CS 446/ECE 449: Machine Learning

Shenlong Wang

University of Illinois at Urbana-Champaign, 2024

Principal Component Analysis

Goals of this lecture

- Transition from supervised to unsupervised learning
- Getting to know the Principal Component Analysis (PCA)
- Relating PCA to the Singular Value Decomposition (SVD)

Reading material:

- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 12

Recap: Supervised learning

Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ of data-label pairs, construct a mapping $f(x; w) : \mathcal{X} \to \mathcal{Y}$.

Examples:

- KNN
- Least squares
- Logistic regression
- Support vector machine
- Decision trees
- Deep neural network

What if we don't have labels?

Without labels, we can still find structure in unlabeled data

$$\mathcal{D} = \{(x^{(i)})\}_{i=1}^{N}$$

Goal of unsupervised learning? Find "interesting patterns" in the data
Less clear but generally:

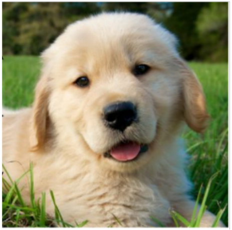
- Recover hidden structure
- Data compression or dimensionality reduction
- Explore or explain data (generate data)
- Construct features for supervised learning (e.g., word embeddings)

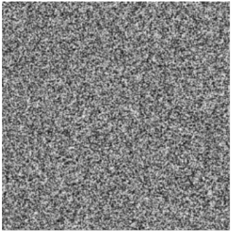
Methods:

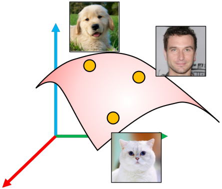
- PCA
- K-means
- Gaussian Mixture Models
- Hidden Markov Models
- Variational Auto-encoders
- Generative Adversarial Nets
- Autoregressive Methods
- Energy-based Models, Diffusion models

Dimension reduction

Background: High-dimensional data often can be described with a small number of degrees of variability







Dimension reduction: find a small number of “directions” in input space that explain variation in input data; re-represent data by projecting along those directions

PCA

Goal: find that lower dimensional **linear** subspace in which the projected data has highest variance

$$\text{(data)}\ \ X = \begin{bmatrix} | & & | \\ x^{(1)} & \cdots & x^{(N)} \\ | & & | \end{bmatrix}$$

$$\text{(centered data)}\ \ \bar{X} = \begin{bmatrix} | & & | \\ x^{(1)} - \mu & \cdots & x^{(N)} - \mu \\ | & & | \end{bmatrix} \text{ where } \mu = \frac{1}{N}\sum_i x^{(i)}$$

Never forget to center data! Symmetric matrix $\Sigma = \frac{1}{N}\bar{X}\bar{X}^T$

$$\max_{w:\|w\|_2^2=1} \mathrm{Var}(w^T\bar{X}) = \max_{w:\|w\|_2^2=1} \mathbb{E}[w^T\bar{X}\bar{X}^T w] = \max_{w:\|w\|_2^2=1} w^T\Sigma w$$

How to solve

$$\max_{w:\|w\|_2^2=1} w^T \Sigma w$$

Lagrangian:

$$L(w, \lambda) = w^T \Sigma w - \lambda(w^T w - 1)$$

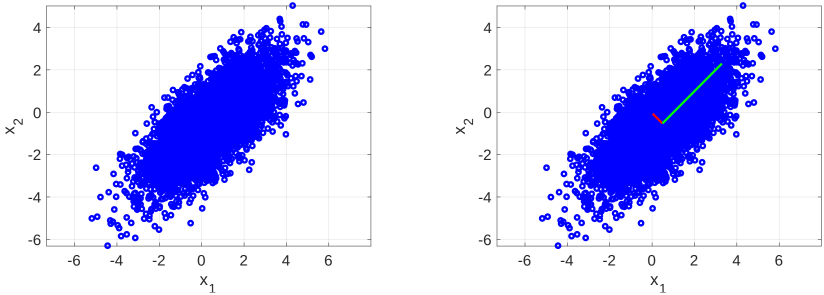
Derivative of L w.r.t. w set to zero:

$$\Sigma w = \lambda w \qquad \text{(eigenvalue problem)}$$

Which eigenvector/eigenvalue should we take?

We want to maximize $w^T \Sigma w = \lambda w^T w = \lambda$. Hence, w is the eigenvector corresponding to the largest eigenvalue.

Example:



What if we want to find the direction with second, third largest variance that's orthogonal to the first, first and second?

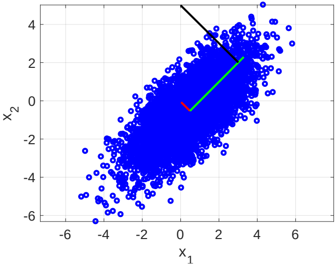
Finding that d-dimensional subspace that captures the largest variance?

$$\max_{w_1,\ldots,w_d : w_i^T w_j = \delta_{ij}} \sum_{i=1}^{d} w_i^T \Sigma w_i$$

Algorithm:

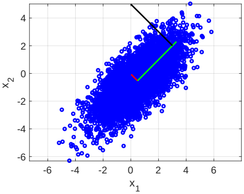
- Work sequentially one vector at a time
- Compute a matrix eigenvalue decomposition

How to project data into low-dimensional space?



1. Collect all subspace directions:

$$U = \begin{bmatrix} | & & | \\ w_1 & \cdots & w_d \\ | & & | \end{bmatrix}$$



2. Project points into subspace (compressed space)

$$\hat{x} = U^T(x - \mu)$$

3. Approximately reconstructed data

$$\tilde{x} = U\hat{x} + \mu$$

Alternative view of PCA:

PCA finds the axis which minimizes the sum of squared distances from points to their orthogonal projections on that axis (we assume $\mu = 0$ for notational convenience):

$$\min_{w:\|w\|_2^2=1} \frac{1}{N}\sum_{i=1}^{N} \|x^{(i)} - ww^T x^{(i)}\|_2^2 \qquad \text{(see previous slide \& lin reg)}$$

Frobenius norm:

$$\|A\|_F^2 = \sum_{i,j} a_{i,j}^2 = \mathrm{Tr}(A^T A)$$

Rewriting the objective:

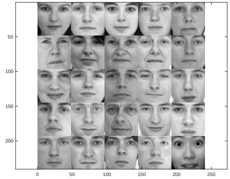
$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\|x^{(i)} - ww^T x^{(i)}\|_2^2 &= \frac{1}{N}\|\bar{X} - ww^T\bar{X}\|_F^2 \\
&= \frac{1}{N}\operatorname{Tr}((P\bar{X})^T(P\bar{X})) \quad \text{where} \quad P = I - ww^T \\
&= \frac{1}{N}\operatorname{Tr}(\bar{X}\bar{X}^T P^T P) \quad \text{since } tr(ABCD) = tr(BCDA) \\
&= \quad \operatorname{Tr}(\Sigma P) \qquad \text{since for projection} \quad P^T P = P
\end{aligned}
$$

Hence:

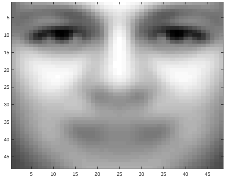
$$
\begin{aligned}
& \arg\min_{w:\|w\|_2^2=1} \frac{1}{N}\sum_{i=1}^{N}\|x^{(i)} - ww^T x^{(i)}\|_2^2 := \operatorname{Tr}(\Sigma) - \operatorname{Tr}(\Sigma ww^T) \\
= \quad & \arg\max_{w:\|w\|_2^2=1} w^T\Sigma w
\end{aligned}
$$

Applications

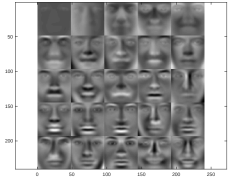
Compressing high-dimensional data



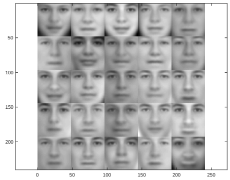
Original x



Mean μ



25 eigenvectors U



Reconstruction $\tilde{x}$

Singular Value Decomposition to compute PCA

Currently we compute the eigenvalues of $\Sigma = \frac{1}{N}\bar{X}\bar{X}^T$
Instead of first computing the outer product and then computing its eigenvalues, we can use the singular value decomposition. How?
Given the singular value decomposition

$$\frac{1}{\sqrt{N}}\bar{X} = USV^T$$

We obtain

$$\Sigma = USV^T VSU^T$$

We obtain

$\Sigma U = USV^T VSU^T U = S^2 U$ since U, V are orthonormal and S is diag

The left singular vectors U of $\frac{1}{\sqrt{N}}\bar{X}$ are needed

Quiz:

- What is PCA?
- What are the two views of PCA?
- Which two approaches can be used to compute principal components?
- How is data compressed and reconstructed?

Important topics of this lecture

- Understanding PCA
- Getting to know different ways to compute PCA

Up next:

- K-means