

Dimensionality Reduction

Dimensionality Reduction

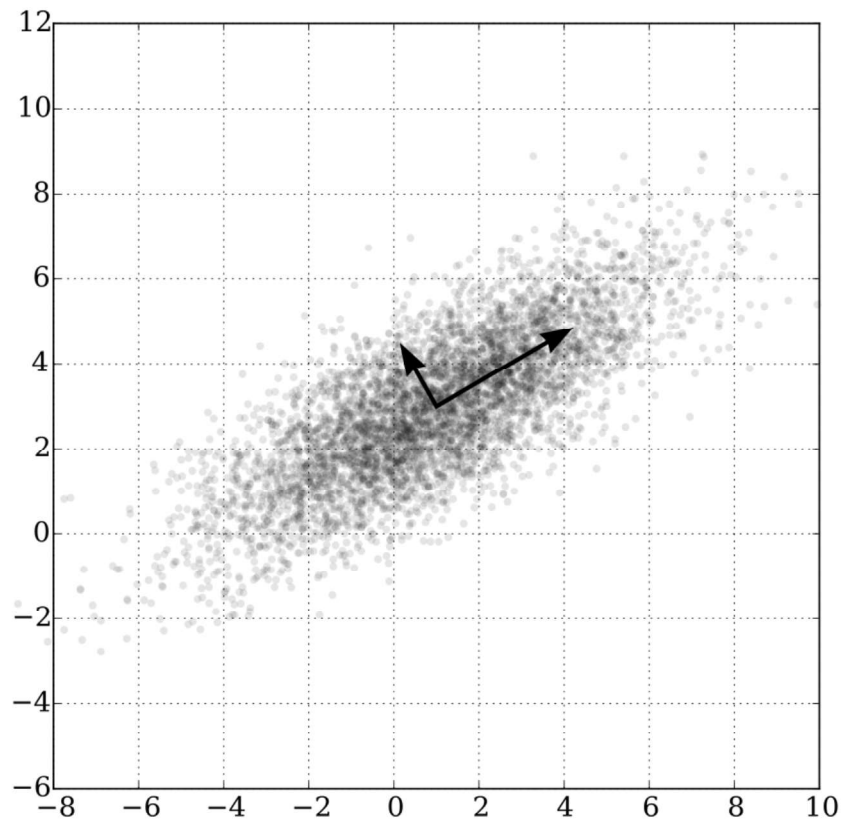
- The process of reducing the number of independent variables
- Reducing dimensionality of independent variables helps in many ways
 - removes multi-collinearity to improve ML model performance
 - helps reduce over fitting
 - decreases computational times for fitting models
 - makes visualization easier
 - decreases storage requirements
 - avoids curse of dimensionality
- Hence dimensionality reduction plays a significant role in analyzing data

Dim. Reduction Techniques

- Feature elimination
 - Simply identify and remove variables (columns) that are not important
 - The disadvantage is that we would gain no insight from those dropped variables and lose any information they contain
- Feature extraction
 - Create a few new variables from the old variables
 - **PCA** Principal Component Analysis: is the most popular feature extraction technique (linear)
 - t-SNE (non-linear)

PCA

- creates new variables using linear combinations of old variables
- is designed to create variables that are independent of one another
- also manages to tell us how important each of these new variables are
- this “importance”, helps us to choose how many variables we will use



- Scale the data and compute the covariance matrix
- Break the covariance matrix into magnitude and direction. Eigen Vectors and the Eigen Values of the covariance matrix can be thought of as the natural axis/directions and magnitudes along those axis, of the data
- The eigen values also can be used to calculate the percentage of variation explained by each component
- Sort in the eigen values in descending order and calculate the cumulative percentage of variation explained
- Pick the number of principal components you will use
- Transform to new variables

