Data reduction

Data reduction means reducing the size of the data while keeping the important information.

Goal

→ Dataset size becomes smaller

→ Result of analysis remains almost same.

Example

* original : 100 features
* Reduced : 10 features
* output remains similar.

Method of Reduction

- Feature Selection
- Dimensionality Reduction

Feature Selection

Remove unnecessary features.

Remove:

- → Irrelevant features
- → Weak features
- → Duplicate features

Example:

house price prediction

→ Area - Useful

→ Location - Useful

→ wall colour - Remove

Dimensionality Reduction

Do not remove the features

Instead:

→ Transform original features

→ create smaller set of features.

Example:

100 features → 5 combined features.

Reduction Techniques

Lossless:

* original data can be perfectly recovered
* NO information loss

Example: Zip file

Lossy

* original data cannot be perfectly recovered
* only approximation is possible.

Example: Image compression.

Note:

PCA uses lossy method.

PCA (Principal component Analysis)

* PCA is a data transformation technique

PCA itself not dimensionality reduction, Dimensionality reduction is one of the application of PCA.

key Idea of PCA

PCA finds new direction in which:

→ Data spread is maximum

→ Information is maximum

These directions are called as Principal components.

Data Representation

Let:

* N = number of samples
* d = number of features

one data point:

$$x_n = [x_{n1}, x_{n2}, x_{n3} \ldots x_{nd}]$$

Orthonormal Vectors:

PCA finds:

* d new directions

these directions are:

→ perpendicular to each other

→ unit length

these are called as **Orthonormal Vectors**.

Projection

Each data point is projected onto each direction

Formula:

$$a_{ni} = q_i^T x_n$$

where,

q_i = i^{th} direction

a_{ni} = i^{th} principal component.

PCA without Reduction

If we use all d direction

$$x_n = a_{n1} q_1 + a_{n2} q_2 + \dots + a_{nd} q_d$$

Results in:

→ Perfect Reconstruction

→ No data loss.

PCA with Reduction

We use only

l direction, where $l << d$

we keep:

→ direction with high variance

we remove:

→ direction with low variance.

Reconstruction Error

Since some directions are removed:

original data cannot be fully recovered.

Error:

$$\text{Error} = \| x_n - \hat{x}_n \|$$

Eigen values:

Eigen values shows how important a direction is:

Large eigen value → More information

Small eigen value → Less information

PCA algorithm steps

Step 1

* Form a data matrix X
* Each column = one data sample.

Step 2

* Compute mean for each feature.

Step 3

* Compute covariance matrix
* Shows relation between features.

Step 4

* Find eigen values and eigen vectors.

Step 5

* Sort eigen values in descending order

Step 6

* Select top k eigen vectors

Step 7

* Project data onto selected eigen vectors.

Choosing Number of components

Total variance

$$\sigma^2_{Total} = \lambda_1 + \lambda_2 + \cdots \lambda_d$$

Cumulative variance

$$\sigma^2_l = \lambda_1 + \lambda_2 + \cdots \lambda_l$$

Rule

Choose l such that

$$\sigma^2_l \geq 0.95 \times \sigma^2_{Total}$$

Properties of PCA output

* New features are uncorrelated
* Data size is reduced
* Noise is reduced.

Limitation of PCA:

PCA focus on:

→ Maximum variance

But in classification we need:

→ Maximum class seperation

So PCA fails for classification tasks.

Fisher Discriminant Analysis (FDA)

FDA is supervised dimensionality reduction technique.

Goal: Find direction where.

→ distance between class means is maximum

→ spread within each class is minimum.

PCA vs FDA

PCA	FDA
unsupervised	supervised
Maximizes variance	Maximizes class seperation
used for compression	used for classification