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 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
    - Drop the attributes that are less correlated with target attribute.

Attribute (Feature) Subset Section

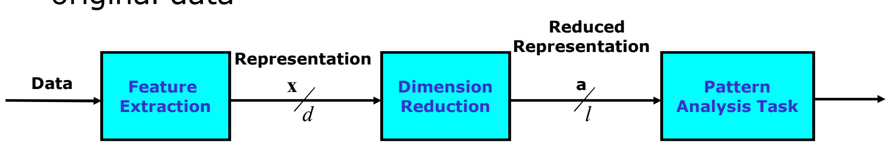
Temperature	Humidity	Pressure	Rain
25.47	82.19	1036.35	6.75
26.19	83.15	1037.60	1761.75
25.17	85.34	1037.89	652.50
24.30	87.69	1036.86	963.00
24.07	87.65	1027.83	254.25
21.21	95.95	1006.92	339.75
23.49	96.17	1006.57	38.25
21.79	98.59	1009.42	29.25
25.09	88.33	991.65	4.50
25.39	90.43	1009.66	112.50
23.89	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

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)

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

Tuple (Data Vector)

- 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

7

Principal Component Analysis (PCA)

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

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

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

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

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

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

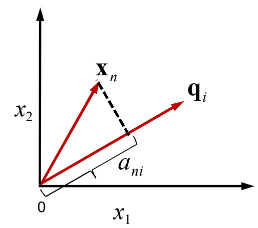
- PCA searches for l orthonormal vectors that can best be used to represent the data, where $l < d$

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

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

– 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

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

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

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

PCA for Dimension Reduction

- Given: Data with N samples, $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathbb{R}^d$
- Remove mean for each attribute (dimension) in data samples (tuples)
- Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathbb{R}^{N \times d}$
  - 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 semideifinite,
    - Each eigenvalues λ_i are distinct and non-negative.
    - Eigenvectors $\mathbf{q}_i$ corresponding to each eigenvalues are orthonormal vectors
    - Eigenvalues indicate the variance or strength of eigenvectors

11

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)

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

- 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}

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

12

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 $|\lambda_1| \geq |\lambda_i|$, $i = 1, 2, \ldots, d$

13

PCA for Dimension Reduction

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

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

- 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)

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

- The eigenvalue λ_i correspond to the variance of projected data

14

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 (stongest 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 $\mathbf{x}_n$

15

PCA: Basic Procedure

- Given: Data with N samples, $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathbb{R}^d$

1. Remove mean for each attribute (dimension) in data samples (tuples)
2. Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples,$\mathbf{X} \in \mathbb{R}^{N \times d}$
   – Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple)
3. Compute a correlation matrix $\mathbf{C}=\mathbf{X}^{\mathsf{T}}\mathbf{X}$
4. 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 is symmetric matrix,
     - Each eigenvalues λ_i are distinct and non-negative
     - Eigenvectors $\mathbf{q}_i$ corresponding to each eigenvalues are orthonormal vectors
     - Eigenvalues indicate the variance or strength of eigenvectors

16

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

5. Rank order the eigenvalues (λ_i's) (sorted order) such that

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

6. Consider the l ($l << d$) eigenvectors corresponding to l significant eigenvalues
7. Project the $\mathbf{x}_n$ onto each of the l directions (eigenvectors) to get reduced dimensional representation

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

17

PCA for Dimension Reduction

8. 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

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

- The new reduced representation $\mathbf{a}_n$ is uncorrelated
- The eigenvalue λ_i correspond to the variance of projected data

18

Illustration: PCA

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

- Atmospheric Data:
  - N = Number tiples (data vectors) = 20
  - d = Number of attributes (dimension) = 5
- Mean of each dimension:

23.42	93.63	1003.55	448.88	14.4

19

Illustration: PCA

Temperature	Humidity	Pressure	Rain	Moisture
2.05	-11.45	32.80	-442.13	-14.37
2.77	-10.49	34.05	1312.88	-8.68
1.75	-8.29	34.34	203.63	-7.52
0.88	-5.95	33.31	514.13	-8.33
0.65	-5.99	24.28	-194.63	16.87
-2.21	2.31	3.37	-109.13	85.63
0.07	2.54	3.02	-410.63	78.83
-1.62	4.96	5.86	-419.63	-8.60
1.68	-5.31	-11.90	-444.38	-10.08
1.98	-3.20	6.11	-336.38	-10.76
0.47	0.90	5.72	286.88	-10.61
-0.91	5.37	6.24	158.63	-10.34
-0.51	4.37	6.34	268.88	-10.54
-1.69	5.37	-7.26	64.13	-11.33
-0.24	5.34	-203.55	-253.13	-11.37
-2.18	5.37	5.65	25.88	-11.32
-1.79	5.37	5.34	-39.38	-11.37
-2.51	5.37	5.34	712.13	-11.18
0.25	4.17	5.83	-448.88	-12.34
1.11	-0.73	5.11	-448.88	-12.57

- Step1: Subtract mean from each attribute

20

Illustration: PCA

- Step2: Compute correlation matrix from the data matrix

50.17	-156.00	268.87	314.10	-183.33
-156.00	666.50	-2224.20	-8746.24	252.92
268.87	-2224.20	47093.53	102982.84	1521.49
314.10	-8746.24	102982.84	4090333.01	-46138.70
-183.33	252.92	1521.49	-46138.70	15811.30

21

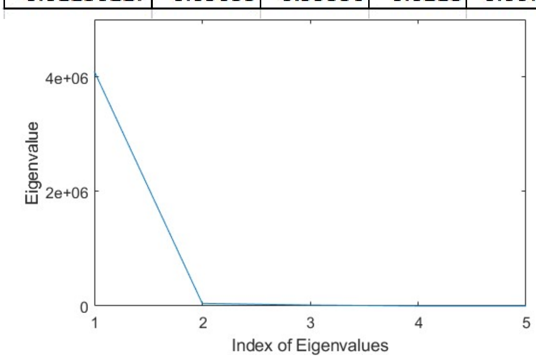
Illustration: PCA

Eigen Values

4093494.12	44809.05	15054.24	587.14	9.95

Eigen Vectors

-7.90E-05	0.00559	-0.01372	0.2496	0.96824
0.00215066	-0.04478	0.02318	-0.967	0.24986
-0.0254375	0.99457	-0.08919	-0.0469	0.00509
-0.99961022	-0.02438	0.01358	-0.0007	0.00042
0.01130117	0.09055	0.99556	0.0218	0.00797



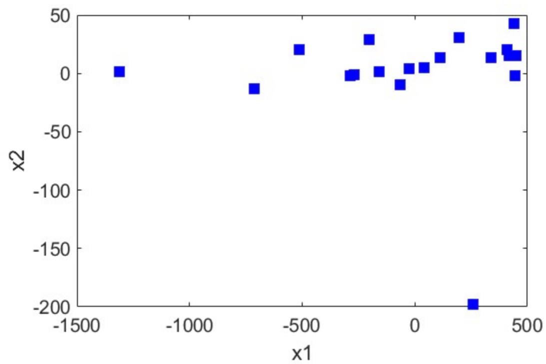
- Step4: Perform Eigen analysis on correlation matrix
  - Get eigenvalues and eigenvectors
- Step5: Sort the eigenvalues in descending order
- Step6: Arrange the eigenvectors in the descending order of their corresponding eigenvalues

22

Illustration: PCA

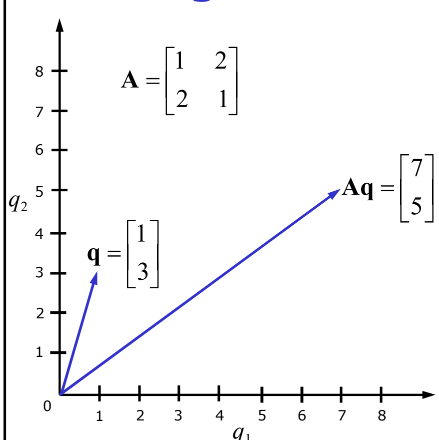
x_1	x_2
440.94	42.62
-1313.36	1.55
-204.53	28.89
-514.88	20.11
194.11	30.69
109.97	13.65
411.29	20.04
419.23	15.05
444.38	-1.67
335.96	13.46
-287.03	-2.30
-158.83	1.16
-269.05	-1.40
-64.04	-10.06
258.09	-197.54
-26.13	3.71
39.11	4.99
-712.10	-13.32
448.43	15.44
448.43	14.93

- Step7: Consider the two leading (significant) eigenvalues and their corresponding eigenvectors
- Step8: Project the mean subtracted data matrix onto the selected two eigenvectors corresponding to leading eigenvalues



23

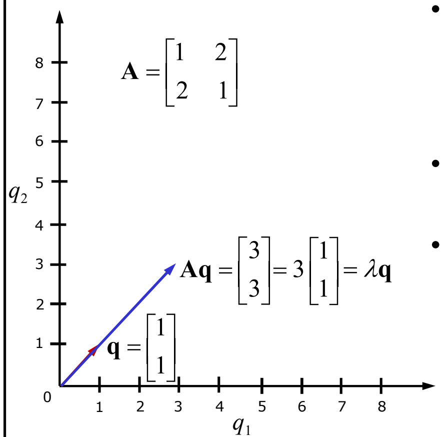
Eigenvalues and Eigenvectors



- What happens when a vector is multiplied with a matrix?
- The vector gets transformed into a new vector
  - Direction changes
- The vector may also get scaled (elongated or shortened) in the process

24

Eigenvalues and Eigenvectors



- For a given square symmetric matrix $\mathbf{A}$, there exist special vectors which do not change direction when multiplied
- These vectors are called eigenvectors
- More formally,

$$\mathbf{Aq} = \lambda\, \mathbf{q}$$

  – λ is eigenvalue
  – Eigenvalue indicate the magnitude of the eigenvector
- The vector will only get scaled but will not change its direction
- *So what is so special about eigenvalues and eigenvectors*?

25

Linear Algebra: Basic Definitions

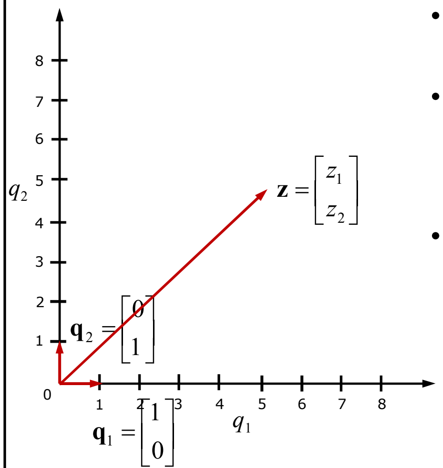
- Basis: A set of vectors $\in \mathbb{R}^d$ is called a basis, if
  – those vectors are linearly independent and
  – every vector $\in \mathbb{R}^d$ can be expressed as a linear combination of these basis vectors
- Linearly independent vectors:
  – A set of d vectors $\mathbf{q}_1$, $\mathbf{q}_2$, . . . , $\mathbf{q}_d$ is linearly independent if no vector in the set can be expressed as a linear combination of the remaining $d-1$ vectors
  – In other words, the only solution to

$$c_1\mathbf{q}_1 + c_2\mathbf{q}_2 + \ldots + c_d\mathbf{q}_d = \mathbf{0} \text{ is } c_1 = c_2 = \ldots = c_d = 0$$

    - Here c_i are scalars

26

Linear Algebra: Basic Definitions



- For example consider the space $\mathbb{R}^2$
- Consider the vectors:

$$\mathbf{q}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \mathbf{q}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Any vector $[z_1 \;\; z_2]^{\mathrm{T}}$ can be expressed as a linear combination of these two vectors

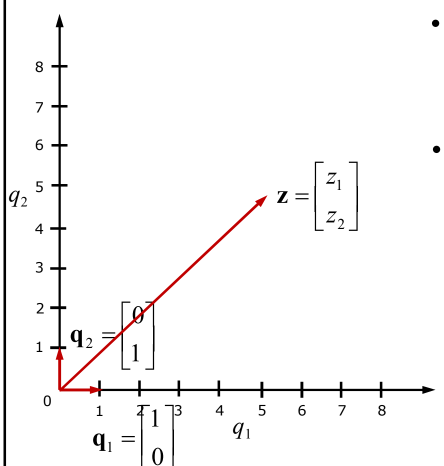
$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = z_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + z_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Further, $\mathbf{q}_1$ and $\mathbf{q}_2$ are linearly independent
  - The only solution to

$$c_1\mathbf{q}_1 + c_2\mathbf{q}_2 = \mathbf{0} \text{ is } c_1 = c_2 = 0$$

27

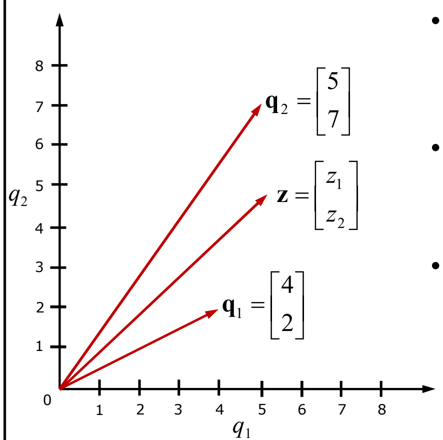
Linear Algebra: Basic Definitions



- It turns out that $\mathbf{q}_1$ and $\mathbf{q}_2$ are unit vectors in the direction of the co-ordinate axes
- And indeed we are used to represent all vectors in $\mathbb{R}^2$ as a linear combination of these two vectors

28

Linear Algebra: Basic Definitions

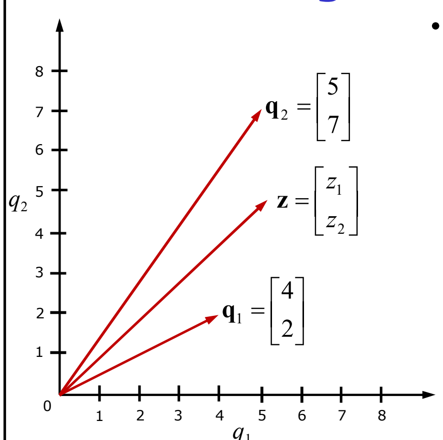


$$z_1 = 4\lambda_1 + 5\lambda_2$$
$$z_2 = 2\lambda_1 + 7\lambda_2$$

- We could have chosen any 2 linearly independent vectors in $\mathbb{R}^2$ as the basis vectors
- For example, consider the linearly independent vectors $[4\ 2]^T$ and $[5\ 7]^T$
- Any vector $\mathbf{z}=[z_1\ \ z_2]^T$ can be expressed as a linear combination of these two vectors
$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \lambda_1 \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \lambda_2 \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$
$$\mathbf{z} = \lambda_1 \mathbf{q}_1 + \lambda_2 \mathbf{q}_2$$
- We can find λ_1 and λ_2 by solving a system of linear equations

29

Linear Algebra: Basic Definitions



- In general, given a set of linearly independent vectors
$$\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_d \in \mathbb{R}^d$$
  - we can express any vector $\mathbf{z} \in \mathbb{R}^d$ as a linear combination of these vectors

$$\mathbf{z} = \lambda_1 \mathbf{q}_1 + \lambda_2 \mathbf{q}_2 + \ldots + \lambda_d \mathbf{q}_d$$

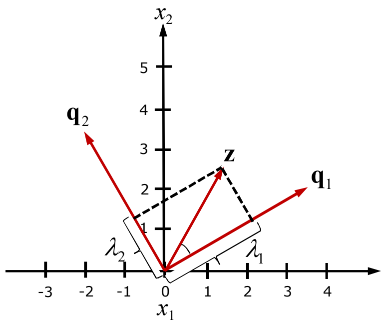
$$\begin{bmatrix} z_1 \\ z_2 \\ . \\ z_d \end{bmatrix} = \lambda_1 \begin{bmatrix} q_{11} \\ q_{12} \\ . \\ q_{1d} \end{bmatrix} + \lambda_2 \begin{bmatrix} q_{21} \\ q_{22} \\ . \\ q_{2d} \end{bmatrix} + \ldots + \lambda_d \begin{bmatrix} q_{d1} \\ q_{d2} \\ . \\ q_{dd} \end{bmatrix}$$

$$\begin{bmatrix} z_1 \\ z_2 \\ . \\ z_d \end{bmatrix} = \begin{bmatrix} q_{11} & q_{21} & \cdots & q_{d1} \\ q_{12} & q_{22} & \cdots & q_{d2} \\ \ldots & \ldots & \ldots & \ldots \\ q_{1d} & q_{2d} & \cdots & q_{dd} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ \lambda_d \end{bmatrix}$$

$$\mathbf{z} = \mathbf{Q} \boldsymbol{\lambda}$$

30

Linear Algebra: Basic Definitions



- Let us see if we have orthonormal basis

  $\mathbf{q}_i^{\mathrm{T}}\mathbf{q}_i = 1$ and $\mathbf{q}_i^{\mathrm{T}}\mathbf{q}_j = 0 \; \forall i \neq j$

- We can express any vector $\mathbf{z} \in \mathbb{R}^d$ as a linear combination of these vectors

  $$\mathbf{z} = \lambda_1\mathbf{q}_1 + \lambda_2\mathbf{q}_2 + \ldots + \lambda_d\mathbf{q}_d$$

  – Multiply $\mathbf{q}_1$ to both sides

  $$\mathbf{q}_1^{\mathrm{T}}\mathbf{z} = \lambda_1\mathbf{q}_1^{\mathrm{T}}\mathbf{q}_1 + \lambda_2\mathbf{q}_1^{\mathrm{T}}\mathbf{q}_2 + \ldots + \lambda_d\mathbf{q}_1^{\mathrm{T}}\mathbf{q}_d$$

  $$\mathbf{q}_1^{\mathrm{T}}\mathbf{z} = \lambda_1$$

- Similarly, $\lambda_2 = \mathbf{q}_2^{\mathrm{T}}\mathbf{z}$

  ...

  $\lambda_d = \mathbf{q}_d^{\mathrm{T}}\mathbf{z}$

- An orthogonal basis is the most convenient basis that one can hope for

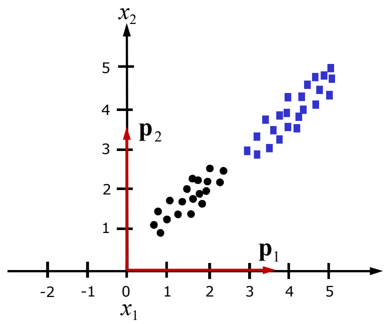
31

Eigenvalues and Eigenvectors

- What does any of this have to do with eigenvectors?
- Eigenvectors can form a basis
- **Theorem 1**: The eigenvectors of a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ having distinct eigenvalues are linearly independent
- **Theorem 2**: The eigenvectors of a square symmetric matrix are orthogonal
- Definition 1: Let $\lambda_1, \lambda_2, \ldots, \lambda_d$, be the eigenvalues of an $d \times d$ matrix $\mathbf{A}$. λ_1 is called the dominant (significant) eigenvalue of $\mathbf{A}$ if $|\lambda_1| \geq |\lambda_i|$, $i = 1, 2, \ldots, d$
- We will put all of this to use for principal component analysis

32

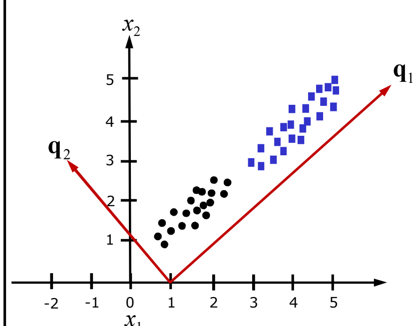
Principal Component Analysis (PCA)



- Each point (vector) here is represented using a linear combination of the x_1 and x_2 axes
- In other words we are using $\mathbf{p}_1$ and $\mathbf{p}_2$ as the basis

33

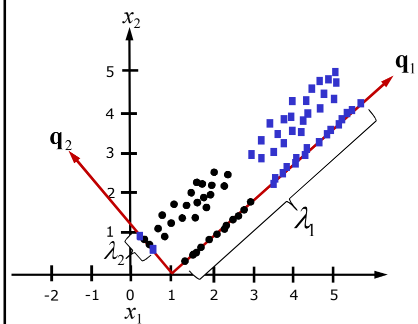
Principal Component Analysis (PCA)



- Lets consider orthonormal vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ as a basis instead of $\mathbf{p}_1$ and $\mathbf{p}_2$ as the basis
- We observe that all the points have a very small component in the direction of $\mathbf{q}_2$ (almost noise)

34

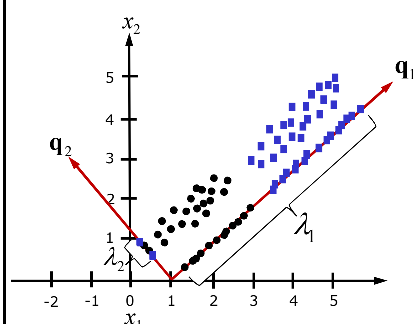
Principal Component Analysis (PCA)



- Lets consider orthonormal vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ as a basis instead of $\mathbf{p}_1$ and $\mathbf{p}_2$ as the basis
- We observe that all the points have a very small component in the direction of $\mathbf{q}_2$ (almost noise)
- Now the same data can be represented in 1-dimension in the direction of $\mathbf{q}_1$ by making a smarter choice for the basis
- Why do we not care about $\mathbf{q}_2$?
  - Variance in the data in this direction is very small
  - All data points have almost the same value in the $\mathbf{q}_2$ direction

35

Principal Component Analysis (PCA)



- If we were to build a classifier on top of this data then $\mathbf{q}_2$ would not contribute to the classier
  - The points are not distinguishable along this direction
- In general, we are interested in representing the data using fewer dimensions such that
  - the data has high variance along these dimensions
  - the dimensions are linearly independent (uncorrelated)
- PCA preserves the geometrical locality of the transformed data with respect to original data

36

PCA: Basic Procedure

- Given: Data with N samples, $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathbb{R}^d$

1. Remove mean for each attribute (dimension) in data samples (tuples)
2. Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathbb{R}^{N \times d}$
   - Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple)
3. Compute a correlation matrix $\mathbf{C}=\mathbf{X}^{\mathsf{T}}\mathbf{X}$
4. 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 is symmetric matrix,
     - Each eigenvalues λ_i are distinct and non-negative
     - Eigenvectors $\mathbf{q}_i$ corresponding to each eigenvalues are orthonormal vectors
     - Eigenvalues indicate the variance or strength of eigenvectors

37

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

5. Rank order the eigenvalues (λ_i's) (sorted order) such that

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$$

6. Consider the l ($l << d$) eigenvectors corresponding to l significant eigenvalues
7. Project the $\mathbf{x}_n$ onto each of the l directions (eigenvectors) to get reduced dimensional representation

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

38

PCA for Dimension Reduction

8. 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

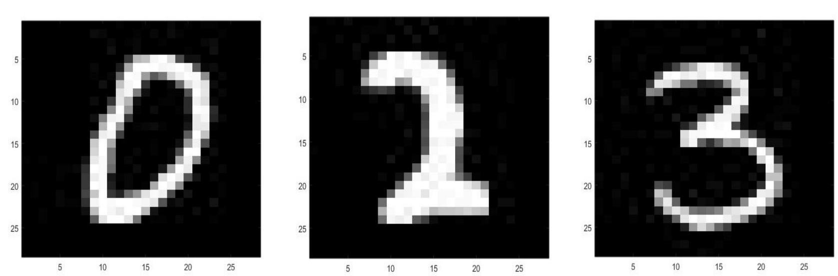
$$\mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \cdots \\ x_{nd} \end{bmatrix} \Rightarrow \mathbf{a}_n = \begin{bmatrix} a_{n1} \\ a_{n2} \\ \cdots \\ a_{nl} \end{bmatrix}$$

- The new reduced representation $\mathbf{a}_n$ is uncorrelated
- The eigenvalue λ_i correspond to the variance of projected data

39

Illustration: PCA

- Handwritten Digit Image [1]:
  - Size of each image: 28 x 28
  - Dimension after linearizing: 784
  - Total number of training examples: 5000 (500 per class)



[1] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Intelligent Signal Processing*, 306-351, IEEE Press, 2001

40

Illustration: PCA

- Handwritten Digit Image:
  - All 784 Eigenvalues

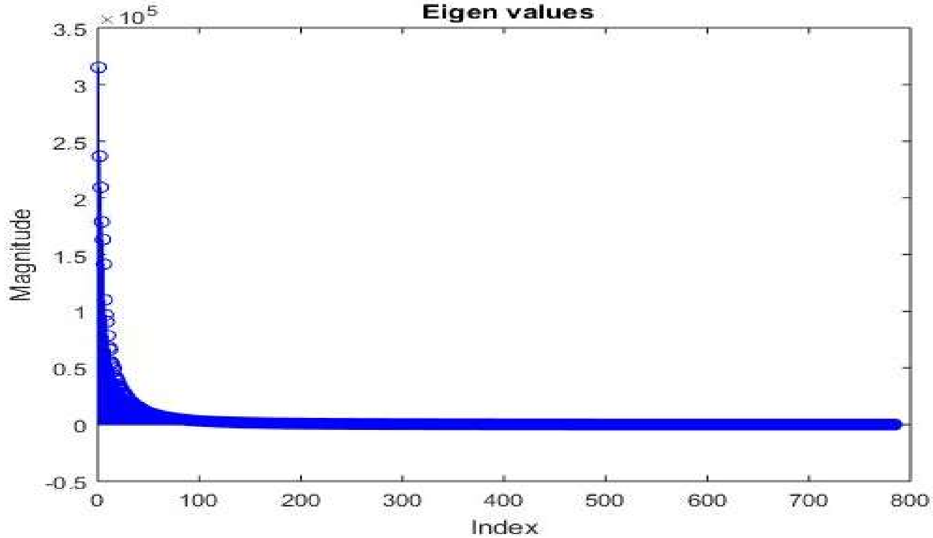


Illustration: PCA

- Handwritten Digit Image:
  - Leading 100 Eigenvalues

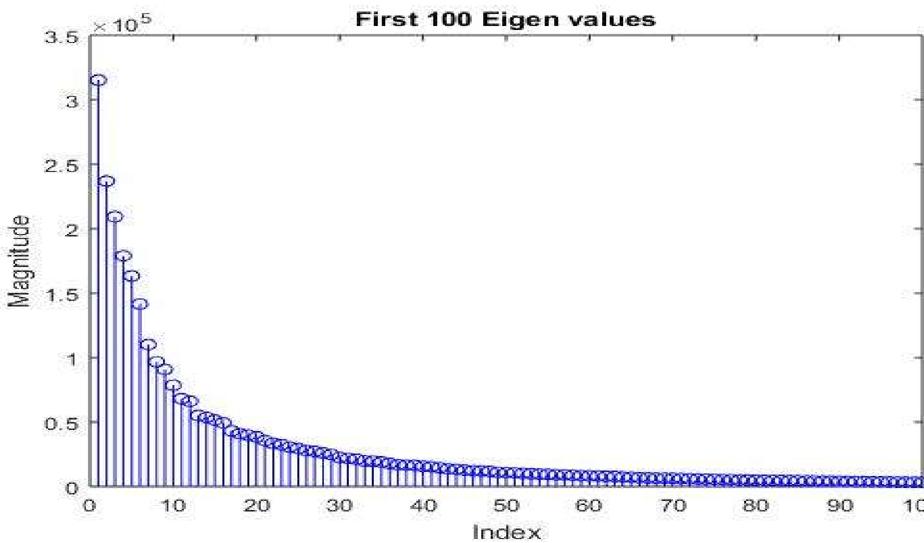
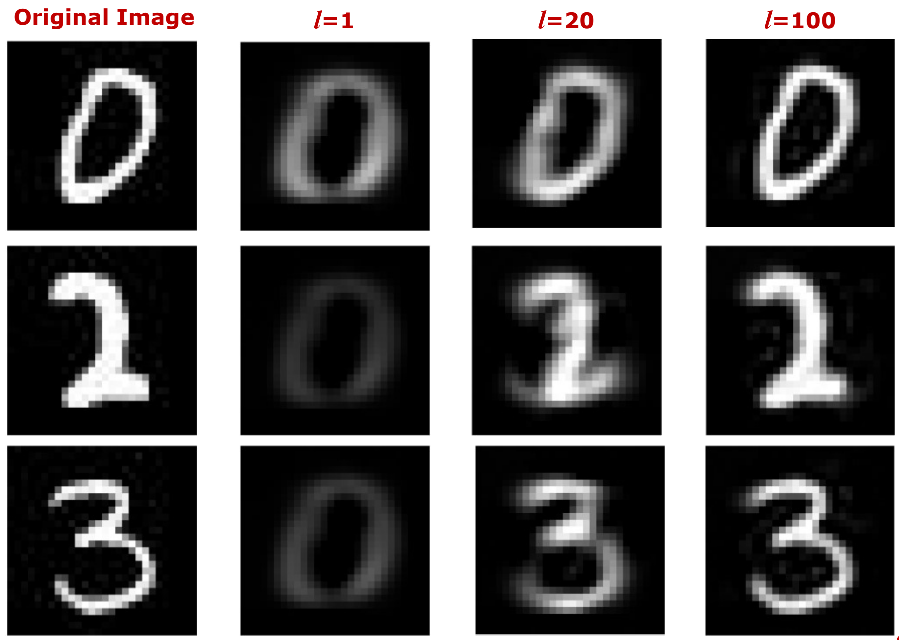


Illustration: PCA-Reconstructed Images



43