Dimensionality reduction techniques

Sanzhar Askaruly (San)

Ulsan National Institute of Science and Technology
Ph.D. Candidate in Biomedical Engineering

CodeSeoul MLA
December 10, 2022

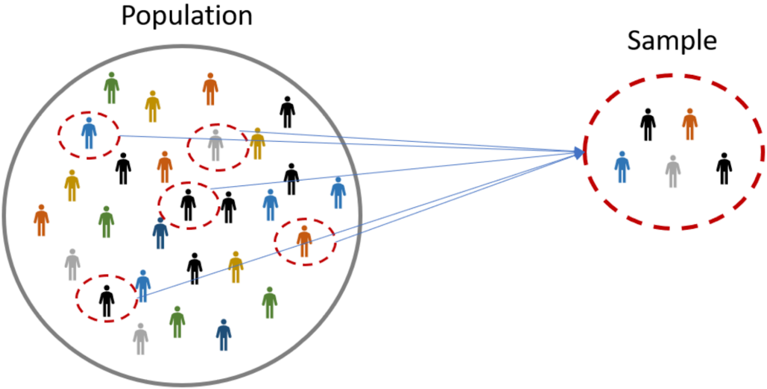
Overview

1. Motivation
2. Background mathematics
   - Population vs sample
   - Mean, standard deviation, variance
   - Covariance, covariance matrix
   - Eigenvectors, eigenvalues
3. Dimensionality reduction techniques
   - PCA (Principal component analysis)
   - SNE (Stochastic neighbor embedding)
4. Visualizations
   - IRIS dataset
   - MNIST dataset
5. Summary
   - Conclusion
   - Practicum

Motivation

- Relationships between data

Population vs sample



Population vs sample

Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

N is number of items in the population

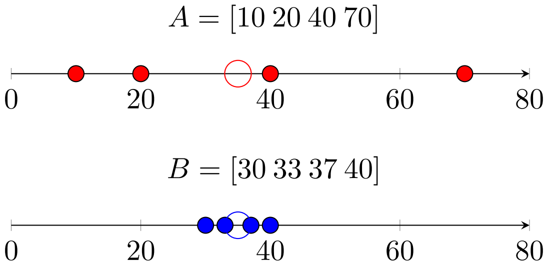
Sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

n is number of items in the sample

Standard deviation

Let's take a look on two samples:



Here, $\bar{A} = \bar{B} = 35$. Unfortunately, mean doesn't tell us a lot except for a middle point.

Standard deviation

For our two sets, $A = [10\ 20\ 40\ 70]$ and $B = [30\ 33\ 37\ 40]$, we would be more interested in the *spread* of the data. So, how do we calculate it?

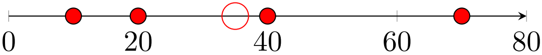
Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}} \tag{3}$$

In plain English, it is the "average distance from the mean of the data set to a point."

Standard deviation

Set 1: $A = [10\ 20\ 40\ 70]$, and $\bar{A} = 35$

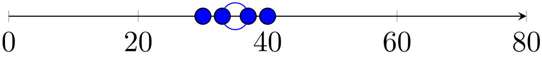


Let's calculate standard deviation:

A	$(A - \bar{A})$	$(A - \bar{A})^2$
10	-25	625
20	-15	225
40	5	25
70	35	1,225
Total		2,100
Divided by (n-1)		700
Square root		**26.4575**

Standard deviation

Set 2: $B = [30\ 33\ 37\ 40]$, and $\bar{B} = 35$



Let's calculate standard deviation:

B	$(B - \bar{B})$	$(B - \bar{B})^2$
30	-5	25
33	-2	4
37	2	4
40	5	25
Total		58
Divided by (n-1)		19.333
Square root		**4.397**

Variance

Similar to standard deviation So, how do we calculate it?

Variance

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} \tag{4}$$

Almost identical to the standard deviation (there is no square root in the formula).

Covariance

Previous measures are 1-dimensional. How do we check relationship between the dimensions?

Covariance

Variance

$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)} \tag{5}$$

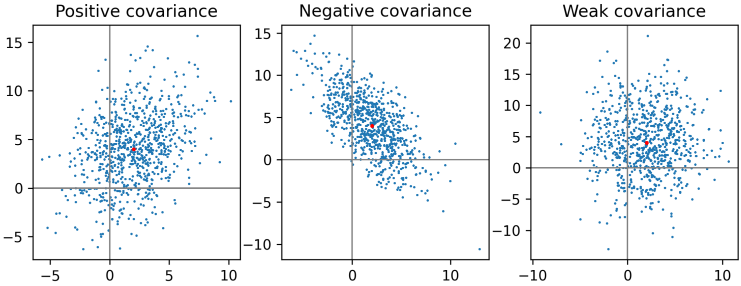
Covariance

$$covar(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \tag{6}$$

How much dimensions vary from the mean *with respect to each other*.

- Covariance between one dimension and itself gives variance.

Covariance


Positive covariance
Negative covariance
Weak covariance
15
10
5
0
−5
−5
0
5
10
15
10
5
0
−5
−10
0
10
20
15
10
5
0
−5
−10
−10
0
10

Covariance

Exercise: Find covariance between 2-dimensional dataset

Item number:	1	2	3	4	5
x	10	39	19	23	28
y	43	13	32	21	20

Ans: -120.55

Covariance matrix

When we have more than two dimensions, covariances between individual dimensions could be described in covariance matrix

Covariance matrix

$$C^{n\times n} = (c_{i,j} : c_{i,j} = cov(Dim_i, Dim_j)) \tag{7}$$

where $C^{m\times n}$ is a matrix with n rows and m columns. Example for three dimensions:

$$c = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix} \tag{8}$$

Since $cov(a,b) = cov(b,a)$, the matrix is symmetrical about the main diagonal.

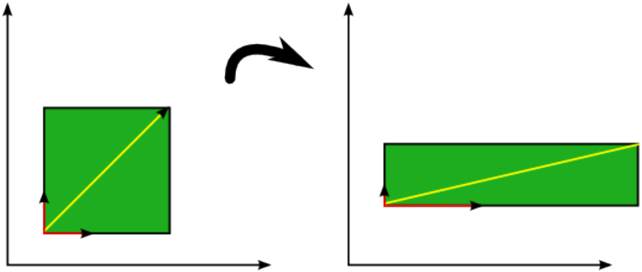
Covariance matrix

Exercise: Find covariance matrix for the 3-dimensional dataset

Item number:	1	2	3
x	1	-1	4
y	2	1	3
z	1	3	-1

Ans: $$\begin{pmatrix} 4.222 & 1.667 & -3.333 \\ 1.667 & 0.667 & -1.333 \\ -3.333 & -1.333 & 2.667 \end{pmatrix}$$

Eigenvectors



- Eigenvectors (red) do not change direction when a linear transformation (e.g. scaling) is applied to them. Other vectors (yellow) do.

Eigenvectors

An example of non-eigenvector:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \boxed{\begin{pmatrix} 1 \\ 3 \end{pmatrix}} = \begin{pmatrix} 11 \\ 5 \end{pmatrix} \tag{9}$$

An example of eigenvector:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \boxed{\begin{pmatrix} 3 \\ 2 \end{pmatrix}} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} \tag{10}$$

Eigenvalues

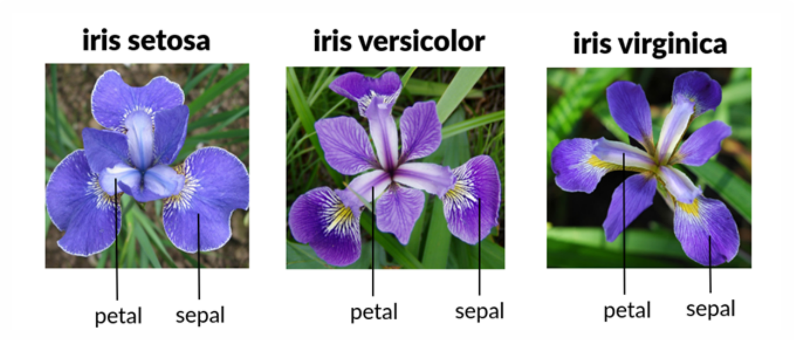
The amount by which the original vector was scaled after multiplication by the square matrix

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = \boxed{4} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} \tag{11}$$

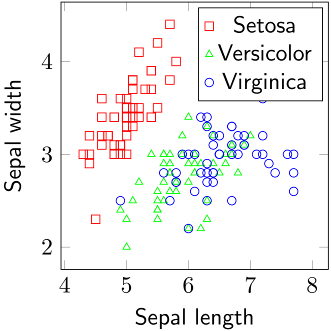
PCA

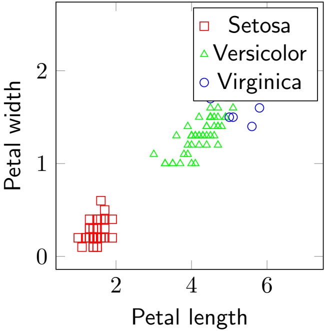
Iris flower dataset

Iris flower dataset

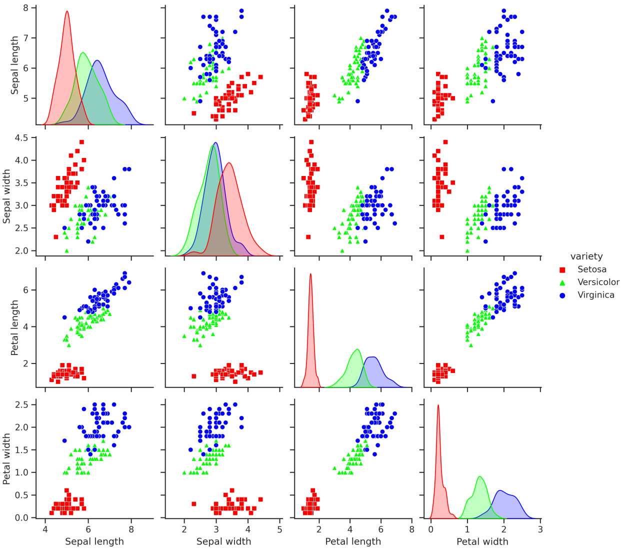


Iris flower dataset





Iris flower dataset



Adagrad

Experiment

Experiment

Experiment

Conclusion

Practicum

Thank you for your attention!

- Workshop contents:
  https://github.com/CodeSeoul/machine-learning/tree/master/221210-pca
- Follow-up QA?
  http://discord.com/users/tuttelikz

References

Ning Qian. "On the momentum term in gradient descent learning algorithms". In: *Neural networks* 12.1 (1999), pp. 145–151.

John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization.". In: *Journal of machine learning research* 12.7 (2011).

Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (2016).

Mykola Novik. *torch-optimizer – collection of optimization algorithms for PyTorch.* Version 1.0.1. Jan. 2020.