

CPMK 07.1 · SEMESTER 5 & 7

Dimensionality Reduction

Feature Extraction & Reduksi Dimensi Data

IF261220 3 SKS · Program Studi Informatika, Fakultas Ilmu Komputer

UPN "Veteran" Jawa Timur · Semester Gasal 2026/2027

Melanjutkan Pertemuan 1 · Konsep Dasar Machine Learning

PERTEMUAN

02

SEBELUM MULAI

Recap Pertemuan 1

Kontrak & CPMK

Tiga CPMK: menjelaskan (07.1), menerapkan (09.1), mengevaluasi (10.1).

Definisi ML

Belajar dari pengalaman E terhadap tugas T, diukur dengan performa P.

Alur Kerja ML

Rumusan masalah → data → pra-pemrosesan → pelatihan → evaluasi → penerapan.

AGENDA PERTEMUAN 2

01

Mengapa dimensi perlu direduksi

02

Feature selection vs feature extraction

03

Metode-metode utama (PCA, LDA, t-SNE, autoencoder)

04

Memilih metode sesuai studi kasus

Kutukan Dimensi (Curse of Dimensionality)

Semakin banyak fitur, semakin jarang data mengisi ruangnya.

Menambah fitur membuat ruang data tumbuh secara eksponensial, sementara jumlah sampel tetap. Data menjadi jarang (sparse), jarak antar-titik menjadi kurang bermakna, dan model menjadi mudah overfitting.

Contoh sederhana: 10 sampel cukup mengisi garis 1 dimensi, tetapi pada ruang 100 dimensi, 10 sampel nyaris tidak berarti apa-apa.

Komputasi membengkak

Waktu latih dan kebutuhan memori naik tajam seiring bertambahnya fitur.

Risiko overfitting naik

Model menghafal noise/derau pada fitur yang tidak relevan, bukan pola sesungguhnya.

Visualisasi yang mustahil

Manusia tak dapat membayangkan ruang berdimensi lebih dari tiga secara langsung.

Apa itu Dimensionality Reduction?

Proses mengurangi jumlah fitur (variabel) dalam sebuah dataset, sambil mempertahankan informasi penting yang dikandungnya.

01

Menyederhanakan Model

Fitur lebih sedikit membuat model lebih cepat dilatih dan lebih mudah diinterpretasi.

02

Mengurangi Redundansi

Fitur yang saling berkorelasi tinggi sering membawa informasi yang tumpang tindih.

03

Mempermudah Visualisasi

Data berdimensi tinggi diproyeksikan ke 2 atau 3 dimensi agar dapat diamati.

04

Mengurangi *Noise*

Fitur yang kurang informatif dan mengandung derau dapat dibuang atau digabung.

Dimensionality Reduction merupakan bagian dari tahap pra-pemrosesan, sebelum data masuk ke tahap pelatihan model.

Feature Selection vs Feature Extraction

Feature Selection

Memilih sebagian fitur asli yang paling relevan, sisanya dibuang.



F1, F3, F6 dipertahankan · F2, F4, F5 dibuang

- Fitur hasil tetap dapat dibaca dan dimaknai secara langsung.
- Pendekatan: filter, wrapper, dan embedded.

Contoh: memilih 10 dari 50 gejala klinis paling berkorelasi dengan diagnosis.

Feature Extraction

Membentuk fitur baru hasil transformasi dari fitur-fitur asli.



6 fitur asli dikombinasikan menjadi 2 fitur baru (Z1, Z2)

- Fitur baru adalah kombinasi matematis, bukan fitur asli.
- Pendekatan: PCA, LDA, t-SNE, autoencoder.

Contoh: menggabungkan ratusan piksel citra menjadi beberapa komponen utama.

Tiga Pendekatan Feature Selection

Filter

Menilai fitur secara independen dari model, memakai ukuran statistik seperti korelasi atau uji chi-square.

- Cepat dan murah secara komputasi
- Tidak bergantung pada algoritma tertentu
- Contoh: korelasi, chi-square, information gain

Wrapper

Mencoba berbagai kombinasi subset fitur dan menilai performanya langsung menggunakan model.

- Hasil lebih akurat untuk model tertentu
- Lebih mahal karena melatih model berulang kali
- Contoh: forward selection, backward elimination

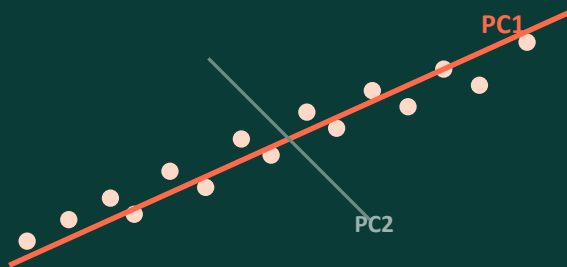
Embedded

Seleksi fitur menyatu di dalam proses pelatihan model itu sendiri.

- Menyeimbangkan biaya dan akurasi
- Fitur dipilih sebagai bagian dari optimisasi model
- Contoh: Lasso (L1), feature importance pada Decision Tree

Principal Component Analysis

Metode feature extraction paling populer dan menjadi dasar banyak metode lainnya.



PC1 menangkap variasi terbesar; PC2 menangkap sisanya secara tegak lurus.

Ide dasar: mencari arah (komponen) yang menyimpan variasi data paling besar, lalu memproyeksikan data ke arah-arah tersebut.

01 Standardisasi Data

Menyamakan skala tiap fitur agar tidak ada yang mendominasi karena satuannya.

02 Hitung Matriks Kovarians

Mengukur bagaimana pasangan fitur saling berhubungan satu sama lain.

03 Cari Komponen Utama

Menentukan eigenvector dan eigenvalue sebagai arah dan besar variasi data.

04 Proyeksikan Data

Memilih k komponen dengan variasi terbesar dan memproyeksikan data ke sana.

PCA: Sebelum dan Sesudah Reduksi

SEBELUM

Ruang fitur asli berdimensi tinggi

- Puluhan hingga ratusan fitur, banyak yang saling berkorelasi.
- Sulit divisualisasikan dan mahal secara komputasi.
- Contoh: 784 piksel pada citra digit tulisan tangan 28×28.

VARIANS TERJELASKAN

2 komponen $\approx$ 85%

SESUDAH

Ruang komponen utama, dimensi rendah

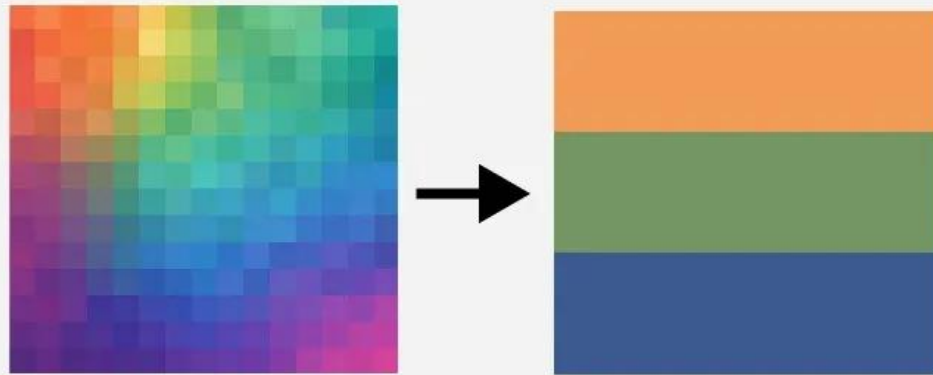
- Direduksi menjadi 2–3 komponen utama saja.
- Sebagian besar informasi (varians) tetap dipertahankan.
- Kini dapat divisualisasikan dan dilatih jauh lebih cepat.

TRADE-OFF

Sedikit informasi hilang

Menentukan jumlah komponen adalah soal menyeimbangkan penyederhanaan dengan informasi yang boleh hilang.

How PCA Works



Multiple Colours

Dominant Colours

- PCA menggunakan aljabar linier untuk transformasi data ke fitur baru, yang disebut *principal component*
- Caranya dengan menghitung *eigenvectors* (arah) dan *eigenvalues* (derajat kepentingan) dari matriks *covariance*
- PCA memilih komponen teratas dengan nilai *eigenvalues* tertinggi dan memproyeksikan data terhadap komponen tersebut untuk menyederhanakan data

Catatan: PCA memprioritaskan arah data yang memiliki variasi terbanyak, karena semakin banyak variasi → semakin banyak informasi berguna

Langkah-langkah PCA:

1. Standarisasi data

$$Z = \frac{X - \mu}{\sigma}$$

μ : mean of independent features $\mu = \{\mu_1, \mu_2, \dots, \mu_m\}$

σ : standard deviation of independent features $\sigma = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$

2. Menghitung matriks covariance

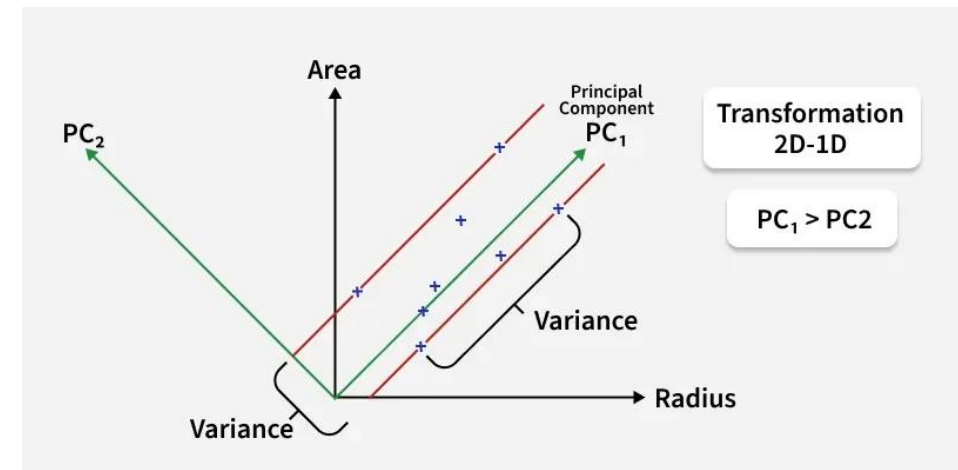
$$\text{cov}(x1, x2) = \frac{\sum_{i=1}^n (x1_i - \bar{x1})(x2_i - \bar{x2})}{n - 1}$$

3. Mencari *principal components* (PC)

1st PC: Arah dari varians paling maksimum (paling tersebar)

2nd PC: Arah terbaik kedua, dst.

4. Memilih arah yang terbanyak dan mentransformasi data

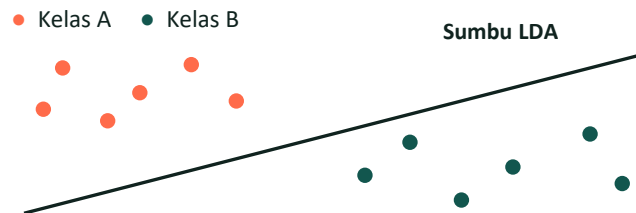


Metode Feature Extraction Lainnya

LDA

LINEAR DISCRIMINANT ANALYSIS

Mencari arah proyeksi yang memaksimalkan pemisahan antar-kelas, bukan hanya variasi data.



PALING COCOK UNTUK

Klasifikasi berlabel

t-SNE

T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

Menjaga kedekatan lokal antar-titik saat diproyeksikan ke ruang rendah, cocok untuk visualisasi.

PALING COCOK UNTUK

Visualisasi eksploratif

Autoencoder

JARINGAN SARAF TIRUAN

Melatih jaringan untuk mengompresi lalu merekonstruksi data; lapisan tengah menjadi representasi ringkas.

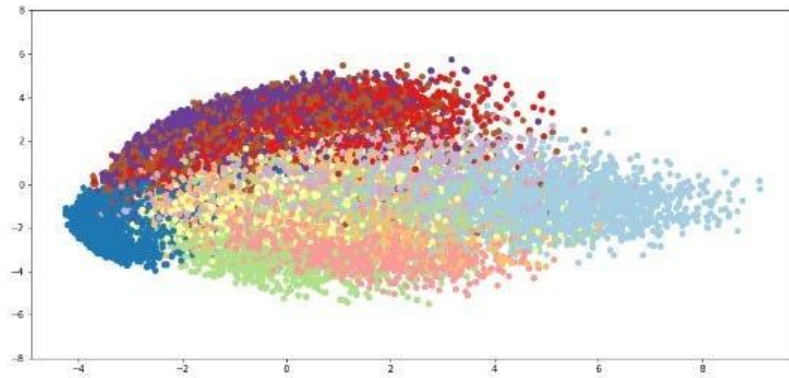
PALING COCOK UNTUK

Data kompleks & nonlinear

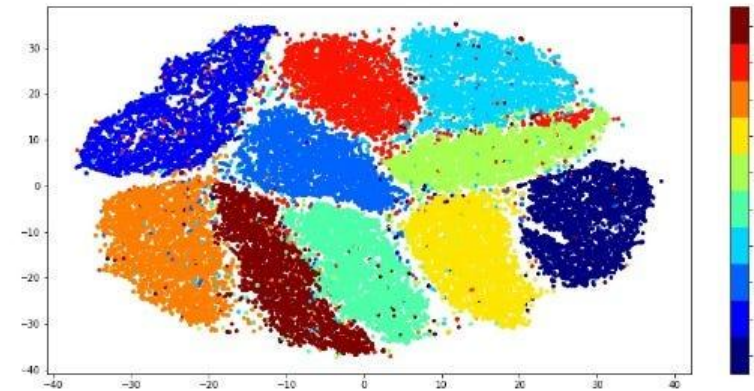
PCA dan LDA bersifat linear; t-SNE dan autoencoder mampu menangkap pola nonlinear yang lebih kompleks.

Hasil partisi PCA vs TSNE pada dataset digit MNIST

MNIST - PCA



MNIST - TSNE



Memilih Metode yang Tepat

01	Apakah fitur asli harus tetap dapat dimaknai?	Ya → feature selection. Tidak → feature extraction terbuka lebih luas.
02	Apakah data memiliki label kelas?	Ya, dan ingin memisahkan kelas → LDA. Tidak ada label → PCA, t-SNE, atau autoencoder.
03	Untuk apa hasilnya dipakai?	Visualisasi eksploratif → t-SNE. Input pelatihan model lanjutan → PCA atau autoencoder.
04	Apakah polanya diduga nonlinear?	Ya, dan data cukup banyak → autoencoder. Hubungan sederhana → PCA sudah memadai.

PRAKTIK

Mengapa Ini Penting untuk Project Anda

Reduksi dimensi sering menjadi langkah pertama yang menentukan kualitas seluruh model.

Citra & Video

Ratusan ribu piksel per gambar direduksi lewat PCA atau autoencoder sebelum diklasifikasikan.

Data Genomik / Kesehatan

Ribuan gen atau fitur klinis diseleksi memakai metode filter dan wrapper sebelum diagnosis.

Teks & Bahasa

Ribuan kata dalam representasi dokumen direduksi melalui embedding atau LSA (varian PCA).

Data Sensor / IoT

Puluhan sinyal sensor yang saling berkorelasi diringkas menjadi beberapa komponen utama.

Menentukan metode reduksi dimensi yang tepat sejak awal akan menghemat banyak waktu pada tahap pelatihan dan evaluasi.

Rangkuman dan Persiapan

SUDAH KITA PAHAM

- Mengapa dimensi tinggi bermasalah bagi model
- Perbedaan feature selection dan feature extraction
- Cara kerja PCA serta metode lain: LDA, t-SNE, autoencoder

CARA MEMILIH METODE

- Pertimbangkan interpretabilitas fitur asli
- Perhatikan ketersediaan label kelas
- Sesuaikan dengan tujuan akhir: visualisasi atau pelatihan

PERSIAPAN PERTEMUAN 03

- Materi: Supervised, Unsupervised, dan Semi-supervised Learning
- Coba jalankan PCA sederhana di Python pada dataset latihan
- Lanjutkan membaca jurnal untuk topik final project