Introduction to Machine Learning

Principal Component Analysis

Varun Chandola

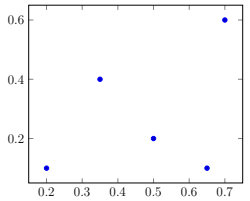
May 3, 2019

Outline

Contents

1 Recap

- Factor Analysis Models
  - **Assumption**: $\mathbf{x}_i$ is a multivariate Gaussian random variable
  - Mean is a function of $\mathbf{z}_i$



  - Covariance matrix is fixed

$$p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

  - $\mathbf{W}$ is a $D \times L$ matrix (loading matrix)
  - $\boldsymbol{\Psi}$ is a $D \times D$ diagonal covariance matrix
- Extensions:
  - *Independent Component Analysis*.
  - If $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$ and $\mathbf{W}$ is orthonormal $\Rightarrow$ FA is equivalent to **Probabilistic Principal Components Analysis** (PPCA)
  - If $\sigma^2 \rightarrow 0$, FA is equivalent to PCA

2 Principal Components Analysis

2.1 Introduction to PCA

- Consider the following data points
- *Embed* these points in 1 dimension
- What is the best way?
  - **Along the direction of the maximum variance**
  - Why?

2.2 Principle of Maximal Variance

- Least loss of information
- Best capture the "spread"
- What is the direction of maximal variance?
- Given any direction ($\hat{\mathbf{u}}$), the projection of $\mathbf{x}$ on $\hat{\mathbf{u}}$ is given by:

$$\mathbf{x}_i^\top \hat{\mathbf{u}}$$

- Direction of maximal variance can be obtained by maximizing

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^\top \hat{\mathbf{u}})^2 &= \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{u}}^\top \mathbf{x}_i \mathbf{x}_i^\top \hat{\mathbf{u}} \\ &= \hat{\mathbf{u}}^\top \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \mathbf{x}_i^\top\right)\hat{\mathbf{u}} \end{aligned}$$

Finding Direction of Maximal Variance

- Find:

$$\underset{\hat{\mathbf{u}}:\hat{\mathbf{u}}^\top\hat{\mathbf{u}}=1}{max}\ \hat{\mathbf{u}}^\top \mathbf{S}\hat{\mathbf{u}}$$

  where:

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \mathbf{x}_i^\top$$

- $\mathbf{S}$ is the sample (empirical) covariance matrix of the mean-centered data

The solution to the above constrained optimization problem may be obtained using the Lagrange multipliers method. We maximize the following w.r.t. $\hat{\mathbf{u}}$:

$$(\hat{\mathbf{u}}^\top \mathbf{S}\hat{\mathbf{u}}) - \lambda(\hat{\mathbf{u}}^\top\hat{\mathbf{u}} - 1)$$

to get:

$$\begin{aligned} \frac{d}{d\hat{\mathbf{u}}}(\hat{\mathbf{u}}^\top \mathbf{S}\hat{\mathbf{u}}) - \lambda(\hat{\mathbf{u}}^\top\hat{\mathbf{u}} - 1) &= 0 \\ \mathbf{S}\hat{\mathbf{u}} - \lambda\hat{\mathbf{u}} &= 0 \\ \mathbf{S}\hat{\mathbf{u}} &= \lambda\hat{\mathbf{u}} \end{aligned}$$

Obviously, the solution to the above equation is an eigen vector of the matrix $\mathbf{S}$. But which $\mathbf{S}$? Note that for the optimal solution:

$$\hat{\mathbf{u}}^\top \mathbf{S}\hat{\mathbf{u}} = (\hat{\mathbf{u}}^\top \lambda\hat{\mathbf{u}}) = \lambda$$

Thus we should choose the largest possible λ which means that the first solution is the eigen vector of $\mathbf{S}$ with largest eigen value.

2.3 Defining Principal Components

- First PC: Eigen-vector of the (sample) covariance matrix with largest eigen-value
- Second PC?
- Eigen-vector with next largest value
- Variance of each PC is given by λ_i
- Variance captured by first L PC ($1 \leq L \leq D$)

$$\frac{\sum_{i=1}^{L}\lambda_i}{\sum_{i=1}^{D}\lambda_i} \times 100$$

- What are eigen vectors and values?

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

  $\mathbf{v}$ is eigen vector and λ is eigen-value for the **square matrix A**

- Geometric interpretation?

For the second PC we need to optimize the variance with a second constraint that the solution should be orthogonal to the first PC. It is easy to show that this solution will be the PC with second largest eigen value, and so on.

2.4 Dimensionality Reduction Using PCA

- Consider first L eigen values and eigen vectors
- Let $\mathbf{W}$ denote the $D \times L$ matrix with first L eigen vectors in the columns (sorted by λ's)
- PC score matrix
$$\mathbf{Z} = \mathbf{X}\mathbf{W}$$
- Each input vector ($D \times 1$) is replaced by a shorter $L \times 1$ vector

2.5 PCA Algorithm

1. *Center* $\mathbf{X}$
$$\mathbf{X} = \mathbf{X} - \hat{\boldsymbol{\mu}}$$
2. Compute sample covariance matrix:
$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}^\top\mathbf{X}$$
3. Find eigen vectors and eigen values for $\mathbf{S}$
4. $\mathbf{W}$ consists of first L eigen vectors as columns
   - Ordered by decreasing eigen-values
   - $\mathbf{W}$ is $D \times L$
5. Let $\mathbf{Z} = \mathbf{X}\mathbf{W}$
6. Each row in $\mathbf{Z}$ (or $\mathbf{z}_i^\top$) is the lower dimensional embedding of $\mathbf{x}_i$

2.6 Recovering Original Data

- Using $\mathbf{W}$ and $\mathbf{z}_i$
$$\hat{\mathbf{x}}_i = \mathbf{W}\mathbf{z}_i$$
- **Average Reconstruction Error**
$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$$

Theorem 1 (Classical PCA Theorem)**.** *Among all possible orthonormal sets of L basis vectors, PCA gives the solution which has the minimum reconstruction error.*

- Optimal "embedding" in L dimensional space is given by $z_i = \mathbf{W}^\top\mathbf{x}_i$

2.7 Eigen Faces

EigenFaces [**?**]

- **Input:** A set of images (of faces)
- **Task:** Identify if a new image is a face or not.

3 Probabilisitic PCA

- Recall the **Factor Analysis** model
$$p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi})$$
- For PPCA, $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$
- Covariance for each observation $\mathbf{x}$ is given by:
$$\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$
- If we maximize the log-likelihood of a data set $\mathbf{X}$, the MLE for $\mathbf{W}$ is:
$$\hat{W} = \mathbf{V}(\boldsymbol{\Lambda} - \sigma^2\mathbf{I})^{\frac{1}{2}}$$
- $\mathbf{V}$ - first L eigenvectors of $\mathbf{S} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$
- $\boldsymbol{\Lambda}$ - diagonal matrix with first L eigen values

3.1 EM for PCA

- PPCA formulation allows for EM based learning of parameters
- $\mathbf{Z}$ is a matrix containing N latent random variables

Benefits of EM

- EM can be faster
- Can be implemented in an online fashion
- Can handle missing data

References