Feature Engineering

[selection from existing] Selection

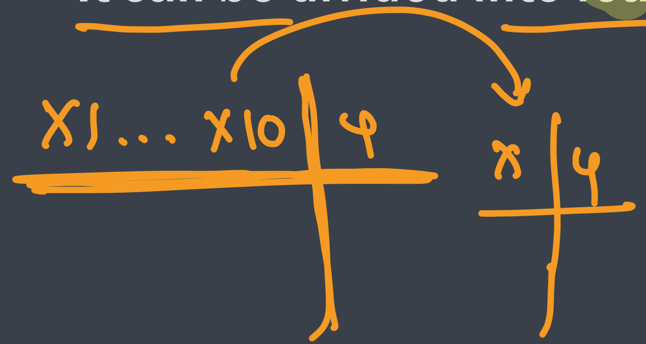
extraction [create new features]

# Dimensionality Reduction

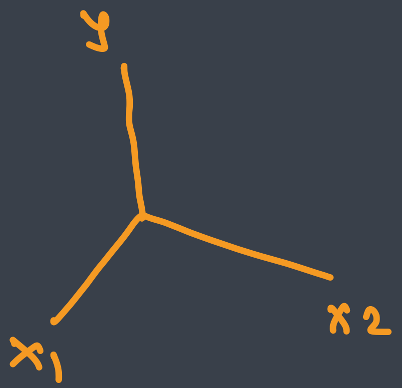
# features

# Dimensionality Issue

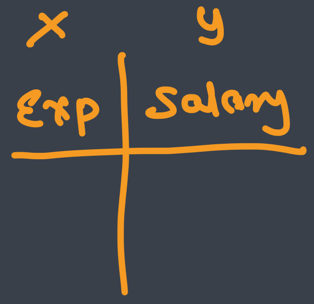
- In machine learning classification problems, there are often too many factors on the basis of which the final classification is done
- These factors are basically variables called features
- The higher the number of features, the harder it gets to visualize the training set and then work on it
- Sometimes, most of these features are correlated, and hence redundant
- This is where dimensionality reduction algorithms come into play
- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables
- It can be divided into feature selection and feature extraction











## Advantages of Dimensionality Reduction

- It helps in data compression, and hence reduced storage space
- It reduces computation time
- It also helps remove redundant features, if any

* Not used for improving model accuracy *

- visualization

# Disadvantages of Dimensionality Reduction

- It may lead to some amount of data loss
- PCA tends to find linear correlations between variables, which is sometimes undesirable
- PCA fails in cases where mean and covariance are not enough to define datasets
- We may not know how many principal components to keep- in practice, some thumb rules are applied

$10 \longrightarrow 2$

# Components of dimensionality reduction

- There are two components of dimensionality reduction:
- **Feature selection:** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  - Filter
  - Wrapper
  - Embedded

correlation coe / cov

- **Feature extraction:** This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

# Methods of Dimensionality Reduction

- Principal Component Analysis (PCA) **
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

PCA

# Overview

- Is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets
- Reduces dimensions by transforming a large set of variables into a smaller one that still contains most of the information in the large set
- Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity
- Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process

# What is PCA?

- This method was introduced by Karl Pearson
- It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum

# Step 1: Standardization → Scaling

- The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis

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

- Once the standardization is done, all the variables will be transformed to the same scale.

# Step 2: Covariance Matrix computation

- The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other
- In other words, to see if there is any relationship between them
- Because sometimes, variables are highly correlated in such a way that they contain redundant information
- So, in order to identify these correlations, we compute the covariance matrix

$(x, y)$

cov cor

$$cov(x, y) = \begin{bmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{bmatrix}$$

## Step 2: Covariance Matrix computation

- The covariance matrix is a $p \times p$ symmetric matrix (where $p$ is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables
- For example, for a 3-dimensional data set with 3 variables $x$, $y$, and $z$, the covariance matrix is a 3×3 matrix of this from

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

## Step 2: Covariance Matrix computation

- Since the covariance of a variable with itself is its variance (Cov(a,a)=Var(a)), in the main diagonal (Top left to bottom right) we actually have the variances of each initial variable
- Since the covariance is commutative (Cov(a,b)=Cov(b,a)), the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.
- if positive then
  - the two variables increase or decrease together (correlated)
- if negative then
  - One increases when the other decreases (Inversely correlated)

if zero the

a Not correlated

## Step 3: Compute eigenvectors eigenvalues

- Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data
- Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables
- These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components
- Organizing information in principal components this way, will allow you to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as your new variables
- the principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables

## Step 4: Feature vector

- choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call Feature vector
- feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep

$\lambda_1$ and $\lambda_2$

# Statistical Calculations

# Example

- Calculate PCA for the following dataset : reduce the dataset to one dimension

| X | Y |
|---|---|
| 4 | 11 |
| 8 | 4 |
| 13 | 5 |
| 7 | 14 |

## Step 1: Calculate Mean

$$\bar{x} = (4+8+13+7)/4 = 32/4 = \underline{8}$$

$$\bar{y} = (11+4+5+14)/4 = 34/4 = \underline{8.5}$$

$$\boxed{\bar{x} = 8} \qquad \boxed{\bar{y} = 8.5}$$

| $x$ | $y$ | $x-\bar{x}$ | $(x-\bar{x})^2$ | $(y-\bar{y})$ | $(y-\bar{y})^2$ | $(x-\bar{x})(y-\bar{y})$ |
|---|---|---|---|---|---|---|
| 4 | 11 | -4 | 16 | 2.5 | 6.25 | -10 |
| 8 | 4 | 0 | 0 | -4.5 | 20.25 | 0 |
| 13 | 5 | 5 | 25 | -3.5 | 12.25 | -17.5 |
| 7 | 14 | -1 | 1 | 5.5 | 30.25 | -5.5 |
| | | | 42 | | 69 | -33 |

$\bar{x} = 8$, $\bar{y} = 8.5$

## Step 2: Calculate Covariance Matrix

$$\begin{bmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) \end{bmatrix} = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

$$\mathrm{cov}(x,x) = \frac{\sum (x-\bar{x})(x-\bar{x})}{N-1} = \frac{\sum (x-\bar{x})^2}{N-1} = \frac{42}{3} = 14$$

$$\mathrm{cov}(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{N-1} = \frac{-33}{3} = -11$$

$$\mathrm{cov}(y,y) = \frac{\sum (y-\bar{y})^2}{N-1} = \frac{69}{3} = 23$$

## Step 3: Calculate eigenvalues of Covariance Matrix

$$|A - \lambda I| = 0$$

$$\left| \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = \begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix} = 0$$

$$(14-\lambda)(23-\lambda) - (-11 \times -11) = 0$$

$$14 \times 23 - 14\lambda - 23\lambda + \lambda^2 - 121 = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

$$\text{roots} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$a = 1, \quad b = -37, \quad c = 201$

$$\lambda = \frac{37 \pm \sqrt{(-37)^2 - 4 \times 201}}{2} = \frac{37 \pm \sqrt{1369 - 809}}{2}$$

$$= \frac{37 \pm \sqrt{565}}{2} = \frac{37 \pm 23.76}{2}$$

$$\lambda = \frac{37 + 23.76}{2} = \underline{\underline{30.38}}, \quad \lambda = \frac{37 - 23.76}{2} = \frac{13.24}{2} = \underline{\underline{6.62}}$$

## Step 4: Calculate eigenvector

$$(A - \lambda I)u = 0$$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \begin{bmatrix} u1 \\ u2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(14-\lambda)u1 - 11u2 = 0 \Rightarrow (14-\lambda)u1 = 11u2$$

$$\frac{u1}{11} = \frac{u2}{14-\lambda}$$

$$-11u1 + (23-\lambda)u2 = 0$$

$$u_1 = 11, \quad u2 = (14-\lambda) = 14-30.38 = -16.38$$

## Step 5: Normalize unit eigenvector

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 11 \\ -16.38 \end{bmatrix},$$

$$\|u\| = \sqrt{(u_1)^2 + (u_2)^2}$$

$$= \sqrt{11^2 + (-16.38)^2}$$

$$= \sqrt{121 + 268.30}$$

$$= \sqrt{389.3}$$

$$= 19.73$$

$$e = \begin{bmatrix} u_1 / \|u\| \\ u_2 / \|u\| \end{bmatrix} =$$

$$\underline{e} = \begin{bmatrix} 11/19.73 \\ -16.38/19.73 \end{bmatrix} = \begin{bmatrix} 0.55 \\ -0.83 \end{bmatrix}$$

# Step 6: Calculate first principal component

$\bar{x} = 8$

$\bar{y} = 8.5$

$$\text{principal component} = e^T \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}$$

1

$$= \begin{bmatrix} 0.55 & -0.83 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$$

$$= 0.55 \times -4 - 0.83 \times 2.5 = -2.2 - 2.07$$

$$= -4.27$$

| $x$ | $y$ | PC |
|---|---|---|
| 4 | 11 | -4.27 |
| 8 | 4 | 3.73 |
| 13 | 5 | 5.65 |
| 7 | 14 | -5.11 |

(2) $0.55 \times 0 - 0.83 \times (-4.5) = 3.73$

(3) $0.55 \times 5 - 0.83 \times (-3.5) = 2.75 + 2.90$

$= 5.65$

(4) $-0.55 \times 1 - 0.83 \times 5.5 = -0.55 - 4.56$

$= -5.11$