A.Z
Data Mining
Dec, 2023

lec 2 PAC and SVD

Abstract

PCA and SVD

1 Dimension reduction

why?

- Curse of dimensionality

- Redundancy in data

Idea: Reduce dimensionality by removing redundancy

Methods: PCA and SVD

1.1 Principal component analysis (PCA)

Intuition: Rotate axes to match highest variance directions and choose the best new axes

Given the two assumptions:

1 High variance reflects important structures of data

2 Data can be presented well as a linear combination of suitable orthogonal basis vectors

Idea: Given $n \times d$ data and suitable $d \times r (r < d)$ matrix $\mathbf{P}_r$, new data will be $n \times r$ matrix $\mathbf{DP}_r$

Note: PCA just a assumption and may not always reflect reality

1.1.1 Eigen-decomposition

Assume $\mathbf{D}$ mean-centered. Convariance matrix $\mathbf{C} = \frac{1}{n-1}\mathbf{D}^T\mathbf{D}$

Since $\mathbf{C}$ positive semidefinite (non-neg, symmetric), it can be diagonalized $\mathbf{C} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$

where $\mathbf{P}$'s columns = $\mathbf{C}$'s orthonormal eigenvectors, $\mathbf{\Lambda}$ is diagonal, $\mathbf{\Lambda}_{ii}$ = eigenvalues

Transformed data $\mathbf{D'} = \mathbf{DP}$

Thus , $\mathbf{\Lambda}$ is the new covariance matrix (diagonal, i.e. , no correlations)

This method essentially rotating and scaling the data into a new coordinate system defined by the eigenvectors.

Keep only r largest eigen values $(\lambda_1, \ldots, \lambda_r)$ and corresponding eigen vectors $\mathbf{P}_r$

Approximate data $\mathbf{D'} = \mathbf{DP}_r$

Further reading: https://arxiv.org/abs/1404.1100

1.2 Singular value decompositions (SVD)

Intuition: Singular values give insight into the 'importance' or 'weight' of each corresponding singular vector in representing the data. A larger singular value indicates that its corresponding singular vector captures more variance (or information) in the data.

Decompositions: Factorize $\mathbf{D}$ as $\mathbf{D} = \mathbf{Q\Sigma P}^T$

where

- $\mathbf{\Sigma}$ diagonal, $\mathbf{\Sigma}_{ii} = \sigma_i$ singular values

- $\mathbf{Q}$'s columns left singular vectors (orthonormal eigenvectors of $\mathbf{DD}^T$)

- $\mathbf{P}$'s columns right singular vectors (orthonormal eigenvectors of $\mathbf{D}^T\mathbf{D}$)

Transformed data $\mathbf{D'} = \mathbf{DP}$

If $\mathbf{D}$ mean-centered, same basis vector as PCA

If $\mathbf{D}$ sparse, non-negative (e.g. document-word matrices), mean-centering often skipped,

Latend semantic analysis (LSA) applies SVD on document-term matrix

1.3 PCA vs SVD

Factorize $\mathbf{C} = \mathbf{P\Lambda P}^T$ or $\mathbf{D} = \mathbf{Q\Sigma P}^T$

Use $\mathbf{P}_r$ (best components of $\mathbf{P}$).

Reduced data $\mathbf{D'} = \mathbf{DP}_r$

further reading: https://jonathan-hui.medium.com/machine-learning-singular-value-decomposition-svd-principal-component-analysis-pca-1d45e885e491