DIMENSIONALITY REDUCTION

CURSE OF DIMENSIONALITY

Dataset with 400 features

The model is overfeeding (that's why the accuracy was decreasing)

TOW WAYS TO REMOVE IT

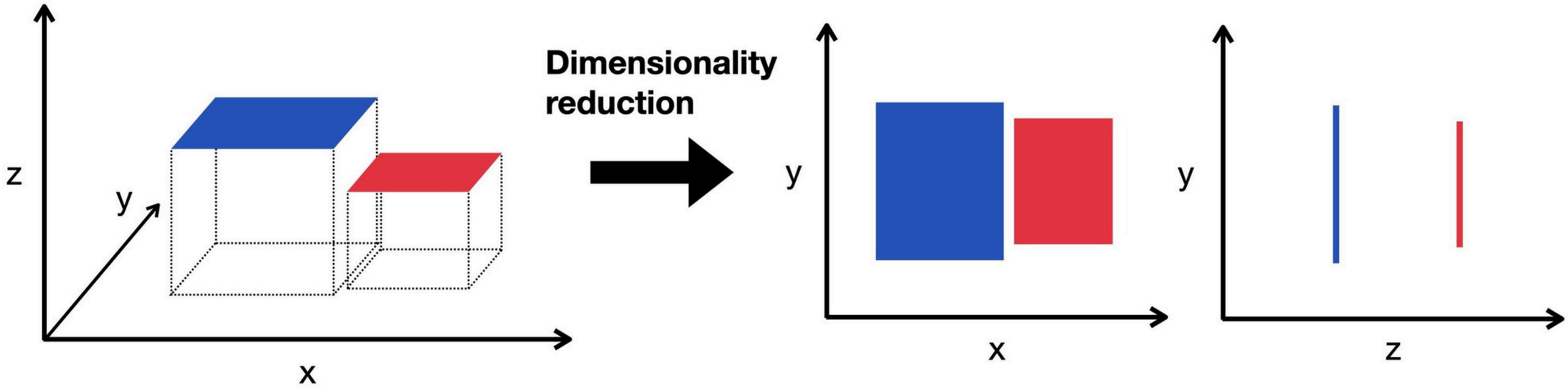
FEATURE SELECTION

Feature selection is the process of selecting the most relevant input features from a dataset that contribute the most to the prediction output. It helps improve model performance by reducing overfitting, computation time, and noise.

DIMENSIONALITY REDUCTION

Dimensionality reduction is the process of reducing the number of input variables (features) in a dataset while preserving as much relevant information as possible. It simplifies models, speeds up computation, and helps visualize high-dimensional data.

INTRODUCTION



Why dimensionality reduction?

- Prevent - Curse of Dimensionality
- Improve the performance of the model
- Visualize the data.

FEATURE SELECTION

X Y

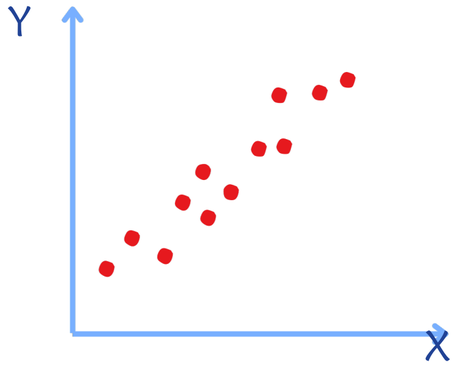
\- -

\- -

\- -

X increase Y Increase

X Decrease Y Decrease



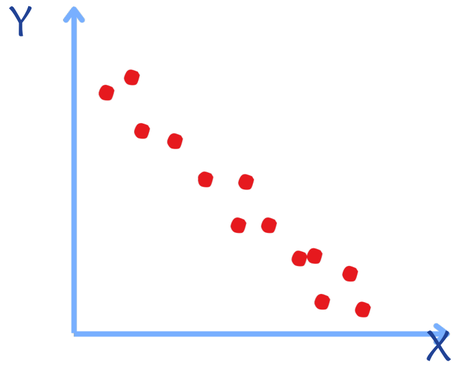
House Size

Fountain Size

Price

X increase Y Decrease

X Decrease Y Increase



House Size

Price

Covariance

Correlation

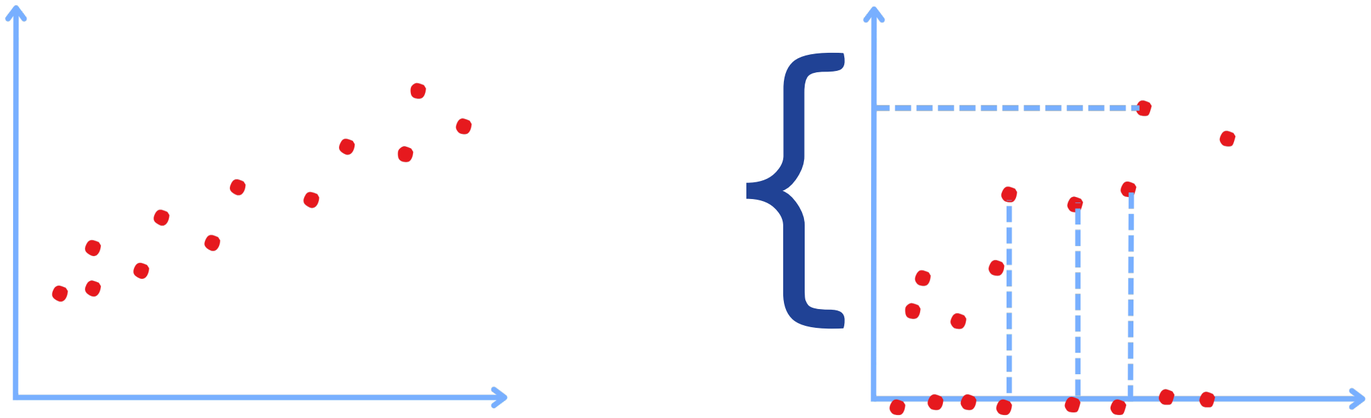
FEATURE EXTRACTION

Room Size No.of Rooms Price ⟶ House Size Price

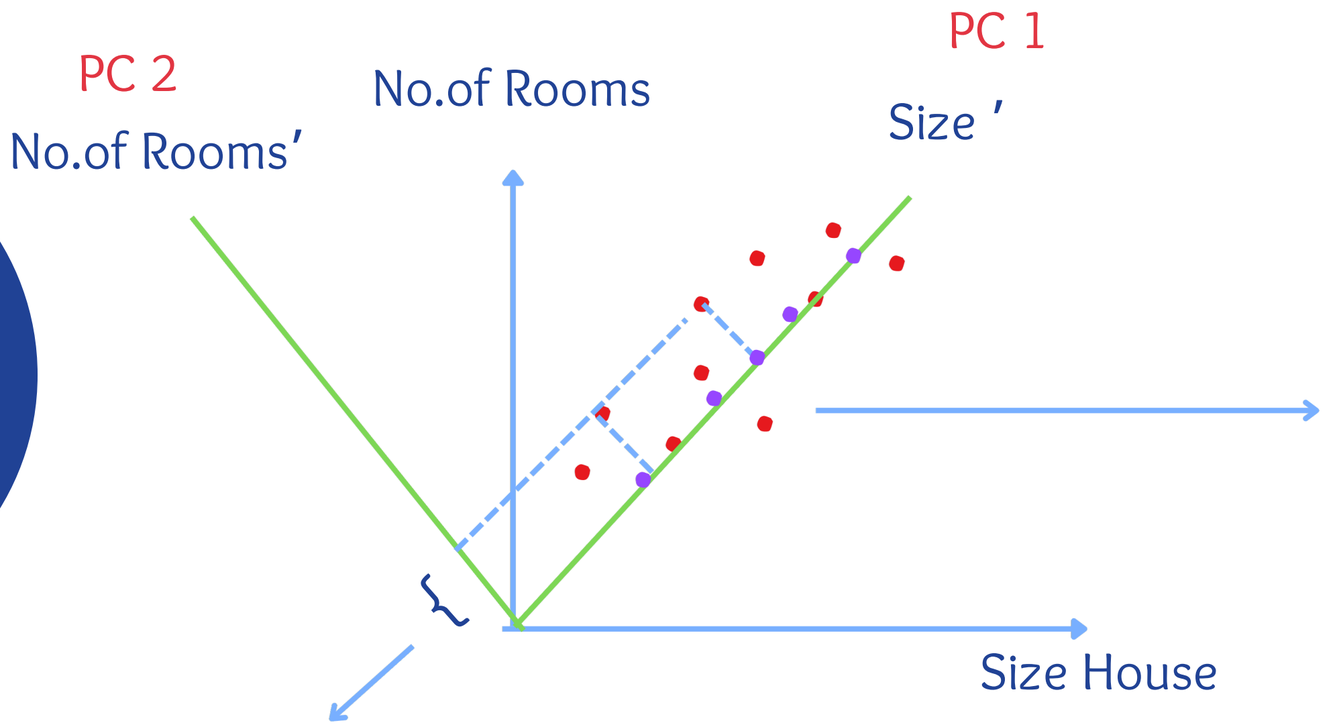
Transformation (to extract new feature)

Example:

Size house No.of Rooms Price



FEATURE EXTRACTION USING PCA

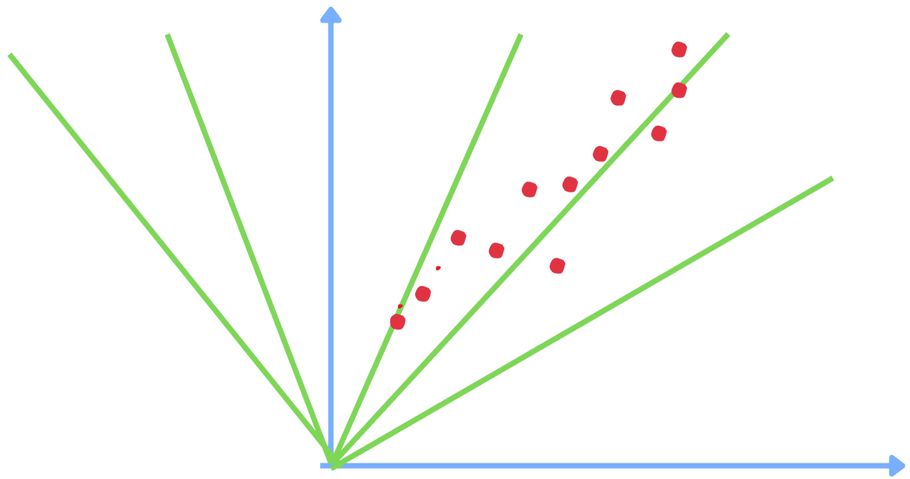


Eigen Decomposition on Matrix

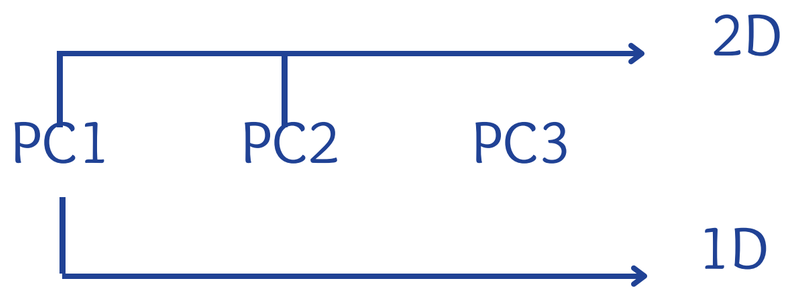
Maximum spreadness captured correctly

3 features → PC1, PC2, PC3

PC1 > PC2 > PC3



3D → 1D



To get the best component which captures maximum variance

MCQ

Q1. What does PCA primarily aim to do?

A) Maximize accuracy of the model

B) Increase the number of features

C) Reduce dimensionality while retaining most variance

D) Select only the most important original features

Q2. Which of the following best describes feature selection?

A) Transforming features into a new space

B) Reducing the number of data samples

C) Selecting a subset of relevant features for the model

D) Combining features to create new ones