpolation