

zenius



Kampus
Merdeka
INDONESIA JAYA

Unsupervised Machine Learning

Day 15

Machine Learning Class

Program Studi Independen Bersertifikat
Zenius Bersama Kampus Merdeka



1. **Type of Unsupervised Machine Learning**
2. **PCA Basics**
3. **PCA for Visualization**
4. **PCA for Reducing Data Size**
5. **Factor Analysis**
6. **K-Means Clustering**
7. **Homework Explanation**

Types of Unsupervised Machine Learning

Unsupervised vs Supervised

Unsupervised ML	Supervised ML
Does not have accuracy metrics	Have accuracy metrics
Does not have labeled data	Can be evaluated with a test set that has labeled target data
No 'absolute source of truth'	There is an 'absolute source of truth' to compare prediction

Unsupervised vs Supervised

Fraud Detection Case 1

You are a Data Scientist tasked to design a Machine Learning model that can identify FRAUDULENT PAYMENTS.

You are given:

- **A dataset consisting of 1000 non-fraud payments and 100 fraud payments**

Is this a supervised or unsupervised ML?



Unsupervised vs Supervised

Fraud Detection Case 2

You are a Data Scientist tasked to design a Machine Learning model that can identify FRAUDULENT PAYMENTS.

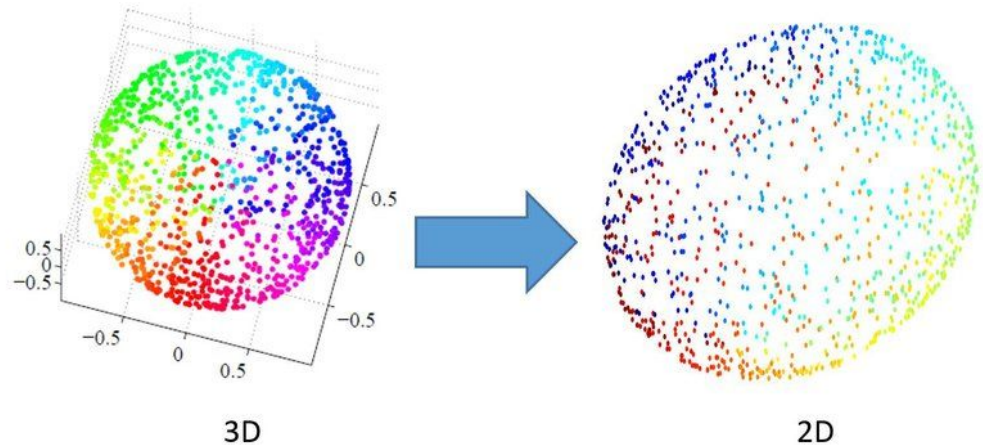
You are given:

- **A dataset consisting of 2000 payments, and are tasked to cluster these payments into 2 groups, fraud and non-fraud. You don't know which of these 2000 are actually fraudulent or not.**

Is this a supervised or unsupervised ML?

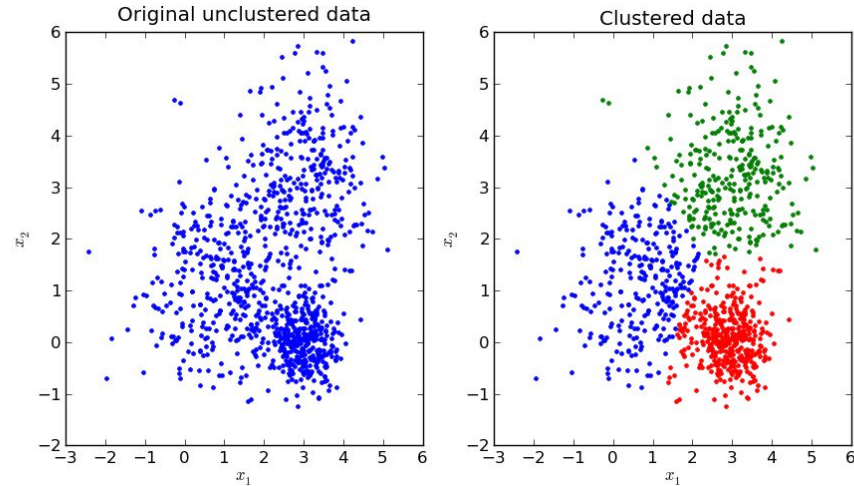
Type 1: Dimensionality Reduction

Given a dataset with a lot of columns, try to reduce the number of features but still retaining as much information as possible.



Type 2: Clustering

Given a bunch of unlabeled data, create clusters that can group similar data together.



PCA Basics

Intuition

The basic type of Dimensionality Reduction.

Suppose you have a dataset that has 50 columns. You want to reduce it to only 2-3 columns. If you delete 40+ columns, you lose 80% of information.

How to still retain the information, but reduce the dataset size? Principal Component Analysis.

Intuition

What for?

1. **Visualization.** If you have data with 5-10 columns, but you want to plot them in a 2D graph (x and y coordinate), you need to reduce them into having 2 principal components.

Intuition

What for?

2. **Reduction in size.** If you have a huge dataset, you might want to reduce the size to speed up model training process.

Intuition

Let's understand what is happening 'behind the scene' when we call the PCA function.

Before that, let's look at the 'animation' / 'visualization' in the following site to get a grasp on PCA better:

<https://setosa.io/ev/principal-component-analysis/>

Simple PCA

F1	F2	F3	F4
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

OK.

Misalkan kita punya dataset seperti ini. 4 Fitur, 5 baris.

Dan kita akan “reduce” the dataset menjadi 2 “Principal Component” saja.

Simple PCA

F1	F2	F3	F4
-1	-0.63246	0	0.2602
0.333	1.26491	1.73205	1.56374
-1	0.63246	-0.57735	-0.17375
0.333	0	-0.57735	-1.04249
1.333	-1.26491	-0.57735	-0.60812

Selanjutnya, kita skalakan dataset tersebut sehingga memiliki rata-rata 0 dan standar deviasi 1. (Standardization)

Simple PCA

Hasil dari perhitungan Covariance Matrix:

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

Simple PCA

Selanjutnya, kita harus mencari **Eigenvalue** dan **Eigenvector** dari matriks kovarians ini.

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

Simple PCA

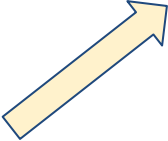
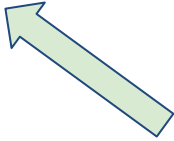
Pada matriks berukuran $n \times n$, maka ada n eigenvalue. Karena matriks kovarian tersebut berukuran 4×4 , maka kita memiliki 4 eigenvalue yang unique:

- **Eigenvalue 1: 2.5158**
- **Eigenvalue 2: 1.0653**
- **Eigenvalue 3: 0.393887**
- **Eigenvalue 4: 0.025**

Simple PCA

Jika kita ingin mengambil 2 Principal Component, maka pilihlah 2 **eigenvalue terbesar**.

Selanjutnya, carilah **eigenvector** untuk masing-masing **eigenvalue**, dan buatlah matriks seperti berikut:

Eigenvector dari eigenvalue 1		0.161960	-0.917059		Eigenvector dari eigenvalue 2
		-0.524048	0.206922		
		-0.585896	-0.320539		
		-0.596547	-0.115935		

Simple PCA

Data awal (setelah di-scale) x matriks eigenvector = Result (Hasil)

f1	f2	f3	f4		e1	e2		nf1	nf2
-1.000000	-0.632456	0.000000	0.260623		0.161960	-0.917059		0.014003	0.755975
0.333333	1.264911	1.732051	1.563740	*	-0.524048	0.206922	=	-2.556534	-0.780432
-1.000000	0.632456	-0.577350	-0.173749		-0.585896	-0.320539		-0.051480	1.253135
0.333333	0.000000	-0.577350	-1.042493		-0.596547	-0.115935		1.014150	0.000239
1.333333	-1.264911	-0.577350	-0.608121					1.579861	-1.228917
			(5,4)		(4,2)			(5,2)	

Data Hasil adalah dataset asal yang telah **direduksi** menjadi hanya 2 kolom saja.

Simple PCA

Untungnya, semua tahap yang kita lakukan sebelumnya telah di **automate** oleh Python.

Sehingga, kita hanya perlu belajar bagaimana cara menginterpretasikan hasil dari PCA yang telah dilakukan Python.

Kelemahan PCA

Cannot be used on categorical/one-hot encoded data. The aim of PCA is to preserve variance within numerical features so it still retains the information. Categorical/one-hot data simply does not have this.

Further read:

<https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6>

PCA For Visualization

PCA For Visualization: Hands On

Sesi Hands On 1 untuk PCA ini akan membahas aplikasi PCA untuk melakukan reduksi pada sebuah dataset agar dataset tersebut dapat divisualisasikan.

PCA For Reducing Data Size

PCA For Reducing Data Size: Hands On

Sesi Hands On 2 untuk PCA akan mendemonstrasikan bahwa selain untuk visualisasi, PCA juga dapat dilakukan untuk mengurangi ukuran file.

Pada sesi ini, kita akan membuat sebuah Logistic Regression dengan sebuah data yang besar.

Kita akan membandingkan:

- Directly fitting a Logistic Regression
- Doing PCA first to reduce data size >> and then doing Logistic Regression

Factor Analysis

Definition

Factor Analysis adalah sebuah teknik untuk melakukan Dimensionality Reduction dalam Unsupervised Machine Learning.

Jikalau dalam PCA kita mencari **principal component**, di Factor Analysis kita mencari **factors**, yaitu *variabel laten* yang mampu menjelaskan hubungan antara **variabel bebas**.

Intuition

Analogi:

IQ siswa diukur dan dicatat dalam sebuah tabel. Kemudian, nilai Bahasa Inggris, German, nilai Matematika, dan nilai Fisika juga dicatat dalam tabel tersebut.

Intuition

Maka, menggunakan FA, kita bisa meng'ekstrak' 2 factor utama yang memengaruhi IQ, yaitu:

- Factor 1: Kemampuan Berbahasa
- Factor 2: Kemampuan Sains

FA akan membuat 2 'variable baru'.

Intuition

Di dalam contoh yang kompleks, mungkin penemuan 'factor' tidak sesederhana contoh di analogi kita. Oleh sebab itu, algoritma ini dapat membantu kita menganalisa apakah terdapat 'factor-factor' yang 'tersembunyi' di dalam data yang multivariate.

Intuition

Ambil contoh analogi IQ dan nilai B.Inggris, German, Matematika, dan Fisika.

If this is PCA...maka...

Principal Component 1 dan Principal Component 2 adalah 'kombinasi linear dari keempat variable'.

$$PC\ 1 = k_1 * nilai_inggris + k_2 * nilai_german + k_3 * nilai_matematika + k_4 * nilai_fisika$$

$$PC\ 2 = k_5 * nilai_inggris + k_6 * nilai_german + k_7 * nilai_matematika + k_8 * nilai_fisika$$

Dengan k_1 sampai k_8 = koefisien dari matriks transformasi

Intuition

Ambil contoh analogi IQ dan nilai B.Inggris, German, Matematika, dan Fisika.

If this is FA...maka...

Nilai b.Inggris	= Factor 1 * k1 + Factor 2 * k2 + error
Nilai b.German	= Factor 1 * k3 + Factor 2 * k4 + error
Nilai matematika	= Factor 1 * k5 + Factor 2 * k6 + error
Nilai fisika	= Factor 1 * k7 + Factor 2 * k8 + error

Factor 1 dan 2 adalah komponen 'penyusun' keempat variable awal. (Algoritma FA akan mencari Factors yang memperkecil error).

Perbedaan PCA vs FA

Principal Component **dibentuk dari** kombinasi linear masing-masing variabel bebas.

Factor adalah **yang membentuk** masing-masing variabel bebas.

Perbedaan PCA vs FA

Kapan pakai Factor Analysis?

- Ketika kita berasumsi bahwa ada 'factor-factor' yang melatarbelakangi banyaknya variable bebas

Kapan kita pakai PCA?

- Ketika tujuan utama kita adalah mereduksi dimensi dengan mempertahankan sebanyak mungkin 'informasi' (variance) dalam data

Factor Analysis: Hands On

Pada Hands-On Factor Analysis, kita akan mencari faktor-faktor yang mungkin muncul dari sebuah dataset HRD yang mencatat personality dari masing-masing pelamar kerja.

K-Means Clustering

Definition

Clustering:

Process of dividing entire data...

**...into groups (known as clusters) based
on similarity and observed patterns.**

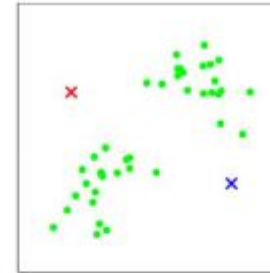
Steps

Step-By-Step K-Means Clustering:

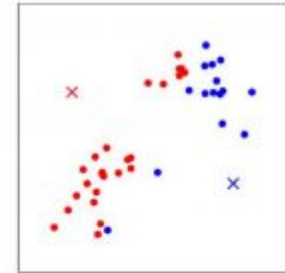
1. Memilih 'k' posisi 'acak' untuk dijadikan 'cluster center'
2. Data points dipisahkan berdasarkan jarak mereka ke masing-masing 'cluster center'
3. Lokasi cluster center diubah dengan mencari 'titik tengah' dari titik-titik yang telah dikelompokkan
4. Karena 'cluster center' berubah lokasi, maka pengelompokan pun akan berubah
5. Lakukan Step 2-3-4 sampai tidak ada titik yang 'berubah kelompok' lagi.



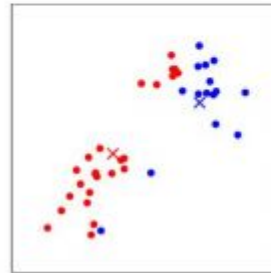
(a)



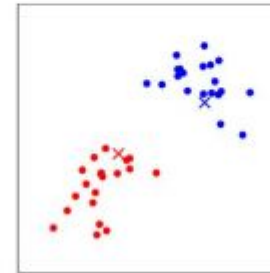
(b)



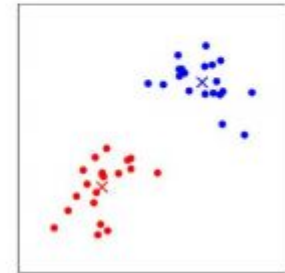
(c)



(d)



(e)

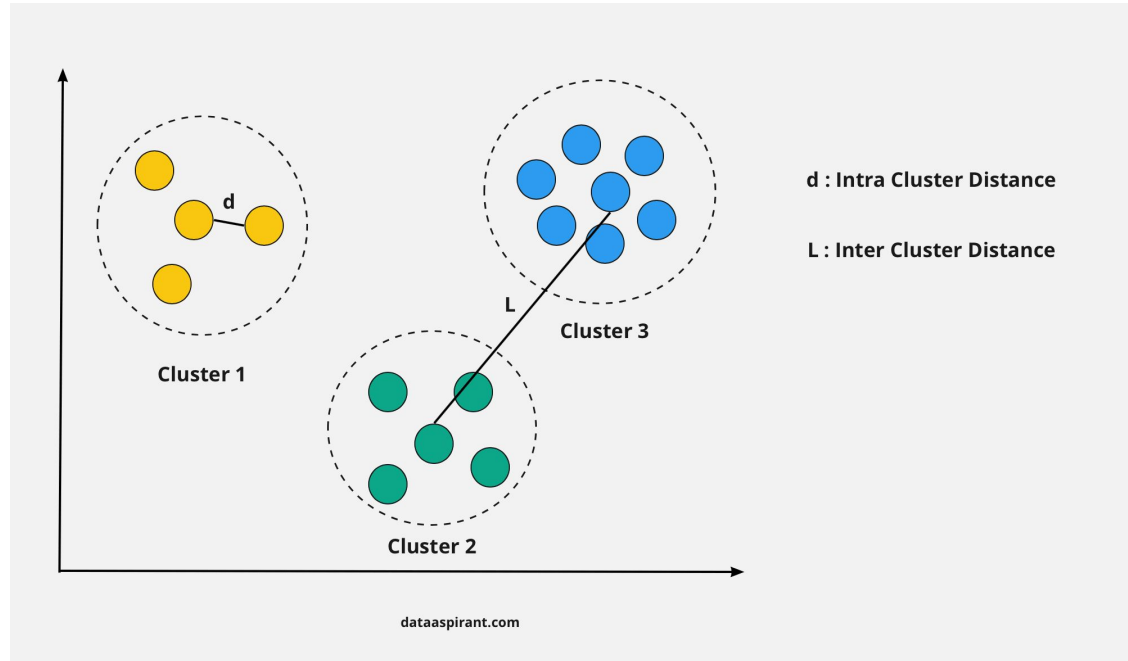


(f)

Principles

K-Means Clustering:

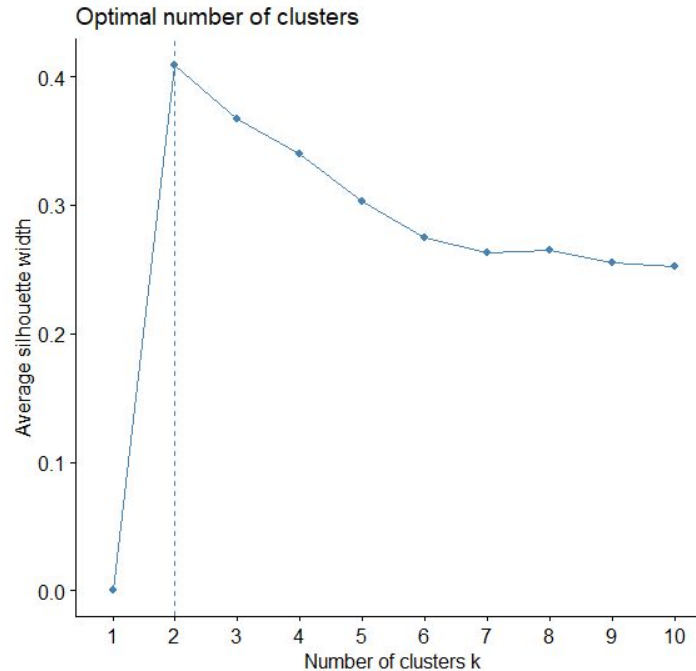
- Meminimalkan 'intra-cluster distance'
- Memaksimalkan 'inter-cluster distance'



Silhouette Method

Memiliki nilai kecil jika intercluster distance kecil.

Memiliki nilai tinggi jika intercluster distance tinggi.



Challenges

- Terkadang 'k' hasil Elbow Method berbeda dengan 'k' optimal hasil Silhouette Method
- Terkadang, stakeholder yang ingin menetapkan 'banyaknya cluster yang harus dibentuk'
- Penentuan 'k' optimal memang proses yang tidak mudah dan memerlukan banyak pertimbangan technical maupun 'business side'

K-Means Clustering: Hands On

Pada K-Means Hands On, kita akan :

Mencoba menentukan jumlah 'k' yang optimal untuk melakukan segmentasi pada mengunjung sebuah mall

Terima kasih!

Ada pertanyaan?

zenius



Kampus
Merdeka
INDONESIA JAYA

Assignment

zenius



Kampus
Merdeka
INDONESIA JAYA



Assignment

Pada assignment kali ini, kita akan:

- Melakukan PCA pada sebuah dataset
- Melakukan k-means terhadap dataset yang telah di-reduksi dimensinya (dari PCA)
- Menentukan 'k' yang optimal dari hasil k-means clustering