Unsupervised Learning

EDVANCER
EDUVENTURES

Copyright 2020 RESTRICTED CIRCULATION

Agenda

Discussion Flow

- What to do in absence of a Target
- Groups in Data and Distances
- Hierarchical Clustering and Limitation
- K-means
- DBScan
- Dimensionality Reduction with PCA

Copyright 2020 RESTRICTED CIRCULATION

When there is nothing to predict

Copyright 2020 RESTRICTED CIRCULATION

Finding Groups in Data

- Create more focused marketing campaigns
- Find Clusters of weather patterns
- Group similar documents
- Product Categorisation
- Detecting anomalies in the data
- Build separate predictive models for different data groups

Copyright 2020 RESTRICTED CIRCULATION

Reducing Dimensionality

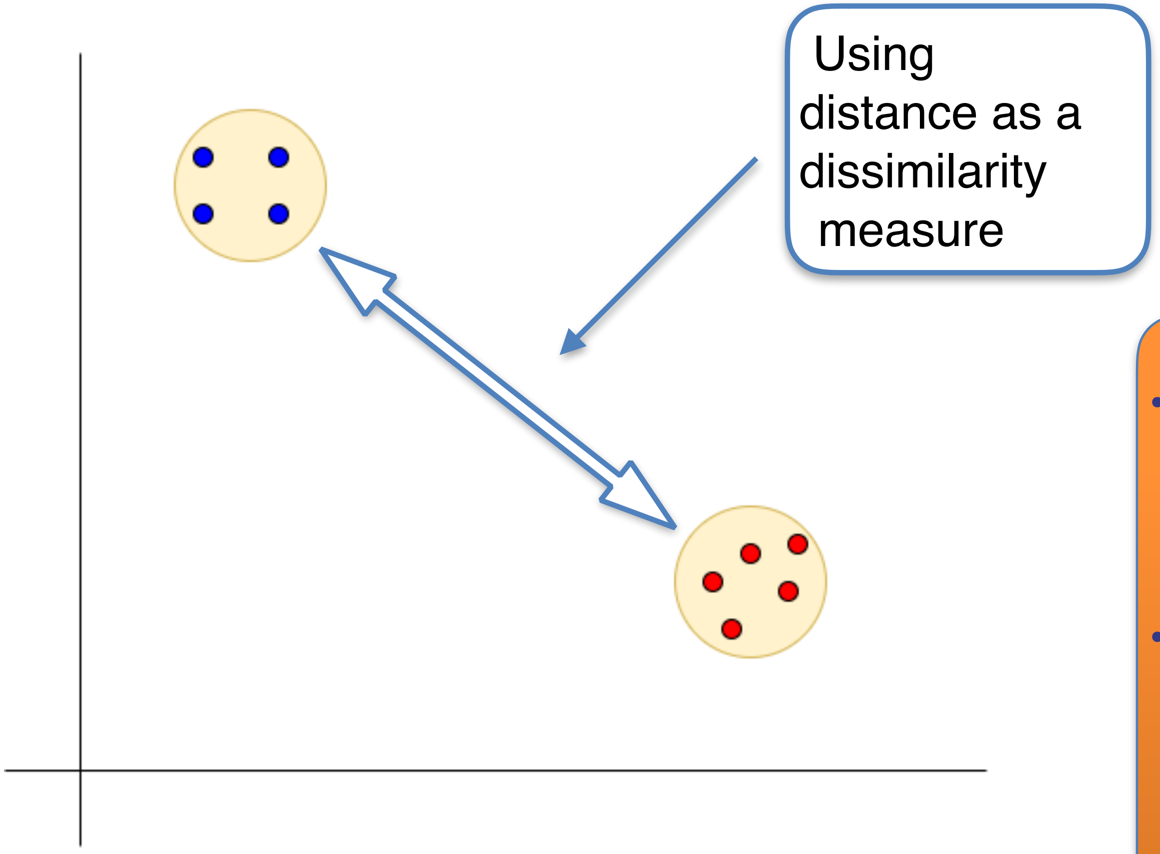
- To visualise data
- To get rid of redundancy in the information
- To get smaller data size for ease of experiments
- To reduce data to its latent factors

Copyright 2020 RESTRICTED CIRCULATION

Groups and distances

Copyright 2020 RESTRICTED CIRCULATION

How to group



Using distance as a dissimilarity measure

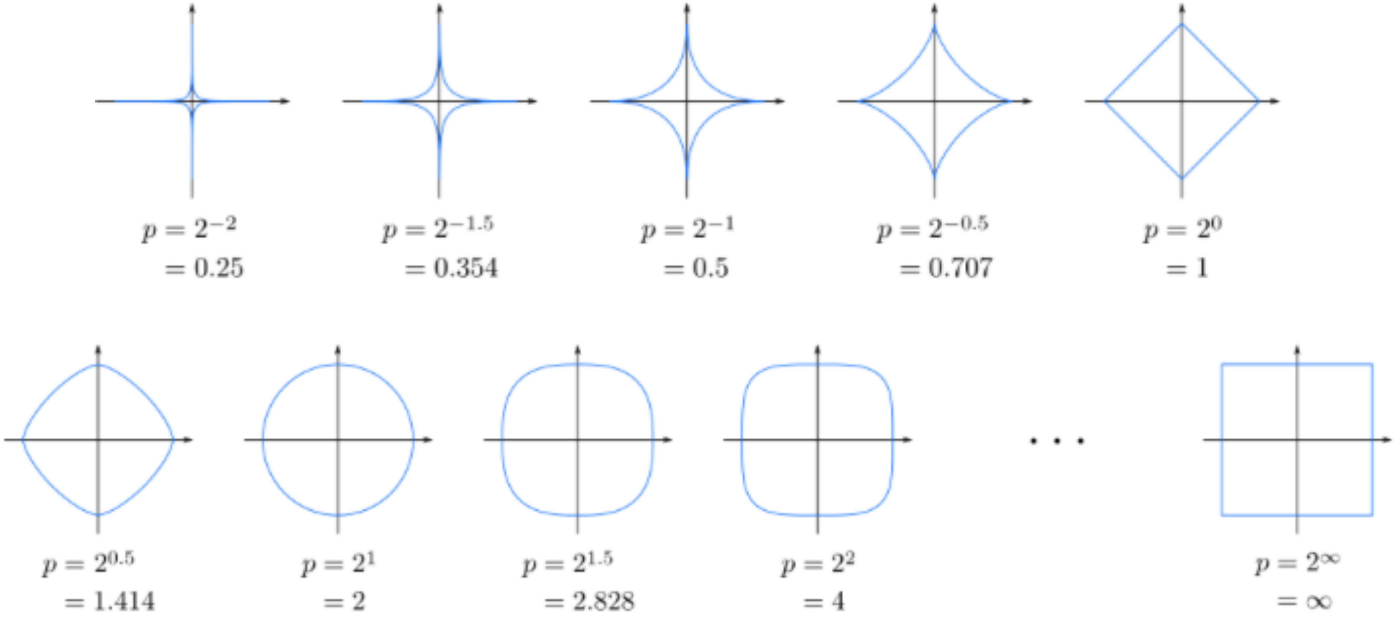
- Without standardising the data , features with high scale , will dominate in grouping determination
- centering with mean and scaling with standard deviation is one of the popular techniques
- You can also centre with median
- Scale with Range, IQR, MAD

Copyright 2020 RESTRICTED CIRCULATION

Distances For Points : general Minkowski distance

$$X = (x_1, x_2, \ldots\ldots, x_n) \; ; \; Y = (y_1, y_2, \ldots\ldots y_n)$$

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$



$p = 2^{-2}$ $= 0.25$

$p = 2^{-1.5}$ $= 0.354$

$p = 2^{-1}$ $= 0.5$

$p = 2^{-0.5}$ $= 0.707$

$p = 2^{0}$ $= 1$

$p = 2^{0.5}$ $= 1.414$

$p = 2^{1}$ $= 2$

$p = 2^{1.5}$ $= 2.828$

$p = 2^{2}$ $= 4$

$\ldots$

$p = 2^{\infty}$ $= \infty$

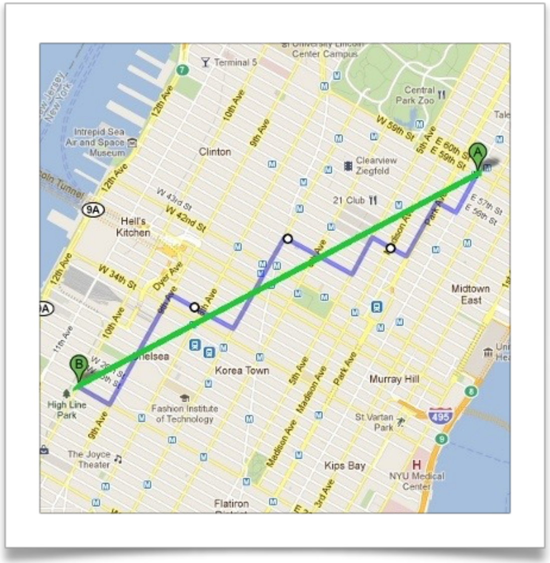
Copyright 2020 RESTRICTED CIRCULATION

EDVANCER EDUVENTURES

Manhattan and Euclidian Distance

p=1

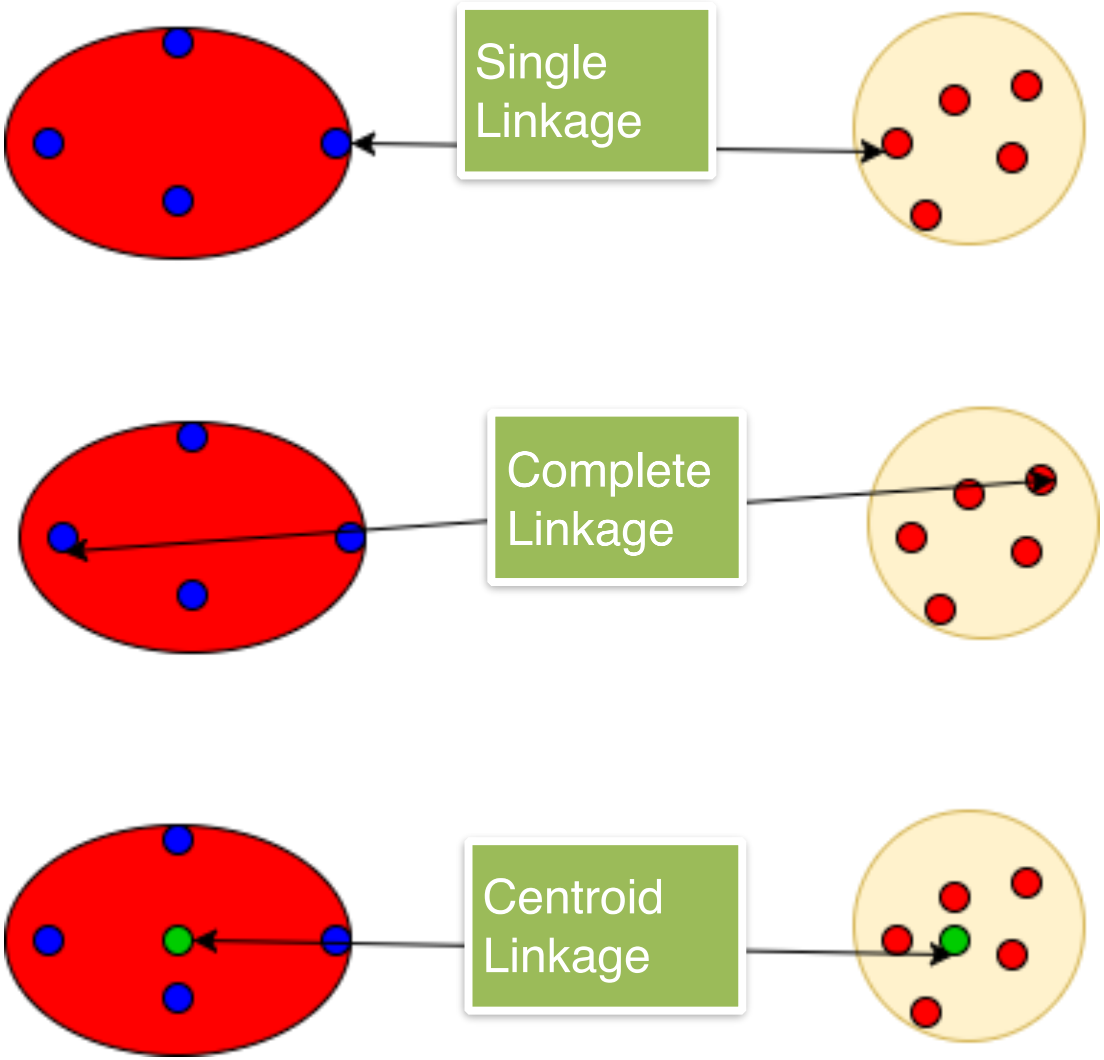
$$\sum_{i=1}^{n} |x_i - y_i|$$



p=2

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Distance between groups



Single Linkage

Complete Linkage

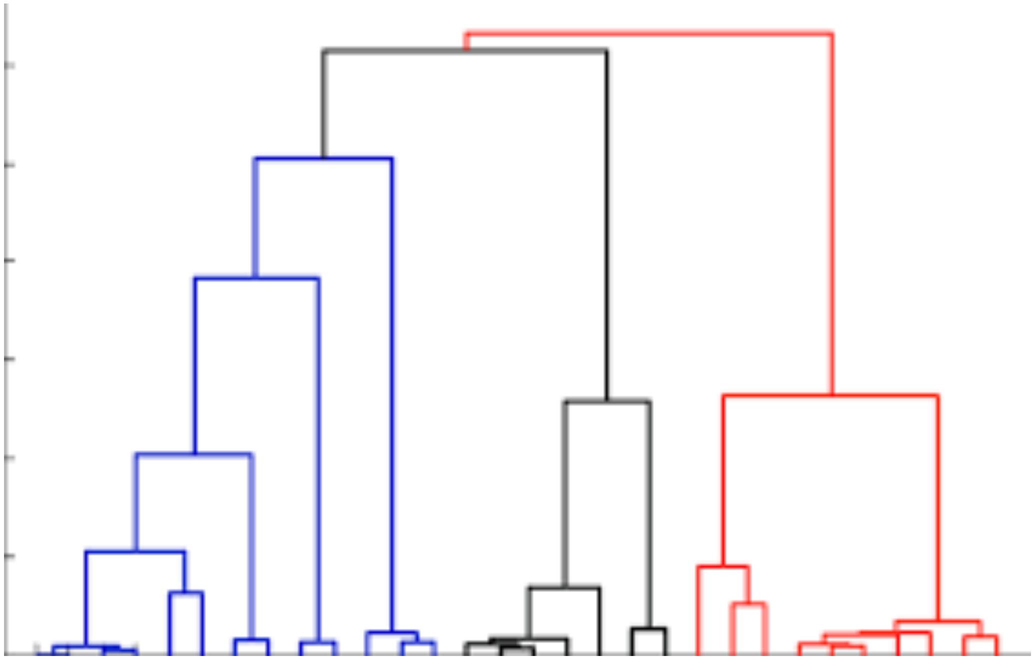
Centroid Linkage

Copyright 2020 RESTRICTED CIRCULATION

Clustering Methods

Copyright 2020 RESTRICTED CIRCULATION

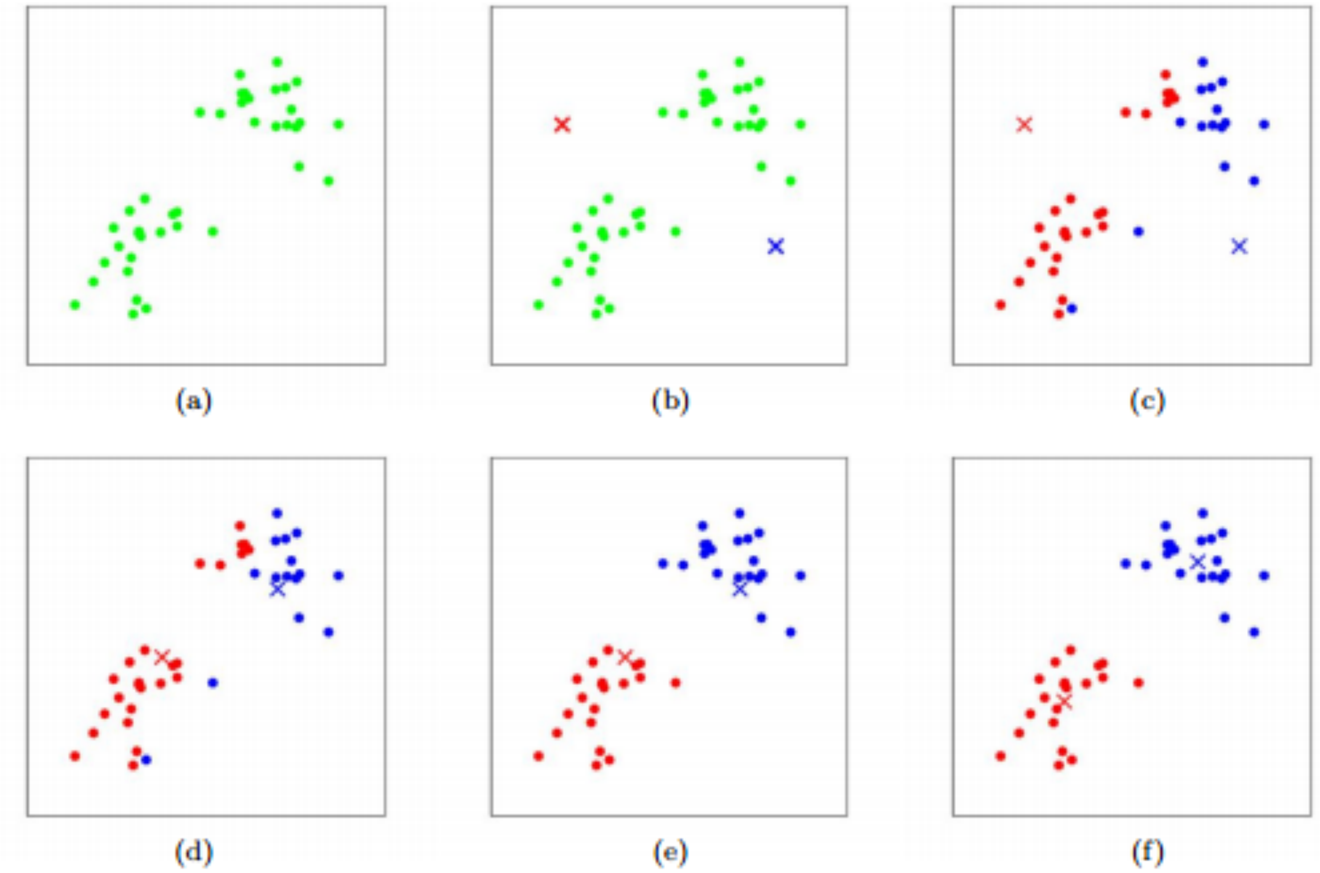
Hierarchical Clustering



- All pairwise distances are calculated
- Observations are clubbed one by one until there is only one group remaining
- This isn't very efficient for even slightly larger datasets

Note : Also Known as Agglomerative

K-means Clustering



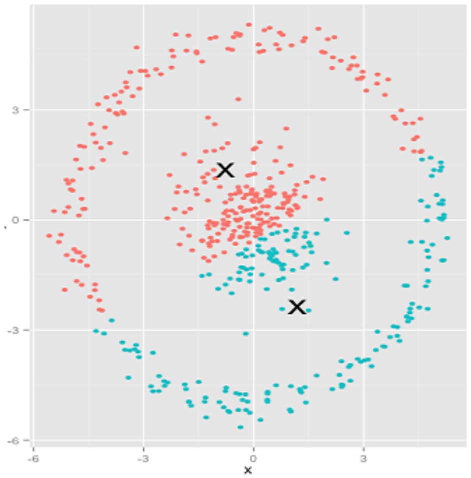
(a) (b) (c)

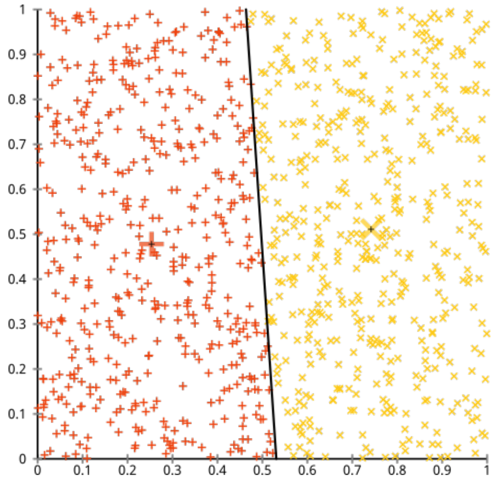
(d) (e) (f)

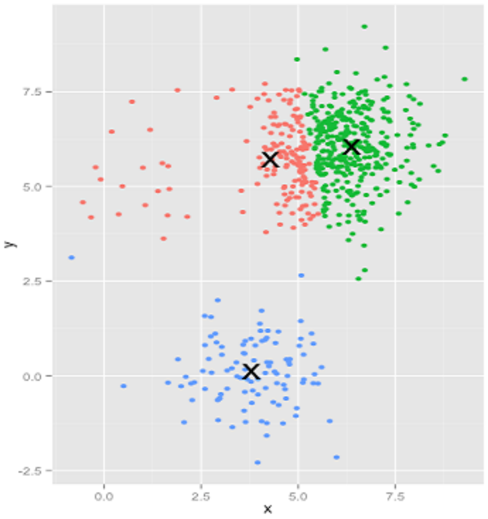
Copyright 2020 RESTRICTED CIRCULATION

A

- Susceptible to extreme values in the data
- Will cluster data into separate groups even if there are no natural separations
- Needs number of clusters as input
- Assumes the clusters to be spherical
- Tends towards equal sized groups

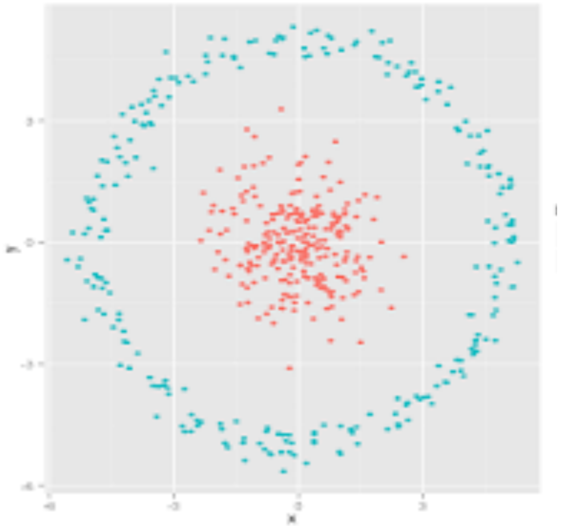


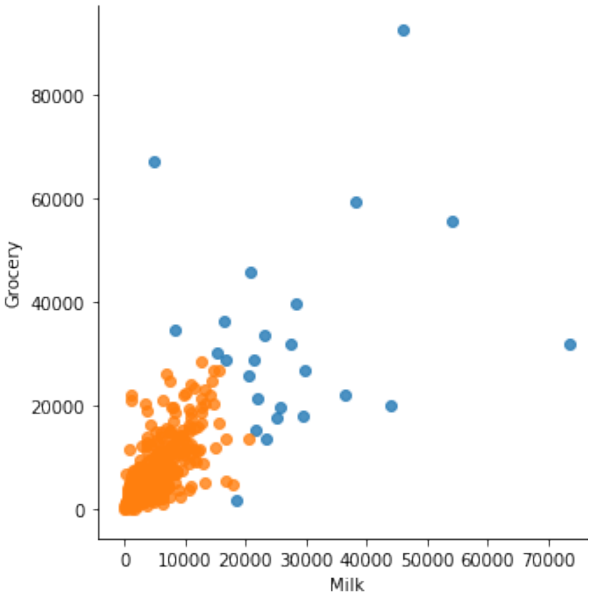




Copyright 2020 RESTRICTED CIRCULATION

DBSCAN



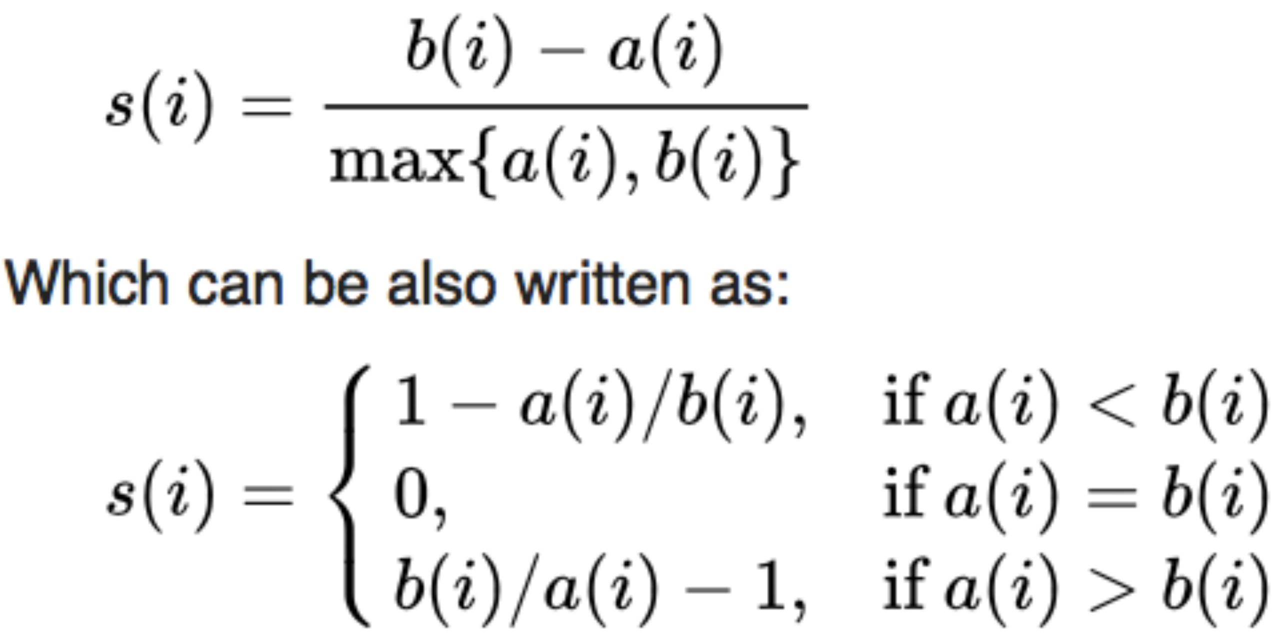


- Takes two parameters :
  - epsilon (neighbourhood size)
  - min pts
- Doesn’t make any assumption about shape of the groups
- There isn’t any good way to measure fitted cluster goodness (DBCV isn’t implemented in sklearn yet)
- Can be used to detect anomalies in the data

Use for visualising dbscan in action :
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Copyright 2020 RESTRICTED CIRCULATION

Silhouette index



$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Which can be also written as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

- Takes values in the range -1 to 1
- For a cluster index of all points can be averaged

Copyright 2020 RESTRICTED CIRCULATION

What to do with Clusters

- Labelling of the groups will come from business context
- Variable selection for the grouping will also be driven by business
- Group wise numeric summaries will be more helpful for making sense of grouped behaviour for higher dimensions
- Group wise means can be used to check how different groups are from each other
- Group wise variances can be used to check how compact or dispersed groups are

Copyright 2020 RESTRICTED CIRCULATION

Dimensionality Reduction

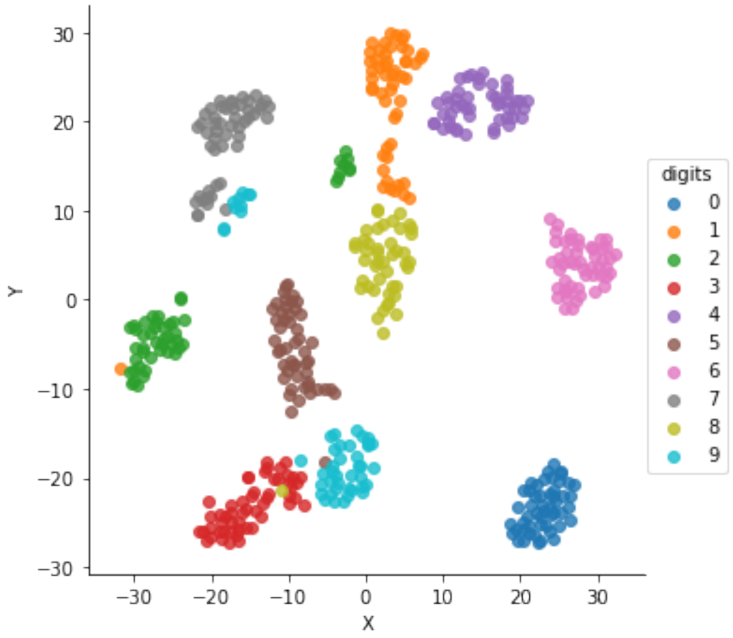
Copyright 2020 RESTRICTED CIRCULATION

Goals

- Treating multi-collinearity
- Reducing data size without information loss for easy experimentation
- Alternate orthogonal representation of the data
- Reducing dimension to visually see groups in the data

Copyright 2020 RESTRICTED CIRCULATION

Visual Dimension Reduction with t-sne



- Reduces dimensions of the data to 2-3 dimensions for easy visualisation
- Converts distances in original data to probabilities
- Multiple runs might yield different results

For details : http://alexanderfabisch.github.io/t-sne-in-scikit-

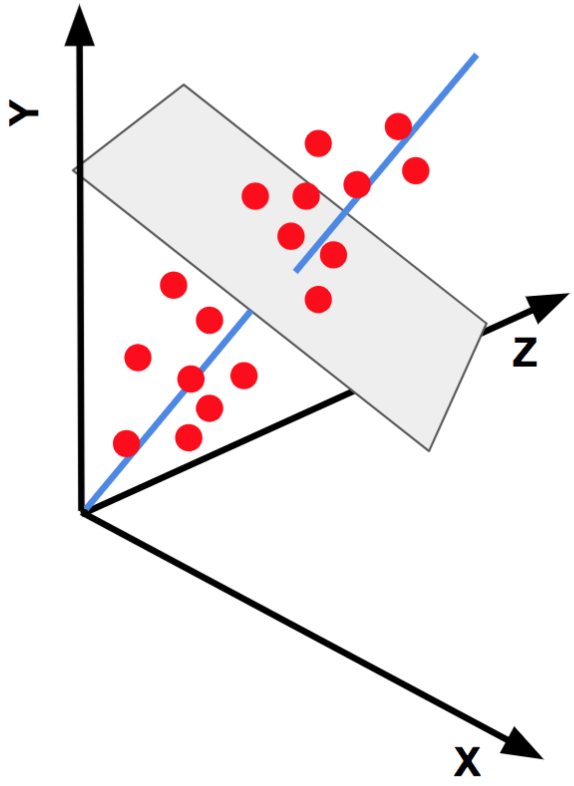
Copyright 2020 RESTRICTED CIRCULATION

PCA (Principal Component Analysis)

- Reduces dimension of the data and retains information if one or more variables in the data are correlated (by means of linear projections)
- New variables are linear combinations of earlier variables
- New variables (Principal Components) are orthogonal to each other (no correlation)

Copyright 2020 RESTRICTED CIRCULATION

Linear Projections



- We start with p dimensional data vectors (observations)
- Dimension is reduced by projecting them onto a q-dimensional (q < p) subspace
- This is done while preserving variance in the data

$$projection\ of\ \vec{x_i}\ on\ to\ subspace\ represented\ by\ unit\ vector\ \vec{w}$$

$$\Rightarrow (\vec{x_i} \cdot \vec{w})\vec{w}$$

Copyright 2020 RESTRICTED CIRCULATION

Minimising loss of Information

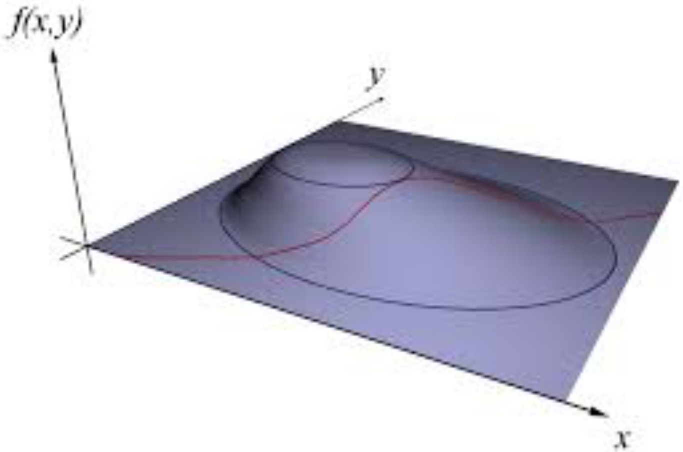
$$min \sum_{i=1}^{n} ||\vec{x_i} - (\vec{x_i} \cdot \vec{w})\vec{w}||^2$$

Unit Vector → $\vec{w}^T \vec{w} = 1$

Centered Data → $\sum_{i=1}^{n} \vec{x_i} = 0$ → $\sum_{i=1}^{n} (\vec{x_i} \cdot \vec{w})\vec{w} = 0$

Copyright 2020 RESTRICTED CIRCULATION

Optimisation with constraints : Lagrange’s Multiplier



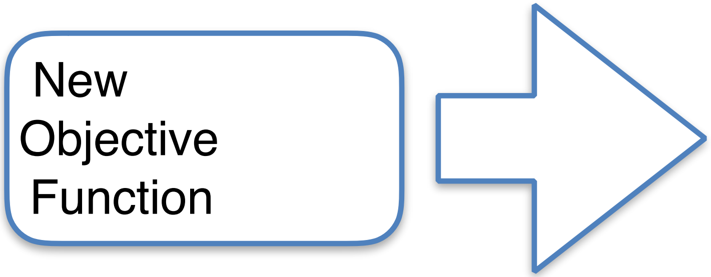
$$\nabla f\left(x, y, z\right) = \lambda \nabla g\left(x, y, z\right)$$

$$g\left(x, y, z\right) = k$$

Copyright 2014 RESTRICTED CIRCULATION

Example

$$find\ values\ of\ x\ and\ y\ for\ which\ (5x-3y)\ takes\ its\ max/min\ value\ under\ constraints\ x^2+y^2=136$$


New Objective Function

$$5x-3y-\lambda*(x^2+y^2-136)$$

$$5=2\lambda x$$
$$-3=2\lambda y$$
$$x^2+y^2=136$$

$$\lambda^2=\frac{1}{16} \quad \Rightarrow \quad \lambda=\pm\frac{1}{4}$$

If $\lambda=-\frac{1}{4}$ we get,

$x=-10$ $\quad y=6$

and if $\lambda=\frac{1}{4}$ we get,

$x=10$ $\quad y=-6$

Copyright 2014 RESTRICTED CIRCULATION

Minimising loss of variance contd..

$$\sum_{i=1}^{n} ||\vec{x_i} - (\vec{x_i} \cdot \vec{w})\vec{w}||^2$$

$$= \sum_{i=1}^{n} (||\overrightarrow{x_i}||^2 + (\overrightarrow{x_i} \,.\, \overrightarrow{w})^2 - 2(\overrightarrow{x_i} \,.\, \overrightarrow{w})^2)$$

$$=> \ \ max \ \frac{1}{n}\sum_{i=1}^{n}(\vec{x_i} \cdot \vec{w})^2$$

$$= (\frac{1}{n}\sum_{i=1}^{n} \vec{x_i} \cdot \vec{w})^2 + var\left[\vec{w} \cdot \vec{x_i}\right]$$

$$=> max \ \ var\left[\vec{w} \cdot \vec{x_i}\right] \ s.t. \ \vec{w}^T\vec{w} = 1$$

Copyright 2020 RESTRICTED CIRCULATION

Contd..

$$obj = W^T V W - \lambda W^T W$$

$$VW = \lambda W$$

- V is variance covariance matrix of X
- Principal Components are eigen vectors of variance covariance matrix
- Eigen vectors of a symmetric matrix are orthogonal to each other

Principal Components

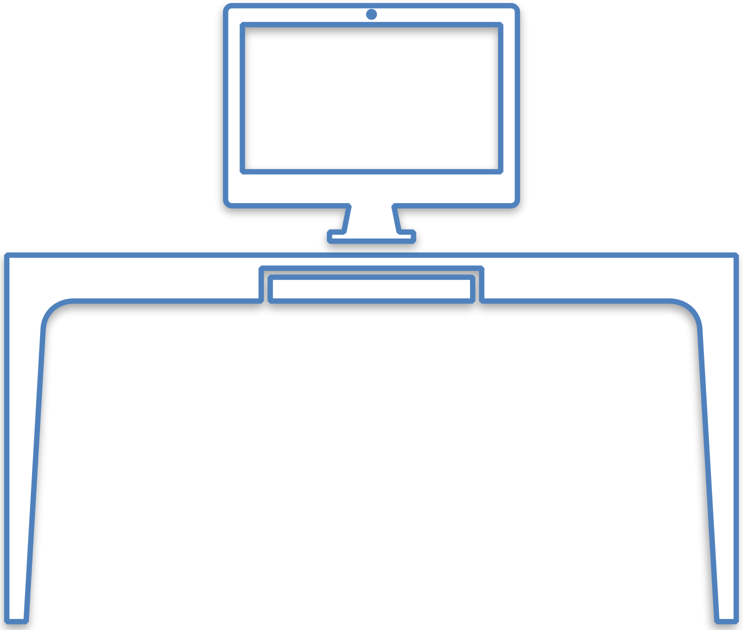
- Eigen vectors of V are the principal components
- First principal component is the eigen vector with largest eigen value
- It means data has highest variance across that
- The second principal components has seconds highest variance and so on

Copyright 2020 RESTRICTED CIRCULATION

Dimensionality Reduction

- If the data really is q dimensional (p-q variables are simple linear combinations of the rest) then p-q eigenvalues will be zero
- If data is near to q dimensional (not perfect linear combination but high correlation) then p-q eigen values will be nearly zero
- We can select top few PCs on the basis of how much variance they represent cumulatively

Lets see it in action in Python



Copyright 2020 RESTRICTED CIRCULATION