Data Science

Predictive and Descriptive Models

- Satishkuar L. Varma

Data Science

Module 1: Introduction to Data Science

Module 2 : Predictive and Descriptive Models

Module 3 : Evaluation and Methodology of Data Science

Module 4: Text Analytics and Recommendation system (RS)

Module 5 : Data Communication and Information Visualization

Module 6 : Scaling with Big Data

Data Science: Predictive and Descriptive Models

- Descriptive Modelling
  - PCA
  - SVD
  - Probabilistic PCA
  - EM Algorithm for PCA
  - ICA
- Predictive Modelling
  - Process
  - Parametric and non-Parametric models
  - BI
  - Challenges of Predictive Analysis
- Time Series Analysis

Data Science: Predictive and Descriptive Models

- Descriptive Modelling
- Data Mining Algorithms: “A data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of models or patterns”
  - “well-defined”: can be encoded in software
  - “algorithm”: must terminate after some finite number of steps

Data Science: Predictive and Descriptive Models

Models

Prediction

- Linear regression
- Piecewise linear
- Nonparametric regression
- Classification

Probability Distributions

- Parametric models
- Mixtures of parametric models
- Graphical Markov models (categorical, continuous, mixed)

Structured Data

- Time series
- Markov models
- Mixture Transition Distribution models
- Hidden Markov models
- Spatial models

Data Science: Predictive and Descriptive Models

- Descriptive Modelling

Patterns

- Global
  - Clustering via partitioning
  - Hierarchical Clustering
  - Mixture Models
- Local
  - Outlier detection
  - Changepoint detection
  - Bump hunting
  - Scan statistics
  - Association rules

Data Science: Predictive and Descriptive Models

- What is a descriptive model?
  - “presents the main features of the data”
  - “a summary of the data”
  - Data randomly generated from a “good” descriptive model will have the same characteristics as the real data
  - Focus on techniques and algorithms for fitting descriptive models to data

- PCA Example

Q Transform following data from 2D space to 1D space using Dimensionality Reduction method (use P.C.A.)

	a_1	a_2
t_1	1	2
t_2	2	1
t_3	3	4
t_4	4	3

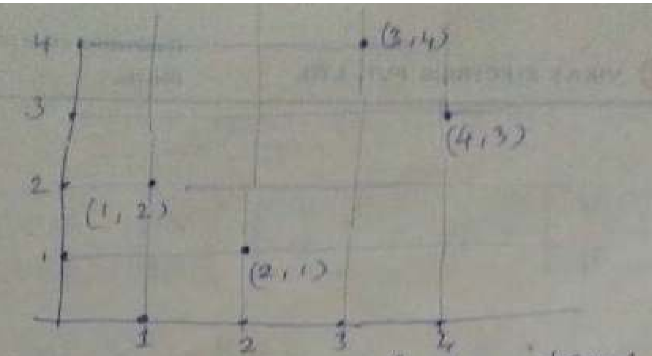
$\Rightarrow$ a_1 & a_2 are two attributes & 4 tuples

Solution:

Drawing the data-items of representing database in Matrix form as Matrix M

Data Science: Predictive and Descriptive Models

- PCA Example



plotting data items

$$M = \begin{vmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{vmatrix}$$

step 1: compute $M^t M$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

step 2: Find Eigenvalues

$$(30-\lambda)(30-\lambda)-(28)(28)=0$$

$$\therefore \lambda = 58, \quad \lambda = 2$$

- PCA Example

Step 3: Calculating EigenVector for $\lambda = 58$

$$\begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 58 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$30x + 28y = 58x$$
$$28x + 30y = 58y$$

giving $x = y$

$\therefore$ assuming $x = y = 1$

So EV for $\lambda = 58 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

(after Normalization)

$\{$ i.e $\sqrt{1^2 + 1^2}$ $\}$

Step 4: Calculating EigenVector for $\lambda = 2$

$$\begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 2 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$30x + 28y = 2x$$
$$28x + 30y = 2y$$

giving $x = -y$ $\therefore$ $x = -1, y = 1$

- PCA Example

So EV for $\lambda = 2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

So $E = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

Step 5: Finding Principal Components.

$$PC = M\,E = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

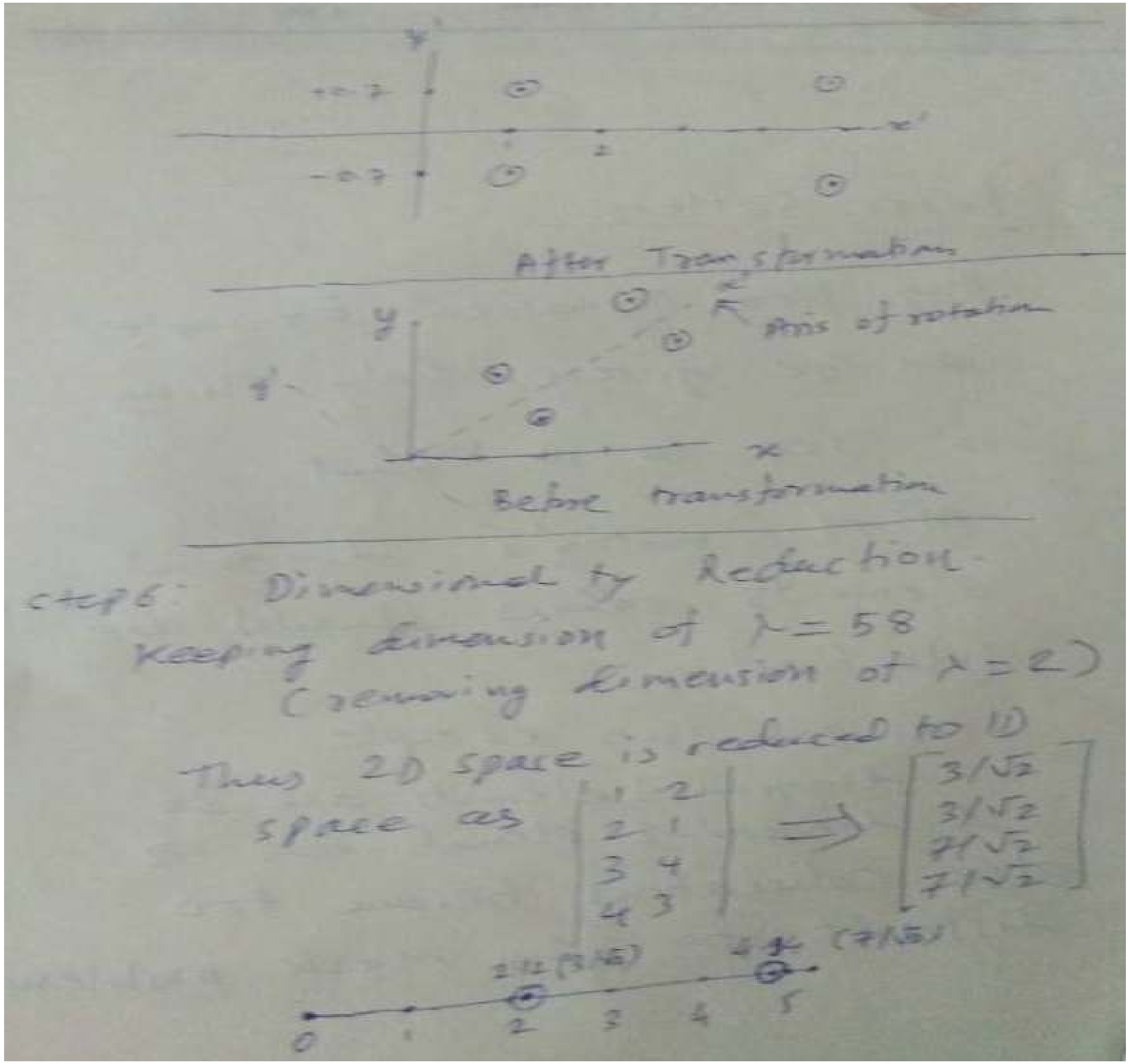
$$\begin{array}{cc} \lambda=58 & \lambda=2 \end{array}$$

$$= \begin{bmatrix} 3/\sqrt{2} & 1/\sqrt{2} \\ 3/\sqrt{2} & -1/\sqrt{2} \\ 7/\sqrt{2} & 1/\sqrt{2} \\ 7/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 2.12 & 0.707 \\ 2.12 & -0.707 \\ 4.94 & 0.707 \\ 4.94 & -0.707 \end{bmatrix}$$

Data Science: Predictive and Descriptive Models

- PCA Example



- Singular Value Decomposition (SVD):
  - A rectangular matrix A_{mn} can be broken down into the product of three matrices
    - an **orthogonal** matrix U,
    - a diagonal matrix S, called **singular** value and
    - the transpose of an **orthogonal** matrix V
    - $A_{mn} = U_{mm} S_{mn} V^T{}_{nn}$
  - Matrix Q is **orthogonal** if its transpose is equal to its inverse $Q^{-1} = Q^T$
  - Unitary if $Q^{-1} = Q^*$ and therefore normal $Q^*Q = QQ^*$
  - Symmetric if $Q = Q^T$
  - $U^TU = I$ and $V^TV = I$
  - Columns of U are orthonormal eigenvectors of AA^T
  - Columns of V are orthonormal eigenvectors of A^TA, and
  - S is a diagonal matrix containing the sqrts of eigenvalues from U or V in descending order

- SVD: Find the singular values of the matrix B

$$B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}. \quad AA^T = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$$

$\lambda^2 - 10\lambda + 9$, so $\lambda = 9$ and $\lambda = 1$ are the eigenvalues

the singular values are 3 and 1

- Example 2

Find the singular values of $A = \begin{bmatrix} 0 & 1 & 1 \\ \sqrt{2} & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$ and find the SDV decomposition

$$AA^T = \begin{bmatrix} 2 & 2 & 2 \\ 2 & 6 & 2 \\ 2 & 2 & 2 \end{bmatrix} \qquad \begin{aligned} -\lambda^3 + 10\lambda^2 - 16\lambda &= -\lambda(\lambda^2 - 10\lambda + 16) \\ &= -\lambda(\lambda - 8)(\lambda - 2) \end{aligned}$$

eigenvalues of AA^T are $\lambda = 8, \lambda = 2, \lambda = 0$

singular values are $\sigma_1 = 2\sqrt{2}, \sigma_2 = \sqrt{2}$ (and $\sigma_3 = 0$).

- SVD: Example 2 (Contd.)
- To give the decomposition, we consider the diagonal matrix of singular values

$$\Sigma = \begin{bmatrix} 2\sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

For $\lambda = 8$, we find an eigenvector $(1, 2, 1)$ - normalizing gives $p_1 = (\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}})$

For $\lambda = 2$ we find $p_2 = (-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}})$, and

For $\lambda = 0$ we get $p_3 = (\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}})$.

This gives the matrix $P = \begin{bmatrix} \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \\ \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$.

- Singular Value Decomposition (SVD): Defintion

Definition

- Any real mxn matrix A can be decomposed uniquely as

$$A = UDV^T$$

U is mxn and column orthogonal (its columns are eigenvectors of AA^T)
$$(AA^T = UDV^T VDU^T = UD^2U^T)$$

V is nxn and orthogonal (its columns are eigenvectors of $A^T A$)
$$(A^T A = VDU^T UDV^T = VD^2 V^T)$$

D is nxn diagonal (non-negative real values called *singular* values)

$$D = diag(\sigma_1, \sigma_2, \ldots, \sigma_n) \quad \text{ordered so that } \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$
(if σ is a singular value of A, it's square is an eigenvalue of $A^T A$)

- If $U = (u_1 \; u_2 \; \cdots \; u_n)$ and $V = (v_1 \; v_2 \; \cdots \; v_n)$, then

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

(actually, the sum goes from 1 to r where r is the rank of A)

- Singular Value Decomposition (SVD) Computation Example
- Step1: To Find U, find AA^T

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \quad A^T = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} \quad AA^T = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$$

- Step2: To find the eigenvalues and corresponding eigenvectors of AA^T

$$A\vec{v} = \lambda\vec{v}, \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$11x_1 + x_2 = \lambda x_1 \qquad x_1 + 11x_2 = \lambda x_2$$

$$(11 - \lambda)x_1 + x_2 = 0 \qquad x_1 + (11 - \lambda)x_2 = 0$$

$$\begin{vmatrix} (11-\lambda) & 1 \\ 1 & (11-\lambda) \end{vmatrix} = 0 \quad (11-\lambda)(11-\lambda) - 1 \cdot 1 = 0 \quad (\lambda - 10)(\lambda - 12) = 0 \quad \lambda = 10, \lambda = 12$$

$$\text{For } \lambda = 10 \quad (11-10)x_1 + x_2 = 0 \quad x_1 = -x_2 \qquad \text{For } \lambda = 12 \quad (11-12)x_1 + x_2 = 0 \quad x_1 = x_2$$

- Eigenvector for λ = 12 is column 1 & eigenvector for λ = 10 is column 2

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

- Singular Value Decomposition (SVD) Computation Example $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
- Step3: Gram-Schmidt orthonormalization process to the column vectors

$$\text{normalize} \quad \vec{u_1} = \frac{\vec{v_1}}{|\vec{v_1}|} = \frac{[1,1]}{\sqrt{1^2+1^2}} = \frac{[1,1]}{\sqrt{2}} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$$

$$\text{Compute} \quad \vec{w_2} = \vec{v_2} - \vec{u_1} \cdot \vec{v_2} * \vec{u_1} =$$

$$[1,-1] - [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \cdot [1,-1] * [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] =$$

$$[1,-1] - 0 * [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] = [1,-1] - [0,0] = [1,-1]$$

$$\text{normalize} \quad \vec{u_2} = \frac{\vec{w_2}}{|\vec{w_2}|} = [\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}] \qquad U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

- Similarly calculate V where V is based on A^TA

$$V = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix} \qquad V^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

- Singular Value Decomposition (SVD) Computation Example
- Step4:
  - S is the sqrts of the non-zero eigenvalues and
  - populate the diagonal with them, putting the largest in s_{11}, the next largest in s_{22} and so on until the smallest value ends up in s_{mn}

$$S = \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}$$

$$A_{mn} = U_{mm} S_{mn} V_{nn}^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sqrt{12}}{\sqrt{2}} & \frac{\sqrt{10}}{\sqrt{2}} & 0 \\ \frac{\sqrt{12}}{\sqrt{2}} & \frac{-\sqrt{10}}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

  - The diagonal entries in *S* are the singular values of *A*, the columns in *U* are called left singular vectors, and the columns in *V* are called right singular vectors

Data Science: Predictive and Descriptive Models

- Application of SVD to document classification

Data Science: Predictive and Descriptive Models

- Descriptive Modelling

References

* **Reference Books:**

1. Davy Cielen,Meysman,Mohamed Ali, “Introducing Data Science”, Dreamtech Press

2. Kevin P. Murphy, “Machine Learning a Probabilistic Perspective”, The MIT Press

3. Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch and George Lapis, “Understanding Big Data: Analytics for Enterprise Class Hadoop and streaming Data”, The McGraw Hill Companies, 2012 "Big Data: The next frontier for innovation, competition, and productivity". Rapporto McKinsey & Company, 2012.

4. Dean Abbott, “Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst”, Wiley, 2014

5. Noel Cressie, Christopher K. Wikle , “Statistics for Spatio-Temporal Data, Wiley

6. Seema Acharya and Subhashini Chellappan, “Big Data and Analytics”, Wiley

7. Rachel Schutt and Cathy O’Neil, “Doing Data Science”, O’Reilly Media

8. Joel Grus, Data Science from Scratch: First Principles with Python, O'Reilly Media

9. EMC Education Services,”Data Science and Big Data Analytics”, Wiley

10. DT Editorial Services, “Big Data Black Book”, Dreamtech Press