Data Preprocessing

Data Reduction

Data Reduction

- Data reduction techniques are applied to obtain a reduced representation of the dataset that is much smaller in volume, yet closely maintain the integrity of the original data
- The pattern mining on the reduced dataset should produce the same or almost same analytical results
- Different strategies:
  - Attribute subset selection (feature selection):
    - Irrelevant, weekly relevant or redundant attributes (dimensions) are detected and removed
  - Dimensionality reduction:
    - Encoding mechanisms are used to reduce the dataset size

Attribute (Feature) Subset Section

- In the context of machine learning, it is termed as feature subset selection
- Irrelevant or redundant features are detected using correlation analysis
- Two strategies:
  - First strategy:
    - Perform the correlation analysis between every pair of attributes
    - Drop one among the two attributes when they are highly correlated
  - Second strategy:
    - Perform the correlation analysis between each attribute and target attribute
      - Classification: Target attribute is class label attribute
      - Regression: Target attribute is attribute corresponding to output variable
    - Drop the attributes that are less correlated with target attribute.

Attribute (Feature) Subset Section

Temperature	Humidity	Pressure	Rain
25.47	82.19	1036.35	6.75
26.19	83.15	1037.60	1761.75
25.17	85.34	1037.89	652.50
24.30	87.69	1036.86	963.00
24.07	87.65	1027.83	254.25
21.21	95.95	1006.92	339.75
23.49	96.17	1006.57	38.25
21.79	98.59	1009.42	29.25
25.09	88.33	991.65	4.50
25.39	90.43	1009.66	112.50
23.89	94.54	1009.27	735.75
22.51	99.00	1009.80	607.50
22.90	98.00	1009.90	717.75
21.72	99.00	996.29	513.00
23.18	98.97	800.00	195.75
21.24	99.00	1009.21	474.75
21.63	99.00	1008.89	409.50
20.91	99.00	1008.89	1161.00
23.67	97.80	1009.38	0.00
24.53	92.90	1008.66	0.00

- Second strategy:
  - Perform the correlation analysis between each attribute and target attribute
  - Drop the attributes that are less correlated with target attribute
- Example:
  - Predicting `Rain` (target attribute) based on `Temperature`, `Humidity` and `Pressure`
  - `Rain` dependent on `Temperature`, `Humidity` and `Pressure`
  - Correlation analysis of `Temperature`, `Humidity`, `Pressure` with `Rain`

Dimensionality Reduction

Tuple (Data Vector) – Attribute (Dimension)

Temperature	Humidity	Pressure	Rain	Moisture
25.47	82.19	1036.35	6.75	0.00
26.19	83.15	1037.60	1761.75	5.69
25.17	85.34	1037.89	652.50	6.85
24.30	87.69	1036.86	963.00	6.04
24.07	87.65	1027.83	254.25	31.24
21.21	95.95	1006.92	339.75	100.00
23.49	96.17	1006.57	38.25	93.20
21.79	98.59	1009.42	29.25	5.77
25.09	88.33	991.65	4.50	4.29
25.39	90.43	1009.66	112.50	3.62
23.89	94.54	1009.27	735.75	3.76
22.51	99.00	1009.80	607.50	4.03
22.90	98.00	1009.90	717.75	3.83
21.72	99.00	996.29	513.00	3.04
23.18	98.97	800.00	195.75	3.00
21.24	99.00	1009.21	474.75	3.05
21.63	99.00	1008.89	409.50	3.00
20.91	99.00	1008.89	1161.00	3.20
23.67	97.80	1009.38	0.00	2.04
24.53	92.90	1008.66	0.00	1.80

- A tuple (one row) is referred as a vector
- Attribute is referred as dimension
- In this example:
  - Number of vectors = number of rows = 20
  - Dimension of a vector = number of attributes = 5
  - Size of data matrix is 20x5

Tuple (Data Vector)

Dimensionality Reduction

- Data encoding or transformations are applied so as to obtain a reduced or compressed representation of the original data

- If the original data can be reconstructed from compressed data without any loss of information, the data reduction is called lossless
- If only an approximation of the original data can be reconstructed from compressed data, then the data reduction is called lossy
- One of the popular and effective methods of lossy dimensionality reduction is principal component analysis (PCA)

Principal Component Analysis (PCA)

- Suppose data to be reduced consist of N tuples (or data vectors) described by d-attributes (d -dimensions)

$$\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$$

$$\mathbf{x}_n = [x_{n1} \;\; x_{n2} \;\; \ldots \;\; x_{nd}]^\mathsf{T}$$

- Let $\mathbf{q}_i$, where $i = 1, 2, \ldots, d$ be the d orthonormal vectors in the d –dimensional space, $\mathbf{q}_i \in \mathrm{R}^d$
  - These are unit vectors that each point in a direction perpendicular to the others

$$\mathbf{q}_i^\mathsf{T}\mathbf{q}_j = 0 \quad \forall i \neq j$$

$$\mathbf{q}_i^\mathsf{T}\mathbf{q}_i = 1$$

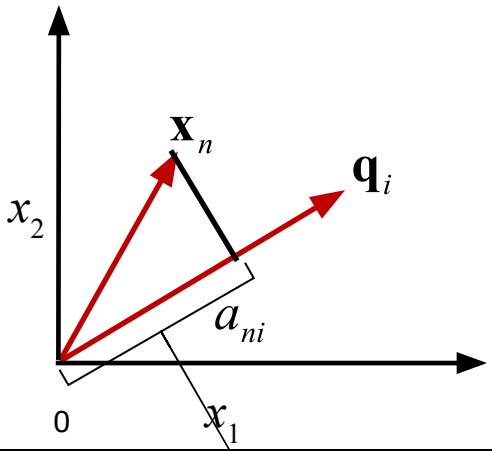
- PCA searches for l orthonormal vectors that can best be used to represent the data, where $l < d$

Principal Component Analysis (PCA)

- These orthonormal vectors are also called as direction of projection
- The original data (each of the tuples (data vectors), $\mathbf{x}_n$) is then projected onto each of the l orthonormal vectors get the principal components

$$a_{ni} = \mathbf{q}_i^{\mathsf{T}} \mathbf{x}_n \quad \forall i = 1, 2, ..., l$$

  - a_{ni} is an i^{th} principal component of $\mathbf{x}_n$
- This transform each of the d – dimensional vectors (i.e. tuples) to l – dimensional vectors



$$\mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \dots \\ x_{nd} \end{bmatrix} \rightarrow \mathbf{a}_n = \begin{bmatrix} a_{n1} \\ a_{n2} \\ \dots \\ a_{nl} \end{bmatrix}$$

- Task:
  - How to obtain the orthonormal vectors?
  - Which l orthonormal vectors to choose?

Principal Component Analysis (PCA)

- Thus the original data is projected onto much smaller space, resulting in dimensionality reduction
- It combines the essence of attributes by creating an alternative, smaller set of variables (attributes)
- It is possible to reconstruct the good approximation of original data, $\mathbf{x}_n$, as linear combination of the direction of projection, $\mathbf{q}_i$, and the principal components, a_{ni}

$$\hat{\mathbf{X}}_n = \sum_{i=1}^{l} a_{ni}\mathbf{q}_i$$

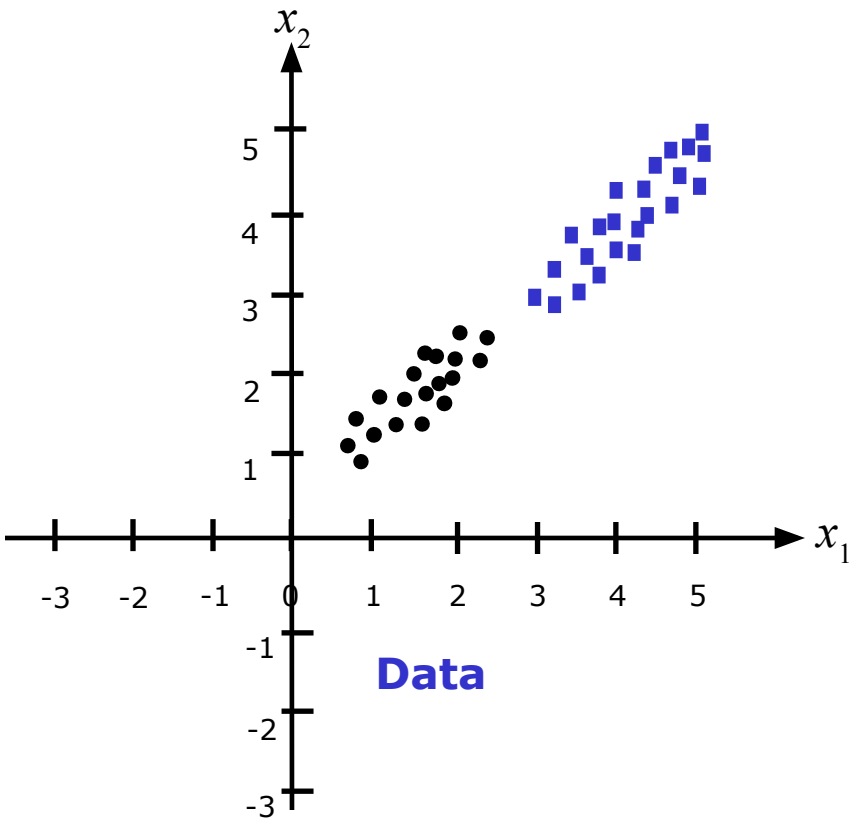
  - $\hat{\mathbf{x}}_n$ is approximation of original tuple $\mathbf{x}_n$

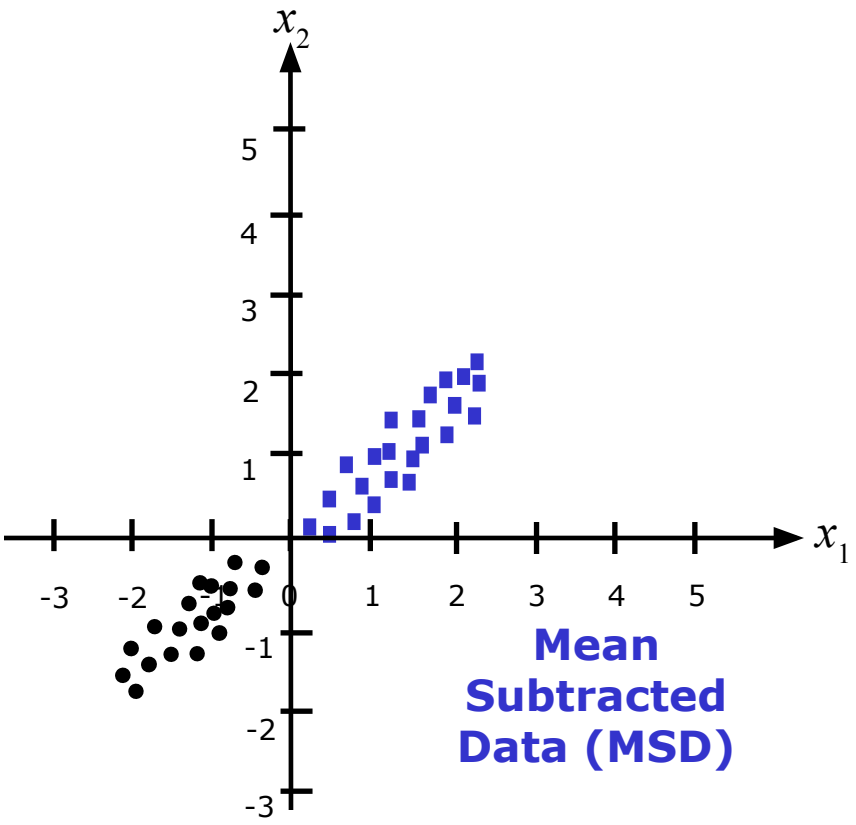
- The Euclidean distance between the original and approximated tuples give the error in reconstruction

$$Error = \|\mathbf{X}_n - \hat{\mathbf{X}}_n\| = \sqrt{\sum_{i=1}^{d}(x_{ni} - \hat{x}_{ni})^2}$$

PCA for Dimension Reduction

- Given: Data with N samples, $\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$
- Remove mean for each attribute (dimension) in data samples (tuples) to obtain the mean subtracted samples





PCA for Dimension Reduction

- Given: Data with N samples, $\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$
- Remove mean for each attribute (dimension) in data samples (tuples) to obtain the mean subtracted samples
- Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathrm{R}^{N \times d}$
  - Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple or a data vector)
- Compute a correlation matrix $\mathbf{C}=\mathbf{X}^{\mathsf{T}}\mathbf{X}$

	A_1	A_2		A_d
$\mathbf{x}_1$	x_{11}	x_{12}	…	x_{1d}
$\mathbf{x}_2$	x_{21}	x_{22}	…	x_{2d}
	.	.	.	.
	.	.	.	.
	.	.	.	.
$\mathbf{x}_N$	x_{N1}	x_{N2}	…	x_{Nd}

PCA for Dimension Reduction

- Given: Data with N samples, $\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$
- Remove mean for each attribute (dimension) in data samples (tuples) to obtain the mean subtracted samples
- Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathrm{R}^{N \times d}$
  - Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple or a data vector)
- Compute a correlation matrix $\mathbf{C}=\mathbf{X}^{\mathsf{T}}\mathbf{X}$

$\mathbf{X}$

$\mathbf{x}_1 - \boldsymbol{\mu}$	$x_{11}-\mu_1$	$x_{12}-\mu_2$	…	$x_{1d}-\mu_d$
$\mathbf{x}_2 - \boldsymbol{\mu}$	$x_{21}-\mu_1$	$x_{22}-\mu_2$	…	$x_{2d}-\mu_d$
	.	.	.	.
	.	.	.	.
	.	.	.	.
$\mathbf{x}_N - \boldsymbol{\mu}_d$	$x_{N1}-\mu_1$	$x_{N2}-\mu_2$	…	$x_{Nd}-\mu_d$

PCA for Dimension Reduction

- Given: Data with N samples, $\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$
- Remove mean for each attribute (dimension) in data samples (tuples) to obtain the mean subtracted samples
- Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathrm{R}^{N \times d}$
  - Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple or a data vector)
- Compute a correlation matrix $\mathbf{C}=\mathbf{X}^T\mathbf{X}$

$\mathbf{X}^T$

$x_{11}-\mu_1$	$x_{21}-\mu_1$	…	$x_{N1}-\mu_1$
$x_{12}-\mu_2$	$x_{22}-\mu_2$	…	$x_{N2}-\mu_2$
…	…	…	…
$x_{1d}-\mu_d$	$x_{2d}-\mu_d$	…	$x_{Nd}-\mu_d$
$(\mathbf{x}_1-\boldsymbol{\mu})^T$	$(\mathbf{x}_2-\boldsymbol{\mu})^T$		$(\mathbf{x}_N-\boldsymbol{\mu})^T$

$\mathbf{X}$

$x_{11}-\mu_1$	$x_{12}-\mu_2$	…	$x_{1d}-\mu_d$	$\mathbf{x}_1-\boldsymbol{\mu}$
$x_{21}-\mu_1$	$x_{22}-\mu_2$	…	$x_{2d}-\mu_d$	$\mathbf{x}_2-\boldsymbol{\mu}$
.	.	.	.	
.	.	.	.	
.	.	.	.	
$x_{N1}-\mu_1$	$x_{N2}-\mu_2$	…	$x_{Nd}-\mu_d$	$\mathbf{x}_N-\boldsymbol{\mu}_d$

PCA for Dimension Reduction

- Compute a correlation matrix $\mathbf{C}=\mathbf{X}^T\mathbf{X}$

$\mathbf{X}^T$

$x_{11}-\mu_1$	$x_{21}-\mu_1$	…	$x_{N1}-\mu_1$
$x_{12}-\mu_2$	$x_{22}-\mu_2$	…	$x_{N2}-\mu_2$
…	…	…	…
$x_{1d}-\mu_d$	$x_{2d}-\mu_d$	…	$x_{Nd}-\mu_d$

$\mathbf{X}$

$x_{11}-\mu_1$	$x_{12}-\mu_2$	…	$x_{1d}-\mu_d$
$x_{21}-\mu_1$	$x_{22}-\mu_2$	…	$x_{2d}-\mu_d$
.	.	.	.
.	.	.	.
.	.	.	.
$x_{N1}-\mu_1$	$x_{N2}-\mu_2$	…	$x_{Nd}-\mu_d$

$\mathbf{X}^T\mathbf{X}$

$-\hat{x}_1$ is approximation of original tuple x_1	$-\hat{x}_1$ is approximation of original tuple x_1	…	$-\hat{x}_1$ is approximation of original tuple x_1
$-\hat{x}_2$ is approximation of original tuple x_2	$-\hat{x}_2$ is approximation of original tuple x_2	…	$-\hat{x}_2$ is approximation of original tuple x_2
…	…	…	…
$-\hat{x}_n$ is approximation of original tuple x_n	$-\hat{x}_n$ is approximation of original tuple x_n	…	$-\hat{x}_n$ is approximation of original tuple x_n

PCA for Dimension Reduction

- Given: Data with N samples, $\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$
- Remove mean for each attribute (dimension) in data samples (tuples) to obtain the mean subtracted samples
- Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathrm{R}^{N \times d}$
  - Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple or a data vector)
- Compute a correlation matrix $\mathbf{C}=\mathbf{X}^{\mathsf{T}}\mathbf{X}$

$\mathbf{X}^{\mathsf{T}}\mathbf{X}$

Var(A1)	COV(A1, A2)	…	COV(A1, Ad)
COV(A2, A1)	Var(A2)	…	COV(A2, Ad)
…	…	…	…
COV(Ad, A1)	COV(Ad, A2)	…	Var(Ad)

PCA for Dimension Reduction

- Given: Data with N samples, $\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$
- Remove mean for each attribute (dimension) in data samples (tuples) to obtain the mean subtracted samples
- Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathrm{R}^{N \times d}$
  - Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple or a data vector)
- Compute a correlation matrix $\mathbf{C}=\mathbf{X}^{\mathsf{T}}\mathbf{X}$
- Perform the eigen analysis of correlation matrix $\mathbf{C}$

$$\mathbf{C}\mathbf{q}_i = \lambda_i \mathbf{q}_i \quad \forall i = 1, 2, ..., d$$

  - As correlation matrix (covariance matrix) is symmetric matrix and positive semidefinite,
    - Each eigenvalues λ_i are distinct and non-negative.
    - Eigenvectors $\mathbf{q}_i$ corresponding to each eigenvalues are orthonormal vectors
    - Eigenvalues indicate the strength of eigenvectors or variance of projected data in the direction of eigenvector

PCA for Dimension Reduction

- Project the $\mathbf{x}_n$ onto each of the directions (eigenvectors) to get the principal components

$$a_{ni} = \mathbf{q}_i^{\mathsf{T}} \mathbf{x}_n \quad \forall i = 1, 2, ..., d$$

  - a_{ni} is an i^{th} principal component of $\mathbf{x}_n$

- Thus, each training example $\mathbf{x}_n$ is transformed to a new representation $\mathbf{a}_n$ by projecting on to d-orthonormal basis (eigenvectors)

$$\mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \dots \\ x_{nd} \end{bmatrix} \rightarrow \mathbf{a}_n = \begin{bmatrix} a_{n1} \\ a_{n2} \\ \dots \\ a_{nd} \end{bmatrix}$$

  - It is possible to reconstruct the original data, $\mathbf{x}_n$, without error as linear combination of the direction of projection, $\mathbf{q}_i$, and the principal components, a_{ni}

$$\mathbf{x}_n = \sum_{i=1}^{d} a_{ni} \mathbf{q}_i$$

PCA for Dimension Reduction

- In general, we are interested in representing the data using fewer dimensions such that the data has high variance along these dimensions
- Idea: Select l out of d orthonormal basis vectors (eigenvectors) that contain high variance of data (i.e. more information content)
- Rank order the eigenvalues (λ_i's) such that

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$$

- Based on the Definition 1, consider the l (l << d) eigenvectors corresponding to l significant eigenvalues
  - Definition 1: Let $\lambda_1, \lambda_2, \ldots, \lambda_d$, be the eigenvalues of an d x d matrix $\mathbf{A}$. λ_1 is called the dominant (significant) eigenvalue of $\mathbf{A}$ if $|\lambda_1| \geq |\lambda_i|, \; i = 1, 2, \ldots, d$

PCA for Dimension Reduction

- Project the $\mathbf{x}_n$ onto each of the l directions (eigenvectors) to get reduced dimensional representation

$$a_{ni} = \mathbf{q}_i^{\mathsf{T}} \mathbf{x}_n \quad \forall i = 1, 2, ..., l$$

- Thus, each training example $\mathbf{x}_n$ is transformed to a new reduced dimensional representation $\mathbf{a}_n$ by projecting on to l-orthonormal basis vectors (eigenvectors)

$$\mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \dots \\ x_{nd} \end{bmatrix} \rightarrow \mathbf{a}_n = \begin{bmatrix} a_{n1} \\ a_{n2} \\ \dots \\ a_{nl} \end{bmatrix}$$

- The eigenvalue λ_i correspond to the variance of projected data

PCA for Dimension Reduction

- Since the strongest l directions are considered for obtaining reduced dimensional representation, it should be possible to reconstruct a good approximation of the original data
- An approximation of original data, $\mathbf{x}_n$, is obtained as linear combination of the direction of projection (strongest eigenvectors), $\mathbf{q}_i$, and the principal components, a_i

$$\hat{\mathbf{X}}_n = \sum_{i=1}^{l} a_i \mathbf{q}_i$$

  - $\hat{\mathbf{x}}_n$ is approximation of original tuple $\mathbf{x}_n$