Applied Analytics and Predictive Modeling

Spring 2021

Lecture-6

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Some of the slides adapted from Intro to Data Mining Tan et al. 2nd edition

Today’s agenda

- Data Quality contd..
- Eigenvalues and eigenvectors
- Principal Component Analysis

Data Quality …

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data

Information Based Measures

- Information theory is a well-developed and fundamental disciple with broad applications

- Some similarity measures are based on information theory
  - Mutual information in various versions
  - Maximal Information Coefficient (MIC) and related measures
  - General and can handle non-linear relationships
  - Can be complicated and time intensive to compute

Information and Probability

- Information relates to possible outcomes of an event
  - transmission of a message, flip of a coin, or measurement of a piece of data

- The more certain an outcome, the less information that it contains and vice-versa
  - For example, if a coin has two heads, then an outcome of heads provides no information
  - More quantitatively, the information is related the probability of an outcome
    - The smaller the probability of an outcome, the more information it provides and vice-versa
  - Entropy is the commonly used measure

Entropy

- For
  - a variable (event), X,
  - with n possible values (outcomes), $x_1, x_2 \ldots, x_n$
  - each outcome having probability, $p_1, p_2 \ldots, p_n$
  - the entropy of X , $H(X)$, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Entropy is between 0 and $\log_2 n$ and is measured in bits
  - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

  - For $p = 0.5$, $q = 0.5$ (fair coin) $H = 1$
  - For $p = 1$ or $q = 1$, $H = 0$

- What is the entropy of a fair four-sided die ?

Entropy for Sample Data: Example

Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Entropy for Sample Data

- Suppose we have
  - a number of observations (m) of some attribute, X, e.g., the gpa (assuming rounded values) of students in the class,
  - where there are n different possible values
  - And the number of observation in the i^{th} category is m_i
  - Then, for this sample

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- For continuous data, the calculation is harder

Mutual Information

- Information one variable provides about another -- it quantifies the "amount of information" obtained about one random variable through observing the other random variable

Formally, $I(X,Y) = H(X) + H(Y) - H(X,Y)$, where

$H(X,Y)$ is the joint entropy of X and Y,

$$H(X,Y) = -\sum_{i}\sum_{j} p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

- For discrete variables, this is easy to compute

- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where n_X (n_Y) is the number of values of X (Y)

Mutual Information Example

Student Status	Count	p	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

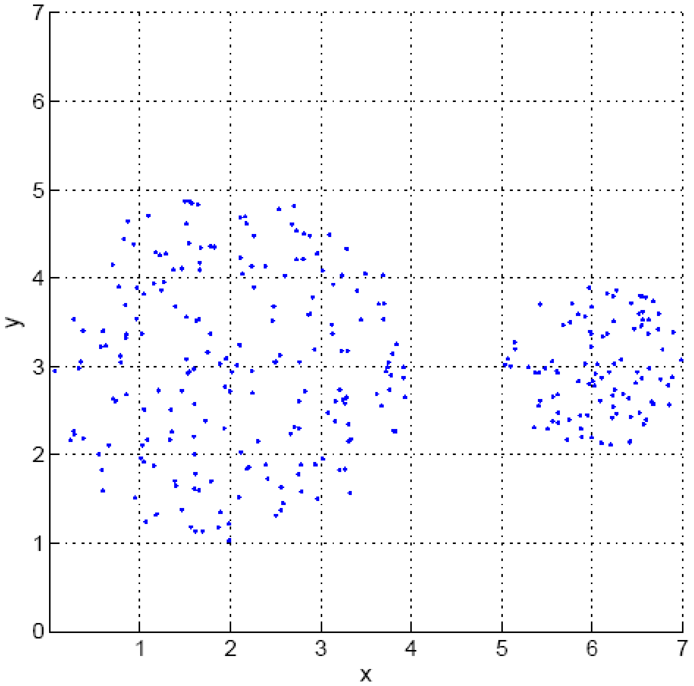
Mutual information of Student Status and Grade = 0.9928 + 1.4406 - 2.2710 = 0.1624

Density

- Measures the degree to which data objects are close to each other in a specified area
- The notion of density is closely related to that of proximity
- Concept of density is typically used for clustering and anomaly detection
- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
    - Estimate what the distribution of the data looks like
  - Graph-based density
    - Connectivity

Euclidean Density: Grid-based Approach

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains



Grid-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Counts for each cell.

Euclidean Density: Center-Based

- Euclidean density is the number of points within a specified radius of the point

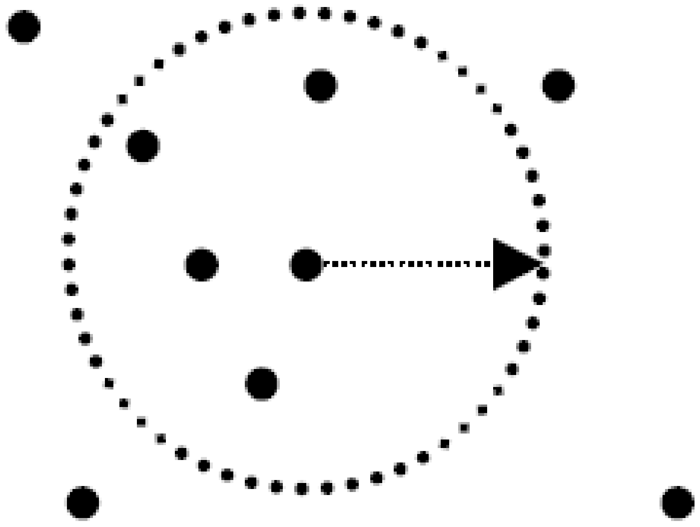
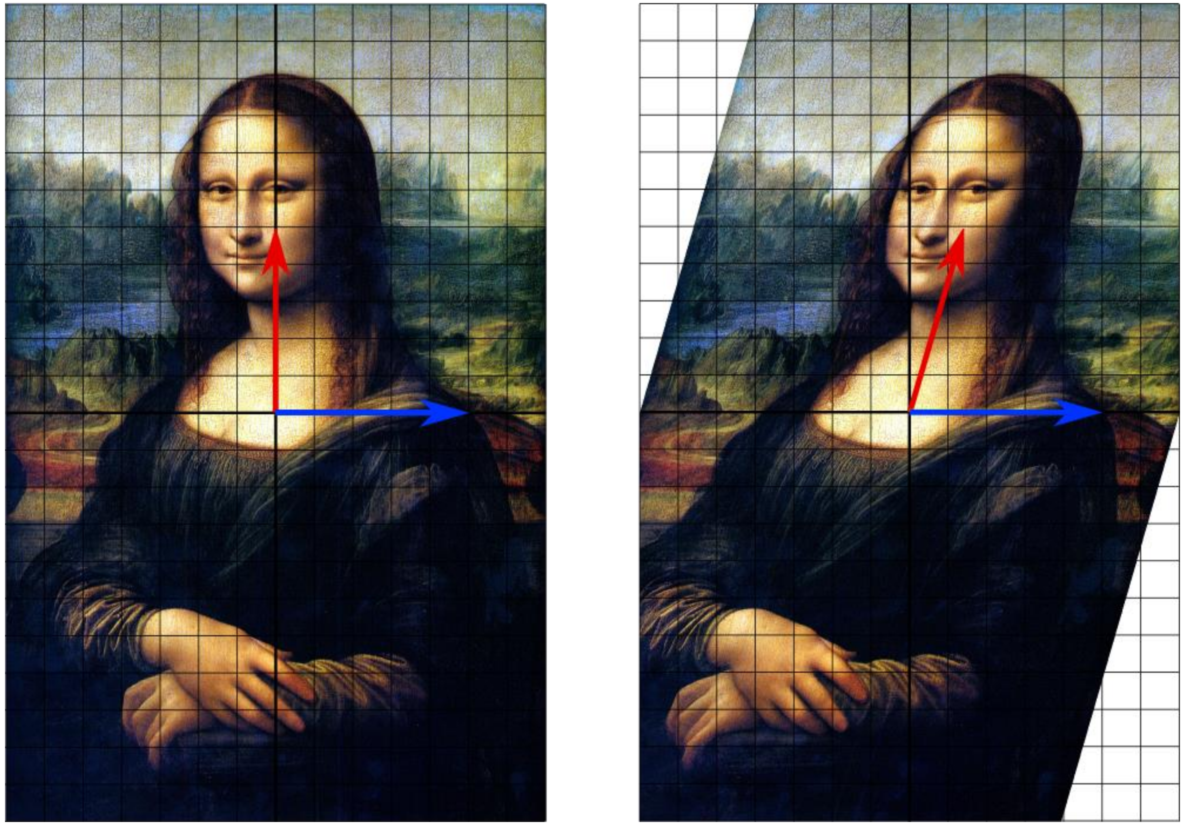


Illustration of center-based density.

Eigenvalues and Eigenvectors

- In the image on the right, when the image is transformed, red arrow changed the direction. But the blue arrow didn't – this is the eigenvector.
- Eigenvector does not change its direction.



Picture credits: By TreyGreer62 - Image:Mona Lisa-restored.jpg, CC0, https://commons.wikimedia.org/w/index.php?curid=12768508

Eigenvalues and Eigenvectors

- Eigenvectors are the **characteristic vectors** that are nonzero vectors.

- Eigenvalues are the scalar values or **factors** with which corresponding eigenvectors are scaled.

- But how do we compute them?

Computing eigenvalues and eigenvectors

- We multiply a matrix with a vector and get the same result when we multiply a scalar by that vector.

We start by finding eigenvalue.

$$Av = \lambda v$$

$$AV = \lambda I V$$

$$AV - \lambda I V = 0$$

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

v is the non-zero eigenvector correspondi to the eigenvalue λ.

Example: Computing eigenvalues and eigenvectors

If $A = \begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}$, Compute eigenvalues and their Corresponding eigenvectors.

Start with: $|A - \lambda I| = 0$ $\longrightarrow$ Finding the determinant.

$$\left| \begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = \left| \begin{bmatrix} -6-\lambda & 3-0 \\ 4-0 & 5-\lambda \end{bmatrix} \right| \quad \text{— ①}$$

$$\begin{vmatrix} -6-\lambda & 3 \\ 4 & 5-\lambda \end{vmatrix} = 0 \quad \text{— ②}$$

$$(-6-\lambda)(5-\lambda) - (3)(4) = 0 \quad \text{— ③}$$

$$-30 + 6\lambda - 5\lambda + \lambda^2 - 12 = 0 \quad \text{— ④}$$

$$\lambda^2 + \lambda - 42 = 0$$

$$(\lambda+7)(\lambda-6) = 0$$

$$\lambda = -7 \text{ or } 6.$$

We found eigenvalues.

Now compute corresponding eigenvectors

If $A = \begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}$, Compute eigenvalues and their Corresponding eigenvectors.

Start with: $|A - \lambda I| = 0$ $\longrightarrow$ Finding the determinant.

$$\left|\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right| = \left|\begin{bmatrix} -6-\lambda & 3-0 \\ 4-0 & 5-\lambda \end{bmatrix}\right| \quad \text{—} ①$$

$$\begin{vmatrix} -6-\lambda & 3 \\ 4 & 5-\lambda \end{vmatrix} = 0 \quad \text{—} ②$$

$$(-6-\lambda)(5-\lambda) - (3)(4) = 0 \quad \text{—} ③$$

$$-30 + 6\lambda - 5\lambda + \lambda^2 - 12 = 0 \quad \text{—} ④$$

$$\lambda^2 + \lambda - 42 = 0$$

$$(\lambda + 7)(\lambda - 6) = 0$$

$$\lambda = -7 \text{ or } 6.$$

Case-1: eigenvalue=6

Case 1: $\lambda = 6$:

$$AV = \lambda V$$

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 6 \begin{bmatrix} x \\ y \end{bmatrix} \quad (\to \text{Multiply})$$

$$\left.\begin{aligned} -6x + 3y &= 6x \\ 4x + 5y &= 6y \end{aligned}\right\} \text{——} \ ①$$

$$-12x + 3y = 0$$

$$4x - y = 0$$

$$\Downarrow$$

$$4x = y \quad \text{or} \quad y = 4x.$$

So, Eigenvector is any non-zero multiple of

$$\begin{bmatrix} 1 \\ 4 \end{bmatrix}.$$

Case-2:
eigenvalue=-7

Case-2: $\lambda = -7$: $\quad Av = \lambda v$

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = (-7) \begin{bmatrix} x \\ y \end{bmatrix}$$

Multiplying these matrices:

$$-6x + 3y = -7x \quad \text{———} \quad (1)$$

$$4x + 5y = -7y$$

$$x + 3y = 0 \quad \text{———} \quad (2).$$

$$4x + 12y = 0$$

$$\Downarrow$$

$$x = -3y \quad \text{or} \quad y = \left(\frac{-1}{3}\right)x.$$

$$\begin{bmatrix} -3 \\ 1 \end{bmatrix}$$ → Eigenvector is any non-zero multiple of this vector.

Lets case-2's eigenvector and multiply with the original matrix

Replace Case-2's eigenvector to multiply with the original matrix.

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} -3 \\ 1 \end{bmatrix} = \begin{bmatrix} (-6)(-3) + (3)(1) \\ (4)(-3) + (5)(1) \end{bmatrix}$$

$$= \begin{bmatrix} 18 + 3 \\ -12 + 5 \end{bmatrix} = \begin{bmatrix} 21 \\ -7 \end{bmatrix}$$

$$\Downarrow$$

$$(-7) \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

Eigenvalue

Eigenvector.

Example-2: eigenvalues and eigenvectors

Matrix is:

$$A = \begin{pmatrix} 2 & 2 \\ 5 & -1 \end{pmatrix}$$

Principal Component Analysis

- Step-1: Standardization
- Step-2: Compute covariance matrix
- Step-3: Compute the eigenvalues and eigenvectors of the covariance matrix
- Step-4: Sort the eigenvalues in a decreasing order
- Step-5: Choose the top-k eigenvectors which are the principal components – these will be the transformed feature vectors