Machine learning - Dimensionality Reduction

Yonghoon Dong

May 18, 2024

Why we have to reduce dimension? : Curse of Dimension

Normally, datasets are typically **high-dimensional**. In other words, the dimension of data space is normally high.

Why we have to reduce dimension? : Curse of Dimension

As dimension grows, our model could be suffered from following reasons

1. fewer observations per region
2. higher computation cost
3. suffered from overfitting problem

Manifold Hypothesis

1. The manifold hypothesis posits that many high-dimensional data sets that occur in the real world actually lie along low-dimensional latent manifolds inside that high-dimensional space.
2. It is suggested that this principle underpins the effectiveness of machine learning algorithms in describing high-dimensional data sets by considering a few common features.

`https://en.wikipedia.org/wiki/Manifold_hypothesis`

Dimensionality Reduction

Dimensionality reduction aims to transform data from a high-dimensional space into a low-dimensional space.

1. Linear Dimensionality reduction
    1. Principal component analysis (PCA)
    2. Factor Analysis
    3. Mixture of factor analysis
2. Non-linear dimensionality reduction
    1. Kernel Principal component analysis (Kernel PCA)
    2. t-SNE

Dimensionality Reduction

This algorithm is commonly used for

1. Feature extraction
2. Data compression
3. Data visualization (e.g. t-SNE)

Formal definition of Dimensionality Reduction

In dimensionality reduction, we want to learn a map f from high-dimensional visible space (or data space) $x \in \mathbb{R}^D$ to a low-dimensional latent space (sometime it is called hidden variable) $z \in \mathbb{R}^L$.

Formal definition of Dimensionality Reduction

This map can be

1. parametric model $z = f(x; \theta)$: it can be used as a preprocessing step for other kinds of algorithm
2. non-parametric model : mostly used for data visualization

Principal components analysis (PCA)

1. PCA is simplist and most widely used form of dimensionality reduction algorithm.
2. Idea : find a **linear** and **orthogonal** projection of the high dimensional data to a low dimensional subspace

Therefore, we want to find a **low dimensional representation** that represents a best approximation to the original data.

Definition of "good approximation"

Then the fundamental question is that what is criteria of determining a best approximation? There is two different but actually same criteria.

1. Variance perspective : maximize the variance
2. Residual perspective : minimize the distance (l_2-norm)

Important

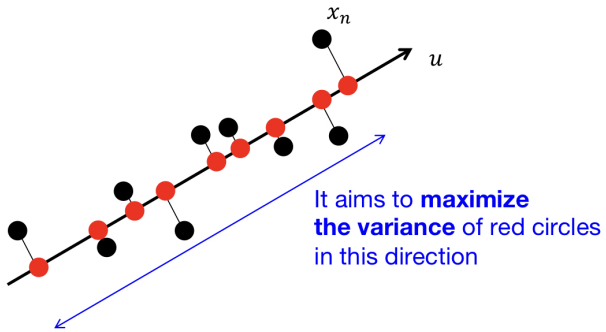
It is really important to know that these perspective are actually same.

Caution

From now on, every vector will be treated as a column vector.

Variance perspective : Maximum Variance

Given a dataset $\{x_n\}_{n=1}^{N}$ where $x_n \in \mathbb{R}^D$, the goal is to project the data onto a space having dimensionality $M < D$ while **maximizing the variance** of the projected data



Variance perspective : M-dimensional Principal Subspace

A projection onto M-dimensional subspace of $\mathbb{R}^D$ can be described by linear combination of orthonormal basis $\mathcal{B} = \{u_1, u_2, \ldots, u_M\}$ where $u_i \in \mathbb{R}^D$.

1. $\|u_i\| = 1$ for all $i \in \{1, \ldots, M\}$
2. $\langle u_i, u_j \rangle = 0$ if $i \neq j$ (i.e. $u_i^T u_j = 0$)
3. $\langle u_i, u_j \rangle = 1$ if $i = j$ (i.e. $u_i^T u_j = 1$)

Intuitive Explanation

You can intuitively think of these vectors as coordinates. In other words, we want to express the data with respect to these vectors.

Variance perspective : M-dimensional Principal Subspace

Let $S = \text{Span}\{u_1, \ldots, u_M\}$. Consider the coordinate representation of $\text{Proj}_S(x)$ where $x \in \mathbb{R}^D$,

$$[\text{Proj}_S(x)]_{\mathcal{B}} = (\lambda_1, \ldots, \lambda_M) \tag{1}$$

where $\text{Proj}_S(x) = \lambda_1 u_1 + \cdots + \lambda_M u_M$, and $\lambda_i = \langle x, u_1 \rangle$ for all $i \in \{1, \ldots, M\}$

Intuitive Explanation

Let $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$. Then $x = (a, b, c)$ could be expressed as $x = ae_1 + be_2 + ce_3$. Therefore, the fundamental idea is we just want to express x with respect to given basis.

Variance perspective : M-dimensional Principal Subspace

We want to maximize the variance for each coordinate. Let $\text{Proj}_S^n(x) = u_n^T x$. Then our objective is

$$\max_U \sum_{m=1}^{M} \mathbb{E}[(\text{Proj}_S^m(X) - \text{Proj}_S^m(\bar{X}))^2] \tag{2}$$

$$\Rightarrow \max_U \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} (u_m^T x_n - u_m^T \bar{x})^2 \tag{3}$$

$$\Rightarrow \max_U \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} (u_m^T (x_n - \bar{x}))(u_m^T (x_n - \bar{x}))^T \tag{4}$$

$$\Rightarrow \max_U \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} u_m^T (x_n - \bar{x})(x_n - \bar{x})^T u_m \tag{5}$$

Variance perspective : M-dimensional Principal Subspace

We want to maximize the variance for each coordinate. Let $\text{Proj}_S^n(x) = u_n^T x$. Then our objective is

$$\Rightarrow \max_U \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} u_m^T (x_n - \bar{x})(x_n - \bar{x})^T u_m \tag{6}$$

$$\Rightarrow \max_U \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} \langle u_m, (x_n - \bar{x})(x_n - \bar{x})^T u_m \rangle \tag{7}$$

$$\Rightarrow \max_U \sum_{m=1}^{M} \langle u_m, \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T u_m \rangle \tag{8}$$

$$\Rightarrow \max_U \sum_{m=1}^{M} u_m^T S u_m \tag{9}$$

Variance perspective : M-dimensional Principal Subspace

Where

$$\bar{x} := \frac{1}{N} \sum_{n=1}^{N} x_n \tag{10}$$

$$S := \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T \tag{11}$$

Variance perspective : M-dimensional Principal Subspace

Therefore, our objective can be formulated as

$$\max_{U} \sum_{n=1}^{N} u_m^T S u_m \tag{12}$$

subject to $u_m^T u_m = 1$ for all $m \in \{1, \ldots, M\}$

Variance perspective : M-dimensional Principal Subspace

Since we have equality constraints, we can apply Lagrange multipliers

$$\mathcal{L}(u, \lambda) = \sum_{n=1}^{N} u_m^T S u_m - \sum_{m=1}^{M} \lambda_m (u_m^T u_m - 1) \tag{13}$$

Then $\frac{\partial \mathcal{L}}{\partial u_m} = 0$ for all $m \in \{1, \ldots, M\}$. In other words,

$$S u_m = \lambda_m u_m \text{ for all } m \in \{1, \ldots, M\} \tag{14}$$

Note

Therefore, **optimal** value of u_m is related to the eigenvector of S and λ_m is related to the eigenvalue corresponds to u_m.

Variance perspective : M-dimensional Principal Subspace

Therefore,

$$(u_m^*)^T S(u_m^*) = (u_m^*)^T \lambda_m u_m^* \tag{15}$$
$$= \lambda_m (u_m^*)^T u_m^* \tag{16}$$
$$= \lambda_m \tag{17}$$

for all $m \in \{1, \ldots, M\}$

Note

Note that u_m^* is a optimal value of u_m

Variance perspective : M-dimensional Principal Subspace

So PCA is equivalent to calculating the eigenvectors of the data covariance matrix S corresponding to the 1st to M'th largest eigenvalues.

Spectral Theorem

Then the natural question is whether we can always find such eigenvalues or eigenvectors always. The answer is always yes.

Theorem (Spectral theorem)

Let V is a real-vector space. If A is symmetric on V, then there exists an orthonormal basis of V consisting of eigenvectors of A. Each eigenvalue of A is real.

So, every symmetric matrix in a real-vector space is always diagonalizable, and its eigenvalues are all real.

Characteristic of Covariance matrix

Every covariance matrix has following properties

1. Positive semi-definite : $v^T S v \geq 0$ for all $v \in \mathbb{R}^D$ (All eigenvalues are positive)
2. Symmetric : By spectral theorem, all eigenvalues are real

Residual perspective : Minimum Error

Consider orthonormal basis $\mathcal{B}' = \{u_1, u_2, \ldots, u_D\}$ where $u_i \in \mathbb{R}^D$. For given x_n, we want to find the best approximation of x_n that belongs to $span\{u_1, \ldots, u_M\}$. In other words,

$$\mathcal{J} := \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2 \tag{18}$$

where $\tilde{x}_n$ is a approximation of x_n belongs to $span\{u_1, \ldots, u_M\}$

Residual perspective : Minimum Error

As you might guess, if the error $\tilde{x}_n - x_n$ perpendicular to $span\{u_1, \ldots, u_M\}$, it is optimal choice. In other words,

$$\tilde{x}_n - x_n \perp u_m \quad m = 1, \ldots, M \tag{19}$$

Residual perspective : Minimum Error

In this perspective, our objective can be written as

$$\min_{U} \sum_{i=M+1}^{D} u_i^T S u_i \tag{20}$$

subject to $u_m^T u_m = 1$ for all $m \in \{M+1, \ldots, D\}$

Extended understanding : SVD in Linear Algebra

Define a matrix X such that

$$X = \begin{bmatrix} x_1 - \bar{x} & \cdots & x_N - \bar{x} \end{bmatrix} \tag{21}$$

where $X \in \mathbb{R}^{D \times N}$ and $\bar{x}$ is the mean of x_1 to x_N. Then the column space or range of X represents the scaled data space.

Extended understanding : SVD in Linear Algebra

Note that $X^T X$ is equivalent to the covariance matrix S.

$$X^T X = \begin{bmatrix} (x_1 - \bar{x})^T \\ \vdots \\ (x_N - \bar{x})^T \end{bmatrix} \begin{bmatrix} x_1 - \bar{x}, & \cdots & , x_N - \bar{x} \end{bmatrix} \tag{22}$$

$$= \begin{bmatrix} (x_1 - \bar{x}_1)^T (x_1 - \bar{x}_1) & \cdots & (x_1 - \bar{x}_1)^T (x_N - \bar{x}_1) \\ \vdots & \ddots & \vdots \\ (x_N - \bar{x}_N)^T (x_1 - \bar{x}_1) & \cdots & (x_N - \bar{x}_N)^T (x_N - \bar{x}_N) \end{bmatrix} \tag{23}$$

$$= S \tag{24}$$

Extended understanding : SVD in Linear Algebra

1. Principle component analysis (PCA) : PCA aims to maximize the variance in the data to find the principal components. It is used for dimensionality reduction, uncovering the main patterns in data, and facilitating data analysis and visualization.
2. Singular value decomposition (SVD) : SVD is primarily used for low-rank approximation of a matrix, which is useful in data compression, noise reduction, and data reconstruction.

They are fundamentally linked. The principal components in PCA can be derived using SVD