PCA and VAE

Sunday, 8 February 2026 09:46

Dimensonality Reduction techs.

PCA

- a statistical technique used to reduce the dimensoinality of large data sets while preserving as much variance as possible.
  - reduces number of features in a data set while keeping the most important information
  - it changes complex data sets by transforming corelated features into smaller sets of uncorelated components.

- it helps us remove redundancy and improve computational efficieny while making the data easier to visulise.

- It uses linear algebra to transform data into principal components
- it does this by calculating eigen vectors (directions) and eigen values (importance) from the covariance matrix.

- Step 1 standardise the data
- Step 2 calculate covariance matrix
- Step 3 find the principal components.
- Step 4 Pick the top Directions and Transform Data

- Can be done in python using Sk learn

+ ve

1. multicollinearity Handling: Creates NEW uncorelated variables to adress issues when original features are highly corelated.
2. Noise Reduction: reduces components with low variance thus increasing data clarity.
3. Data Compression: Reduces data size
4. Outlier detection: identifies outliers.

-ve

1. Interpretation challenges: principal components are combinations so can be hard to explain
2. Data scaling sensitivity: Requires proper scaling of data or results will be misleading.
3. Information Loss: may lead to loss if too few components are kept.