DBSCAN

Density based special clustering with noise.

- Dimentionality reduction means reducing noise.

Reducing noise

↓

Reducing unnecessary features.
(Not related to our problems)

Feature selection

(Finding new features after transforming Data) Feature Extraction

(Feature / Dimentions)
?

linear combination of original variables

Principle Component Analysis:
New set of variable (principle components)
orthogonal components.

Such that greatest variance project is on the

First principle component.
Variance : How spread out the data is.

project direction

Maximizing the variance

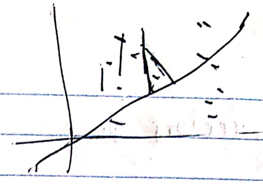
1) Standardise the Data
Standard deviation — How far it's from mean

Eigen decomposition

2) Eigen Pairs [characteristic pairs]
Eigen vector does not change it's value / direction

Co-variance matrix (symetric)
two change w.r.t each other.

Discord Eigen values ↓



projection matrix [to k eigen values]

PCA Data pre-processing [Mean normalisation / feature scaling]

line on lower dimentional subspace

↓

So that it will reduce the variance

↓

Find a vector in a direction with most points and variance.

$x^{(i)} \leftarrow \mathbb{R}^2 \longrightarrow z^{(i)} \in \mathbb{R}$

→ Covariance matrix

$$\Sigma = \frac{1}{M} \sum x^{(j)} (x^{(i)}$$

Difference betw covariance = Differential matrix.

$\boxed{\text{Random to — linear / planar}}$

Find vectors pair $\longrightarrow$ to reduce projection error.

$n \longrightarrow k$

1) Variance matrix

$$\Sigma = \frac{1}{m} \sum_{j=1}^{n} \underset{n\times1}{(x^{i})} \underset{1\times n}{(x^{i})}^{T}$$

2) Compute * eigen vectors of this matrix

$\rightarrow \boxed{U, S, V} = \text{svd}(\Sigma)_{n\times n}$ eig

* singular value decomposition.

$$U = \begin{bmatrix} u^{(1)} & u^{(2)} & u^{(3)} \end{bmatrix}_{n\times n}$$

PCA vectors

First few k will give you reduced dimensional

dimensionality reduced.

k vector

$$\star \underset{k\times1}{Z} = \begin{bmatrix} u \text{ reduced} \end{bmatrix}^{T}_{k\times n} X_{n\times1}$$

How to select k
minimise average square projection error.

$$\frac{\text{Average square projection error}}{\text{Total variance}} \leq 0.01$$

i.e. 99% of variance is retained.

$$S = \begin{bmatrix} \cdot & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \end{bmatrix}_{n \times n}$$

diagonal matrix

$$1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}} \leq 0.01$$

hence k can be decided easily.

UMAP Uniform manifold approximation.

1) Matrix Factorization
global structure.

2) Neighbour graphs
local structure
t - SNE

- Simplicial complex.
  to recover all the topology of the space.
- Riemannian metric on the manifold to make it well distributed over space

Multidimentional Scaling.

Reconstruct a map that preserves the distancy.

- classical
- Metric
- Non-Metric.

t - SNE

Depends upon attraction & repulsion betw^n relative points.

GMM

1) distance similarity value is in the normal distribu^n the point is close.

or else its not.

↓

Matrix of similarity score.

1) Random projects

↓

t distrubution

↓

Trying getting near Normal distrubution.