Machine Learning
Dimensionality Reduction

Dr. Farshid Alizadeh-Shabdiz
Spring 2021

1

Why Dimensionality Reduction

- Large number of features makes training slow
- Makes it harder to find a good solution – curse of dimensionality
  - Higher dimensionality, larger distance between training points, harder prediction
- Generally - We lose some information so there will be degradation (small one)
  - In some cases, this will reduce noise and increase performance

2

Curse of Dimensionality

- As dimension increase more and more points are at the border
  - Example:
    - In a unit square only 0.4% of points are 0.001 from a border
    - In 10,000 dimensional unit, this probability is almost 1
- Distance between points increase
  - Example
    - Distance of two points in a 1D unit is 0.33
    - Distance of two points in a 2D unit is 0.52
    - Distance of two points in 3D unit is 0.66
    - Distance of two points in 10D unit is 1.26
  - So high-dimensional datasets can be very sparse

3

Dimensionality Reduction Meaning

- Selecting M lines in N dimension space in which M<<N.
- Simple example, selecting every other pixel in MNIST pictures!

4

Principal Component Analysis
PCA

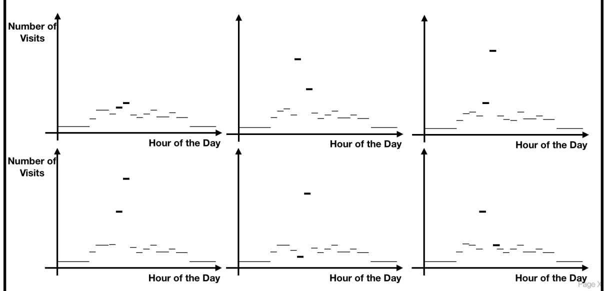
5

PCA is an Unsupervised Learning

- Unsupervised Learning
  - 01 There is no prediction or response parameter
  - 02 The data is unlabeled
- The goal can be
  - 01 To discover "interesting" parts of the observations
  - 02 To discover subgroups among variables/parameters or among the observations
  - 03 Extract patterns among the observations

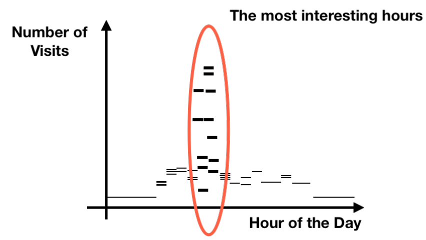
6

Number of Visits to a Coffee Shop



7

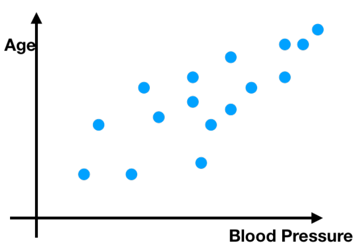
Number of Visits to a Coffee Shop



8

Example

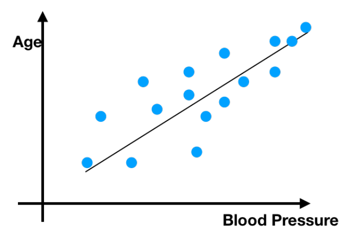
n x d : n points in d dimensional space
Example - *(16 x 2)*- 16 patients in two dimensional space



9

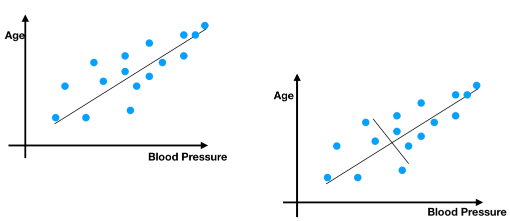
Example

n x d : n points in d dimensional space
Example - *(16 x 2)*- 16 patients in two dimensional space



10

Example



- Variation in the direction of principle component is the highest.
- Distance of the points to the line is the shortest

11

Principal Component Analysis (PCA)

- With preservation of dimensionality, PCA transforms a data set from a set of feature dimension into a linearly uncorrelated dimension of the features that have maximal variance. So mathematically
  - 00 After data preparation:
  - 01 if the data set is X, and it is *(n x p)*. *p* number of features and (*n*>1), the first principle component is written as follows
    $z_1 = c_{11}X_{i1} + c_{21}X_{i2} + \ldots + c_{p1}X_{ip}\ ,\ i \in [1,n]$
  - 02 That z_1 has the largest variance, given the coefficients c_{i1} are normalized to one, which means
    $$\sum_{i=1}^{p} c_{i1}^2 = 1$$

12

Data Preparation

- No data is perfect - so the first step is data cleansing and maybe also removing outliers and abnormal values
- Each of the features in X gets centered around mean - the column means of X become zero, since we are interested in variance.
- Scaling - Scale features and make the ranges comparable.

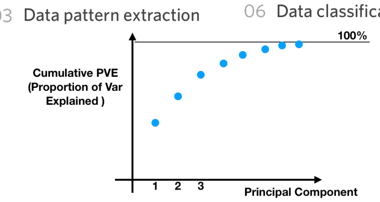
  Dividing by standard deviation is the most common way.
  Do we need to have scaling? Not always. We don't need it when
  - Parameters are in the same unit
  - The magnitude of the features matter

13

PCA Application

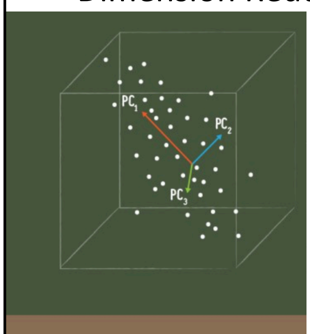
- PCA has been widely used, for example

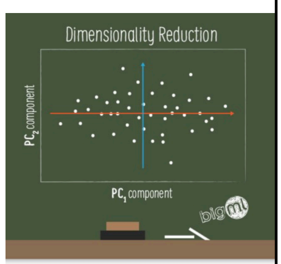
01 Data compression
02 Data visualization
03 Data pattern extraction
04 Dimension reduction
05 Data signature extraction
06 Data classification



14

Dimension Reduction Example



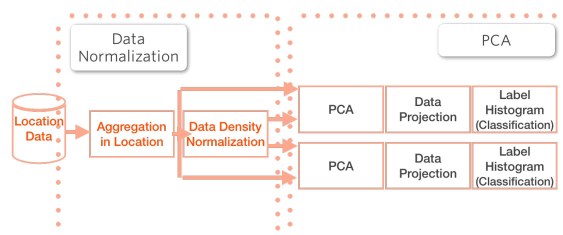


Reference: Mathanraj Sharma "Guide to PCS" https://medium.com/analytics-vidhya/guide-to-principal-component-analysis-ab04a8a9c305

15

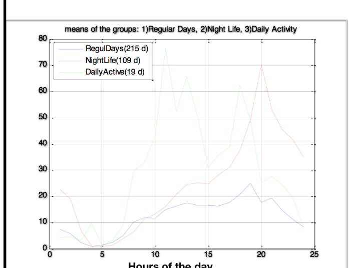
Behavioral Analysis Using PCA

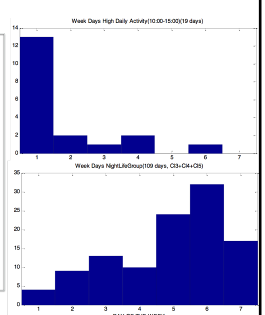
- It is an unsupervised learning



16

Example of Users Behavior Analysis Using PCA

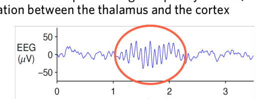




17

Example of a Research Paper: A Data-Driven Bayesian Algorithm for Sleep Spindle Detection

- The sleep spindle is a transient pulse of high frequency waves (12-14Hz) on the EEG, emerging from communication between the thalamus and the cortex

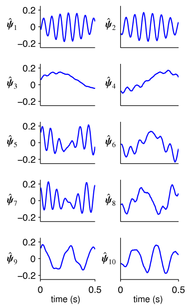


- Sleep spindle has been implicated in:
  - active memory consolidation
  - general cognitive ability
  - sleep stability
  - psychiatric and neurological disorders
- Automatic detection and quantification of sleep spindles is very important in the analysis of sleep studies involving several hours of recorded data.

Behtash Babadi, Scott M. McKinney, Vahid Tarokh, and Jeffrey M. Ellenbogen, 'A data driven bayesian algorithm for sleep spindle detection', IEEE Transactions on Biomedical Engineering, 59(2), 483–493, 2012.

18

Sleep Spindle Eigenvectors



The first 10 spindle basis elements obtained from a pool of 1231 sample spindles (sampling frequency of 200Hz).

19

Variance and Covariance

- Variance and standard deviation
  - A measure of how spread out the data is
  - It can be used to examine each dimension of the data independently

$$Var[R] = E[(R - \mu_R)^2]$$

- Covariance
  - A measure of relationship between two dimensions
  - Considering general trend of two dimensions, positive covariance means both increase together and negative means that they move in different direction, i.e. as one increases, the other one decreases

$$E[(R - \mu_R)(Q - \mu_Q)]$$

20

Covariance Matrix

- Assume a data set with N number of dimensions. For example age, weight, height, blood pressure from many patients.
- Compose covariance matrix of data, which captures relationship between each pair of dimensions of the data

$$\begin{pmatrix} cov(x_1,x_1) & cov(x_1,x_2) & \cdots & cov(x_1,x_n) \\ cov(x_2,x_1) & cov(x_2,x_2) & \cdots & cov(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_n,x_1) & cov(x_n,x_2) & \cdots & cov(x_n,x_n) \end{pmatrix}$$

Note that

- ***$cov(x_i,x_j)=cov(x_j,x_i)$.***
- Covariance matrix is a square matrix, and it is symmetric with respect to the main diagonal
- Components on the main diagonal are variances

21

Eigenvectors and Eigenvalues

- Eigenvector is a non-zero vector satisfying following equation

$$Av = \lambda v$$

- The vector v is eigenvector of the data matrix A, and λ is corresponding eigenvalue of the eigenvector.
- Eigenvalues are calculated solving following equation

$det[A-\lambda I]=0$

22

Principal Component Analysis Intro

- Principal component analysis or PCA is also called Karhunen-Loeve transform (KLT)
- PCA has been introduced by Karl Pearson in 1901
- PCA is a multivariate data analysis - many measurements with multiple parameters

23

PCA - Simple Eigenvector Proof

- For the simple two dimensional case, with orthogonal unit vectors *v,u*.
- X can be written as a sum of projection on v and u
  $X=(u^TX)u+(v^TX)v$
- Therefore, variance of data at the direction of *u* will be
  $\Rightarrow f(u) = E[((u^TX)u)^2]=E[u^T(u^TX)^T(u^TX)u)]=E[u^TX^Tuu^TXu)]$
  $= E[u^TX^TXu)]=u^TE[X^TX]u$
- Maximizing *f(u)* given unit matrix *u*. Therefore, applying Lagrange and finding derivative of *u*:
  $\Rightarrow f'(u)-\lambda(u^Tu)' = 2u^TE[X^TX]-2\lambda(u^T)=0 \Rightarrow \boldsymbol{E[X^TX]u=\lambda u}$
- So, *u* is the eigenvector of covariance matrix of *X* and the optimum value of *f(u)* is equal to the eigenvalue.

24

PCA Calculation – Step by Step

1. Calculate mean value for each data dimension
2. Subtract mean value of each dimension from the data
3. Create covariance matrix of the data
4. Calculate eigenvectors and eigen-values of the covariance matrix
5. Order eigenvectors according to the corresponding eigen-values
6. Analyzing data in the new dimensions according to eigenvectors and eigen-values

25

How to Compute PCA

- Given *(n x p)* data set, X.
  - For example, n=11 patients with p=2 measurements of age and blood pressure.
  - Ages are
    - [18 , 20 , 22 , 30 , 45 , 46 , 50 , 51 , 60 , 64 , 66], and
  - Corresponding blood pressures are
    - [94 , 96 , 96 , 92 , 102 , 115 , 120, 260 , 126 , 120 , 130]
- With n>1, find dimensions $z_j = c_{1j} x_{i1} + c_{2j} x_{i2} + \dots + c_{pj} x_{ip}\ ,\ i \in [1,n]\ ,\ j \in [1,p]$
- which best represent the data in terms of variance

$$max_{c11,\dots,cp1} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} c_{i1} x_{ij} \right)^2$$

- Given the coefficients c_{i1} are normalized $\sum_{i=1}^{p} c_{i1}^2 = 1$
- **Solution:** Eigenvectors from Singular Value Decomposition (SVD) in linear algebra

26

SVD Overview

$$E[X^T X]u = \lambda u$$

- Problem statement: with data X *(nxd)*, *n* data points in *d* dimensional space, finding the best k-dimensional subspace *(k<d)* in terms of minimizing the sum of the squares of distance of the points to the subspace.
  Special case is a line through the origin (k=1)
- Solution: SVD of a data matrix X (square & symmetric) is factorization of X into

$$X = VDV^T \quad Or \quad Xv_i = \lambda v_i$$

  - Columns of *V* are orthonormal vectors (Eigenvectors) and are the solutions.
  - *D* is diagonal with positive, real entries (Eigenvalues). The values are from large to small and they capture squared of projection of data into Eigenvectors

27

Example – Three dimensional data

Assume a three dimensional data, which captures age, blood pressure and weight of 10 people

Ages are

[18 , 20 , 22 , 30 , 45 , 46 , 50 , 60 , 64 , 66], and

Corresponding blood pressures are

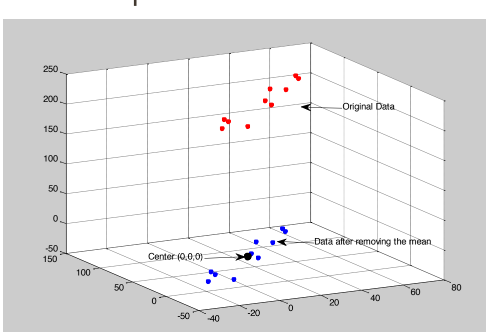
[94 , 96 , 96 , 92 , 102 , 115 , 120 , 126 , 120 , 130]

The corresponding weights are as follows

[161 , 175 , 170 , 160 , 185 , 185 , 201 , 193 , 212 , 211]

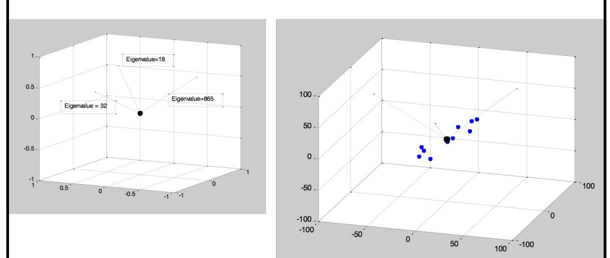
28

Example – Data Visualization



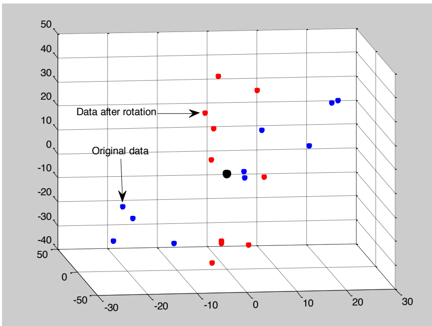
29

Example – Eigenvectors + data



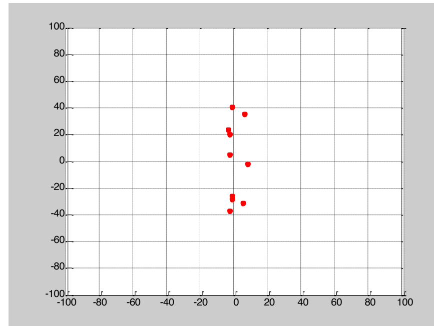
30

Example – Data Before & After Rotation



31

Example – Reduced Data to 2 Dimensions



32

Select Number of Components

Cross validation is the best way to choose number of PCA components, and also how much of the variation has been captured by the components.

33

PCA limitations

- PCA is very popular and it has been an effective tool for many applications
- PCA is computationally heavy
- PCA is based on linear combination of features
  - 01 Solution is Kernel PCA - same idea as Kernel SVM

34

Kernel PCA

35

Kernel PCA

- Kernel PCA is an extension of PCA which uses a Kernel to expand linear PCA to non-linear domain.

36

Kernel PCA - Example

- For example second degree polynomial $\phi\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}x_1^2\\\sqrt{2}x_1x_2\\x_2^2\end{pmatrix}$
- Dot product $\phi(a)\phi(b)=\phi\begin{pmatrix}a_1\\a_2\end{pmatrix}\phi\begin{pmatrix}b_1\\b_2\end{pmatrix}=\begin{pmatrix}a_1^2\\\sqrt{2}a_1a_2\\a_2^2\end{pmatrix}\begin{pmatrix}b_1^2\\\sqrt{2}b_1b_2\\b_2^2\end{pmatrix}$

$$=(a_1b_1+a_2b_2)^2=(a^Tb)^2$$

- Dot product of a=(a_1,a_2) and b=(b_1,b_2) is
  - a.b = (a_1b_1,a_2b_2)

37

Kernel PCA

- Kernel provides a way to calculate dot product of vectors without transforming the vectors.

$$Kernel(a,b)=\phi(a)^T\phi(b)$$

- Dot product is all we need, since we need covariance matrix of the data
- Common kernels

d dimensional Polynomial : $\left(1+\sum_{i=1}^{K}x_iy_i\right)^d$

$Polynomial:(\gamma a^Tb+\lambda)^d$

$GaussianRBF:\exp(-\gamma\,\|a-b\|^2)$

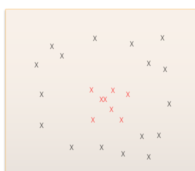
$Sigmoid:Tanh(\gamma a^Tb+\lambda)$

- RBF – Radial Basis Function

38

Kernel PCA - Example

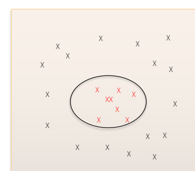
- Separate black and red dots



39

Kernel PCA - Example

- Separate green and red dots



- Problem can be become linear by following Kernel $(X_1,X_2)\rightarrow(X_1^2,\sqrt{2}X_1X_2,X_2^2)$

40

Kernel PCA

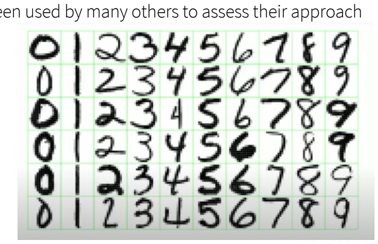
- Transforming data to the higher dimension using a kernel
- Extracting principal components in the higher dimension
- Since it is an unsupervised learning, there is no metric to measure goodness of the kernel

Note: remember data has to be centered after transformation to the new space.

41

PCA Example – MNIST Data

- MNIST - a database of 70,000 images of hand-written digits
  - It is publicly available
  - It can be processed fast
  - It has been used by many others to assess their approach



42