Artificial Intelligence for Medical Data with Python

10 SAMPLE SLIDES

5th session –
Matrix Factorization for Health Data
(Chapter 11 of MMDS book)

UNIVERSITY OF THE AEGEAN

SCHOOL OF ENGINEERING

DEPARTMENT OF INFORMATION AND COMMUNICATION SYSTEMS ENGINEERING

Presenter: Panagiotis Symeonidis

Associate Professor

http://panagiotissymeonidis.com

psymeon@aegean.gr

Eigen Decomposition

- Let $\boldsymbol{A}$ be a square and diagonalizable matrix and let $\boldsymbol{\lambda}$ be a constant and $\boldsymbol{e}$ a column nonzero vector with the same number of rows as $\boldsymbol{A}$. Then $\boldsymbol{\lambda}$ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{e}$ is the corresponding eigenvector of $\boldsymbol{A}$ if:

$$A \cdot e = \lambda \cdot e$$

- For a matrix $\boldsymbol{A}$ of rank $\boldsymbol{r}$, we can group the $\boldsymbol{r}$ nonzero eigenvalues in a $r \times r$ diagonal matrix $\boldsymbol{\Lambda}$ and their eigenvectors in a n$\times r$ matrix $\boldsymbol{E}$. So, we have:

$$A \cdot E = E \cdot \Lambda$$

- In case that the rank $\boldsymbol{r}$ of the matrix $\boldsymbol{A}$ is equal to its dimension n, then $\boldsymbol{A}$ can be factorized as:

$$A = E \cdot \Lambda \cdot E^{-1}$$

This diagonalization is **similar to SVD**, which will be described later!

Principal Components Analysis (PCA)

- How can we find the direction with largest variance?
  - By the **eigenvector** for the **covariance matrix** of the data
  - Suppose there are 3 dimensions, denoted as X, Y, Z. The covariance matrix is

$$COV = \begin{bmatrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & cov(Z,Z) \end{bmatrix}$$

where

$$cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

where N is the number of the observations.

- Note the **diagonal** is the covariance of each dimension with respect to itself, which is just the **variance** of each random variable
- Also, $cov(X,Y) = cov(Y,X)$, hence matrix is **symmetric** about the diagonal
- d-dimensional data will result in a $d \times d$ covariance matrix

PCA steps

- Centering Data
- Calculation of the covariance matrix of
  - ✓ drugs,
  - ✓ side effects
  - ✓ diseases
- Application of PCA to the covariance matrix
  - ✓ calculation of eigenvalues
  - ✓ calculation of eigenvectors
- Finding the highest eigenvalues together with their eigenvectors
- Data Visualization

Toy Example

Question: Can we bring to the surface the latent associations between the adverse side effects of drug-drug interactions?

Patient	Insulin	Anticoangulants	Hypoglykemia	Bleeding	Diseases
1	360	9	270	2	Diabetes
2	366	8	274	1	Diabetes
3	145	10	119	3	Diabetes
4	138	8	112	4	Diabetes
5	32	360	28	350	Heart
6	22	358	11	352	Heart
7	11	112	2	102	Heart
8	13	113	3	100	Heart

Toy Example

Patient	Insulin	Anticoangulants	Hypoglykemia	Bleeding	Diseases
1	360	9	270	2	Diabetes
2	366	8	274	1	Diabetes
3	145	10	119	3	Diabetes
4	138	8	112	4	Diabetes
5	32	360	28	350	Heart
6	22	358	11	352	Heart
7	11	112	2	102	Heart
8	13	113	3	100	Heart

- Insulin ↔ side effect of Hypoglycemia
- Anticoagulants ↔ side effect of Bleeding

- two distinct data sets:
  - insulin and hypoglycemia
  - anticoagulants and bleeding

- The steps to perform the PCA are as follows:
  1. Calculation of covariance matrix
  2. Calculation of the eigenvalues – eigenvectors
  3. Selection of principal components

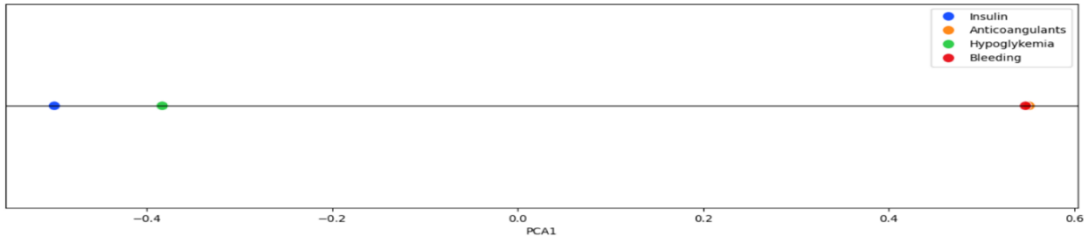
Covariance Matrix Computations

- From the example table:
  - from the 4 columns → centered covariance matrix

A/A	Insulin	Anticoangulants	Hypoglykemia	Bleeding
1	224.125	-113.25	167.625	-112.25
2	230.125	-114.25	171.625	-113.25
3	9.125	-112.25	16.625	-111.25
4	2.125	-114.25	9.625	-110.25
5	-103.875	237.75	-74.375	235.75
6	-113.875	235.75	-91.375	237.75
7	-124.875	-10.25	-100.375	-12.25
8	-122.875	-9.25	-99.375	-14.25

- Application of PCA → calculating the eigenvalues and eigenvectors
- The first 2 eigenvectors of the covariance matrix in our example are shown in the table below
- In a horizontal scatter plot we plot the values of the first principal component

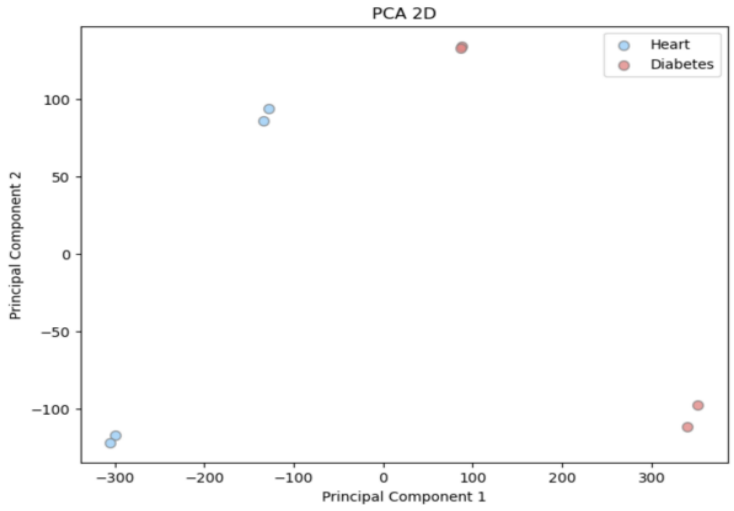
1st eigenvector	2nd eigenvector
-0.499443	0.622078
0.551398	0.441734
-0.383077	0.465651
0.547511	0.448396

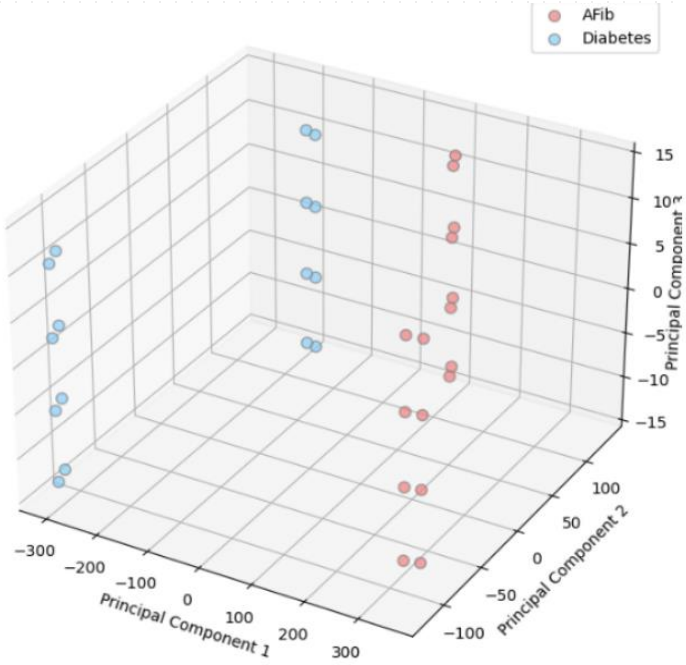


New dimensional space

1st Pr. Comp.	2nd Pr. Comp.	Diseases
-300.074613	-117.121182	Diabetes
-305.695474	-121.814171	Diabetes
-133.736505	86.042010	Diabetes
-128.107132	94.101243	Diabetes
340.534492	-111.501690	Heart
352.040467	-97.369043	Heart
88.478676	134.453103	Heart
86.560089	133.209730	Heart

- Multiplying the centred covariance matrix by the first 2 eigenvectors gives the new matrix with the new coordinates
- The clustering of patients after applying PCA to our example data can be seen in the graph.





Singular-Value Decomposition (SVD)

- SVD allows an exact representation of any matrix that
  - eases the elimination of less important parts of that representation
  - produces an approximate representation with any desired number of dimensions.
- The fewer the dimensions we choose, the less accurate will be the approximation

Definition

$$\boldsymbol{A}_{[m\times n]} = \boldsymbol{U}_{[m\times r]}\Sigma_{[r\times r]}\left(\boldsymbol{V}_{[n\times r]}\right)^{\boldsymbol{T}}$$

$\boldsymbol{A}$: **Input data matrix**
$[\boldsymbol{m}\times\boldsymbol{n}]$ matrix
(e.g., *m* documents, *n* terms)

$\boldsymbol{U}$: **Left singular vectors**
$[\boldsymbol{m}\times\boldsymbol{r}]$ matrix
(*m* documents, *r* concepts)

Σ : **Singular values**
$[\boldsymbol{r}\times\boldsymbol{r}]$ diagonal matrix
(strength of each 'concept')
($\boldsymbol{r}$: rank of the matrix $\boldsymbol{A}$)

$\boldsymbol{V}$: **Right singular vectors**
$[\boldsymbol{n}\times\boldsymbol{r}]$ matrix
($\boldsymbol{n}$ terms, $\boldsymbol{r}$ concepts)

CUR Decomposition

- Goal: Express A as a product of matrices C,U,R
  - Make $\|A - C \cdot U \cdot R\|_F$ small
- "Constraints" on C and R:
  - C columns are "randomly" selected from matrix A