Principal Component Analysis

Exploring the Concepts and Applications

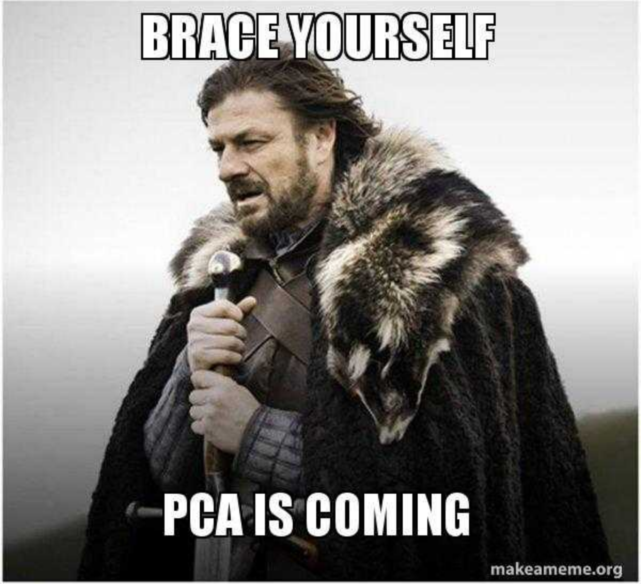
Aniket Nath,

National Institute of Science Education and Research Bhubaneswar

February 3, 2024

Outline

Introduction to PCA



A Terrible Curse

The Curse of Dimensionality

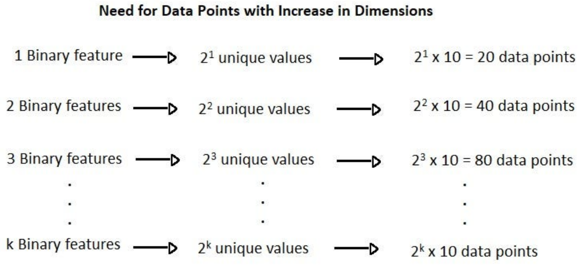


Figure: Scaling of datapoints with dimensions[1]

- Higher dimensional data needs more computational effort
- As the dimensionality increases, number of minimum data points for nominal analysis increases.

[1] https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb

Curse of Dimensionality

- Are dimensions to be discarded then?

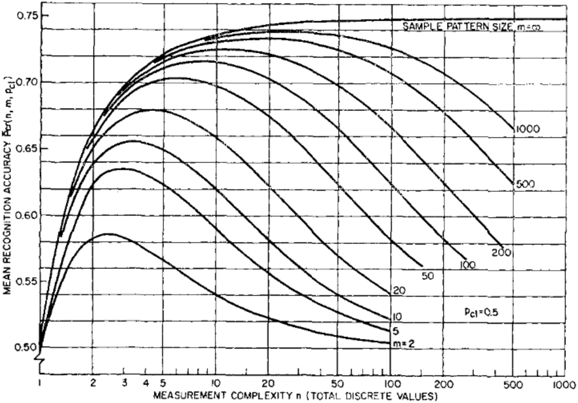
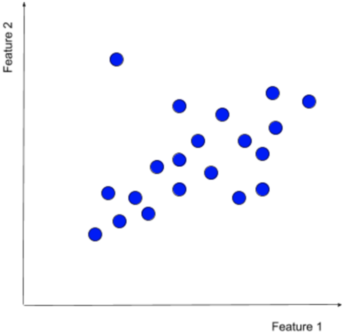
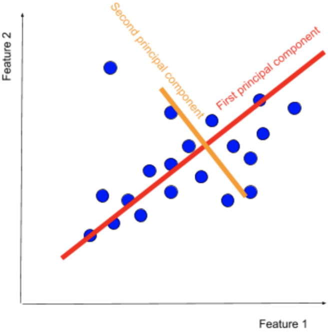


Figure: The tradeoff between dimensionality and number of datapoints (Hughes 1968).

Principal Component Analysis

Principal Component Analysis (PCA) is an algorithm to find best set of basis.





Principal Components are new variables that are constructed from linear combinations of the old feature dimensions.
PCA tries to assign maximum information in the first Principal Component, and gradually decreases over the later ones.
Geometrically, it represents directions which explain maximal amount of variance.


>Barges into any discussion or argument
>PCA is just the eigenvectors of the covariance matrix
>Refuses to elaborate further
>Leaves

- **Standardization**: This basically sets the scales in the data, so that each data contributes equally to it.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \tag{1}$$

- **Covariance Matrix Computation**: Calculate the covariance matrix over all the feature dimensions.

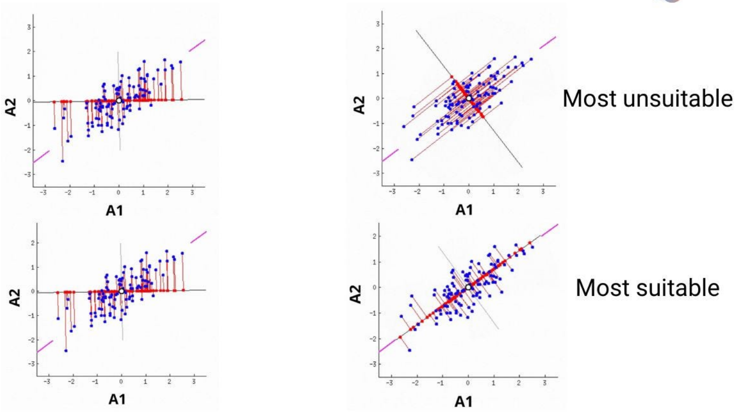
$$\text{covar}(x, y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n} \tag{2}$$

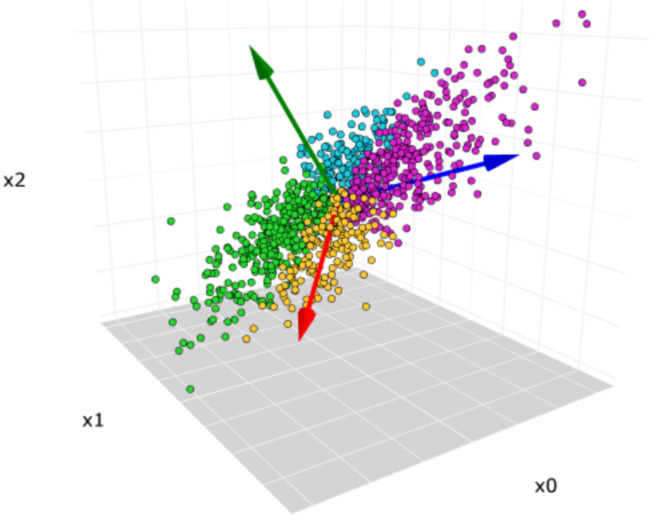
- **Compute the eigenvectors and eigenvalues of Covariance Matrix**: Then pick dominant eigenvectors, as per requirement.

Principal Component Analysis

Does not look very promising, for a two dimensional case, but this is extremely useful in higher dimensions.

enjoy algorithms




x2
x1
x0

Seems all Statistics!!

Applications of PCA

- Image compression: Calculate the PCA of the image, and remove components with less information.
- Feature reduction in machine learning.
- Anomaly detection.

PCA in Machine Learning

- PCA can be used to reduce the dimensions, and re-cast the data to newer dimensions. If the data size is lesser than dimensionality, we can try to do some trade-off.

Disadvantages

- Interpretability
- Information Loss
- Outliers affect PCA strongly
- Computationally expensive for big datasets.


PRESENTATION FINISHED !!
ANY QUESTIONS ?

Thank You

References I

Hughes, G. (Jan. 1968). "On the mean accuracy of statistical pattern recognizers". In: *IEEE Transactions on Information Theory* 14.1. Conference Name: IEEE Transactions on Information Theory, pp. 55–63. ISSN: 1557-9654. DOI: 10.1109/TIT.1968.1054102. URL: https://ieeexplore.ieee.org/document/1054102/citations#citations (visited on 02/03/2024).