Unsupervised Learning

Dimension Reduction

Why Dimension Reduction?

- So many dimensions - Curse of dimensionality
- How to summarize it without going into much details
- Eg 1. Group customers by purchasing patterns
- Eg 2. Compress the data (Dimension Reduction)
- Compression of data improves speed (much faster)
- Removes Noise Features

How can we reduce the features?

1. Multicollinearity : Many features are correlated hence same information (remove all but one)

2. Low Variance : If some feature has almost same data (very low variance) remove that

3. Most discriminative features : eyes in a face

Dimension reduction #1 : PCA

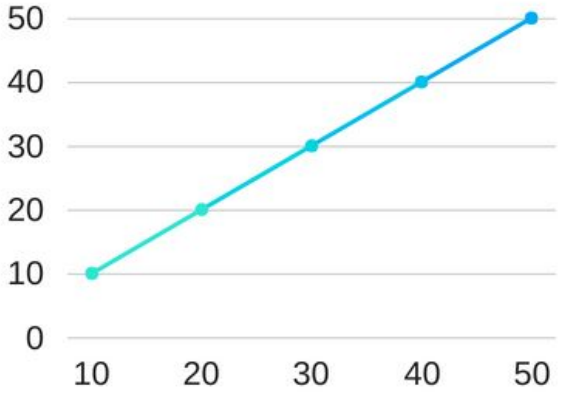
- PCA = Principal Component Analysis
- Finds Principal components of a data
- Does in 2 steps
  - De corelation : Doesn’t change data
  - Dimension reduction : compresses the data
- Decorrelation :
  - Transforms (rotation) the samples so that they align with axis
  - Shifts (translation) the samples to get the means as 0
  - No information is lost in this step

To Code and Beyond

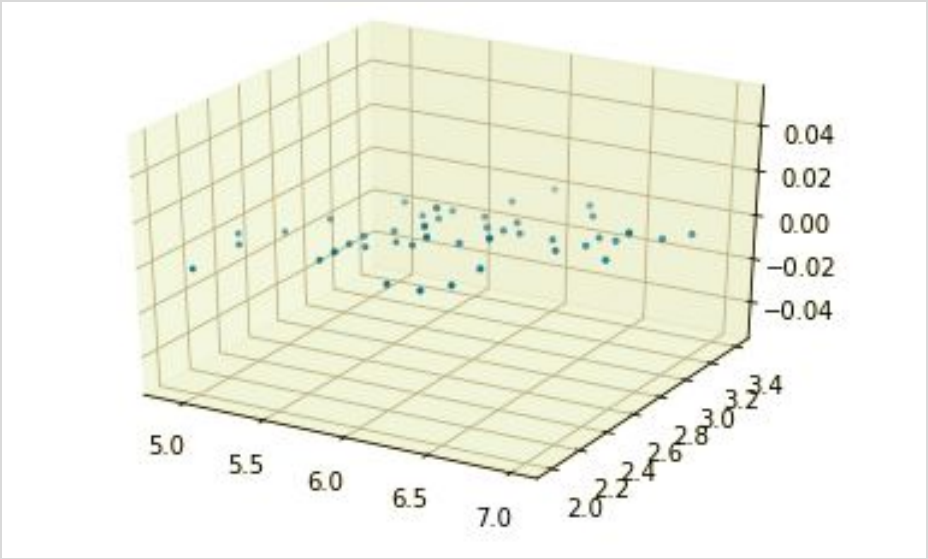
Intrinsic Dimension

- Intrinsic dimension = number of features needed to approximate the dataset
- Essential idea behind dimension reduction
- What is the most compact representation of the samples?
- Can be detected with PCA

Examples



Can be approximated using 1 feature

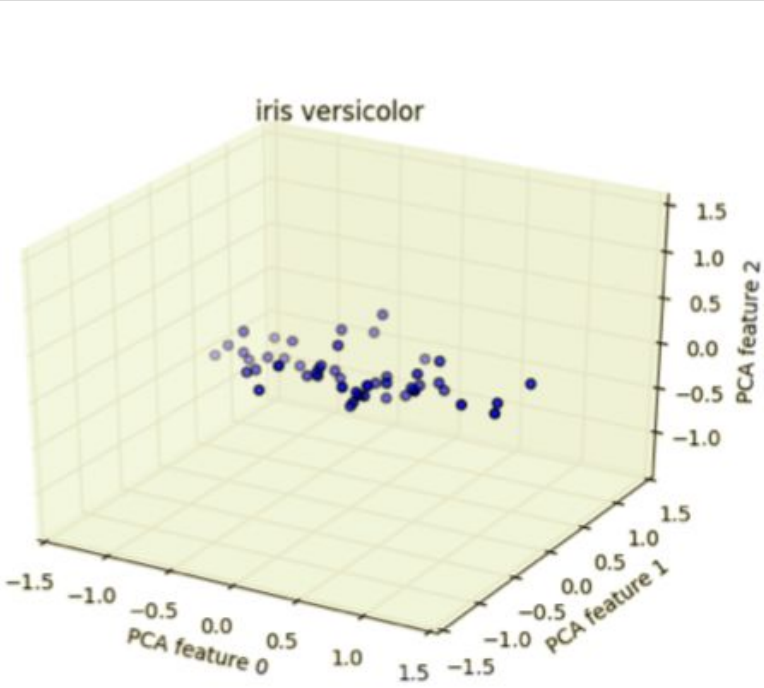


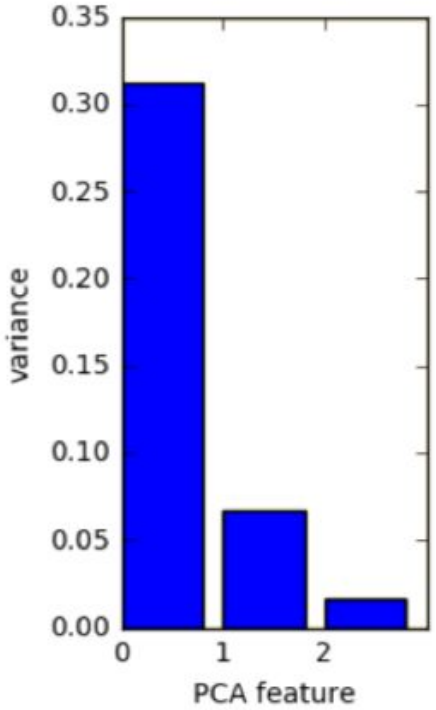
How many dimensions here?

PCA can find Intrinsic Dimension

- Humans can look at 2-d 3-d plots and identify
- What if there are many features
- PCA will help in that
- Intrinsic dimension = number of PCA features with significant variance

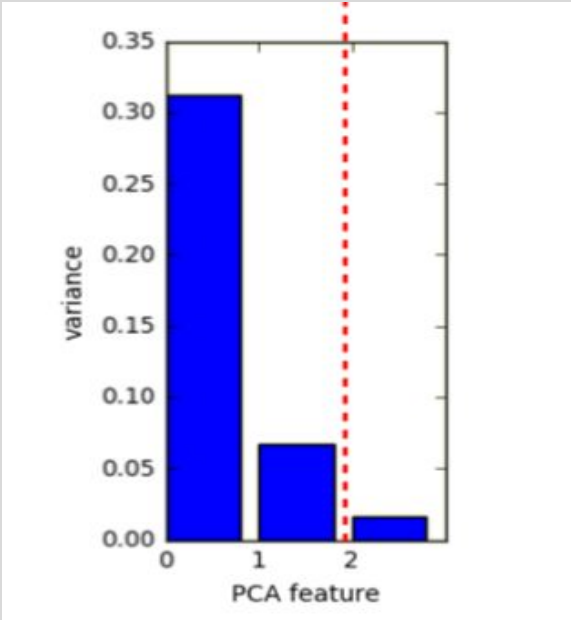
PCA features ordered by variances





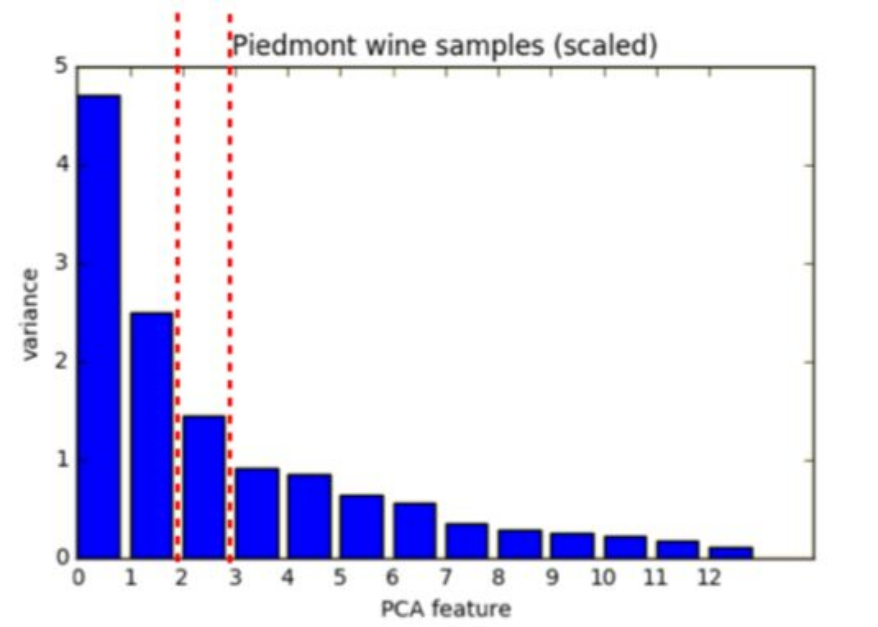
Variance and Intrinsic dimension

- Intrinsic Dimenision = Number of features with significant variance
- Last Example : 2
- Intrinsic dimension = 2



Intrinsic Dimension is ambiguous

- There is no correct answers but an approximation
- Example wine dataset



Dimension Reduction using PCA

Dimension Reduction

- Represents same data with less features
- Very highly used in ML Pipelines
- Can be done using PCA

Dimension Reduction with PCA

- PCA features are in decreasing order of variance
- Assumes low variance features are noise
- High Variance features are important

Steps For PCA

1. Specify how many features to keep (n_components)
    a. Intrinsic dimension is a good choice.
2. This will keep only those many features
3. Fit and Transform

To Code and Beyond

Dimension Reduction with PCA

- Assumes high variance features are important
- Assumption generally holds true
- Features can’t be get back after transformation
- Information is not lost

Assignment :

Cluster wikipedia article with TF-IDF and PCA