Machine Learning Course
Department of Computer Sc. & Engg.
IIT Patna

Principal Component Ananlysis[a]

Joydeep Chandra
joydeep@iitp.ac.in

December 10, 2018

[a]The contents of the slides are mostly adapted from Jonathon Shlens (PCA) as well as Gilbert Strang book

Content

Characterizing these wines

- Color (C)
- Odour (O)
- Bottle Shape (B)
- Alcohol Content (Al)
- Age (Ag)
- Acidity (Ac)
- ***And many more . . .***

Principal Component Analysis

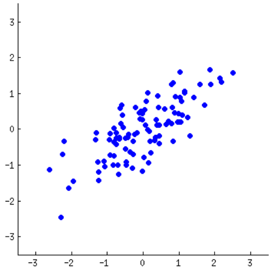
- Many of the properties are related!!!
- So how to summarize the wines with less characteristics?

Summarizing Data Characteristics

Developing New Characteristics

- Look for properties (characteristics) that strongly differs across wines
  - A property like *Alcohol Content* will make all wines look similar
- Look for properties that will allow to predict or reconstruct the original property
  - Selecting a property set like *Acidity, Age* will not help in predicting the *Color*

Changing the Basis



- Figure shows 2 correlated properties (x, y)
- Construct a new property $w_1 x + w_2 y$
- What does this linear transformation signify geometrically?
  - If $\begin{bmatrix} w_1 & w_2 \end{bmatrix}$ represents a vector, then
  - $\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is a measure of projection of x onto line w
- So what ideally should the vector w_1, w_2 be??

- In the figure if x varies then y also varies, so both x and y are required to represent the points
- It would have been ideal if we can transform x and y into x' and y' such that with variation in x', y' did not vary at all and hence could be dropped from consideration.

Representing the data in vector format

- If $C(i), O(i), B(i), Al(i), Ag(i), Ac(i)$ represents the data collected from the i^{th} bottle

$$\text{Let } X_i = \begin{bmatrix} C(i) \\ O(i) \\ B(i) \\ Al(i) \\ Ag(i) \\ Ac(i) \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ X_1 & X_2 & \cdots & X_n \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}$$

Basis for a Vector Space

Definition

A basis for a vector space **V** is a sequence of vectors that hold 2 properties

- Vectors are linearly independent
- Spans the space **V**

A Naive Example

$$\mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}$$

Dimension of a Vector Space

The number of basis vectors is the dimension of the vector space

Change of Basis

Searching for Alternate Basis

Q: Is there another basis **P**, that is a linear combination of the original basis that can best represent our data set?

Change of Basis

- If **P** is a transformation matrix, then
  - **PX=Y** re-represents dataset **X** in form of **Y**

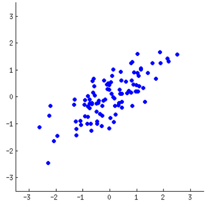
Geometric Interpretation

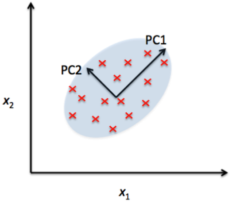
- If $\mathbf{p_i}$ are the row vectors of **P**
- X_i are the column vectors of **X**
- Y_i are the column vectors of **Y**
- $\mathbf{p_i} \cdot X_i$ is a projection of X_i on $\mathbf{p_i}$

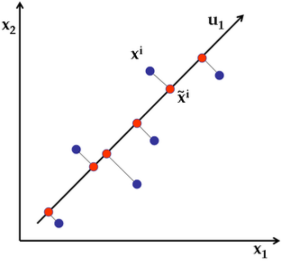
$$\mathbf{PX} = \begin{bmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{bmatrix} \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix} = \mathbf{Y}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{p_1} \cdot X_1 & \cdots & \mathbf{p_1} \cdot X_n \\ \vdots & \ddots & \vdots \\ \mathbf{p_m} \cdot X_m & \cdots & \mathbf{p_m} \cdot X_n \end{bmatrix}$$

The PCA Concept: Diagrammatically







The PCA Objective

The objective of PCA is to find a new basis that minimizes the co-variance of the data between 2 features

Understanding the relations between features

Variance and Covariance

Consider two sets of measurements with zero mean $Ac = \{a_1, a_2, \dots, a_n\}$ and $Ag = \{b_1, b_2, \dots, b_n\}$

- Variance $\sigma^2_{Ac} = \langle a_i a_i \rangle_i = \frac{1}{n-1} \sum_i a_i^2$
- Covariance $\sigma^2_{AcAg} = \langle a_i b_i \rangle_i = \frac{1}{n-1} \sum_i a_i b_i$

Important Facts about Covariance

- $\sigma^2_{AcAg} = 0$ if Ac & Ag are entirely un-correlated
- $\sigma^2_{AcAg} = \sigma^2_{Ac}$ if $Ac = Ag$

Covariance as Dot Product of 2 vectors

- Suppose $\mathbf{x_1}$ and $\mathbf{x_2}$ are two row vectors representing data for 2 features
- Obtained from n instance/trials (in our case different bottles of wine)

$$\text{Covariance} = \frac{1}{n-1} \mathbf{x_1} \mathbf{x_2}^T$$

The Covariance Matrix

- Row vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}$
  - Denotes the data obtained for each of the m features in n trials (instances)

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_m} \end{bmatrix}$$

Covariance matrix $\mathbf{S_x} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T$

Properties of Covariance Matrix

- $\mathbf{S_x}$ is a square symmetric $m \times m$ matrix
- Diagonal terms of $\mathbf{S_x}$=Variances
- Non-diagonal terms of $\mathbf{S_x}$ = Covariances

Goals

- Transform $\mathbf{X}$ to $\mathbf{Y}$ such that
  - All non-diagonal entries of Covariance matrix $\mathbf{S_Y}$ = ZERO

Formally

Find a orthonormal matrix $\mathbf{P}$ where $\mathbf{Y} = \mathbf{PX}$ such that $\mathbf{S_Y} = \frac{1}{n-1}\mathbf{YY}^T$ is diagonalized.

Solving PCA

- Let $\mathbf{Y} = \mathbf{PX}$

$$
\begin{aligned}
\mathbf{S_Y} &= \frac{1}{n-1}\mathbf{YY}^T \\
&= \frac{1}{n-1}(\mathbf{PX})(\mathbf{PX})^T \\
&= \frac{1}{n-1}(\mathbf{PXX}^T\mathbf{P}^T) \\
&= \frac{1}{n-1}\mathbf{P}(\mathbf{XX}^T)\mathbf{P}^T) \\
&= \frac{1}{n-1}\mathbf{PAP}^T
\end{aligned}
$$

- Where $\mathbf{A} = \mathbf{XX}^T$ (Symmetric Matrix).

Solving PCA Contd...

Property of Symmetric Matrix

For a symmetric matrix $\mathbf{A}$

$$\mathbf{A} = \mathbf{EDE}^T$$

- $\mathbf{D}$ is a diagonal matrix
- $\mathbf{E}$ is the orthonormal Eigen Vectors of $\mathbf{A}$ arranged in columns

Choosing $\mathbf{P} = \mathbf{E}^T$

- If $\mathbf{P} = \mathbf{E}^T$, then $\mathbf{A} = \mathbf{P}^T\mathbf{DP}$ and $\mathbf{P}^T = \mathbf{P}^{-1}$ (Property of Orthonormality)
- Then

$$\mathbf{S_Y} = \frac{1}{n-1}\mathbf{PAP}^T = \frac{1}{n-1}\mathbf{P}(\mathbf{P}^T\mathbf{DP})\mathbf{P}^T = \frac{1}{n-1}\mathbf{D}$$

- Thus principal components of $\mathbf{X}$ are
  - Eigen vectors of $\mathbf{XX}^T$

Singular Value Decomposition

- Suppose $\mathbf{X}$ is a $n \times m$ matrix of rank r (Convention Reversed)
- Then $\mathbf{X}^T\mathbf{X}$ is of rank r and dimension $n \times n$
- Suppose $\{\mathbf{v_1}, \mathbf{v_2}, \dots, \mathbf{v_r}\}$ be the set of orthonormal eigen vectors of $\mathbf{X}^T\mathbf{X}$
- Let $\{\lambda_1, \lambda_2, \dots, \lambda_r\}$ be the corresponding eigen values

$$(\mathbf{X}^T\mathbf{X})\mathbf{v_i} = \lambda_i \mathbf{v_i}$$

- Let $\sigma_i = \sqrt{\lambda_i}$ (Positive real and are called Singular Values)
- Let $\mathbf{u_i} = \frac{1}{\sigma}\mathbf{X}\mathbf{v_i}$ be orthonormal vectors of dimension $n \times 1$

Properties of $\mathbf{u_i}$ and $\mathbf{v_i}$

- Property 1: $\mathbf{u_i} \cdot \mathbf{u_j} = \delta_{ij}$
- Property 2: $||\mathbf{X}\mathbf{v_i}|| = \sigma_i$

- Let $\mathbf{V} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_m}]$
- Let $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_n}]$
- Let $\Sigma = Diag(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, 0, \ldots)$ such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$

The Matrix Version of SVD

$$\begin{aligned} \mathbf{XV} &= \mathbf{U}\Sigma \\ \text{or } \mathbf{X} &= \mathbf{U}\Sigma V^T \end{aligned}$$

Limitations of SVD

- Difficult to interpret the negative values of $\Sigma\mathbf{V}^T$ and $\mathbf{U}\Sigma$
- Assumes a Gaussian distribution of the data