Data Pre-Processing-IV

(Feature Extraction: PCA)

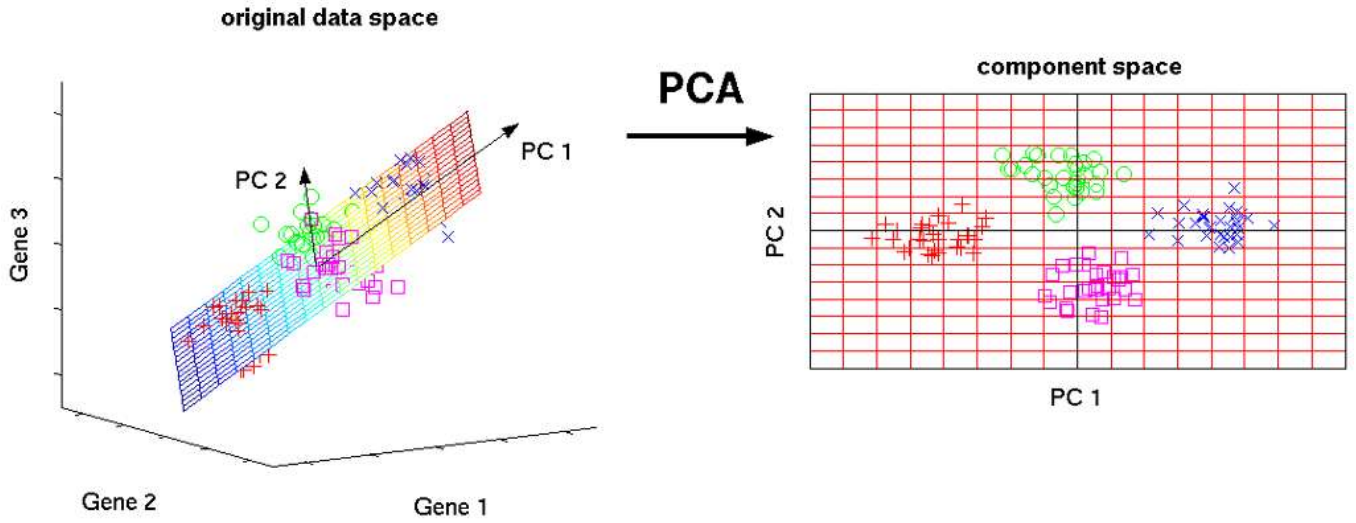
Dr. JASMEET SINGH
ASSISTANT PROFESSOR, CSED
TIET, PATIALA

Feature Extraction

- Feature extraction, creates new features from a combination of original features.
- For a given Feature set F_i (F_1, F_2, F_3,……..F_n), feature extraction finds a mapping function that maps it to new feature set F_i ' (F_1', F_2', F_3',…….F_m') such that F_i'=f(F_i) and m <n.
- For instance F_1'= k_1 F_1 + k_2F_2
- Some commonly used methods are:
  - Principal Component Analysis (PCA)
  - Singular Valued Decomposition (SVD)
  - Linear Discriminant Analysis (LDA)

Principal Component Analysis

Principal Component Analysis (PCA): It is a technique of dimensionality reduction which performs the said task by reducing the higher-dimensional feature-space to a lower-dimensional feature-space. It also helps to make visualization of large dataset simple.



Principal Component Analysis

Some of the major facts about PCA are:

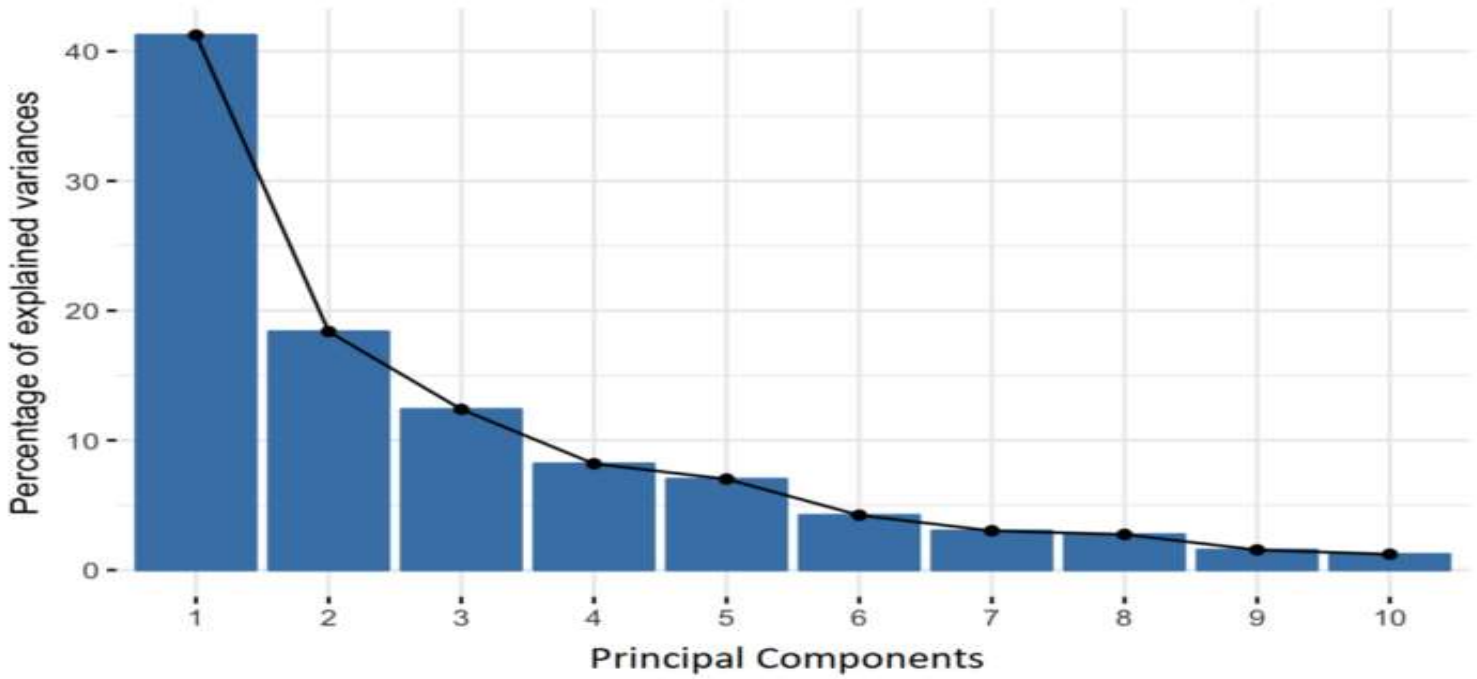
- Principal components are new features that are constructed as a linear combinations or mixtures of the initial feature set.
- These combinations is performed in such a manner that all the newly constructed principal components are uncorrelated.
- Together with reduction task, PCA also preserving as much information as possible of original data set.

Principal Component Analysis

Some of the major facts about PCA are:

- Principal components are usually denoted by PCi, where i can be 0, 1, 2, 3,. ….,n (depending on the number of feature in original dataset).
- The major proportion of information about original feature set can be alone explained by first principal component i.e. PC1.
- The remaining information can be obtained from other principal components in a decreasing proportion as per increase in value of i.

Principal Component Analysis



PCA- Geometrical Interpretation

- **Geometrically ,** it can be said that principal components are lines pointing the directions that captures maximum amount of information about the data.
- Principal components also aims to minimize the error between the true location of the data points (in original feature space) and the projected location of the data points (in projected feature space).
- The larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has.

Simply, principal components are new axes to get better data visibility with clear difference in observations.

PCA- Geometrical Interpretation

- Suppose we have the following standardized data (as shown in figure 1).
- If suppose we have to choose 1 feature out of X1 and X2, we will choose feature X1. (The one which explains the maximum variation in the data).
- This is exactly what PCA does. It finds the features which have maximum spread and drop the others with the aim to **minimize information loss**.
- Let's take a slightly complex example where we can not simply drop one feature (as shown in figure 2).
- Here, both the features X1 and X2 have equal spread. So we can't tell which feature is more important.
- But if we try to find a direction (or axis) which explains the variation in data we can find a line which fits the data very well. So if we rotate our axis slightly by theta, we get f1 and f2 (perpendicular to f1). We can then drop f2 and say f1 is the most important feature. This is what PCA does.

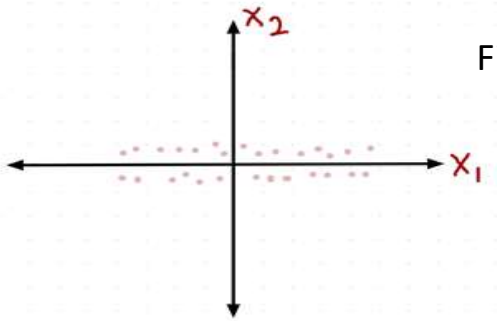


Fig 1

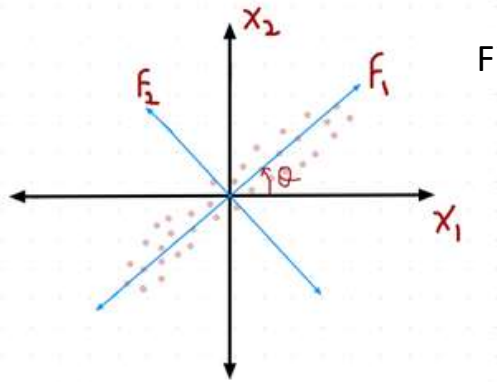


Fig 2

Mathematics behind PCA

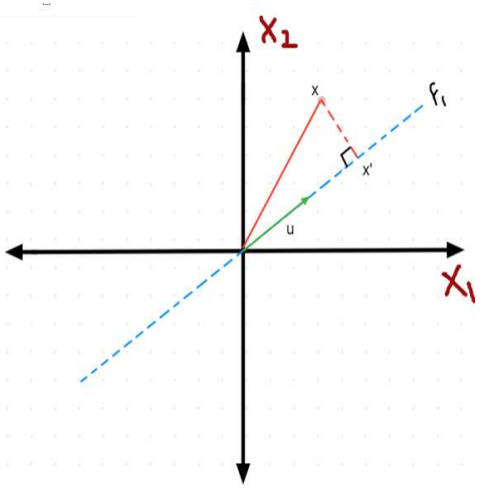
Let X be a feature matrix centered around mean of size $n \times k$, where n denote number of examples and k denote number of features.

Let F_1 be the direction along which there is maximum variance of data and $\vec{u}$ be a unit vector in the direction of F_1 (for simplicity, two dimensional data is shown in figure).

Let x_i be any data point from X.

Projection of x_i on $\vec{u} = |\overrightarrow{x_i}|\cos\theta = \frac{\overrightarrow{x_i}.\vec{u}}{|\vec{u}|} = \overrightarrow{x_i}.\vec{u}$

We have to find $\vec{u}$, such that Variance (projected $\underset{u}{\rightarrow} x_i$) is maximum for all x_i



Mathematics behind PCA (Contd…..)

Total variance of all data points x_i

$$= \frac{1}{n}\sum_{i=1}^{n}(\overrightarrow{x_i} \,.\, \vec{u} - \bar{x}\,\vec{u})^2$$

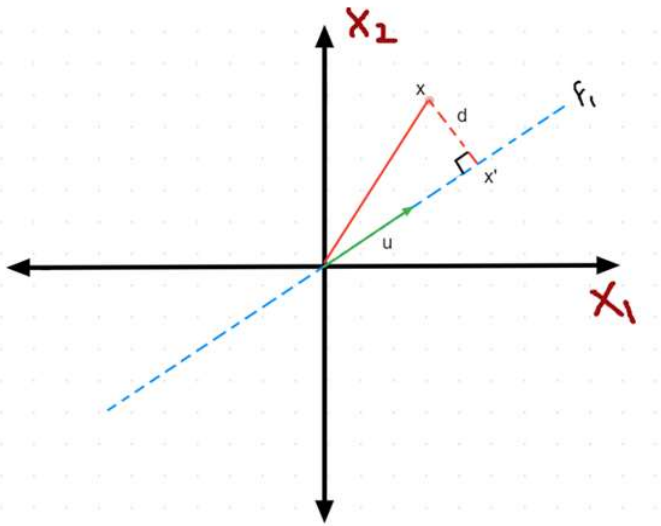
Since the feature vector is centered around mean, therefore, $\bar{x}$ =0 (sum of deviations from mean is zero)

Therefore, total variance = V = = $\frac{1}{n}\sum_{i=1}^{n}(\overrightarrow{x_i} \,.\, \vec{u})^2$

This approach is called *variance maximization approach*

Mathematics behind PCA (Contd.....)

Another way to think of PCA is that it fits the best line that passes through our data with an aim to minimize the projection error 'd' for each point. This approach is called distance minimization approach.



$$d^2 = |x_i|^2 - (\vec{x_i} \cdot \vec{u})^2$$

So our optimization problem becomes

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

Notice that both the optimization problems, though look different, are same. Since the $|x_i|^2$term is independent of u so in order to minimize the function we have to maximize $(u^T x_i)^2$ which is same as our first optimization problem.

Mathematics behind PCA (Contd…..)

So, using both objectives, our optimization problem is,

$$max_u \frac{1}{n}\sum_{i=1}^{n}(\overrightarrow{x_i} \cdot \vec{u})^2$$

Writing out all the summations grows tedious, so let's do our algebra in matrix form. If we stack our n data vectors into an n × k matrix, X, then the projections are given by Xu, which is an n ×1 matrix.

Total variance = $\frac{1}{n}|Xu|^2 = \frac{1}{n}\,(Xu)^T\,(Xu) = \frac{1}{n}u^T X^T Xu = u^T \frac{X^T X}{n} u = u^T Su$

Where S is the covariance matrix, given by $\frac{X^T X}{n}$ where X is centered around mean

Mathematics behind PCA (Contd…..)

- The given optimization problem is solved using Lagrange Optimization (which is used for constrained optimization)
- The method can be summarized as follows: in order to find the maximum or minimum of a function f(x) subjected to the equality constraint g(x)=0, form the Lagrangian function

$$L(x, \lambda) = f(x) - \lambda\, g(x)$$

and find the stationary points of L considered as a function of x and the Lagrange multiplier λ

Mathematics behind PCA (Contd…..)

$$L(\lambda, u) = u^T S u - \lambda(u^T u - 1)$$

$$\frac{\partial L(\lambda,u)}{\partial \lambda} = -(u^T u - 1)$$

$$\frac{\partial L(\lambda,u)}{\partial u} = 2Su - 2\lambda u = 2(Su - \lambda u)$$

For L to be maximum or minimum substitute $\frac{\partial L(\lambda,u)}{\partial \lambda} = 0$ and $\frac{\partial L(\lambda,u)}{\partial u} = 0$

Therefore, $u^T u = 1 \; and \; Su = \lambda u$

Thus, u is eigen vectors corresponding to the eigen values of covariance matrix S.

Since, S is of k X k , hence there will k eigen vectors.

$Let \; \lambda_1 > \lambda_2 > \cdots \ldots \ldots \ldots \ldots \ldots \lambda_k \; be \; k \; eigen \; values \; of \; S$

Mathematics behind PCA (Contd…..)

$$\frac{\partial^2 L(\lambda,u)}{\partial \lambda^2} = 0$$

$$\frac{\partial^2 L(\lambda,u)}{\partial u^2} = 2(S - \lambda I)$$

$$\frac{\partial^2 L(\lambda,u)}{\partial u^2}_{\lambda_1} = 2(S - \lambda_1 I)$$

Eigen values of $2(S - \lambda_1 I)$ will be $\lambda_i - \lambda_1$ where $1 \leq i \leq k$

(because If $\lambda_1, \lambda_2 \ldots\ldots \lambda_n$ are the eigen values of A, then λ_1-k, λ_2-k…….λ_n-k are eigen values of A-kI)

Hence all eigen values of $2(S - \lambda_1 I)$ will be negative or zero.

There L is maximum when $\lambda = \lambda_1$

Thus, the principal components that captures maximum variance of input data points are the eigen vectors of the covariance matrix of the input feature matrix corresponding to largest eigen values.

Principal Component Analysis- Step-wise Working

Step 1: Construction of covariance matrix named as S.

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, or in other words, to see if there is any relationship between them.

$$S = Cov\ Matrix = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_d) \\ Cov(X_1, X_2) & Var(X_2) & & \vdots \\ \vdots & & \ddots & \\ Cov(X_1, X_d) & \dots & & Var(X_d) \end{bmatrix}_{d \times d}$$

where

$$Cov(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})$$

Alternatively, Covariance matrix, S, can be computed as

$$S = \frac{1}{n} X^T X,$$

where X is a feature matrix with training examples as rows and features as columns and X is centered around mean

Principal Component Analysis- Step-wise Working

Step 2: Computation of eigenvalues for covariance matrix , using following equation:

$$\mathbf{det}(\boldsymbol{S - \lambda I})=0$$

The eigenvalues are simply the coefficients attached to eigenvectors, which give the *amount of variance carried in each Principal Component*.

Step 3: Sort the eigenvectors in decreasing order of eigenvalues and choose k eigenvectors with the largest eigenvalues

Step 4: Compute eigenvectors corresponding to every eigenvalue obtained in step 2

$$[\boldsymbol{S - \lambda I}]\, \boldsymbol{X} = \mathbf{O}$$

The eigenvectors of the Covariance matrix are actually *the directions of the axes where there is the most variance* (most information*) and that we call Principal Components*

Step 5: Transform the data along the principal component axis.

PCA –Numerical Example

Check (mathematically) whether the following two-dimensional datapoints can be transformed to one dimension using Principal Component Analysis.
If yes, determine the magnitude, percentage variance captured along the new principal components and the new principal component.
Data points (x, y): {(2, 1), (3, 5), (4, 3), (5, 6), (6, 7), (7, 8)}

PCA –Numerical Example (Solution)

Step 1: Compute covariance matrix, S:

$$\text{Feature matrix=X=}\begin{pmatrix}2 & 1\\ 3 & 5\\ 4 & 3\\ 5 & 6\\ 6 & 7\\ 7 & 8\end{pmatrix}$$

$$\text{Mean Vector=}\mu = \begin{pmatrix}4.5 & 5\end{pmatrix}$$

$$\text{Feature Vector centered around mean=}X = \begin{pmatrix}-2.5 & -4\\ -1.5 & 0\\ -0.5 & -2\\ 0.5 & 1\\ 1.5 & 2\\ 2.5 & 3\end{pmatrix}$$

PCA –Numerical Example (Solution)

Covariance Matrix=S=$\frac{X^T X}{n-1}$

$$=\frac{1}{5}\begin{pmatrix}-2.5 & -1.5 & -0.5 & 0.5 & 1.5 & 2.5\\ -4 & 0 & -2 & 1 & 2 & 3\end{pmatrix}\begin{pmatrix}-2.5 & -4\\ -1.5 & 0\\ -0.5 & -2\\ 0.5 & 1\\ 1.5 & 2\\ 2.5 & 3\end{pmatrix}$$

$$=\frac{1}{5}\begin{pmatrix}17.5 & 22\\ 22 & 34\end{pmatrix}=\begin{pmatrix}3.5 & 4.4\\ 4.4 & 6.8\end{pmatrix}$$

Alternately covariance between each pair of variables can be computed using following equation (as shown in next slide):

$$Cov(X,Y)=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\overline{X})(Y_i-\overline{Y})$$

PCA –Numerical Example (Solution)

Step 1 (second method)

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
2	1	-2.5	-4	6.25	16	10
3	5	-1.5	0	2.25	0	0
4	3	-0.5	-2	0.25	4	1
5	6	0.5	1	0.25	1	0.5
6	7	1.5	2	2.25	4	3
7	8	2.5	3	6.25	9	7.5
$\bar{x}$=4.5	$\bar{y}$=5			$var(x) = \frac{1}{n-1}\sum(x-\bar{x})^2$ =3.5	$var(y) = \frac{1}{n-1}\sum(y-\bar{y})^2$ =6.8	$cov(x,y) = \frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y})$ =4.4

$$S = \begin{pmatrix} 3.5 & 4.4 \\ 4.4 & 6.8 \end{pmatrix}$$

PCA –Numerical Example (Solution)

Step 2: Find eigen values of covariance matrix

Characteristic Equation=|S-λI|=0

$$\begin{vmatrix} 3.5-\lambda & 4.4 \\ 4.4 & 6.8-\lambda \end{vmatrix} = 0$$

$$(3.5-\lambda)(6.8-\lambda) - 19.36 = 0$$

$$23.8 - 6.8\lambda - 3.5\lambda + \lambda^2 - 19.36 = 0$$

$$\lambda^2 - 10.3\lambda + 4.44 = 0$$

$$\lambda = \frac{10.3 \mp \sqrt{106.09 - 17.76}}{2} = \frac{10.3 \mp 9.43}{2} = 9.865, 0.435$$

Step 3: Magnitude of variance captured along first principal components= 9.865

Percentage of variance captured along first principal components= $\frac{9.865}{9.865+0.435} \times 100\% = 95.78\%$

Yes, it can be transformed to one dimension because maximum variance is captured in first dimension.

PCA –Numerical Example (Solution)

Step 4: First Principal Component i.e. eigen vector for λ_1= 9.86

$$(S- \lambda I)X=O$$

$$\left(\begin{bmatrix} 3.5 & 4.4 \\ 4.4 & 6.8 \end{bmatrix} - 9.86 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) X = O$$

$$\begin{bmatrix} -6.36 & 4.4 \\ 4.4 & -3.06 \end{bmatrix} X = O$$

$$\begin{bmatrix} 1 & -0.69 \\ 4.4 & -3.06 \end{bmatrix} X = O$$

$$X = \begin{bmatrix} 0.47 \\ 0.68 \end{bmatrix} (length\ normalized)$$